frontiers | Research Topics

Integrative analysis of single-cell and/or bulk multi-omics sequencing data

Edited by
Geng Chen, Xingdong Chen, Rongshan Yu and Zhichao Liu

Published in
Frontiers in Genetics

ISSN 1664-8714
ISBN 978-2-83251-332-3
DOI 10.3389/978-2-83251-332-3

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

Integrative analysis of single-cell and/or bulk multi-omics sequencing data

Topic editors

Geng Chen — Stemirna Therapeutics Co., Ltd., China

Xingdong Chen — Fudan University, China

Rongshan Yu — Xiamen University, China

Zhichao Liu — Boehringer Ingelheim (United States), United States

Citation

Chen, G., Chen, X., Yu, R., Liu, Z., eds. (2023). *Integrative analysis of single-cell and/or bulk multi-omics sequencing data*. Lausanne: Frontiers Media SA.

doi: 10.3389/978-2-83251-332-3

Table of contents

frontiers | Frontiers in Genetics

TYPE Editorial
PUBLISHED 04 January 2023
DOI 10.3389/fgene.2022.1121999

OPEN ACCESS

EDITED AND REVIEWED BY
Quan Zou,
University of Electronic Science and Technology of China, China

*CORRESPONDENCE
Geng Chen,
chengeng66666@outlook.com

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

RECEIVED 12 December 2022
ACCEPTED 13 December 2022
PUBLISHED 04 January 2023

CITATION
Chen G, Yu R and Chen X (2023), Editorial: Integrative analysis of single-cell and/or bulk multi-omics sequencing data.
Front. Genet. 13:1121999.
doi: 10.3389/fgene.2022.1121999

Editorial: Integrative analysis of single-cell and/or bulk multi-omics sequencing data

Geng Chen[1]*, Rongshan Yu[2] and Xingdong Chen[3]

[1]Stemirna Therapeutics Co., Ltd., Shanghai, China, [2]Department of Computer Science, School of Informatics, Xiamen University, Xiamen, China, [3]State Key Laboratory of Genetic Engineering, Human Phenome Institute, School of Life Sciences, Fudan University, Shanghai, China

KEYWORDS

multi-omics, single-cell sequencing, bulk sequencing, integrative analysis, data integration

Editorial on the Research Topic
Integrative analysis of single-cell and/or bulk multi-omics sequencing data

Introduction

Each type of omics data including genomics, epigenomics, transcriptomics, proteomics, metabolomics, and metagenomics mainly provides the profile of one particular layer for a cell or sample (Hasin et al., 2017). Integrative analysis of multi-omics data could enable a more comprehensive dissection from different perspectives, which may facilitate a better and deeper understanding of the underlying molecular functions and mechanisms (Li et al., 2021). With the innovation and development of sequencing technologies, various single-cell and bulk profiling technologies have been developed and applied to a diversity of biological and clinical research (Lei et al., 2021; Li et al., 2021; Jiang et al., 2022). Bulk sequencing approaches allow the elucidation of each sample at the cell-population level, providing the averaged profile of a multitude of cells. By contrast, single-cell sequencing methods can interrogate thousands of cells at single-cell resolution for a given sample simultaneously. Joint analysis of multi-omics data generated from bulk and single-cell sequencing protocols could effectively facilitate the translation of basic science to practical applications (Stuart and Satija, 2019; Leng et al., 2022). On the other hand, the sample/cell scale and data size are growing rapidly in biomedical investigation. Thus, novel bioinformatics approaches are also in urgent need to more efficiently and robustly integrate distinct types of omics data.

Since multi-omics strategies could be more powerful than single omics, combining different types of single-cell or bulk sequencing data for a more comprehensive exploration has become increasingly popular and important (Figure 1). In this Research Topic on Integrative Analysis of Single-Cell and/or Bulk Multi-omics

Sequencing Data, we planned to collect novel findings and methods related to analyzing bulk and single-cell multi-omics or multimodal data with a systematic strategy. In total, 12 original research articles and one case report were published in this Research Topic, covering multi-omics-based cancer dissection, comparison of different data integration methods, and database construction for expression examination in various tissues. Here we concisely summarize and discuss the main results revealed in these studies.

Studies published in this research topic

Guo et al. found that the mutations in TP53 and KRAS were significantly associated with the poor prognosis of intrahepatic cholangiocarcinoma (ICC). They further classified the ICC patients into different subgroups based on the mutation feature of TP53 and KRAS, which could benefit the clinical management of ICC. Johann et al. uncovered that the mutations of AKT1 and TP53 signaling pathways were closely associated with the pulmonary sclerosing pneumocytoma (PSP) through integrative analysis of genomic, transcriptomic, radiomic, and pathomic data. The insights into the underlying etiology and molecular behavior of PSP gained in this study may benefit corresponding therapy. Gao et al. constructed an effective prognostic model for breast cancer using the differentially expressed genes among distinct glycosylation patterns. Their results highlight the value and importance of risk score characterization based on glycosylation patterns for predicting the overall survival and immune infiltration of breast cancer patients. Hao et al. identified two subgroups of MYC signaling inhibition and activation for lung adenocarcinoma (LUAD) through joint analysis of genomics, transcriptomics, and single-cell sequencing data from multiple cohorts. The two LUAD subgroups discovered by them exhibited significant differences in terms of prognosis, genomic variations, immune microenvironment, as well as clinical features. Additionally, Jiang et al. built and validated a model for predicting the prognosis of LUAD by integrating bulk and single-cell RNA-seq data. They also detected two distinct subtypes of LUAD patients that differed in prognosis and immune characteristics. Sun et al. systematically analyzed the transcriptome of synovial sarcoma in terms of gene expression, alternative splicing, gene fusion, and circular RNAs. Their integrative analysis provided new insights into the transcriptomic profile and the underlying molecular mechanism of synovial sarcoma. Wang et al. constructed a clinical diagnostic map and a cluster prediction model for glioblastoma based on the methylation, expression, and single-cell sequencing data. The classification method developed by them could potentially promote the analysis of methylation heterogeneity for the promoter CpG islands in glioblastoma. Zhao et al. revealed high cellular heterogeneity in both malignant and immune cells of diffuse large B-cell lymphoma (DLBCL). They provided novel insights into the transcriptional dynamics of the tumor microenvironment for DLBCL. Zhang et al. established a prognostic model based on eight genes (DEFB1, AICDA, TYK2, CCR7, SCARB1, ULBP2, STC2, and LGR5) for

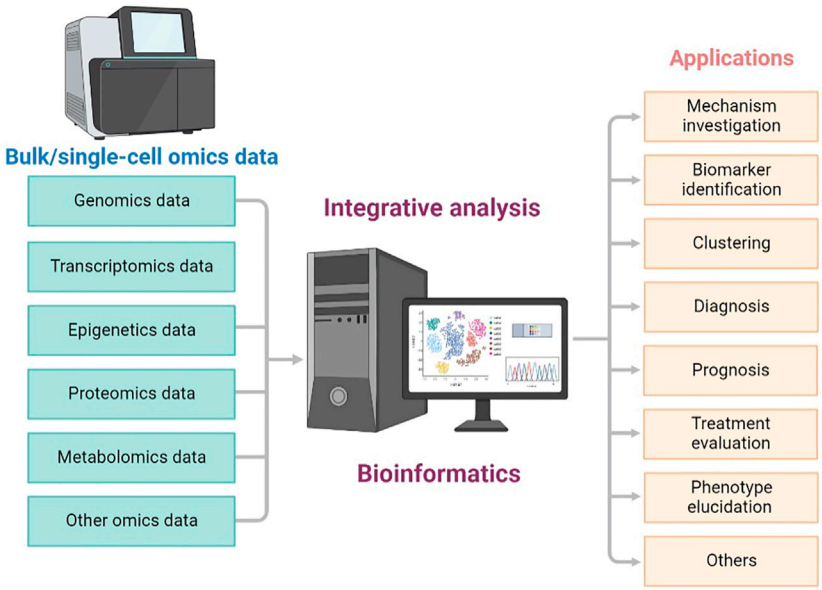


FIGURE 1
Overview of integrative analysis of multi-omics data generated from different bulk and single-cell sequencing technologies.

predicting the overall survival of head and neck squamous cell carcinoma (HNSCC) patients. The low-risk and high-risk groups of HNSCC separately showed higher and lower immune scores, thus those eight gene signatures have the potential to be used in the clinical management of HNSCC. Li et al. classified hepatocellular carcinoma patients into high-necroptosis and low-necroptosis groups, which had a significant difference in survival time. They found that the high-necroptosis patients were with an enriched expression of immune checkpoint-related genes and could benefit from certain immunotherapy. Wang et al. uncovered four dysregulated oncogenic signaling pathways and identified related potential prognostic biomarkers for pan-cancer through systematic analysis of the TCGA multi-omics data. Their results could facilitate a better understanding of the function of oncogenic signaling pathways in human pan-cancer. Kujawa et al. systematically evaluated the influence of six different data integration methods on single-cell analysis. They found that ComBat-seq (Zhang et al., 2020), limma (Leek et al., 2012), and MNN (Haghverdi et al., 2018) could effectively reduce batch effects and preserve the differences between distinct biological conditions. Deng et al. constructed a gene expression omnibus database named ECO (https://heomics.shinyapps.io/ecodb/) for mouse endothelial cells based on the sequencing data of 203 samples from 71 different conditions. ECO could enable researchers to friendly explore endothelial expression profiles of diverse tissues in conditions of certain genetic modifications, disease models, and other stimulations *in vivo*.

Summary and perspectives

The studies published on this Research Topic discovered meaningful results and offered new insights into corresponding biomedical research. As we all know that the cost of sequencing technologies is gradually decreasing, which can facilitate the conduction of multi-omics investigations. Bulk and single-cell protocols have their own advantages and limitations. Compared to single-cell sequencing methods, bulk approaches do not need living cells and the experimental procedures are usually simpler (Li et al., 2021). Dissecting large-scale samples is more affordable for bulk strategies, but bulk data can not effectively provide cellular heterogeneity information. Single-cell sequencing allows a better understanding of cell-to-cell variations and molecular dynamics at single-cell resolution. However, existing single-cell technologies for generating different types of omics data still suffer lower capture efficiency and higher technical noise compared to traditional bulk protocols (Mustachio and Roszik, 2022; Wen and Tang, 2022). Therefore, bulk and single-cell approaches are complementary, the combination of bulk and single-cell data is valuable for getting both cell-population and single-cell level perspectives (Li et al., 2021). For example, the proportion of cell subtypes for large-scale bulk data could be deconvoluted with the cell-type-specific signatures inferred from the single-cell data of a small number of samples (Aibar et al., 2017; Wang et al., 2019; Zaitsev et al., 2019; Decamps et al., 2020; Lin et al., 2022). The biomarkers identified in single-cell sequencing data can be further correlated to the outcomes of patients to assess their potential clinical value using corresponding bulk data from public databases such as The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013).

Collectively, joint analysis of bulk and single-cell multi-omics data can help us gain a more comprehensive and systematic view of biological and clinical samples. The innovation of various omics profiling technologies and related machine learning methods for integrating different types of data will further make multi-omics exploration more feasible and easier. We hope the studies published on this Research Topic will inspire related biomedical researchers to better understand the benefit and value of multi-omics strategies.

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 supported by the project of mRNA Innovation and Translation Center, Shanghai, China; and the project of Shanghai Strategic Emerging Industry Development Special Fund (ZJ640070216).

Conflict of interest

Author GC was employed by the company Stemirna Therapeutics Co., Ltd.

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

Publisher's note

References

Aibar, S., Gonzalez-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., et al. (2017). Scenic: single-cell regulatory network inference and clustering. *Nat. Methods* 14, 1083–1086. doi:10.1038/nmeth.4463

Decamps, C., Privé, F., Bacher, R., Jost, D., Waguet, A., Houseman, E. A., et al. (2020). Guidelines for cell-type heterogeneity quantification based on a comparative analysis of reference-free DNA methylation deconvolution software. *BMC Bioinforma.* 21, 16. doi:10.1186/s12859-019-3307-2

Haghverdi, L., Lun, A. T. L., Morgan, M. D., and Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* 36, 421–427. doi:10.1038/nbt.4091

Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biol.* 18, 83. doi:10.1186/s13059-017-1215-1

Jiang, Y., Wang, J., Sun, M., Zuo, D., Wang, H., Shen, J., et al. (2022). Multi-omics analysis identifies osteosarcoma subtypes with distinct prognosis indicating stratified treatment. *Nat. Commun.* 13, 7207. doi:10.1038/s41467-022-34689-5

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28, 882–883. doi:10.1093/bioinformatics/bts034

Lei, Y., Tang, R., Xu, J., Wang, W., Zhang, B., Liu, J., et al. (2021). Applications of single-cell sequencing in cancer research: progress and perspectives. *J. Hematol. Oncol.* 14, 91. doi:10.1186/s13045-021-01105-2

Leng, D., Zheng, L., Wen, Y., Zhang, Y., Wu, L., Wang, J., et al. (2022). A benchmark study of deep learning-based multi-omics data fusion methods for cancer. *Genome Biol.* 23, 171. doi:10.1186/s13059-022-02739-2

Li, Y., Ma, L., Wu, D., and Chen, G. (2021). Advances in bulk and single-cell multi-omics approaches for systems biology and precision medicine. *Brief. Bioinform.* 22, bbab024. doi:10.1093/bib/bbab024

Lin, Y., Li, H., Xiao, X., Zhang, L., Wang, K., Zhao, J., et al. (2022). DAISM-DNN(XMBD): Highly accurate cell type proportion estimation with *in silico* data augmentation and deep neural networks. *Patterns (N Y)* 3, 100440. doi:10.1016/j.patter.2022.100440

Mustachio, L. M., and Roszik, J. (2022). Single-cell sequencing: Current applications in precision onco-genomics and cancer Therapeutics. *Cancers (Basel)* 14, 3433. doi:10.3390/cancers12113433

Stuart, T., and Satija, R. (2019). Integrative single-cell analysis. *Nat. Rev. Genet.* 20, 257–272. doi:10.1038/s41576-019-0093-7

Wang, X., Park, J., Susztak, K., Zhang, N. R., and Li, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* 10, 380. doi:10.1038/s41467-018-08023-x

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., et al. (2013). The cancer Genome Atlas pan-cancer analysis project. *Nat. Genet.* 45, 1113–1120. doi:10.1038/ng.2764

Wen, L., and Tang, F. (2022). Recent advances in single-cell sequencing technologies. *Precis. Clin. Med.* 5, pbac002. doi:10.1093/pcmedi/pbac002

Zaitsev, K., Bambouskova, M., Swain, A., and Artyomov, M. N. (2019). Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures. *Nat. Commun.* 10, 2209. doi:10.1038/s41467-019-09990-5

Zhang, Y., Parmigiani, G., and Johnson, W. E. (2020). ComBat-seq: batch effect adjustment for RNA-seq count data. *Nar. Genom Bioinform.* 2, lqaa078. doi:10.1093/nargab/lqaa078

frontiers in Genetics

ORIGINAL RESEARCH
published: 27 January 2022
doi: 10.3389/fgene.2022.833797

Integration of Single-Cell RNA Sequencing and Bulk RNA Sequencing Data to Establish and Validate a Prognostic Model for Patients With Lung Adenocarcinoma

Aimin Jiang, Jingjing Wang, Na Liu, Xiaoqiang Zheng, Yimeng Li, Yuyan Ma, Haoran Zheng, Xue Chen, Chaoxin Fan, Rui Zhang, Xiao Fu and Yu Yao**

Department of Medical Oncology, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China

OPEN ACCESS

Edited by:
Geng Chen,
GeneCast Biotechnology Co., Ltd., China

Reviewed by:
Ming Yi,
Huazhong University of Science and Technology, China
Chunhou Zheng,
Anhui University, China

***Correspondence:**
Xiao Fu
15829793085@126.com
Yu Yao
13572101611@163.com

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

Received: *12 December 2021*
Accepted: *14 January 2022*
Published: *27 January 2022*

Citation:
Jiang A, Wang J, Liu N, Zheng X, Li Y, Ma Y, Zheng H, Chen X, Fan C, Zhang R, Fu X and Yao Y (2022) Integration of Single-Cell RNA Sequencing and Bulk RNA Sequencing Data to Establish and Validate a Prognostic Model for Patients With Lung Adenocarcinoma. Front. Genet. 13:833797. doi: 10.3389/fgene.2022.833797

Background: Lung adenocarcinoma (LUAD) remains a lethal disease worldwide, with numerous studies exploring its potential prognostic markers using traditional RNA sequencing (RNA-*seq*) data. However, it cannot detect the exact cellular and molecular changes in tumor cells. This study aimed to construct a prognostic model for LUAD using single-cell RNA-*seq* (scRNA-*seq*) and traditional RNA-*seq* data.

Methods: Bulk RNA-*seq* data were downloaded from The Cancer Genome Atlas (TCGA) database. LUAD scRNA-*seq* data were acquired from Gene Expression Omnibus (GEO) database. The uniform manifold approximation and projection (UMAP) was used for dimensionality reduction and cluster identification. Weighted Gene Correlation Network Analysis (WGCNA) was utilized to identify key modules and differentially expressed genes (DEGs). The non-negative Matrix Factorization (NMF) algorithm was used to identify different subtypes based on DEGs. The Cox regression analysis was used to develop the prognostic model. The characteristics of mutation landscape, immune status, and immune checkpoint inhibitors (ICIs) related genes between different risk groups were also investigated.

Results: scRNA-*seq* data of four samples were integrated to identify 13 clusters and 9cell types. After applying differential analysis, NK cells, bladder epithelial cells, and bronchial epithelial cells were identified as significant cell types. Overall, 329 DEGs were selected for prognostic model construction through differential analysis and WGCNA. Besides, NMF identified two clusters based on DEGs in the TCGA cohort, with distinct prognosis and immune characteristics being observed. We developed a prognostic model based on the expression levels of six DEGs. A higher risk score was significantly correlated with poor survival outcomes but was associated with a more frequent *TP53* mutation rate, higher tumor mutation burden (TMB), and up-regulation of *PD-L1*. Two independent external validation cohorts were also adopted to verify our results, with consistent results observed in them.

Conclusion: This study constructed and validated a prognostic model for LUAD by integrating 10× scRNA-*seq* and bulk RNA-*seq* data. Besides, we observed two distinct subtypes in this population, with different prognosis and immune characteristics.

Keywords: ScRNA-seq, prognosis, prognostic model, NMF, lung adenocarcinoma

INTRODUCTION

Lung cancer is one of the most common incident cancers and the leading cause of cancer-related death worldwide (Chen et al., 2016). As the most predominant pathological subtype, lung adenocarcinoma (LUAD) makes up more than 40% of lung cancer cases (Travis et al., 2015; Neal et al., 2019). Although promizing progress has been made in the screening, diagnosis, and management of LUAD patients in recent decades, it remains a lethal disease because a significant fraction of patients is diagnosed at the advanced disease stage (Denisenko et al., 2018; Lurienne et al., 2020). It is reported that more than 60% of newly diagnosed patients present locoregional or distant metastases at the time of detection (Brozos-Vázquez et al., 2021), with overall survival (OS) less than 5 years (Denisenko et al., 2018). With the rapid development of cancer genomics in recent decades, more and more gene alteration has been identified as an effective treatment target for LUAD. The majority of LUAD patients with driver gene mutation can benefit from molecular targeted therapy, such as epidermal growth factor receptor (*EGFR*)- tyrosine kinase inhibitors (TKIs), anaplastic lymphoma kinase (*ALK*)-TKIs (Yi et al., 2021a), and recently *KRAS* (Uras et al., 2020) and c-*MET* (Zhang et al., 2018) inhibitors. However, there is still part of patients who cannot get rid of the fate of resistance to these drugs due to secondary mutation in tumors. Recently, immune checkpoint inhibitors (ICIs) that target cytotoxic T lymphocyte-associated protein 4 (*CTLA4*), programmed death 1 (*PD1*), and programmed death-ligand 1 (*PD-L1*) have shown promising effects in various malignancies, including LUAD (Chen Y. et al., 2021; Huang et al., 2021). Unfortunately, not all patients can benefit from ICIs intervention, with a lower overall response rate observed in clinical practice. Therefore, there is an urgent need to identify potential prognostic and predictive biomarkers that could precisely stratify patients and recognize patients who will respond to treatment.

In recent decades, a growing body of studies explored potential prognostic markers of LUAD using traditional RNA sequencing (RNA-*seq*) data and have improved our understanding of tumor occurrence and development (Chen et al., 2020). For instance, Yi et al. developed a prognostic model to predict LUAD patients' survival and response to immunotherapy based on 17 immune-related genes (Yi et al.). Liang et al. also constructed a prognostic model for these patients based on seven ferroptosis-related genes (Liang et al.). Besides, our previous study also identified an autophagy-related long non-coding RNA signature as a prognostic biomarker for LUAD patients (Jiang et al., 2021). Despite the promising predictive power has been observed in the above studies, these prognostic signatures are based on traditional RNA-*seq*, which cannot detect the exact cellular and molecular changes in tumor cells because it mainly concentrates on the "average" expression of all cells in a sample (Chen et al., 2020).

Recently, single-cell RNA-*seq* (scRNA-*seq*) has been used to investigate the transcriptome of different cell types as an innovative technology (Chen et al., 2020). It uses optimized next-generation sequencing technologies to define the global gene expression profiles of single cells, thus facilitating dissection of the previously hidden heterogeneity in cell populations (Liang et al., 2021). Given this advantage, numerous studies have focused on identifying novel biomarkers for malignancies by integrating scRNA-*seq* and traditional RNA-*seq* (Zhang et al., 2019; Chen et al., 2020; Liang et al., 2021). This study aimed to construct a prognostic model for patients with LUAD by integrating scRNA-*seq* and traditional RNA-*seq* data, with two external validation cohorts being adopted to verify its risk stratification ability. Besides, we also identified two different population subtypes using non-negative matrix factorization (NMF), with distinct prognosis and immune characteristics observed. We believe our findings will provide potential prognostic biomarkers and therapeutic targets for LUAD.

MATERIALS AND METHODS

Raw Data Acquisition

10× scRNA-*seq* data of two LUAD samples (T1 and T2) and two normal samples (N1 and N2) were downloaded from the GSE149655 series, which included 2,642 cells, 3,203 cells, 4,243 cells, and 2,466 cells for each sample. LUAD bulk RNA-*seq* data, mutation data, and clinicopathological characteristics were downloaded from the TCGA database. Besides, we also downloaded progression-free survival (PFS) records of these patients from UCSC Xena (https://xena.ucsc.edu/). The human. gtf file was adopted to raw matrix annotation. Furthermore, GSE31210 and GSE13213 cohorts were also acquired from the Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/) database to serve as independent external cohorts for risk model validation. The detailed clinical characteristics of patients in the TCGA and GEO cohorts are summarized in **Supplementary Table S1**.

scRNA-*Seq* Data Processing and Analysis

The 10× scRNA-*seq* data were processed according to the following steps: 1) R software, "Seurat" package (Macosko et al., 2015) was adopted to convert 10× scRNA-*seq* data as a Seurat object; 2) quality control (QC) of the raw counts by calculating the percentage of mitochondrial or ribosomal genes and excluding low-quality cells; 3) the "FindVariableFeatures" function was adopted to filter the top 2000 highly variable genes

after QC; 4) principal component analysis (PCA) was performed based on the 2000 genes, and uniform manifold approximation and projection (UMAP) (Becht et al., 2018) was used for dimensionality reduction and cluster identification; 5) the "Find All Markers" function was exploited to identify significant marker genes for different clusters by setting $\log_2$ [Foldchange (FC)] as 0.3 and min.pct as 0.25; and 6) R software, "SingleR" package (Aran et al., 2019) was applied to cluster annotation to recognize different cell types. Next, we performed Fisher's exact test to identify potential significant cell types between tumor and normal samples. We calculated the FC value of each cell type in tumor and normal samples and determined the cell types with FC> 4 or FC <0.25, p-value < 0.05 as the key cell types. Furthermore, we performed functional enrichment analysis for the identified hub cell types using R software, "ReactomeGSA" package (Griss et al., 2020). We used the "analyze_sc_clusters" function for enrichment analysis and extracted the results through the "pathways" function. R software, "monocle" package (Borcherding et al., 2019) was adopted to cell trajectory and pseudo-time analysis, with the method "DDRTree" being used for dimensionality reduction. Subsequently, the statistical method "BEAM" was used to calculate the contribution of genes during cell development, and the top 100 genes were selected for visualization. Ultimately, R software, "CellChat" (Jin S. et al., 2021) and "patchwork" packages were adopted for cell-cell communication analysis and network visualization.

Differentially Expressed Genes Identification and Functional Enrichment Analysis

Differential expression analysis was performed to filter differentially expressed genes (DEGs) in the TCGA cohort by using the R software, "limma" package, with $|\log_2 FC| > 1.0$ and false discovery rate (FDR) < 0.05 being used as cut-off value. The volcano plot was generated to visualize the distribution of the identified DEGs. Subsequently, Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) analyses were exploited to investigate the most significantly enriched pathways and biological processes of the DEGs using R software, "clusterProfiler" package.

Weighted Gene Correlation Network Analysis

Weighted Gene Correlation Network Analysis (WGCNA) was utilized to filter hub genes in DEGs via R software, "WGCNA" package. WGCNA is divided into expression cluster and phenotypic correlation analyses (Langfelder and Horvath, 2008). It mainly includes four steps: calculation of correlation coefficient between genes, determination of gene modules, co-expression network, and correlation between modules and traits (Langfelder and Horvath, 2008). In the process of co-expression network construction, soft thresholding power β was selected as the lowest power with which fit index of scale-free topology reached 0.90. The modules were presented together via dendrogram after the process of clustering. Subsequently, the module-trait heatmap was generated to further identify the most significant DEGs in LUAD development by comparing their correlation coefficients and p values. Ultimately, we selected the intersection genes among the marker genes and DEGs found in WGCNA for further analysis.

Sample Clustering Using Non-Negative Matrix Factorization Algorithm

Non-negative matrix factorization (NMF) was carried to divide patients into different subtypes according to the following steps: 1) the univariate Cox regression analysis was performed to identify potential prognostic DEGs *via* R software, "survival" package; 2) sample clustering through "brunet" method in R software, "NMF" package; 3) according to parameters such as cophenetic, dispersion, and silhouette, the optimal number of the cluster was identified to classify patients into different subtypes; and 4) the consensus heatmap was generated in accordance with the above optimal cluster number to view the distribution characteristic among different subtypes. Then, we also explored the relationship between different clusters and OS and PFS. Besides, the MCPcounter algorithm was adopted to estimate the infiltration of the immune cells between different clusters. We also investigated the association between clusters and six immune subtypes identified in a previously published study (Tamborero et al., 2018).

Prognostic Model Construction and Validation

First, the univariate Cox regression analysis was performed to identify potential prognostic DEGs. Variables with a p-value < 0.01 were selected into the Least Absolute Shrinkage and Selection Operator (LASSO) regression analysis to reduce the number of genes in the final risk model through R software, "glmnet" package. Ultimately, genes in the LASSO regression were selected into the multivariate Cox regression analysis and therefore constructed the prognostic model according to the following formula:

$$\text{risk score} = \sum_{i=1}^{k} \beta i * exp i \quad (1)$$

In the formula, "βi" represents the coefficient of the selected genes in the multivariate Cox analysis and "expi" refers to its expression value. All patients were divided into high- and low-risk groups according to the median value of risk score. Survival curves and risk plots were generated to visualize the survival difference and status for each patient via R software, "survminer" and "ggrisk" packages. Besides, we used R software, "timeROC" package to draw the receiver operating characteristic (ROC) curves to evaluate the performance of risk score in predicting 1-, 3-, and 5 years OS of LUAD patients. Additionally, GSE31210 and GSE13213 cohorts were used as independent external cohorts to validate the utility of the prognostic model.

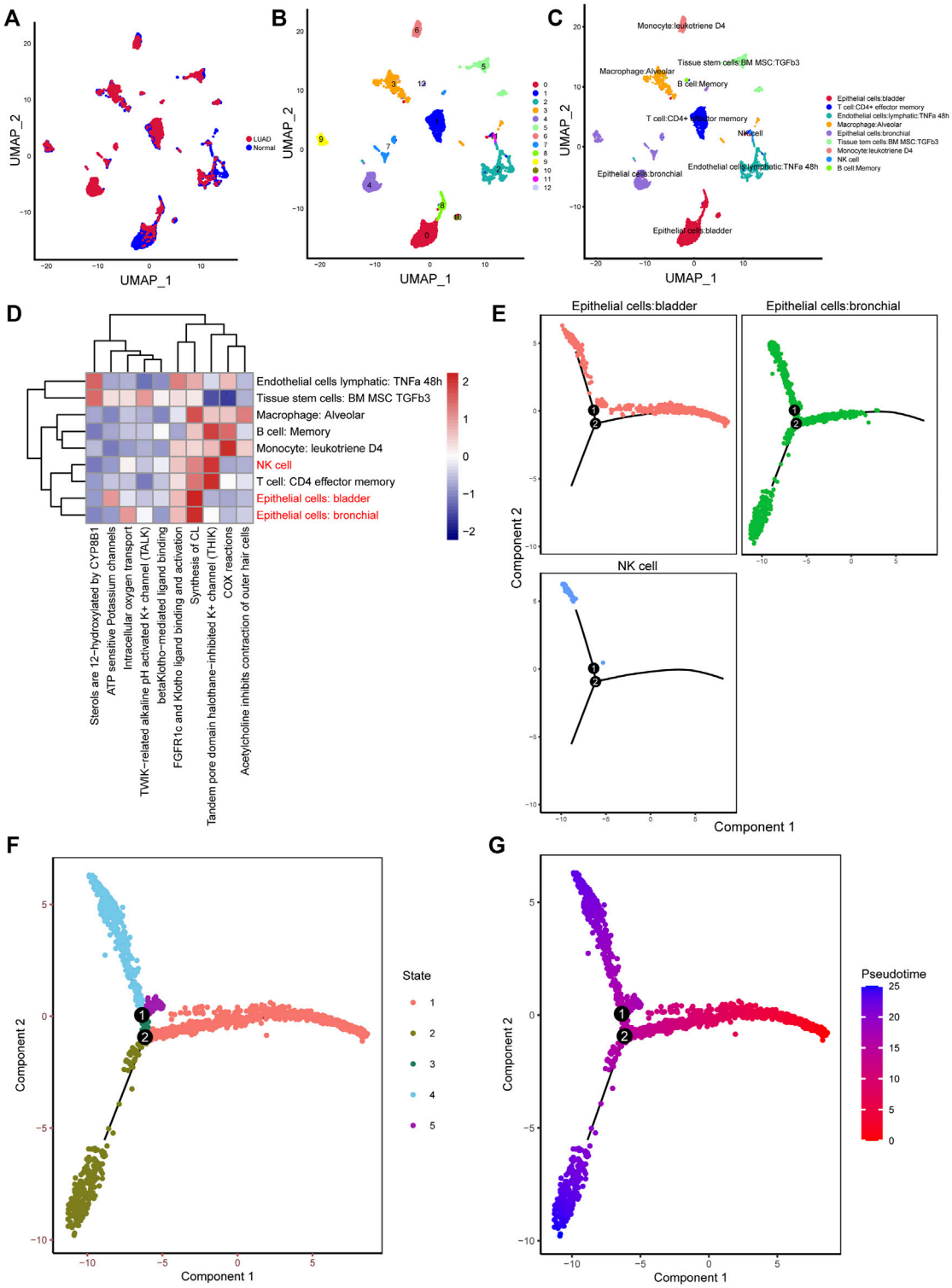


FIGURE 1 | Different clusters annotation and cell types identification in LUAD 10× scRNA-*seq* data. **(A–C)** Clusters annotation and cell types identification via UMAP; **(D)** Functional enrichment analysis for the identified hub cell types using "ReactomeGSA" package; **(E–G)** Cell trajectory and pseudo-time analysis for the identified hub cell types. LUAD, lung adenocarcinoma; scRNA-*seq*, single-cell RNA sequencing; UMAP, uniform manifold approximation and projection.

Clinical Relevance, Mutation Landscape, and Enrichment Analysis Between High- and Low-Risk Groups

Next, we investigated the association between the risk score and clinicopathological characteristics of patients in the TCGA cohort. Furthermore, we adopted Cox regression analysis to determine whether the risk score could be an independent prognostic factor for LUAD patients via R software, "survcomp" package. At the same time, R software, "forestplot" package was used to draw forest plots of the univariate and multivariate Cox regression analyses. Gene set enrichment analysis (GSEA) was then performed to identify the most significantly enriched pathways between high- and low-risk groups through R software, "org.Hs.eg.db," "clusterProfiler," and "enrichplot" packages. In addition, two waterfall plots were generated to explore the detailed gene mutation characteristics between high- and low-risk groups *via* "oncoplot" function in R software, "maftools" package.

Immune Cells Infiltration and Immune Function Status Between High- and Low-Risk Groups

Then, single-sample gene set enrichment analysis (ssGSEA) (Rooney et al., 2015) was adopted to estimate the infiltrating score of immune cells and the activity of immune-related pathways using R software, "GSVA" and "GSEABase" packages. The Wilcoxon rank-sum test was used to compare the statistical difference between high- and low-risk groups. Besides, we also investigated the correlation between risk score and immune checkpoint inhibitors (ICIs) related genes expression levels and tumor mutation burden (TMB), with R software, "ggplot2" package being adopted for visualization.

Statistical Analysis

The non-parameter Wilcoxon rank-sum test was used to examine the relationship of continuous variables between the two groups. The LASSO regression and Cox regression analyses were used for predictive model development. Kaplan-Meier survival analysis was used to test the survival difference between different risk groups. A log-rank test was adopted to examine the statistical difference. A two-sided p-value < 0.05 was considered significant. All analyses were conducted in R software (version 4.1.1) for windows 64.0.

RESULTS

scRNA-*Seq* and Cell Typing of Normal and Lung Adenocarcinoma Lung Samples

10× scRNA-*seq* data of two LUAD and two normal samples were downloaded from the GSE149655 dataset. A total of 8,170 cells were identified after QC, as shown in **Supplementary Figure S1A**. We visualized the top 20 highly variable genes in **Supplementary Figure S1B**. Thirteen distinct clusters were identified after PCA and UMAP analysis (**Figures 1A,B**). Then "SingleR" package was adopted to cluster annotation, with UMAP being used to visualize the cell types after dimensionality reduction. Overall, we identified nine cell types in this step, including bladder epithelial cells, $CD4^+$ effector memory T cell, lymphatic endothelial cells, alveolar macrophage, bronchial epithelial cells, tissue stem cells, monocyte, NK cells, and memory B cell (**Figure 1C**). Of these, NK cells, bladder epithelial cells, and bronchial epithelial cells were identified as significant cell types. ReactomeGSA functional enrichment analysis suggested that these cell types mainly are involved in intracellular oxygen transport, FGFR1c and Klotho ligand binding and activation, and synthesis of cardiolipin (CL) (**Figure 1D**). Then, "monocle" package was exploited to analyze the cell trajectory and pseudo-time of the identified three significant cell types. We observed that NK cell only corresponds to state 4, while bronchial epithelial cells occurred in the whole state (**Figures 1E–G**). We then calculated the contribution of genes during cell development, and the top 100 genes were selected for visualization (**Supplementary Figure S2A**). We investigated the cell-cell communication network by calculating communication probability (**Supplementary Figure S2B**). Furthermore, we inferred the cell-cell communication network based on specific pathways and ligand-receptors. We identified that SEMA4D—PLXNB2 (**Figure 2A**), HLA–DPA1—CD4 (**Figure 2B**), and C3—C3AR1 (**Figure 2C**) play crucial roles in the communication network.

Identification of Differentially Expressed Genes in Bulk RNA-*Seq* Data

A total of 1971 genes were identified as DEGs after differential expression analysis (**Figure 2D**). Of these, 902 were up-regulated genes, while 1,069 were down-regulated (**Figure 2D**). GO analysis revealed that the DEGs were mainly enriched in the biological processes of the humoral immune response, complement activation, and protein activation (**Figure 2E**). KEGG analysis indicated that the DEGs were mainly enriched in cell adhesion molecules, cell cycle, and complement and coagulation cascades (**Figure 2F**). Next, we performed WGCNA to identify DEGs involved in LUAD development and progression. In the process of co-expression network construction, we observed that the soft thresholding power β was 5 when the fit index of scale-free topology reached 0.90 (**Figure 3A**). Nine modules were identified based on the average linkage hierarchical clustering and the soft thresholding power (**Figure 3B**). We observed that the turquoise module was significantly correlated with LUAD development according to the correlation coefficient and p-value (**Figure 3C**). Ultimately, 329 common genes, which are both marker genes and WGCNA module genes, were selected to construct an expression matrix for further analysis.

Different Molecular Subtypes Identification

All patients were divided into two clusters according to relevant parameters after NMF (**Figure 4A**; **Supplementary Figure S3**). It showed that patients in cluster 2 were correlated with poor OS and PFS than patients in cluster 1 (**Figure 4B**). The MCPcounter

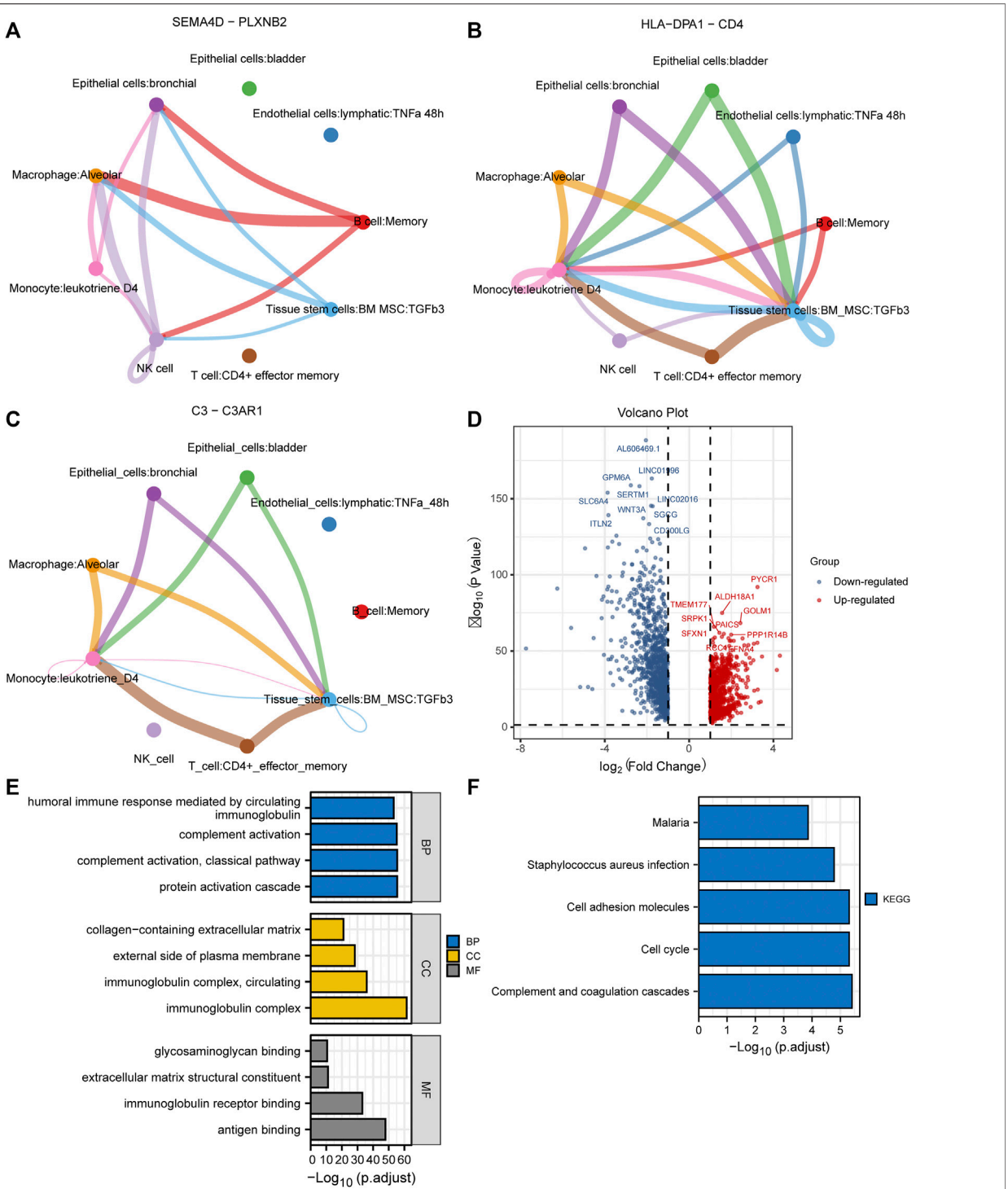


FIGURE 2 | Cell-cell communication network and identification of DEGs in TCGA cohort. **(A–C)** Cell-cell communication network identified that SEMA4D—PLXNB2, HLA–DPA1—CD4, and C3—C3AR1 play crucial roles in the communication network; **(D)** The volcano plot to show the up-regulated and down-regulated DEGs in TCGA cohort; **(E,F)** GO and KEGG enrichment analysis of the identified DEGs. DEGs, differentially expressed genes; TCGA, The Cancer Genome Atlas; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes.

algorithm was used to estimate the infiltration of the immune cells in different clusters. We found that the infiltration levels of endothelial cells, myeloid dendritic cells, and neutrophils were significantly higher in cluster 1 (**Figure 4C**). However, cluster 2 had higher infiltration levels of B lineage, cytotoxic lymphocytes, fibroblasts, and NK cells (**Figure 4C**). Besides, the Sankey plot

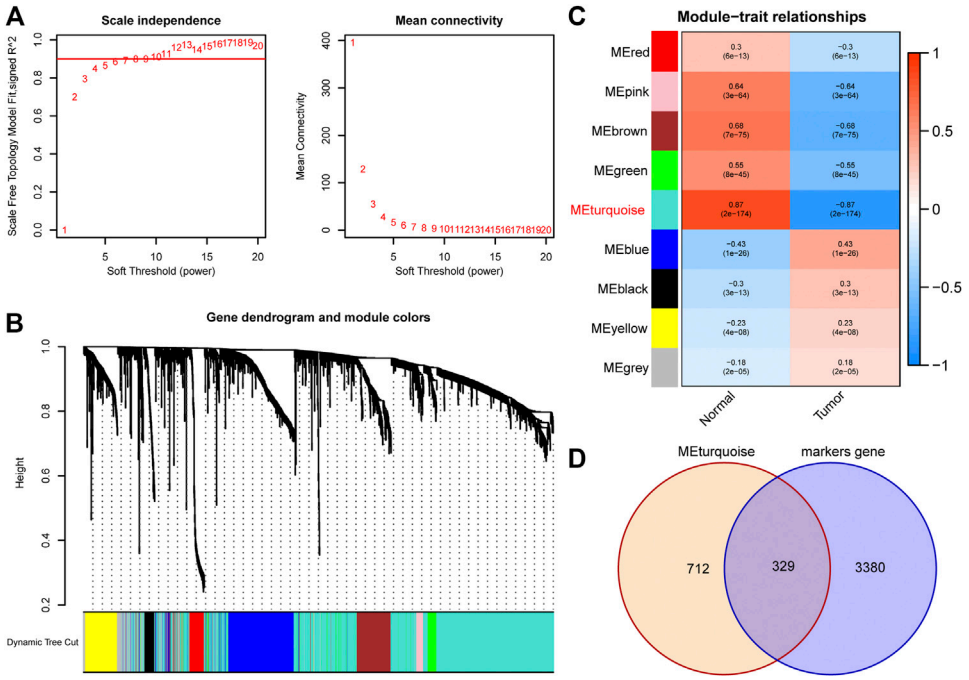


FIGURE 3 | Identification of hub DEGs that participate in LUAD development through WGCNA. **(A)** The scale-free fit index for soft thresholding powers. The soft thresholding power β in the WGCNA was determined based on a scale-free R^2 (R^2 = 0.90). The left panel illustrates the relationship between β and R^2. The right panel illustrates the relationship between β and mean connectivity. **(B)** A dendrogram of the DEGs clustered based on different metrics. **(C)** A heatmap illustrates the correlation between different gene modules and clinical traits (normal vs. tumor); **(D)** The Venn plot to identify common DEGs between WGCNA module genes and marker genes. DEGs, differentially expressed genes; LUAD, lung adenocarcinoma; WGCNA, Weighted Gene Correlation Network Analysis.

was also applied to investigate the relationship between different immune subtypes and clusters. It showed that patients in cluster 1 are mainly classified into Immune C3 (inflammatory) subtype (**Figure 4D**). However, patients in cluster 2 are mainly classified into Immune C1 (wound healing), Immune C2 (IFN-gamma dominant), and Immune C6 (TGF-beta dominant) subtypes (**Figure 4D**).

Prognostic Model Construction and Validation

We performed univariate Cox regression analysis to identify potential prognostic DEGs for LUAD in the TCGA cohort. Seven genes were identified as prognostic DEGs. Then, LASSO regression analysis was performed to reduce the number of DEGs in the final risk model, with six genes were identified through this step (**Figure 5A**). Ultimately, six genes were recognized as independent prognostic DEGs via multivariate Cox analysis, including *CP*, *GOLM1*, *CYP4B1*, *DAPK2*, *NFIX*, and *FHL2*. According to their coefficients, we calculated the risk score according to the following formula: risk score= expression level of *CP* * 0.088 + expression level of *GOLM1** 0.15 + expression level of *CYP4B1* * (−0.064) + expression level of *DAPK2* * (−0.082) + expression level of *NFIX* *(−0.059) + expression level of *FHL2* * 0.086. All patients were divided into high- and low-risk groups according to the median value of risk score. The survival curve showed that patients in the high-risk group were associated with the worse OS when compared with patients in the low-risk group (**Figure 5B**). Besides, it revealed that the risk score had good performance in predicting the OS in these individuals in the TCGA cohort (AUC for 1-, 3-, and 5 years OS: 0.669, 0.674, and 0.642; **Figure 5B**). Consistently, we observed similar results in the GSE31210 cohort and GSE13213 cohort (**Figures 5C,D**). The risk plots were generated to show detailed survival outcomes of each patient in the TCGA cohort and external validation cohorts (**Figures 5E–G**).

Clinical Relevance, Enrichment Analysis, and Mutation Landscape Between High- and Low-Risk Groups

Next, we investigated the relationship between the risk score and clinicopathological characteristics, suggesting that younger patients, males, current smokers, and positive lymph nodes status were correlated with higher risk scores (**Figure 5H**). We also performed single factor and multi-factor Cox analyses to determine whether the risk score could be an independent prognostic factor for LUAD patients compared with other common clinicopathological parameters. We observed that the risk score could serve as an independent prognostic factor for these individuals (**Figures 6A,B**). Furthermore, we performed GSEA analysis to identify the most significantly enriched pathways between the two groups. We found that genes in the

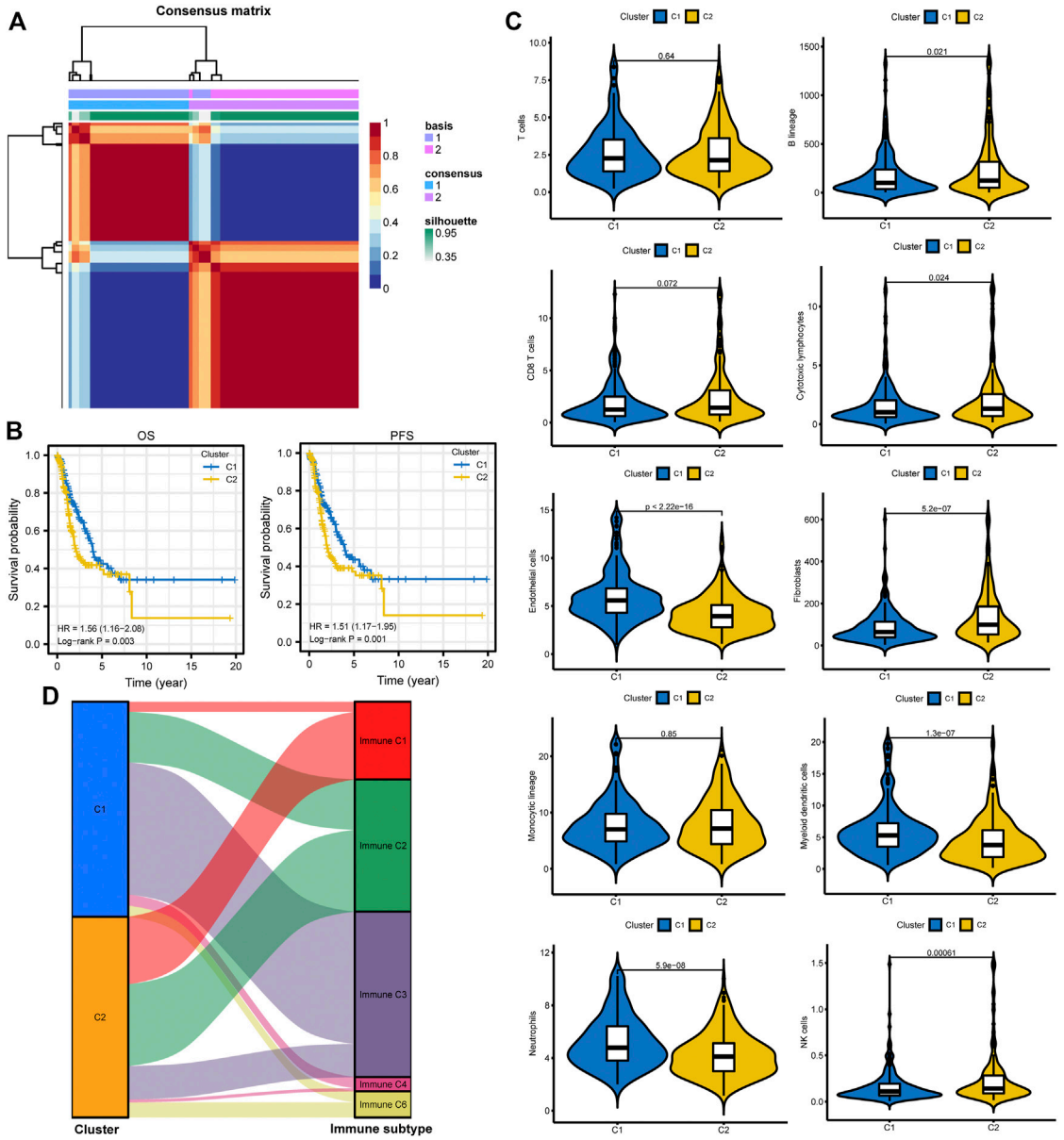


FIGURE 4 | Different subtype identification and clinical relevance analysis. **(A)** Two different subtypes were identified via the NMF algorithm. **(B,C)** The relationship between different subtypes and OS and PFS of LUAD. **(D)** TME composition between different subtypes. **(E)** Sankey plot to show the association between different subtypes and immune subtypes. NMF, non-negative Matrix Factorization; OS, overall survival; PFS, progression-free survival; LUAD, lung adenocarcinoma; TME, tumor microenvironment.

high-risk group significantly enriched in cell cycle and DNA replication (**Figure 6C**). However, genes in the low-risk group significantly enriched in arachidonic acid metabolism (**Figure 6D**). Afterward, we generated two waterfall plots to explore the detailed gene mutation characteristics between high- and low-risk groups. We identified that *TP53*, *TTN*, and *MUC16* were the most frequently mutated genes in high- and low-risk groups (**Figures 6E,F**). Besides, we also observed that the high-risk group harbored a more frequent *TP53* mutation rate than the low-risk group (**Figures 6E,F**).

The Immune Function Between High- and Low-Risk Groups

We then adopted ssGSEA to estimate the infiltrating score of immune cells and the activity of immune-related pathways in different risk groups. The results demonstrated that the infiltration levels of DCs, B cells, Mast cells, NK cells, T helper cells, and TIL were significantly different in the two groups (**Figure 7A**). Meanwhile, the two groups also had different scores of MHC class I, parainflammation, and Type II IFN response (**Figure 7A**). Subsequently, we investigated the

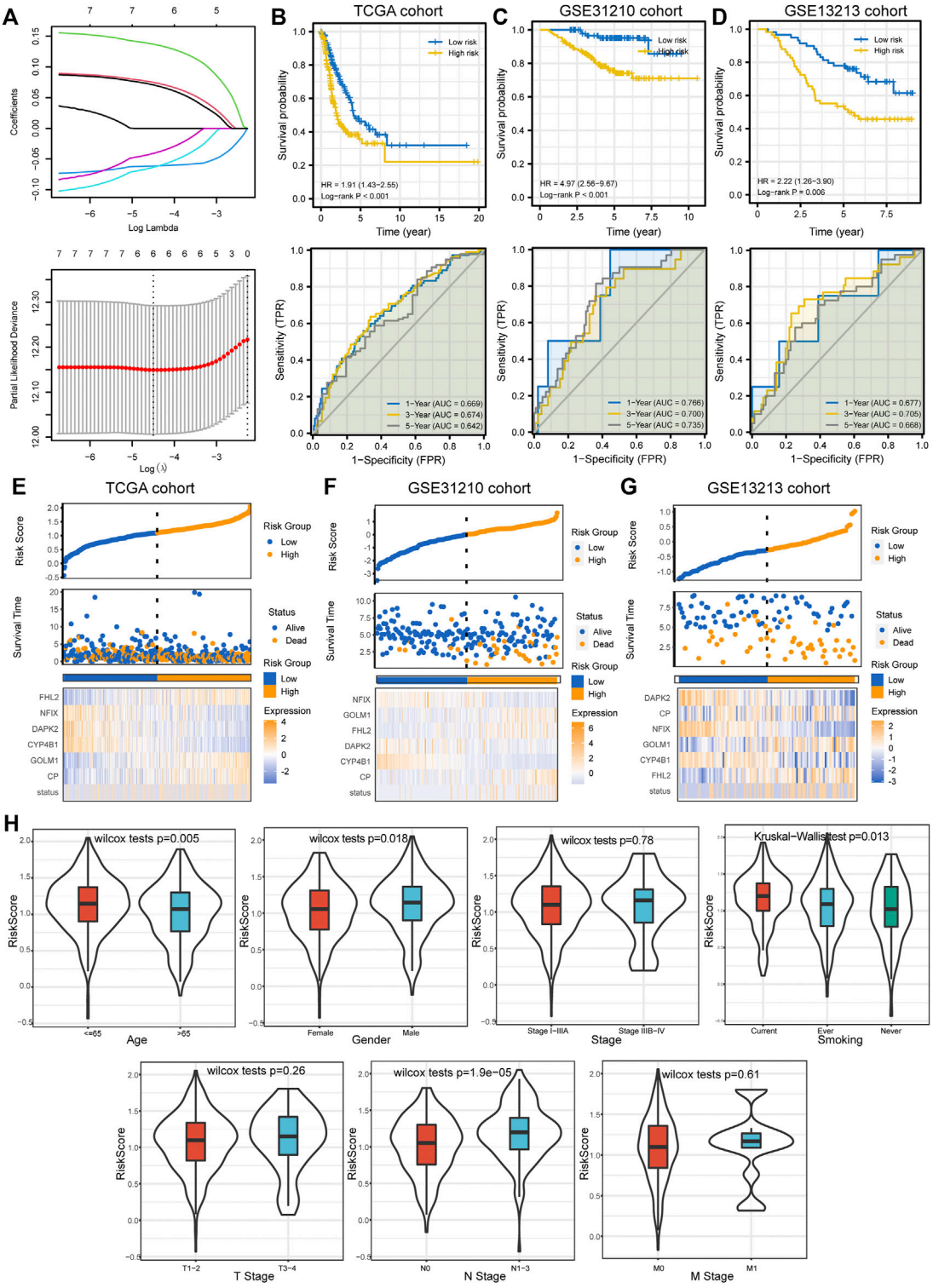


FIGURE 5 | Prognostic model establishment and validation for patients with LUAD. **(A)** Six DEGs were selected for multivariate analysis via LASSO regression analysis. **(B–D)** Survival curves and ROC curves evaluate the risk stratification ability and predictive ability of the constructed risk model in the TCGA**,** GSE31210, and GSE13213 cohorts. **(E–G)** Risk plots to illustrate the survival status of each sample in the TCGA**,** GSE31210, and GSE13213 cohorts. **(H)** The relationship between risk score and common clinicopathological characteristics of LUAD. LUAD, lung adenocarcinoma; DEGs, differentially expressed genes; LASSO, Least Absolute Shrinkage and Selection Operator; ROC, receiver operating characteristic curve; TCGA, The Cancer Genome Atlas.

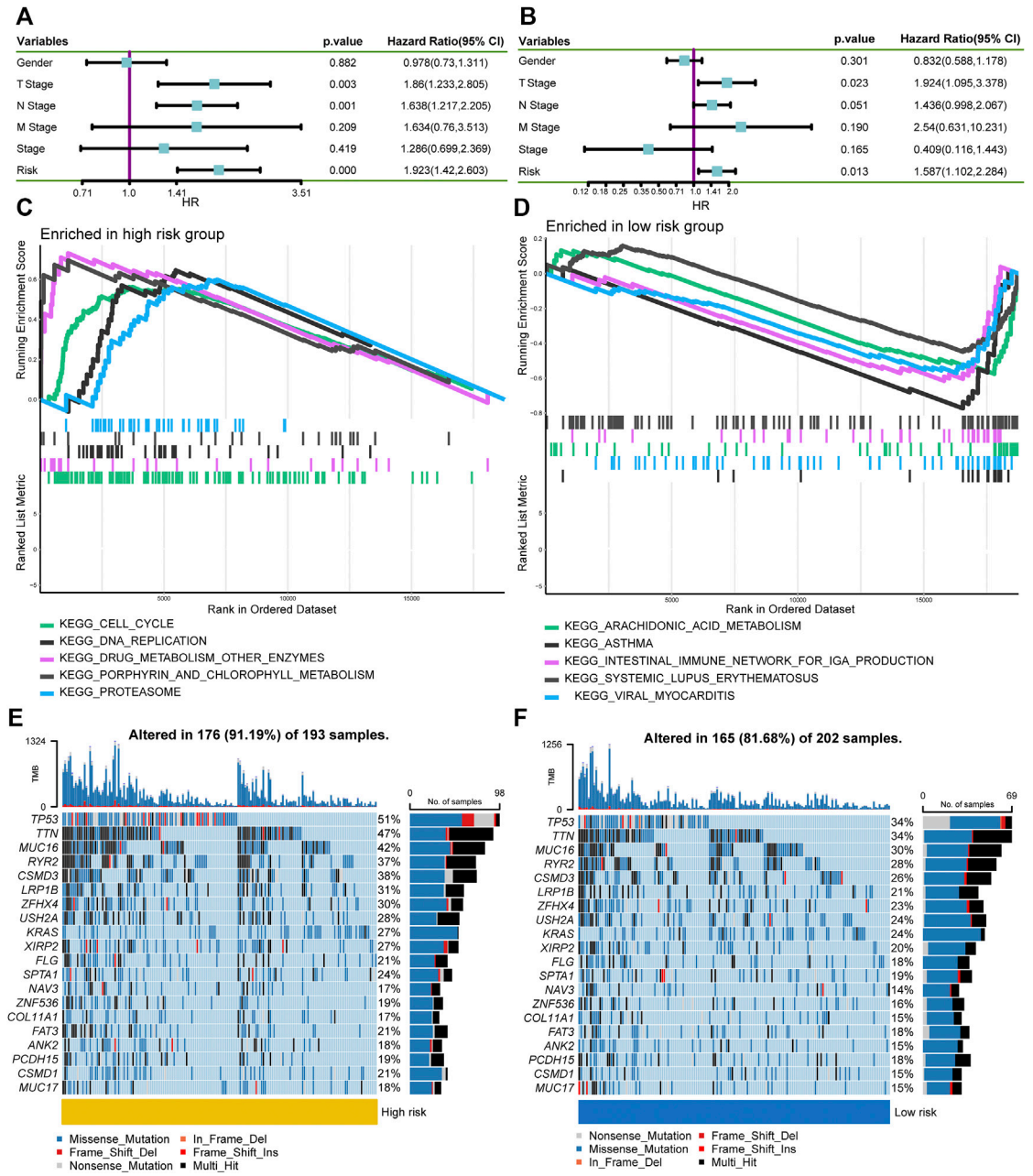


FIGURE 6 | Independent prognostic ability evaluation, pathway enrichment analysis, and gene mutation landscape analysis. **(A,B)** The univariate and multivariate Cox regression analysis demonstrates the risk score's independent prognostic ability. **(C,D)** GSEA to investigate the biological processes and pathways enriched in high- and low-risk groups. **(E,F)** Waterfall plots summarize the gene mutation landscape in high- and low-risk groups. GSEA, Gene Set Enrichment Analysis.

correlation between the risk score and the expression level of common ICIs related genes. The results revealed that a higher risk score was significantly associated with up-regulation of *CD274* (*PD-L1*) (**Figure 7B**). Nevertheless, there was no significant statistical difference between the risk score and *PDCD1* (**Figure 7C**), *CTLA4* (**Figure 7D**), *LAG3* (**Figure 7E**), and *TIGIT* (**Figure 7F**) expression. Besides, we also observed that a higher risk score was positively correlated with a higher TMB value (**Figure 7G**).

DISCUSSION

This study developed a prognostic model for LUAD patients by integrating 10× scRNA-*seq* and bulk RNA-*seq* data. We found that the constructed prognostic model can effectively stratify patients into high- and low-risk groups in the TCGA and GEO cohorts. Furthermore, we also explored the clinical relevance, mutation landscape, and tumor immune microenvironment (TME) in different groups. We noticed that

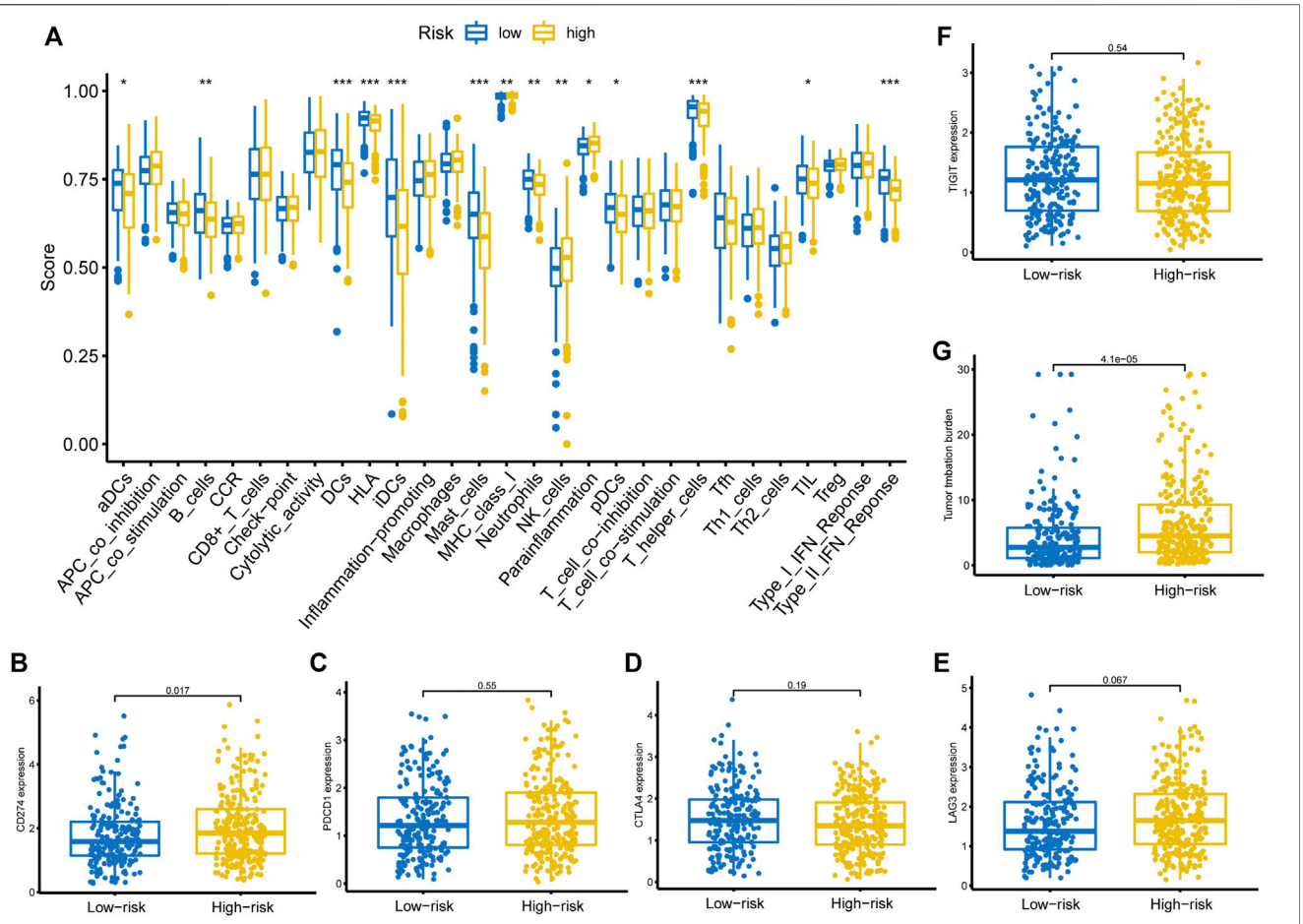


FIGURE 7 | Immune function, ICIs related genes expression pattern, and TMB between different risk groups. **(A)** Immune cells infiltration score and immune-related pathways activity in the low- and high-risk groups estimated by ssGSEA. **(B–F)** The correlation between the risk score and the expression level of *CD274*, *PDCD1*, *CTLA4*, *LAG3*, and *TIGIT*. **(G)** The relationship between the risk score and TMB. ICIs, immune checkpoint inhibitors; TMB, tumor mutation burden, ssGSEA, single-sample gene set enrichment analysis.

a higher risk score was significantly correlated with a more frequent *TP53* mutation rate, up-regulation of *PD-L1*, and higher TMB value. These results support that patients with higher risk scores could have potential clinical benefits from immunotherapy. Moreover, we identified two distinct subtypes using the NMF algorithm. We observed that different clusters have distinct prognoses and TME components. Cluster 2 was correlated with worse clinical outcomes and high infiltration levels of fibroblasts. Accumulating studies have shown that cancer-associated fibroblasts (CAFs) could transfer lipid to the TME to support cancer cell growth (Lopes-Coelho et al., 2018; Gong et al., 2020; Ma and Zhang, 2021). Recently, Gong et al. elucidated that reprogramming of lipid metabolism in CAFs potentiates migration of colorectal cancer cells through *in vivo* and *in vitro* experiments (Gong et al., 2020). Furthermore, we found that patients in cluster 2 are mainly classified into Immune C1, Immune C2, and Immune C6 subtypes, which are correlated with more aggressive immune infiltrates and worse prognosis (Tamborero et al., 2018; Zhang et al., 2020). On the contrary, patients in cluster 1 are mainly classified into the Immune C3 subtype, associated with a more favorable immune composition and better clinical outcomes (Tamborero et al., 2018; Zhang et al., 2020).

We identified six hub genes to develop the prognostic model through LASSO and Cox regression analyses, including *CP*, *GOLM1*, *CYP4B1*, *DAPK2*, *NFIX*, and *FHL2*. Ceruloplasmin (*CP*) is a multicopper ferroxidase that mainly utilizes the redox activity of copper to oxidize ferrous iron, facilitating iron efflux *via FPN1* (Chen F. et al., 2021). A previous study reported that *CP* is up-regulated in LUAD samples and correlated with poor clinical stage and survival outcome in these patients (Matsuoka et al., 2018). *GOLM1* belongs to the Golgi-associated protein and is a crucial promoter of liver cancer growth and metastasis (Mao et al., 2010). Numerous studies indicated that *GOLM1* is up-regulated in LUAD and can serve as an unfavorable prognostic factor (Liu et al., 2018; Yang et al., 2018; Zhao M. et al., 2021; Song et al., 2021). Song et al. reported that overexpression *GOLM1* enhances lung cancer aggressiveness via inhibiting the formation of *P53* tetramer (Song et al., 2021). Although *GOLM1* has been previously regarded as a diagnostic marker of liver

cancer, it is an independent prognostic factor for liver cancer (Mao et al., 2010). In a recent study, Ye et al. revealed that *GOLM1* could drive hepatocellular carcinoma metastasis by modulating EGFR /growth-factor-responsive receptor tyrosine kinase (RTK) cell-surface recycling (Ye et al., 2016). *CYP4B1* is a drug-metabolizing enzyme gene. Several studies detected the mRNA expression level of *CYP4B1* in lung cancer samples and its corresponding paraneoplastic samples (Czerwinski et al., 1994; Tamaki et al., 2011). Tamaki et al. indicated that *CYP4B1* polymorphism is not correlated with lung cancer risk. Therefore, further studies need to be performed to evaluate the expression level of *CYP4B1* in LUAD and its prognostic significance. Death-associated protein kinase (*DAPK*) is the Ser/Thr kinases family member. It has been reported that *DAPK* family proteins play vital roles in mediating apoptosis and function as tumor suppressors in various malignancies (Chen et al., 2014; Jin M. et al., 2021). Interestingly, Jin et al. elucidated that cigarette smoking induces aberrant N6-methyladenosine of *DAPK2* to promote lung cancer progression by activating NF-κB pathway (Jin M. et al., 2021). Nuclear factor IX (*NFIX*) serves as a master regulator, and its expression is associated with 17 genes involved in the migration and invasion pathways, including interleukin-6 receptor subunit β (*IL6ST*), metalloproteinase inhibitor 1 (*TIMP1*), and integrin β-1 (*ITGB1*) (Rahman et al., 2017). In a recent study, Zhao et al. indicated that long non-coding RNA *SNHG3* promotes the development of lung cancer via the miR-1343-3p/*NFIX* pathway (Zhao L. et al., 2021). The four and a half LIM domains 2 (*FHL2*) is a multifunctional scaffolding protein regulating signaling cascades and gene transcription (Wang et al., 2020). Numerous studies have revealed that *FHL2* is an adverse prognostic factor of gynecological malignancies (Wang et al., 2020). However, no study reported the expression level and prognostic significance of *FHL2* in lung cancer.

Subsequently, all patients were divided into low- and high-risk groups by integrating the six hub genes. Two external validation cohorts were also used to verify its predictive ability, with consistent results were observed in these two cohorts. Besides, we identified that the constructed prognostic model has independent predictive ability in predicting the OS of LUAD patients. We then investigated the gene mutation landscape and immune function in different risk groups. We identified that the high-risk group harbored a more frequent *TP53* mutation rate than the low-risk group. Numerous studies identified that *TP53* mutation is closely correlated with treatment resistance and terminal prognosis in lung cancer (Steels et al., 2001; Viktorsson et al., 2005; Xu et al., 2020). However, many studies revealed that *TP53* mutation was significantly correlated with remarkable clinical benefit from PD-1 inhibitors for patients with LUAD since it increases TMB, up-regulates *PD-L1* expression, and remodels TME (Dong et al., 2017; Skoulidis and Heymach, 2019; Xu et al., 2020). Hence, we investigated the relationship between the risk score and TMB value and *PD-L1* expression level. Not surprisingly, it indicated that a higher risk score was significantly correlated with higher TMB value and *PD-L1* expression level. Recently, Yi et al. investigated the regulation of *PD-L1* expression in the TME, suggesting that the expression of *PD-L1* is regulated by numerous factors, including inflammatory stimuli and oncogenic pathways at the levels of transcription, post-transcription, and post-translation (Yi et al., 2021b). Besides, they indicated that a comprehensive framework containing multiple surrogate markers such as TMB would be valuable for selecting patients and predicting outcomes (Yi et al., 2021b). Taken together, patients with higher risk scores could have a potential survival benefit from immune checkpoint blockades treatment. The constructed prognostic model might be a potential predictive biomarker for patients who received immunotherapy. To our knowledge, this is the first study that constructed and validated a prognostic model for LUAD by integrating 10× scRNA-*seq* and bulk RNA-*seq* data. Besides, two external validation cohorts were also used to verify its performance in predicting the OS of these patients. Nevertheless, there are several inevitable shortcomings in our study. First, all these results were obtained from the bioinformatic analysis, and experimental validation needs to be performed in the future. Second, searching for effective prognostic and predictive biomarkers for patients with malignancy is an arduous task for us and needs a long way to go. Our study developed a novel biomarker and provided potential insights in this area. However, well-designed prospective studies are warranted in the future to address this issue.

CONCLUSION

This study constructed and validated a prognostic model for LUAD by integrating 10× scRNA-*seq* and bulk RNA-*seq* data. Besides, we identified two distinct subtypes in this population, with different prognosis and immune characteristics being observed in them. The higher risk score was correlated with poor survival outcomes but associated with a more frequent *TP53* mutation rate, higher TMB value, and up-regulation of *PD-L1*. Our prognostic model might be a potential biomarker for LUAD patients' risk stratification and treatment response prediction. Well-designed prospective studies are warranted in the future to verify our findings.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

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

with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

Conception/design: YY,XF, and AJ; Provision of study material: AJ, JW, NL, and YL; Collection and/or assembly of data: AJ, YM, HZ, and XZ; Data analysis and interpretation: AJ,XC, CF, and RZ; Manuscript writing: AJ; Final approval of manuscript: YY and XF. All authors read and approved the final manuscript and agree to be accountable for all aspects of the research in ensuring that the accuracy or integrity of any part of the work is appropriately investigated and resolved.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.833797/full#supplementary-material

REFERENCES

Aran, D., Looney, A. P., Liu, L., Wu, E., Fong, V., Hsu, A., et al. (2019). Reference-based Analysis of Lung Single-Cell Sequencing Reveals a Transitional Profibrotic Macrophage. *Nat. Immunol.* 20 (2), 163–172. doi:10.1038/s41590-018-0276-y

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., et al. (2018). Dimensionality Reduction for Visualizing Single-Cell Data Using UMAP. *Nat. Biotechnol.* 37, 38–44. doi:10.1038/nbt.4314

Borcherding, N., Voigt, A. P., Liu, V., Link, B. K., Zhang, W., and Jabbari, A. (2019). Single-Cell Profiling of Cutaneous T-Cell Lymphoma Reveals Underlying Heterogeneity Associated with Disease Progression. *Clin. Cancer Res.* 25 (10), 2996–3005. doi:10.1158/1078-0432.Ccr-18-3309

Brozos-Vázquez, E. M., Díaz-Peña, R., García-González, J., León-Mateos, L., Mondelo-Macía, P., Peña-Chilet, M., et al. (2021). Immunotherapy in Nonsmall-Cell Lung Cancer: Current Status and Future Prospects for Liquid Biopsy. *Cancer Immunol. Immunother.* 70 (5), 1177–1188. doi:10.1007/s00262-020-02752-z

Chen, F., Han, B., Meng, Y., Han, Y., Liu, B., Zhang, B., et al. (2021). Ceruloplasmin Correlates with Immune Infiltration and Serves as a Prognostic Biomarker in Breast Cancer. *Aging* 13 (16), 20438–20467. doi:10.18632/aging.203427

Chen, H.-Y., Lee, Y.-R., and Chen, R.-H. (2014). The Functions and Regulations of DAPK in Cancer Metastasis. *Apoptosis* 19 (2), 364–370. doi:10.1007/s10495-013-0923-6

Chen, W., Zheng, R., Baade, P. D., Zhang, S., Zeng, H., Bray, F., et al. (2016). Cancer Statistics in China, 2015. *CA: A Cancer J. Clinicians* 66 (2), 115–132. doi:10.3322/caac.21338

Chen, Y., Li, Z.-Y., Zhou, G.-Q., and Sun, Y. (2021). An Immune-Related Gene Prognostic Index for Head and Neck Squamous Cell Carcinoma. *Clin. Cancer Res.* 27 (1), 330–341. doi:10.1158/1078-0432.ccr-20-2166

Chen, Z., Zhao, M., Li, M., Sui, Q., Bian, Y., Liang, J., et al. (2020). Identification of Differentially Expressed Genes in Lung Adenocarcinoma Cells Using Single-Cell RNA Sequencing Not Detected Using Traditional RNA Sequencing and Microarray. *Lab. Invest.* 100 (10), 1318–1329. doi:10.1038/s41374-020-0428-1

Czerwinski, M., McLemore, T. L., Gelboin, H. V., and Gonzalez, F. J. (1994). Quantification of CYP2B7, CYP4B1, and CYPOR Messenger RNAs in normal Human Lung and Lung Tumors. *Cancer Res.* 54 (4), 1085–1091.

Denisenko, T. V., Budkevich, I. N., and Zhivotovsky, B. (2018). Cell Death-Based Treatment of Lung Adenocarcinoma. *Cell Death Dis* 9 (2), 117. doi:10.1038/s41419-017-0063-y

Dong, Z.-Y., Zhong, W.-Z., Zhang, X.-C., Su, J., Xie, Z., Liu, S.-Y., et al. (2017). Potential Predictive Value of TP53 and KRAS Mutation Status for Response to PD-1 Blockade Immunotherapy in Lung Adenocarcinoma. *Clin. Cancer Res.* 23 (12), 3012–3024. doi:10.1158/1078-0432.ccr-16-2554

Gong, J., Lin, Y., Zhang, H., Liu, C., Cheng, Z., Yang, X., et al. (2020). Reprogramming of Lipid Metabolism in Cancer-Associated Fibroblasts Potentiates Migration of Colorectal Cancer Cells. *Cel Death Dis* 11 (4), 267. doi:10.1038/s41419-020-2434-z

Griss, J., Viteri, G., Sidiropoulos, K., Nguyen, V., Fabregat, A., and Hermjakob, H. (2020). ReactomeGSA - Efficient Multi-Omics Comparative Pathway Analysis. *Mol. Cell Proteomics* 19 (12), 2115–2125. doi:10.1074/mcp.TIR120.002155

Huang, M.-Y., Jiang, X.-M., Wang, B.-L., Sun, Y., and Lu, J.-J. (2021). Combination Therapy with PD-1/PD-L1 Blockade in Non-small Cell Lung Cancer: Strategies and Mechanisms. *Pharmacol. Ther.* 219, 107694. doi:10.1016/j.pharmthera.2020.107694

Jiang, A., Liu, N., Bai, S., Wang, J., Gao, H., Zheng, X., et al. (2021). Identification and Validation of an Autophagy-Related Long Non-coding RNA Signature as a Prognostic Biomarker for Patients with Lung Adenocarcinoma. *J. Thorac. Dis.* 13 (2), 720–734. doi:10.21037/jtd-20-2803

Jin, M., Li, G., Liu, W., Wu, X., Zhu, J., Zhao, D., et al. (2021). Cigarette Smoking Induces Aberrant N6-Methyladenosine of DAPK2 to Promote Non-small Cell Lung Cancer Progression by Activating NF-κB Pathway. *Cancer Lett.* 518, 214–229. doi:10.1016/j.canlet.2021.07.022

Jin, S., Guerrero-Juarez, C. F., Zhang, L., Chang, I., Ramos, R., Kuan, C.-H., et al. (2021). Inference and Analysis of Cell-Cell Communication Using CellChat. *Nat. Commun.* 12 (1), 1088. doi:10.1038/s41467-021-21246-9

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

Liang, L., Yu, J., Li, J., Li, N., Liu, J., Xiu, L., et al. (2021). Integration of scRNA-Seq and Bulk RNA-Seq to Analyse the Heterogeneity of Ovarian Cancer Immune Cells and Establish a Molecular Risk Model. *Front. Oncol.* 11, 711020. doi:10.3389/fonc.2021.711020

Liu, X., Chen, L., and Zhang, T. (2018). Increased GOLM1 Expression Independently Predicts Unfavorable Overall Survival and Recurrence-Free Survival in Lung Adenocarcinoma. *Cancer Control* 25 (1), 107327481877800. doi:10.1177/1073274818778001

Lopes-Coelho, F., André, S., Félix, A., and Serpa, J. (2018). Breast Cancer Metabolic Cross-Talk: Fibroblasts Are Hubs and Breast Cancer Cells Are Gatherers of Lipids. *Mol. Cell Endocrinol.* 462, 93–106. doi:10.1016/j.mce.2017.01.031

Lurienne, L., Cervesi, J., Duhalde, L., de Gunzburg, J., Andremont, A., Zalcman, G., et al. (2020). NSCLC Immunotherapy Efficacy and Antibiotic Use: A Systematic Review and Meta-Analysis. *J. Thorac. Oncol.* 15 (7), 1147–1159. doi:10.1016/j.jtho.2020.03.002

Ma, K., and Zhang, L. (2021). Overview: Lipid Metabolism in the Tumor Microenvironment. *Adv. Exp. Med. Biol.* 1316, 41–47. doi:10.1007/978-981-33-6785-2_3

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 161 (5), 1202–1214. doi:10.1016/j.cell.2015.05.002

Mao, Y., Yang, H., Xu, H., Lu, X., Sang, X., Du, S., et al. (2010). Golgi Protein 73 (GOLPH2) Is a Valuable Serum Marker for Hepatocellular Carcinoma. *Gut* 59 (12), 1687–1693. doi:10.1136/gut.2010.214916

Matsuoka, R., Shiba-Ishii, A., Nakano, N., Togayachi, A., Sakashita, S., Sato, Y., et al. (2018). Heterotopic Production of Ceruloplasmin by Lung Adenocarcinoma Is Significantly Correlated with Prognosis. *Lung Cancer* 118, 97–104. doi:10.1016/j.lungcan.2018.01.012

Neal, R. D., Sun, F., Emery, J. D., and Callister, M. E. (2019). Lung Cancer. *Bmj* 365, l1725. doi:10.1136/bmj.l1725

Rahman, N. I. A., Abdul Murad, N. A., Mollah, M. M., Jamal, R., and Harun, R. (2017). NFIX as a Master Regulator for Lung Cancer Progression. *Front. Pharmacol.* 8, 540. doi:10.3389/fphar.2017.00540

Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G., and Hacohen, N. (2015). Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity. *Cell* 160 (1-2), 48–61. doi:10.1016/j.cell.2014.12.033

Skoulidis, F., and Heymach, J. V. (2019). Co-occurring Genomic Alterations in Non-small-cell Lung Cancer Biology and Therapy. *Nat. Rev. Cancer* 19 (9), 495–509. doi:10.1038/s41568-019-0179-8

Song, Q., He, X., Xiong, Y., Wang, J., Zhang, L., Leung, E. L.-H., et al. (2021). The Functional Landscape of Golgi Membrane Protein 1 (GOLM1) Phosphoproteome Reveal GOLM1 Regulating P53 that Promotes Malignancy. *Cell Death Discov.* 7 (1), 42. doi:10.1038/s41420-021-00422-2

Steels, E., Paesmans, M., Berghmans, T., Branle, F., Lemaitre, F., Mascaux, C., et al. (2001). Role of P53 as a Prognostic Factor for Survival in Lung Cancer: a Systematic Review of the Literature with a Meta-Analysis. *Eur. Respir. J.* 18 (4), 705–719. doi:10.1183/09031936.01.00062201

Tamaki, Y., Arai, T., Sugimura, H., Sasaki, T., Honda, M., Muroi, Y., et al. (2011). Association between Cancer Risk and Drug-Metabolizing Enzyme Gene (CYP2A6, CYP2A13, CYP4B1, SULT1A1, GSTM1 and GSTT1) Polymorphisms in Cases of Lung Cancer in Japan. *Drug Metab. Pharmacokinet.* 26 (5), 516–522. doi:10.2133/dmpk.dmpk-11-rg-046

Tamborero, D., Rubio-Perez, C., Muiños, F., Sabarinathan, R., Piulats, J. M., Muntasell, A., et al. (2018). A Pan-Cancer Landscape of Interactions between Solid Tumors and Infiltrating Immune Cell Populations. *Clin. Cancer Res.* 24 (15), 3717–3728. doi:10.1158/1078-0432.ccr-17-3509

Travis, W. D., Brambilla, E., Nicholson, A. G., Yatabe, Y., Austin, J. H. M., Beasley, M. B., et al. (2015). The 2015 World Health Organization Classification of Lung Tumors. *J. Thorac. Oncol.* 10 (9), 1243–1260. doi:10.1097/jto.0000000000000630

Uras, I. Z., Moll, H. P., and Casanova, E. (2020). Targeting KRAS Mutant Non-Small-Cell Lung Cancer: Past, Present and Future. *Ijms* 21 (12), 4325. doi:10.3390/ijms21124325

Viktorsson, K., De Petris, L., and Lewensohn, R. (2005). The Role of P53 in Treatment Responses of Lung Cancer. *Biochem. Biophysical Res. Commun.* 331 (3), 868–880. doi:10.1016/j.bbrc.2005.03.192

Wang, C., Lv, X., He, C., Davis, J. S., Wang, C., and Hua, G. (2020). Four and a Half LIM Domains 2 (FHL2) Contribute to the Epithelial Ovarian Cancer Carcinogenesis. *Ijms* 21 (20), 7751. doi:10.3390/ijms21207751

Xu, F., Lin, H., He, P., He, L., Chen, J., Lin, L., et al. (2020). A TP53-Associated Gene Signature for Prediction of Prognosis and Therapeutic Responses in Lung Squamous Cell Carcinoma. *Oncoimmunology* 9 (1), 1731943. doi:10.1080/2162402x.2020.1731943

Yang, L., Luo, P., Song, Q., and Fei, X. (2018). DNMT1/miR-200a/GOLM1 Signaling Pathway Regulates Lung Adenocarcinoma Cells Proliferation. *Biomed. Pharmacother.* 99, 839–847. doi:10.1016/j.biopha.2018.01.161

Ye, Q.-H., Zhu, W.-W., Zhang, J.-B., Qin, Y., Lu, M., Lin, G.-L., et al. (2016). GOLM1 Modulates EGFR/RTK Cell-Surface Recycling to Drive Hepatocellular Carcinoma Metastasis. *Cancer Cell* 30 (3), 444–458. doi:10.1016/j.ccell.2016.07.017

Yi, M., Li, A., Zhou, L., Chu, Q., Luo, S., and Wu, K. (2021a). Immune Signature-Based Risk Stratification and Prediction of Immune Checkpoint Inhibitor's Efficacy for Lung Adenocarcinoma. *Cancer Immunol. Immunother.* 70 (6), 1705–1719. doi:10.1007/s00262-020-02817-z

Yi, M., Niu, M., Xu, L., Luo, S., and Wu, K. (2021b). Regulation of PD-L1 Expression in the Tumor Microenvironment. *J. Hematol. Oncol.* 14 (1), 10. doi:10.1186/s13045-020-01027-5

Zhang, C., He, H., Hu, X., Liu, A., Huang, D., Xu, Y., et al. (2019). Development and Validation of a Metastasis-Associated Prognostic Signature Based on Single-Cell RNA-Seq in clear Cell Renal Cell Carcinoma. *Aging* 11 (22), 10183–10202. doi:10.18632/aging.102434

Zhang, P., Li, S., Lv, C., Si, J., Xiong, Y., Ding, L., et al. (2018). BPI-9016M, a C-Met Inhibitor, Suppresses Tumor Cell Growth, Migration and Invasion of Lung Adenocarcinoma via miR203-DKK1. *Theranostics* 8 (21), 5890–5902. doi:10.7150/thno.27667

Zhang, X., Klamer, B., Li, J., Fernandez, S., and Li, L. (2020). A Pan-Cancer Study of Class-3 Semaphorins as Therapeutic Targets in Cancer. *BMC Med. Genomics* 13 (Suppl. 5), 45. doi:10.1186/s12920-020-0682-5

Zhao, L., Song, X., Guo, Y., Ding, N., Wang, T., and Huang, L. (2021). Long Non-coding RNA SNHG3 P-romotes the D-evelopment of N-on-small C-ell L-ung C-ancer via the miR-1343-3p/NFIX P-athway. *Int. J. Mol. Med.* 48 (2). doi:10.3892/ijmm.2021.4980

Zhao, M., Li, X., and Chen, X. (2021). GOLM1 Predicts Poor Prognosis of Patients with NSCLC and Is Associated with the Proliferation and Chemosensitivity of Cisplatin in NSCLC Cells: Bioinformatics Analysis and Laboratory Validation. *J. Bioenerg. Biomembr* 53 (2), 177–189. doi:10.1007/s10863-021-09875-7

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

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 04 March 2022
doi: 10.3389/fgene.2022.844544

ECO: An Integrated Gene Expression Omnibus for Mouse Endothelial Cells *In Vivo*

*Xiangyi Deng[1†], Fan Yang[1†], Lei Zhang[2,3†], Jianhao Wang[1†], Boxuan Liu[3], Wei Liang[3], Jiefu Tang[4], Yuan Xie[2] and Liqun He[1,5]**

[1]Department of Neurosurgery, Tianjin Medical University General Hospital, Tianjin Neurological Institute, Key Laboratory of Post-Neuroinjury Neuro-Repair and Regeneration in Central Nervous System, Ministry of Education and Tianjin City, Tianjin, China, [2]Key Laboratory of Ministry of Education for Medicinal Plant Resource and Natural Pharmaceutical Chemistry, National Engineering Laboratory for Resource Developing of Endangered Chinese Crude Drugs in Northwest of China, College of Life Sciences, Shaanxi Normal University, Xi'an, China, [3]Precision Medicine Center, the Second People's Hospital of Huaihua, Huaihua, China, [4]Trauma Center, First Affiliated Hospital of Hunan University of Medicine, Huaihua, China, [5]Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden

Endothelial cell (EC) plays critical roles in vascular physiological and pathological processes. With the development of high-throughput technologies, transcriptomics analysis of EC has increased dramatically and a large amount of informative data have been generated. The dynamic patterns of gene expression in ECs under various conditions were revealed. Unfortunately, due to the lack of bioinformatics infrastructures, reuse of these large-scale datasets is challenging for many scientists. Here, by systematic re-analyzing, integrating, and standardizing of 203 RNA sequencing samples from freshly isolated mouse ECs under 71 conditions, we constructed an integrated mouse EC gene expression omnibus (ECO). The ECO database enables one-click retrieval of endothelial expression profiles from different organs under different conditions including disease models, genetic modifications, and clinically relevant treatments *in vivo*. The EC expression profiles are visualized with user-friendly bar-plots. It also provides a convenient search tool for co-expressed genes. ECO facilitates endothelial research with an integrated tool and resource for transcriptome analysis. The ECO database is freely available at https://heomics.shinyapps.io/ecodb/.

Keywords: endothelial cells, gene expression, RNAseq, database, integration

OPEN ACCESS

Edited by:
Geng Chen, GeneCast Biotechnology Co., Ltd., China

Reviewed by:
Ting Li, National Center for Toxicological Research (FDA), United States
Artem Kasianov, Vavilov Institute of General Genetics (RAS), Russia

****Correspondence:***
Liqun He
liqun.he@igp.uu.se

[†]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: *28 December 2021*
Accepted: *15 February 2022*
Published: *04 March 2022*

Citation:
Deng X, Yang F, Zhang L, Wang J, Liu B, Liang W, Tang J, Xie Y and He L (2022) ECO: An Integrated Gene Expression Omnibus for Mouse Endothelial Cells In Vivo. Front. Genet. 13:844544. doi: 10.3389/fgene.2022.844544

INTRODUCTION

Endothelial cells (ECs) are single-layered squamous cells distributed on the inner surface of the vasculature, constructing a barrier between the vasculature and tissues and controlling the exchange of substances and fluids (Krüger-Genge et al., 2019). ECs are involved in many essential physiological functions, such as regulating vasoconstriction and vasodilation, blood coagulation, paracrine action, angiogenesis, and constitute barriers (Reglero-Real et al., 2016; Wong et al., 2017; Paone et al., 2019). Dysfunction of EC is the driving factor for many diseases, including atherosclerosis, cancer, hypertension, glomerular disease, and inflammation (Goveia et al., 2014; Li et al., 2019). Uncovering the molecular mechanism of endothelial cells in these pathological conditions is essential to understand the occurrence and treatment of diseases.

With the rapid development of high-throughput sequencing technologies in the last decades, especially the wide use of RNA sequencing, the molecular level analysis of EC has increased significantly and a variety of EC transcriptomics datasets have been accumulated in the public domain (Khan et al., 2019; Munji et al., 2019). Their raw RNAseq data generated by high-throughput sequencing are deposited in the public databases, such as Gene Expression Omnibus (GEO) (Barrett et al., 2007) and ArrayExpress (Parkinson et al., 2007), but unfortunately, it is difficult for researchers without bioinformatics skills to process these raw data and extract the desired information. In some other fields, there are already some databases that provide practical functions to greatly promote the development of this field, such as the Allen Brain Atlas (Lein et al., 2007) for neuroscience and ONCOMINE (Rhodes et al., 2004) for oncology. For EC data, the effort of integrating has been initiated, for example, EndoDB, which has made a collection of EC data (Khan et al., 2019). However, there is still a lack of database integrating all latest RNAseq data and also providing user-friendly analysis functions and visualization tools.

Here, we integrated all freshly isolated EC bulk RNA sequencing data from public sequence databases, processed them with a standardized pipeline, and constructed a user-friendly online database, ECO. It provides a one-click search tool for *in vivo* EC profiles for each gene in various conditions including pathological alterations, genetic modifications, and other treatment conditions, in the form of easily understandable bar-plots. Also, the database provides a search function to find genes with similar expression profiles, which may generate interesting hypothesis for future research.

METHODS

Retrieval of EC RNA Sequencing Datasets

We first conducted a systematic literature search for murine *in vivo* EC bulk RNAseq studies in PubMed, the NCBI GEO database, and the ArrayExpress database. It resulted in 19 RNA studies for EC under various conditions. They include 71 EC conditions. Each condition has multiple replicated samples, and in total, there are 203 samples. The raw sequence data for each condition, including the raw data for its exact control group, were obtained from the NCBI Short Read Archive (SRA) or ArrayExpress database.

Data Preprocessing on Galaxy

The raw sequence data obtained from SRA and ArrayExpress were preprocessed with the Galaxy online server (Jalili et al., 2020) (https://usegalaxy.eu/, version: 20.09) using a standardized procedure for all datasets. The detailed procedure is described in the Galaxy RNA-seq analysis instruction (https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/rna-seq-reads-to-counts/tutorial.html).

The sequence data were uploaded in two ways: for the data available in SRA, the SRA-tools (Leinonen et al., 2011) (version: 2.10.8) in Galaxy were used to upload these datasets reads in the FASTA/Q format from the NCBI; for the other datasets from ArrayExpress, the ArrayExpress FTP download links were used. The FASTQ sequence files were then aligned to the referenced mouse genome assembly (GRCm38/mm10) obtained from the UCSC Genome Browser database (Navarro Gonzalez et al., 2021) using the HISAT2 tool (Kim et al., 2015) (version: 2.1.0) on Galaxy. The gene annotation file GTF (2020, ncbiRefSeq, mm10) was also obtained from the UCSC Genome Browser database, which was consistent with the genome sequence file. The alignment bam files were then input to the featureCounts tool (Liao et al., 2014) (version: 2.0.1, with default parameters) to get the raw read counts for each genes (feature count files). In total, 203 samples were quantified and their count data were processed in R (version: 4.0.3) for downstream analysis.

Data Normalization

In order to compare the EC expression level among different samples in different conditions, all the raw count data were normalized using rpkm function in the edgeR package (version: 3.32.0). The FPKM values for each sample were calculated, and then, the average expressions and standard deviations for each of the 32 conditions (71 bars in the FPKM plot) were calculated in R. The result for each gene was visualized in bar-plot using the ggplot2 package (version: 3.3.2).

Differential Expression Gene Analysis

The gene expression raw count files were imported into the limma package (version: 3.46.0) in R, and the voom function was used to compare the gene expression between two groups (treated versus control) with the default parameters. To remove low-expression genes in each sample, the genes which were detected in only one sample were filtered out. To visualize the differential expression profiles among the 40 comparison groups, the fold changes and the standard deviations for each gene were visualized in bar-plots.

Correlation Analysis

To search for the genes with similar expression profiles with a query gene, the corr.test function from psych package (version: 2.0.12) was applied. The correlation coefficient and the p values were calculated. The sorted result was stored in a table and is available for download through our ECO database. In addition, to better illustrate the correlation result, we chose the 10 most correlated genes to the query gene and generated a heatmap with the pheatmap package (version: 1.0.12).

ECO Web Tool Construction

Our ECO database, an interactive web application, is built mainly using the R Shiny package (version: 1.6.0), as well as the other auxiliary packages including shinythemes (version: 1.2.0), ggplot2 (version: 3.3.3), and ggh4x (version: 0.1.2.1). The ECO database is available for free at https://heomics.shinyapps.io/ecodb/.

RESULTS

Construction of ECO

In order to construct a comprehensive omnibus of mouse *in vivo* EC RNAseq profiles, we performed literature mining and identified 19 currently available RNA studies (**Supplementary Table S1**), which

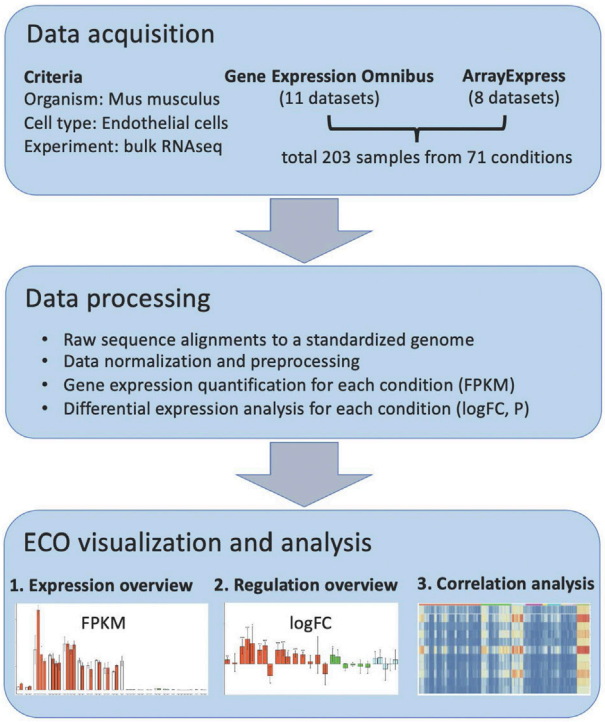


FIGURE 1 | ECO workflow.

cover EC in a variety of pathological alterations, genetic modifications, and other stimulated conditions. In these studies, freshly isolated ECs were analyzed with RNA sequencing. In total, there are 203 samples covering 71 *in vivo* conditions from 10 organs. These data composite the base for ECO, and they were processed as shown in the workflow (**Figure 1**). First, their raw data were obtained from the GEO database or ArrayExpress database, respectively. The sequence data were aligned to a standardized mouse genome assembly (GRCm38/mm10), and gene expression in each sample was quantified using the Galaxy analysis platform (Jalili et al., 2020). The gene expression levels in each condition were then summarized (average FPKM and standard deviation) and available for bar-plot visualization in the ECO database (https://heomics.shinyapps.io/ecodb). Also, the gene expression in each condition was compared with its respective control by differential expression analysis, and log scaled fold change (logFC) and *p* values were calculated, which are also illustrated with the bar-plot in the database. Besides the display of expression profiles in ECs, ECO can further identify the genes which showed similar expression profiles with the queried gene by using correlation analysis. The results were shown both as a heatmap and table.

Investigation for Inter-Organ Heterogeneity of ECs by Using ECO

ECs in different tissues have heterogeneous phenotypes for their distinct physiological needs (Kalucka et al., 2020). For instance, brain ECs form tight junctions and express active transporters to restrict diffusion, known as the blood–brain barrier (BBB) (Daneman and Prat, 2015). In contrast, ECs in the kidney are associated with fenestrae to allow efficient passage of high-volume fluids and formation of urine (Dumas et al., 2021). EC profiles from 10 organs, including the brain, lung, bone, kidney aorta, liver, eye, muscles, lymph node, and embryo, were cataloged in ECO. Users can access and download the expression of the gene of their interest in ECs of different organs in ECO by simple one-click of FPKM button. Also, users can input a customized gene list to analyze their overall gene expression enrichment pattern in a heatmap.

We use *Slc2a1* as an example to explore the inter-organ heterogeneity of a given gene. *Slc2a1*, encoding Glut1, which is highly expressed in BBB ECs but not peripheral ECs and facilitates glucose transport over BBB (Zheng et al., 2010). When we access *Slc2a1* expression by pressing the FPKM button after entering the gene symbol in the query interface, we get the normalized data bar-plot visualization for 32 sub-groups from 10 organs. As expected, *Slc2a1* is highly expressed in ECs from the brain, but almost absent in other organs (**Figure 2**).

Exploring EC Gene Alterations in Response to Disease, Genetic Manipulations, or Other Stimulations *In Vivo* by Using ECO

ECs participate in the regulation of multiple processes including angiogenesis, coagulation, and inflammation. Endothelial dysfunction is associated with many pathological alterations and aggravates progression of multiple life-threatening diseases including cancers, cardiovascular disease, diabetes mellitus, and renal disorders. In ECO, we collected EC transcriptomes from eleven mice disease models (cerebral cavernous malformation (CCM), epilepsy, experimental autoimmune encephalomyelitis (EAE), stroke, traumatic brain injury (TBI), diabetic nephropathy, Alport syndrome, liver cancer, non-alcoholic steatohepatitis (NASH), experimental autoimmune uveitis (EAU), and facioscapulohumeral muscular dystrophy (FSHD)), seven gene-modified animal models (*Jnk1/2/3* EC-specific deficient, *Cpt1a* EC-specific deficient, *Tsc2* mesenchyme cell-specific deficient, *Zmpste24* deficient, adrenomedullin (AD) EC-specific deficient, *Tankl* stroma cell-deficient, and EC-specific *Notch1* mutants), and two clinically relevant treatments (VEGF stimulations and chemo/radiotreatment) (**Supplementary Table S1**). The users can access the alteration of the genes of their interest in response to the abovementioned conditions compared to their control by clicking the logFC button. The result is illustrated in a bar-plot with 40 columns; each column represents the log2 scaled fold change, and its statistical significance (*p* value range) is indicated by asterisks (**Figure 3**).

We use *Sele* as an example to demonstrate the exploration of its regulation in different pathological conditions, genetic modifications, and treatments *in vivo*. E-selectin, encoded by *Sele*, is upregulated in ECs in response to pro-inflammatory signals, promoting the rolling and adherence of immune cells to ECs for their diapedesis (Jubeli et al., 2012). Inflammation is closely linked in the EC dysfunction in multiple diseases (Steyers and Miller, 2014). As shown in **Figure 3**, ECO provides a comprehensive portrait for *Sele* in different pathological

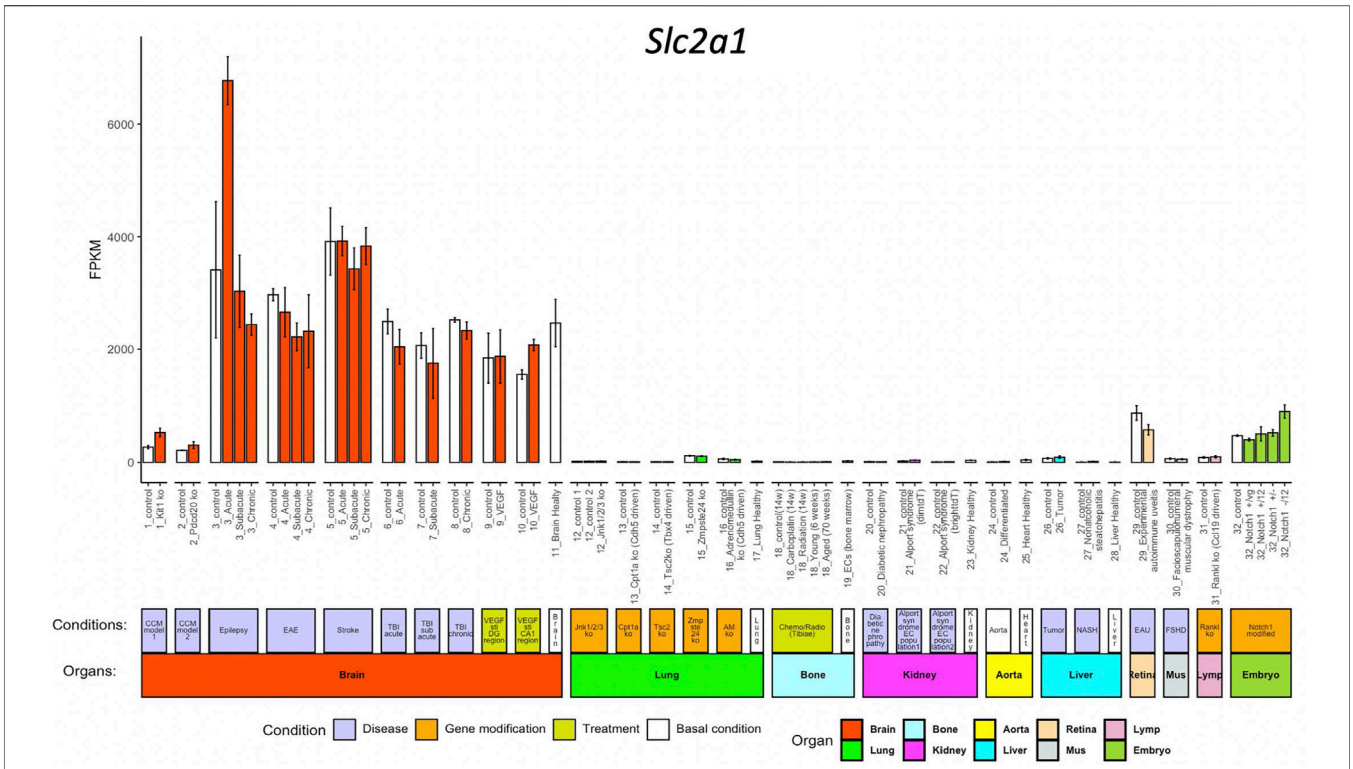


FIGURE 2 | Bar-plot of *Slc2a1* expression in different conditions. The *x*-axis shows the 71 conditions, and the *y*-axis shows the normalized FPKM values. Each bar shows the average expression (+/− standard deviation) in each condition. The bars are colored according to their organ origins, and all basal/control conditions are shown with white color (high-resolution image is available in the ECO online database).

conditions. *Sele* was upregulated in ECs from eight disease models including CCM, EAE, stroke, TBI, epilepsy, AOD, diabetic nephropathy, Alport syndrome, and NASH, highlighting the broad role of *Sele* in multiple disease progressions (Silva et al., 2017).

Predicting the Function of Poorly Characterized Genes Based on Correlation Analysis Using ECO

The correlation analysis identifies the genes which have similar expression profiles, and those co-expressed genes may have similar function. In ECO, it provides correlation analysis for the query gene to all other genes in all cataloged EC groups, as well as in individual organs which have relatively large number of samples. This facilitates uncovering the function of novel or not much characterized genes based on correlation analysis.

For example, we used a gene named *C330027C09Rik* as an example. *C330027C09Rik* did not yet have a clear gene name at the time of the gene assembly from the Ensembl database and was named after the full-length cDNA sequences from the RIKEN project (Hayashizaki, 2003). Among the top correlated genes, a list of well-known cell cycle-related genes appears, for example, *Mki67* and *Cenpf*, indicating that this gene maybe related with cell cycle (**Figure 4**). Interestingly, in the NCBI gene database, *C330027C09Rik* has been formally named as cell proliferation-regulating inhibitor of protein phosphatase 2A (*Cip2a*) (https://www.ncbi.nlm.nih.gov/gene/?term=C330027C09Rik). This confirmed the prediction from the correlation analysis.

DISCUSSION

Uncovering the EC transcriptional profile is critical to understand the EC functions in various vascular disease conditions. Previously, we have analyzed EC transcriptomes in normal mice brain (Vanlandewijck et al., 2018) and lung (He et al., 2018). It has improved the understanding of EC in these individual organs, while, on the public domain, many transcriptional profiling studies by different labs have accumulated extensive datasets for EC. However, using bioinformatics technologies to analyze these transcriptome data is a challenging task for many researchers. As such, it is of a great value to provide ECO, a user-friendly EC database, to explore expression profiles. Compared with the previously published EndoDB database (Khan et al., 2019), we have included all nine RNAseq studies in EndoDB, as well as eleven studies which were not presented there. ECO is a user-friendly web-based tool making the ever-increasing amount of EC transcriptome data easily accessible to non-bioinformatics researchers, as well as specialists as a resource of curated data.

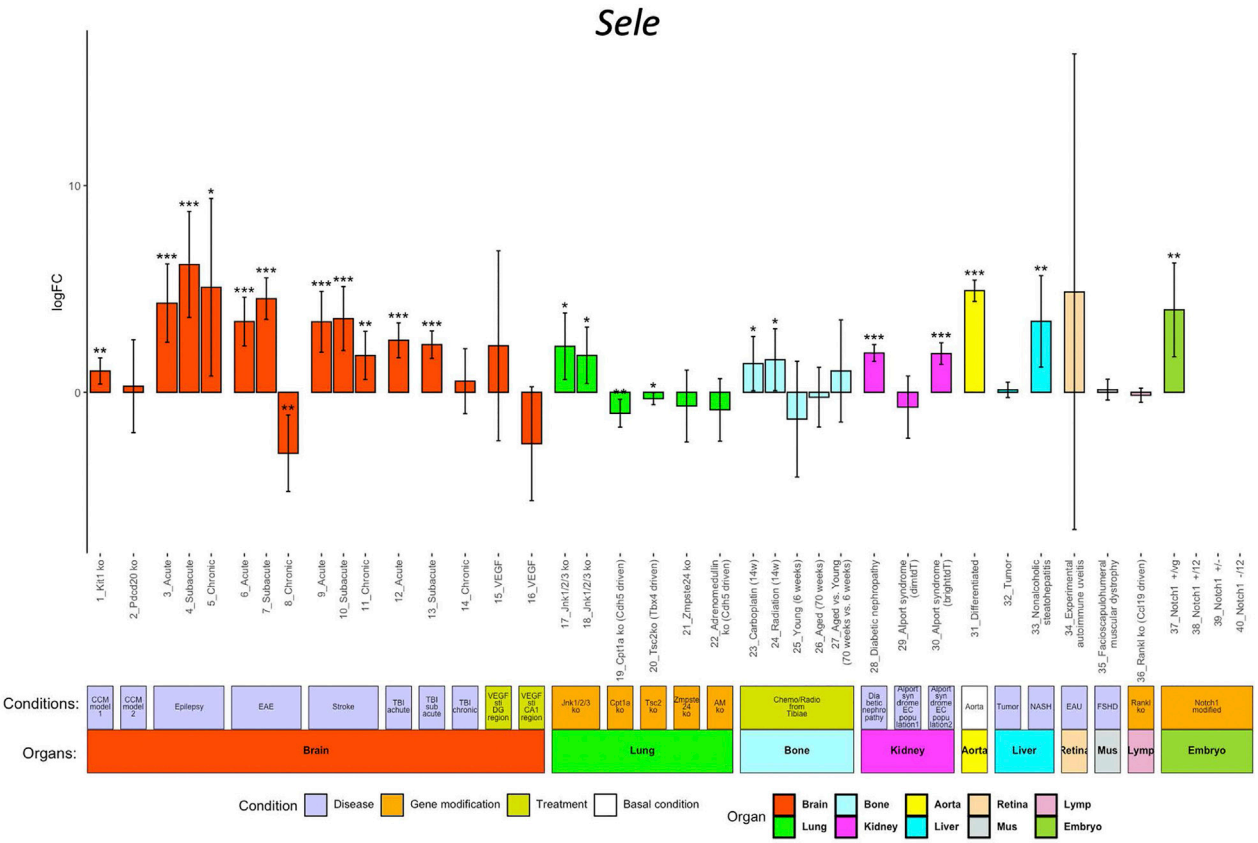


FIGURE 3 | *Sele* regulation in different conditions. The *x*-axis shows the 40 conditions, and the *y*-axis shows the log2 scaled regulation fold change (logFC). Each bar shows the average fold change with confidence interval in each condition. The bars are colored according to their organ origins (high-resolution image is available in the ECO online database).

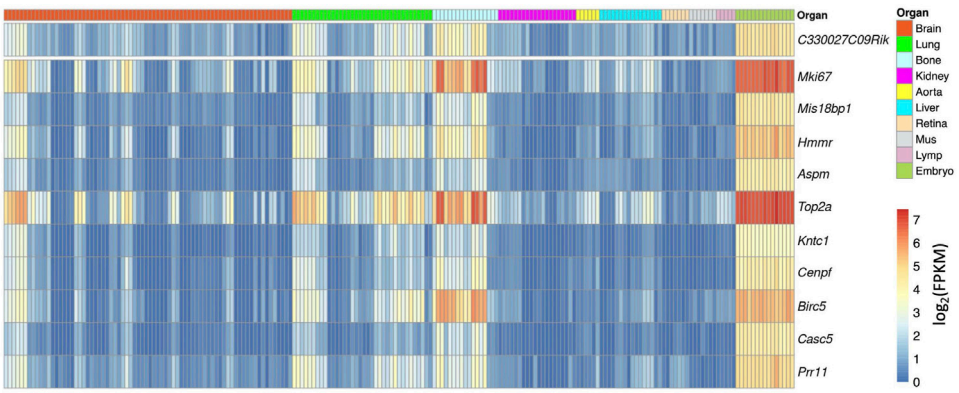


FIGURE 4 | Heatmap overview of the top correlated genes to *C330027C09Rik*. Each column shows one RNAseq sample, and its organ origin is colored on the top of the heatmap. The top 10 correlated genes are visualized. The heatmap color shows the expression level in each sample (log2 scaled FPKM) (high-resolution image is available in the ECO online database).

The core feature of ECO is one-click access to EC gene expression in different organs and alterations under different conditions for all genes on the genome. Unlike other databases, ECO dedicates to curate bulk RNAseq data from purified mouse EC under different conditions. All the data are processed using a standardized method for cross comparisons, and the results are visualized with easily understandable bar-plots. To make the users readily obtain the figures from ECO for presentation or publication usage, all the figures can be downloaded in the high-resolution PDF format.

ECO facilitates endothelial research with an integrated tool and resource for transcriptome analysis. With the friendly interactive interface, users can easily explore the published endothelial datasets from a variety of conditions, which may save some unnecessary animal experiments for vascular researchers. Also, ECO maximizes the value of published datasets by integrating them under a standardized platform. It may reveal potential global patterns which cannot be overserved from individual analysis. We expect that ECO will be a useful tool for researchers in the vascular community.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

LH conceived the project. XD, FY, and LH constructed the database. XD, FY, LZ, JW, BL, WL, JT, YX, and LH analyzed data. XD, LZ, JW, and LH wrote the manuscript. All authors reviewed and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Nos. 81870978 (LH), 81911530166 (LZ), 81702489 (LZ), and 82002659 (YX)), Tianjin Natural Science Foundation (No. 18JCYBJC94000 (LH)), Natural Science Foundation of Shaanxi Province (Nos. 2021KW-46 (LZ) and 2020JQ-429 (YX)), Fundamental Research Funds for the Central University (Nos. GK202003050 (LZ) and GK202003048 (YX)), Natural Science Foundation of Huaihua City (2020R3118 (JT) and 2020R3116 (WL)), and Natural Science Foundation of Hunan Province (No. 2020JJ4071 (B.L.)).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.844544/full#supplementary-material

REFERENCES

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., et al. (2007). NCBI GEO: Mining Tens of Millions of Expression Profiles-Ddatabase and Tools Update. *Nucleic Acids Res.* 35 (Database issue), D760–D765. doi:10.1093/nar/gkl887

Daneman, R., and Prat, A. (2015). The Blood-Brain Barrier. *Cold Spring Harb. Perspect. Biol.* 7, a020412. doi:10.1101/cshperspect.a020412

Dumas, S. J., Meta, E., Borri, M., Luo, Y., Li, X., Rabelink, T. J., et al. (2021). Phenotypic Diversity and Metabolic Specialization of Renal Endothelial Cells. *Nat. Rev. Nephrol.* 17, 441. doi:10.1038/s41581-021-00411-9

Goveia, J., Stapor, P., and Carmeliet, P. (2014). Principles of Targeting Endothelial Cell Metabolism to Treat Angiogenesis and Endothelial Cell Dysfunction in Disease. *EMBO Mol. Med.* 6 (9), 1105–1120. doi:10.15252/emmm.201404156

Hayashizaki, Y. (2003). The Riken Mouse Genome Encyclopedia Project. *C R. Biol.* 326 (10-11), 923–929. doi:10.1016/j.crvi.2003.09.018

He, L., Vanlandewijck, M., Mäe, M. A., Andrae, J., Ando, K., Del Gaudio, F., et al. (2018). Single-cell RNA Sequencing of Mouse Brain and Lung Vascular and Vessel-Associated Cell Types. *Sci. Data* 5, 180160. doi:10.1038/sdata.2018.160

Jalili, V., Afgan, E., Gu, Q., Clements, D., Blankenberg, D., Goecks, J., et al. (2020). The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2020 Update. *Nucleic Acids Res.* 48 (W1), W395–W402. doi:10.1093/nar/gkaa434

Jubeli, E., Moine, L., Vergnaud-Gauduchon, J., and Barratt, G. (2012). E-selectin as a Target for Drug Delivery and Molecular Imaging. *J. Control. Release* 158 (2), 194–206. doi:10.1016/j.jconrel.2011.09.084

Kalucka, J., de Rooij, L. P. M. H., Goveia, J., Rohlenova, K., Dumas, S. J., Meta, E., et al. (2020). Single-Cell Transcriptome Atlas of Murine Endothelial Cells. *Cell.* 180 (4), 764–779. doi:10.1016/j.cell.2020.01.015

Khan, S., Taverna, F., Rohlenova, K., Treps, L., Geldhof, V., de Rooij, L., et al. (2019). EndoDB: a Database of Endothelial Cell Transcriptomics Data. *Nucleic Acids Res.* 47 (D1), D736–D744. doi:10.1093/nar/gky997

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a Fast Spliced Aligner with Low Memory Requirements. *Nat. Methods* 12 (4), 357–360. doi:10.1038/nmeth.3317

Krüger-Genge, A., Blocki, A., Franke, R. P., and Jung, F. (2019). Vascular Endothelial Cell Biology: An Update. *Int. J. Mol. Sci.* 20, 4411. doi:10.3390/ijms20184411

Lein, E. S., Hawrylycz, M. J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., et al. (2007). Genome-wide Atlas of Gene Expression in the Adult Mouse Brain. *Nature* 445 (7124), 168–176. doi:10.1038/nature05453

Leinonen, R., Sugawara, H., and Shumway, M. (2011). The Sequence Read Archive. *Nucleic Acids Res.* 39 (Database issue), D19–D21. doi:10.1093/nar/gkq1019

Li, X., Kumar, A., and Carmeliet, P. (2019). Metabolic Pathways Fueling the Endothelial Cell Drive. *Annu. Rev. Physiol.* 81, 483–503. doi:10.1146/annurev-physiol-020518-114731

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features. *Bioinformatics* 30 (7), 923–930. doi:10.1093/bioinformatics/btt656

Munji, R. N., Soung, A. L., Weiner, G. A., Sohet, F., Semple, B. D., Trivedi, A., et al. (2019). Profiling the Mouse Brain Endothelial Transcriptome in Health and Disease Models Reveals a Core Blood-Brain Barrier Dysfunction Module. *Nat. Neurosci.* 22 (11), 1892–1902. doi:10.1038/s41593-019-0497-x

Navarro Gonzalez, J., Zweig, A. S., Speir, M. L., Schmelter, D., Rosenbloom, K. R., Raney, B. J., et al. (2021). The UCSC Genome Browser Database: 2021 Update. *Nucleic Acids Res.* 49 (D1), D1046–D57. doi:10.1093/nar/gkaa1070

Paone, S., Baxter, A. A., Hulett, M. D., and Poon, I. K. H. (2019). Endothelial Cell Apoptosis and the Role of Endothelial Cell-Derived Extracellular Vesicles in the Progression of Atherosclerosis. *Cell. Mol. Life Sci.* 76 (6), 1093–1106. doi:10.1007/s00018-018-2983-9

Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., et al. (2007). ArrayExpress--a Public Database of Microarray

Experiments and Gene Expression Profiles. *Nucleic Acids Res.* 35 (Database issue), D747–D750. doi:10.1093/nar/gkl995

Reglero-Real, N., Colom, B., Bodkin, J. V., and Nourshargh, S. (2016). Endothelial Cell Junctional Adhesion Molecules. *Arterioscler Thromb. Vasc. Biol.* 36 (10), 2048–2057. doi:10.1161/atvbaha.116.307610

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., et al. (2004). ONCOMINE: a Cancer Microarray Database and Integrated Data-Mining Platform. *Neoplasia* 6 (1), 1–6. doi:10.1016/s1476-5586(04)80047-2

Silva, M., Videira, P. A., and Sackstein, R. (2017). E-selectin Ligands in the Human Mononuclear Phagocyte System: Implications for Infection, Inflammation, and Immunotherapy. *Front. Immunol.* 8, 1878. doi:10.3389/fimmu.2017.01878

Steyers, C., 3rd, and Miller, F., Jr (2014). Endothelial Dysfunction in Chronic Inflammatory Diseases. *Int. J. Mol. Sci.* 15 (7), 11324–11349. doi:10.3390/ijms150711324

Vanlandewijck, M., He, L., Mäe, M. A., Andrae, J., Ando, K., Del Gaudio, F., et al. (2018). A Molecular Atlas of Cell Types and Zonation in the Brain Vasculature. *Nature* 554 (7693), 475–480. doi:10.1038/nature25739

Wong, B. W., Marsch, E., Treps, L., Baes, M., and Carmeliet, P. (2017). Endothelial Cell Metabolism in Health and Disease: Impact of Hypoxia. *EMBO J.* 36 (15), 2187–2203. doi:10.15252/embj.201696150

Zheng, P.-P., Romme, E., Spek, P. J. v. d., Dirven, C. M. F., Willemsen, R., and Kros, J. M. (2010). Glut1/SLC2A1 Is Crucial for the Development of the Blood-Brain Barrier *In Vivo*. *Ann. Neurol.* 68 (6), 835–844. doi:10.1002/ana.22318

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: 25 March 2022
doi: 10.3389/fgene.2022.844800

TP53 /KRAS Co-Mutations Create Divergent Prognosis Signatures in Intrahepatic Cholangiocarcinoma

Chunguang Guo[1†], Zaoqu Liu[2†], Yin Yu[3†], Yunfang Chen[4], Hui Liu[5], Yaming Guo[1], Zhenyu Peng[1], Gaopo Cai[1], Zhaohui Hua[1], Xinwei Han[2]* and Zhen Li[1]**

[1]Department of Endovascular Surgery, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China, [2]Department of Interventional Radiology, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China, [3]Department of Pathophysiology, School of Basic Medical Sciences, The Academy of Medical Science, Zhengzhou University, Zhengzhou, China, [4]Department of Oncology, Zhumadian Central Hospital Affiliated to Huanghuai University, Zhumadian, China, [5]Department of Nursing, Zhumadian Central Hospital Affiliated to Huanghuai University, Zhumadian, China

OPEN ACCESS

Edited by:
Zhichao Liu,
National Center for Toxicological Research (FDA), United States

Reviewed by:
Seogsong Jeong,
Seoul National University, South Korea
Ting Li,
National Center for Toxicological Research (FDA), United States

***Correspondence:**
Zhaohui Hua
huazhaohuisfy@163.com
Xinwei Han
fcchanxw@zzu.edu.cn
Zhen Li
fccliz2@zzu.edu.cn

†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: 04 January 2022
Accepted: 07 March 2022
Published: 25 March 2022

Citation:
Guo C, Liu Z, Yu Y, Chen Y, Liu H, Guo Y, Peng Z, Cai G, Hua Z, Han X and Li Z (2022) TP53 /KRAS Co-Mutations Create Divergent Prognosis Signatures in Intrahepatic Cholangiocarcinoma. Front. Genet. 13:844800. doi: 10.3389/fgene.2022.844800

Background: Due to high invasiveness and heterogeneity, the morbidity and mortality of intrahepatic cholangiocarcinoma (ICC) remain unsatisfied. Recently, the exploration of genomic variants has decoded the underlying mechanisms of initiation and progression for multiple tumors, while has not been fully investigated in ICC.

Methods: We comprehensively analyzed 899 clinical and somatic mutation data of ICC patients from three large-scale cohorts. Based on the mutation landscape, we identified the common high-frequency mutation genes (FMGs). Subsequently, the clinical features, prognosis, tumor mutation burden (TMB), and pharmacological landscape from patients with different mutation carriers were further analyzed.

Results: We found TP53 and KRAS were the common FMGs in the three cohorts. Kaplan–Meier survival curves and univariate and multivariate analysis displayed that TP53 and KRAS mutations were associated with poor prognosis. Considering the co-mutation phenomenon of TP53 and KRAS, we stratified patients into "Double-WT," "Single-Hit," and "Double-Hit" phenotypes by mutation status. Patients with the three phenotypes showed significant differences in the mutation landscape. Additionally, compared with "Double-WT" and "Single-Hit" phenotypes, patients with "Double-Hit" presented a dismal prognosis and significantly high TMB. Through chemotherapy sensitivity analysis, we identified a total of 30 sensitive drugs for ICC patients, of which 22 were drugs sensitive to "Double-WT," 7 were drugs sensitive to "Double-Hit," and only one was a drug sensitive to "Single-Hit."

Conclusion: Our study defined a novel mutation classification based on the common FMGs, which may contribute to the individualized treatment and management of ICC patients.

Keywords: intrahepatic cholangiocarcinoma, mutation, TP53, KRAS, prognosis, TMB, chemotherapy

INTRODUCTION

Intrahepatic cholangiocarcinoma (ICC), a primary malignant tumor derived from the bile ducts, has high invasiveness and heterogeneity (Moeini et al., 2016; Rizvi et al., 2018). In recent decades, ICC has attracted increasing global attention due to its difficult diagnosis, high morbidity, and poor prognosis features (Zou et al., 2021). Despite continued advances in the modalities of treatment, there is limited improvement in the overall survival (OS) of ICC patients (Moeini et al., 2016; Sirica et al., 2019; Kelley et al., 2020). The maximum OS of advanced ICC has not exceeded 15 months and the 5-year survival rate of ICC is under 10% (Antwi et al., 2018). The genetic heterogeneity of ICC is an important cause of its high malignancy (Sirica et al., 2019). Therefore, it is necessary to recognize "high-risk" patients based on genomic alterations of ICC, which will facilitate improve prognosis and personalized treatment.

With the development of high-throughput sequencing technologies and bioinformatics, the genomic characteristics of ICC were proved to correlate with prognosis (Lamarca et al., 2020). For example, the extracellular domain in-frame deletions of FGFR2 promoted the progression of cholangiocarcinoma and served as a genomic alteration of targeted therapy (Cleary et al., 2021). Zhou et al. reported that SLIT2 was identified as a driver of ICC dissemination and inflammatory cell infiltration (Zhou et al., 2021). Additionally, tumor mutation burden (TMB) as a novel mutational signature guides the prognosis of multiple solid tumors. Based on the International Cancer Genome Consortium (ICGC) database and the Memorial Sloan Kettering (MSK) Cancer Center, the comprehensive mutational characterization of ICC has been well described. Researchers have made numerous efforts to reveal tumor-associated drivers such as TP53, KRAS, ARID1A, IDH1, and SMAD4. Mutations of these drivers were involved in the progression, prognosis, immunotherapy, and targeted therapy (Liu et al., 2021a). Herein, we conjecture that some high-frequency mutation genes (FMGs) may play an important role in the prognosis of ICC. Compared with the existing prognosis signatures, FMGs do not require a defining cutoff value to stratify patients due to their binary data characteristics, which is more conducive to the cross-platform promotion and clinical application.

In this study, we identified FMGs in ICC patients based on multiple large-scale mutation cohorts. Then, based on the common FMGs (TP53 and KRAS) of three cohorts, we formulate three novel mutation phenotypes ("Double-WT," "Single-Hit," and "Double-Hit"), and the relationship of three mutation phenotypes with TMB and OS was further explored. Finally, we identified multiple chemotherapeutic drugs with specific sensitivity between the three phenotypes. Findings from our work may be conducive to the identification of "high-risk" ICC patients and the application of precise chemotherapy in clinical practice.

MATERIALS AND METHODS

Data Collection and Processing

Somatic gene mutation data of three independent cohorts were collected from the cBioPortal dataset (https://www.cbioportal.org/), including the ICGC dataset, MSK-2021 dataset, and Shanghai dataset (Zou et al., 2014). The inclusion criteria for ICC cohorts and samples were as follows: 1) the sample size of the cohort was over 100; 2) selected the most recent cohort from the same institution for inclusion in the study; 3) have somatic mutation data; and 4) all were intrahepatic cholangiocarcinoma. A total of 899 patients (ICGC: 417; MSK-2021: 379; and SH: 103) meeting the inclusion criteria were included in the study. The baseline clinical data of patients are presented in **Supplementary Table S1**.

Delineate the Mutation Landscape

Somatic mutation and clinical information were processed using R software. The "maftools" R package was further used to visualize the mutation oncoplot (Liu et al., 2021b). For each independent cohort, the mutation oncoplot displayed the genes with top 20 mutation frequency, which were defined as FMGs. The intersection genes of FMGs in the three cohorts were defined as the common FMGs.

Assessment of Tumor Mutation Burden

TMB was defined as the total number of base substitutions, insertions, and deletions in the coding region per megabase (Liu et al., 2021c). Using the "tmb" function in the "maftool" R package, we calculated the TMB of each patient. All based substitutions and indels in the coding region of targeted genomes were retained. In contrast, synonymous mutations failing to contribute to amino acid change were discarded.

Clinical Characteristics and Prognostic Evaluation

Univariate and multivariate Cox regression analyses were used for survival analysis of clinical characteristics of patients, including age, gender, hepatitis B virus (HBV), etc. Kaplan–Meier survival analysis was used to estimate the association between mutation phenotype and OS. Multiple boxplots were used to display differences in TMB among patients with the three phenotypes. In addition, to compare the clinical characteristics of patients with the three phenotypes in ICGC cohorts, we combined some clinical features to facilitate comparison. For example, I, IA, and IB stage (AJCC stages) were collectively referred to as I stage.

Drug-Response Prediction

To explore the therapeutic response of different drugs, we downloaded the gene mutation and drug sensitivity information from the Genomics of Drug Sensitivity in Cancer (GDSC, https://www.cancerrxgene.org/). The sensitivity of different drugs was assessed by the half-maximal inhibitory (IC50), and the higher the IC50, the lower the sensitivity. Using our previous integrated pipeline (Liu et al., 2021c), we compare the drug sensitivity of different phenotypes. A summary is as follows: 1) Kolmogorov–Smirnov tests, a normality test algorithm, indicated that the imputed drug response (IC50) data were not normally distributed ($p < 0.05$). 2) Based on this result, Kruskal–Wallis and Wilcoxon rank-sum tests were utilized to

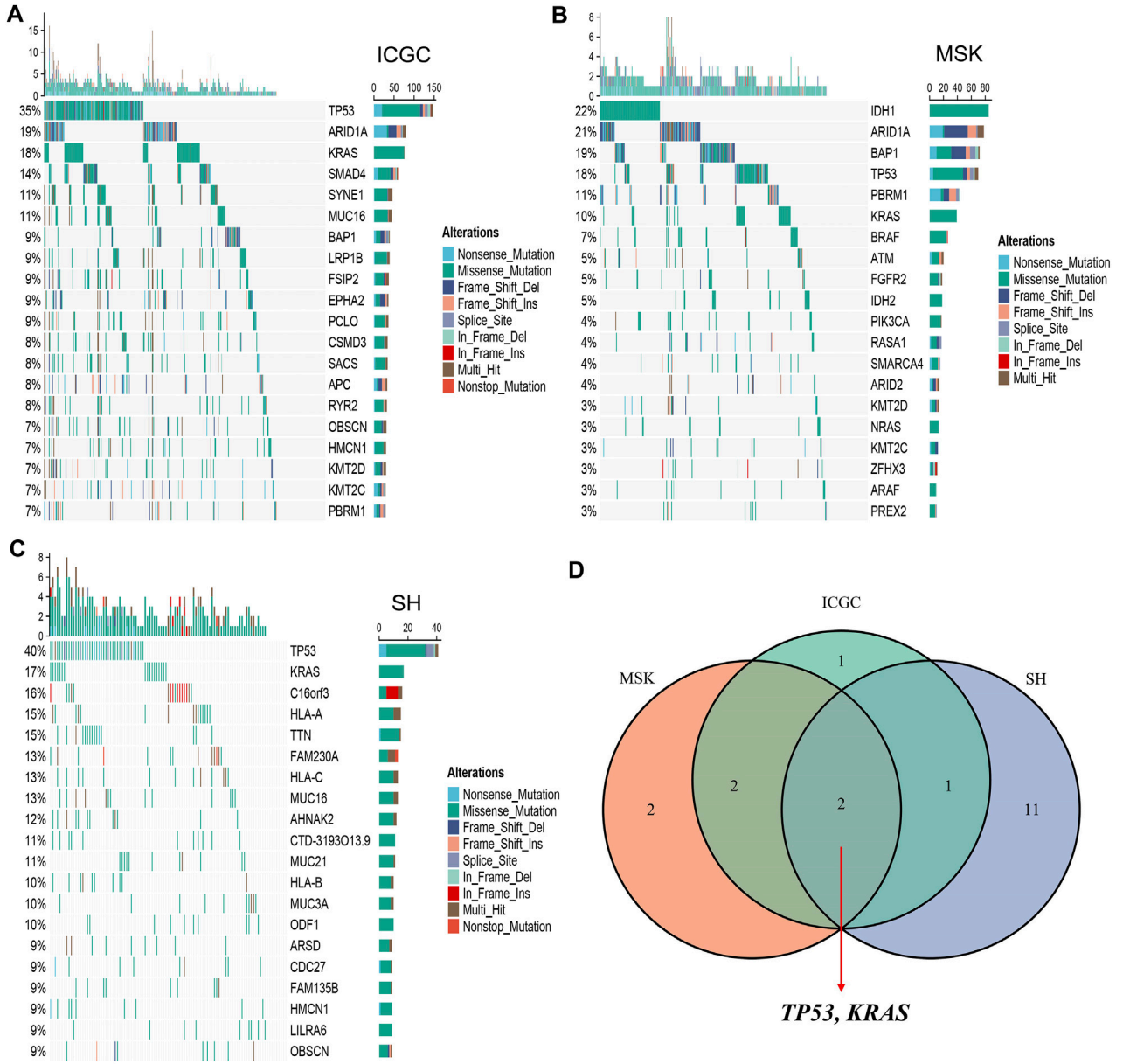


FIGURE 1 | Landscapes of high-frequency mutated genes (FMGs) in intrahepatic cholangiocarcinoma (ICC). **(A–C)** Oncoplot depicts the FMGs of ICC in the ICGC **(A)**, MSK **(B)**, and Shanghai **(C)** cohorts. **(D)** Venn diagram of FMGs covered by the three large-scale cohorts.

calculate the *p*-values and the Benjamini–Hochberg (BH) method was used for multiple testing correction. 3) For each potential drug, if one phenotype was significantly lower than other phenotypes (Wilcoxon rank-sum and Kruskal–Wallis test, false discovery rate (FDR) < 0.05), the phenotype were defined as more sensitive to the drug. 4) The sensitivity of the three phenotypes was designated "Low sensitivity," "Intermediate sensitivity," and "High sensitivity" according to the magnitude of the median IC50 value.

Statistical Analysis

All data processing, statistical analysis, and plotting were performed in R 4.0.5 software. The Wilcoxon rank-sum and Kruskal–Wallis tests were performed to compare the differences of two and multiple groups, respectively. Comparisons between categorical variables using Fisher's exact test or chi-squared test were carried out. The Benjamin–Hochberg method was used to further calculate the FDR. For every analysis, statistical significance was considered at $p < 0.05$.

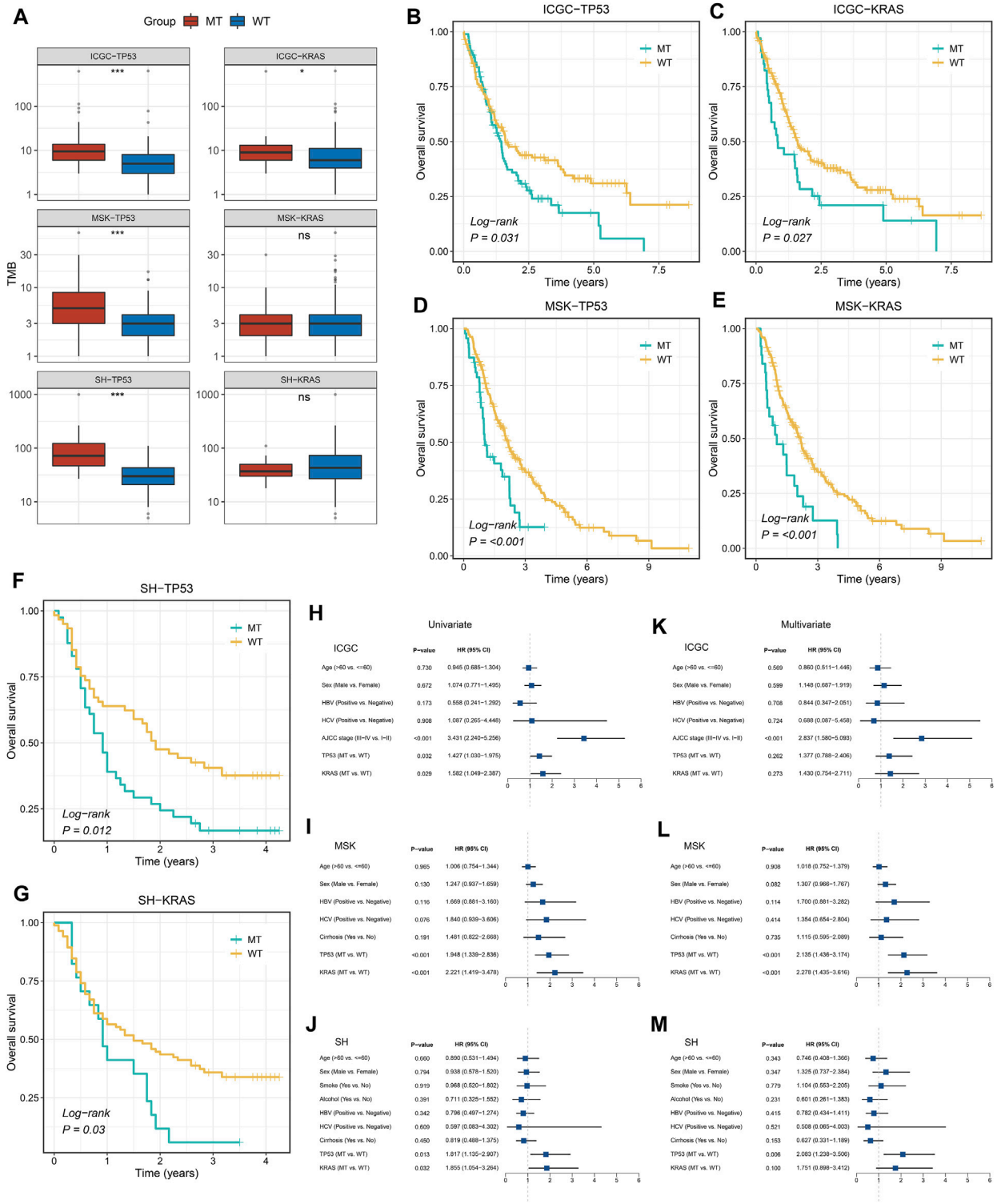


FIGURE 2 | Gene mutations are associated with TMB and clinical prognosis. **(A)** TP53 and KRAS mutations are associated with a higher TMB. **(B–G)** Kaplan–Meier survival analysis of patients with TP53 or KRAS mutations in the three cohorts. **(H–M)** Univariate and multivariate Cox regression analysis. ns $p > 0.05$; *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

RESULTS

Landscape of Somatic Mutations in ICC

The waterfall plot was utilized to describe the landscape of somatic mutations in ICC patients. We defined 20 FMGs in ICC samples from the ICGC cohort, which were TP53 (35%), ARID1A (19%), KRAS (18%), SMAD4 (14%), SYNE1 (11%), MUC16 (11%), BAP1 (9%), LRP1B (9%), FSIP2 (9%), and EPHA2 (9%) (**Figure 1A**). A total of 20 FMGs were also defined in ICC samples from the MSK cohort, including IDH1 (22%), ARID1A (21%), BAP1 (19%), TP53 (18%), PBRM1 (11%), KRAS (10%), BRAF (7%), ATM (5%), FGFR2 (5%), and IDH2 (5%) (**Figure 1B**). In addition, we also defined 20 FMGs in ICC samples from the Shanghai cohort, including TP53 (40%), KRAS (17%), C16orf3 (16%), HLA-A (15%), TTN (15%), FAM230A (13%), HLA-C (13%), MUC16 (13%), AHNAK2 (12%), and CTD-3193O13.9 (11%) (**Figure 1C**). Interestingly, three cohorts shared some common FMGs, including TP53 and KRAS (**Figure 1D**). Consequently, the subsequent analysis focused on TP53 and KRAS mutations.

TP53 and KRAS Mutations Associated With TMB and Survival Prognosis

Among the two common mutated genes, ICC patients with mutation in TP53 demonstrated significantly high TMB in the three cohorts (**Figure 2A**). Nevertheless, compared with patients without mutation in KRAS, patients with a mutation group only presented significantly high TMB in the ICGC cohort, which was not significantly different in the MSK and SH cohorts (**Figure 2A**). Subsequently, the Kaplan–Meier analysis was exploited to identify whether TP53 and KRAS mutations were associated with OS in ICC patients. As illustrated in **Figures 2B–G**, patients with TP53 and KRAS mutations presented a dismal prognosis. Univariate Cox regression analysis displayed that the hazard ratios (HRs) of TP53 and KRAS in the three cohorts (**Figures 2H–J**), respectively, were 1.427 (95% confidence interval [CI]: 1.030–1.975), 1.582 (95% CI: 1.049–2.387), 1.948 (95% CI: 1.339–2.836), 2.221 (95% CI: 1.419–3.478), 1.817 (95% CI: 1.135–2.907), and 1.855 (95% CI: 1.054–3.264) (all $p < 0.05$). Additionally, the multivariate analysis also indicated that TP53 and KRAS mutations remained statistically significant in the MSK cohort (all $p < 0.05$) (**Figure 2L**), and the HRs of TP53 and KRAS were 2.135 (95% CI: 1.436–3.174) and 2.278 (95%CI: 1.435–3.174). In the Shanghai cohort (**Figure 2M**), the HRs of TP53 and KRAS mutations were 2.083 (95% CI: 1.238–3.506, $p < 0.05$) and 1.751 (95% CI: 0.898–3.412, $P = 0.10$). However, TP53 and KRAS were also risk factors for prognosis in the ICGC cohorts, but the results were non-significant (**Figure 2K**).

TP53/KRAS Mutation Phenotypes

Prior studies have suggested that TP53 and KRAS mutation had a co-mutation phenomenon (Chen et al., 2021). Therefore, we suggested that the mutation status of TP53 and KRAS may be associated with clinical outcome and underlying biological characteristics of ICC patients. Based on the above considerations, patients with the double wild-type of TP53 and KRAS were labeled "Double-WT," patients with one mutation (TP53 and KRAS) were labeled "Single-Hit," and patients with the commutation of TP53 and KRAS were labeled "Double-Hit." As showcased in **Figures 3A–C**, there were significant differences among the survival outcome among patients with three mutation subtypes in the three independent cohorts. Notably, patients' OS becomes progressively shorter as TP53 and KRAS mutations accumulated. The "Double-Hit" phenotype patients had the shortest OS and the "Double-WT" phenotype patients had the longest OS, while the OS of "Single-Hit" phenotype patients was intermediate. Additionally, to further evaluate the prognostic values of the three phenotypes, the multivariable-adjusted analysis was utilized. As shown in **Supplementary Figure S2**, the "Double-WT" phenotype was an independent protective factor, while the "Single-Hit" and "Double-Hit" phenotypes were independent risk factors of prognosis. Subsequently, analysis of clinical characteristics in the ICGC cohort showed that there were no statistical differences in age, AJCC stage, and HBV status between the three subtypes (**Figures 3G–I**). In contrast, patients with the "Double-Hit" were more inclined to be female in the ICGC cohort (**Figure 3J**). Further comparison of TMB among the three phenotypes of patients revealed that the "Double-Hit" phenotype was a tendency toward higher, and significant differences were observed between the three phenotypes (**Figures 3D–F**). Waterfall plots of the three phenotypes suggest significant differences in the mutation landscapes of different phenotypes, and the "Double-Hit" phenotype had the lowest proportion (**Figure 4A–I**). Additionally, we calculated the frequencies of genes in the three phenotypes (**Figure 5A**), which were reported to be associated with the invasion and progression of cancer, such as SMAD4, APC, and ERBB4 (Zou et al., 2014; Lee et al., 2016). Noteworthily, patients with "Double-Hit" phenotype have higher mutation frequencies of SMAD4, APC, and AXIN1, which were numbers of the Wnt signaling pathway (**Figure 5A**). Previous study has reported that the Wnt signaling pathway contributed to the progression of cholangiocarcinoma by activating the downstream target genes (Zhang et al., 2020). BRAF mutation has been identified as a risk factor of cholangiocarcinoma (Tannapfel et al., 2003) and was most common in "Double-Hit" phenotype (**Figure 5A**).

Assessment of Chemotherapy Sensitivity

Based on the mutation and drug sensitivity information obtained from the GDSC database, the responses of ICC patients with different phenotypes to 266 chemotherapeutic agents were compared, which contributed to exploring drugs with specific sensitivity to each phenotype. As illustrated in **Figure 5B**, we identified a total of 30 sensitive drugs for ICC patients, of which 22 drugs were sensitive to "Double-WT" (such as Axitinib, Cisplatin, Pazopanib, Lestaurtinib, and PFI-1 et al.), 7 drugs were sensitive to "Double-Hit" (such as Refametinib-2,

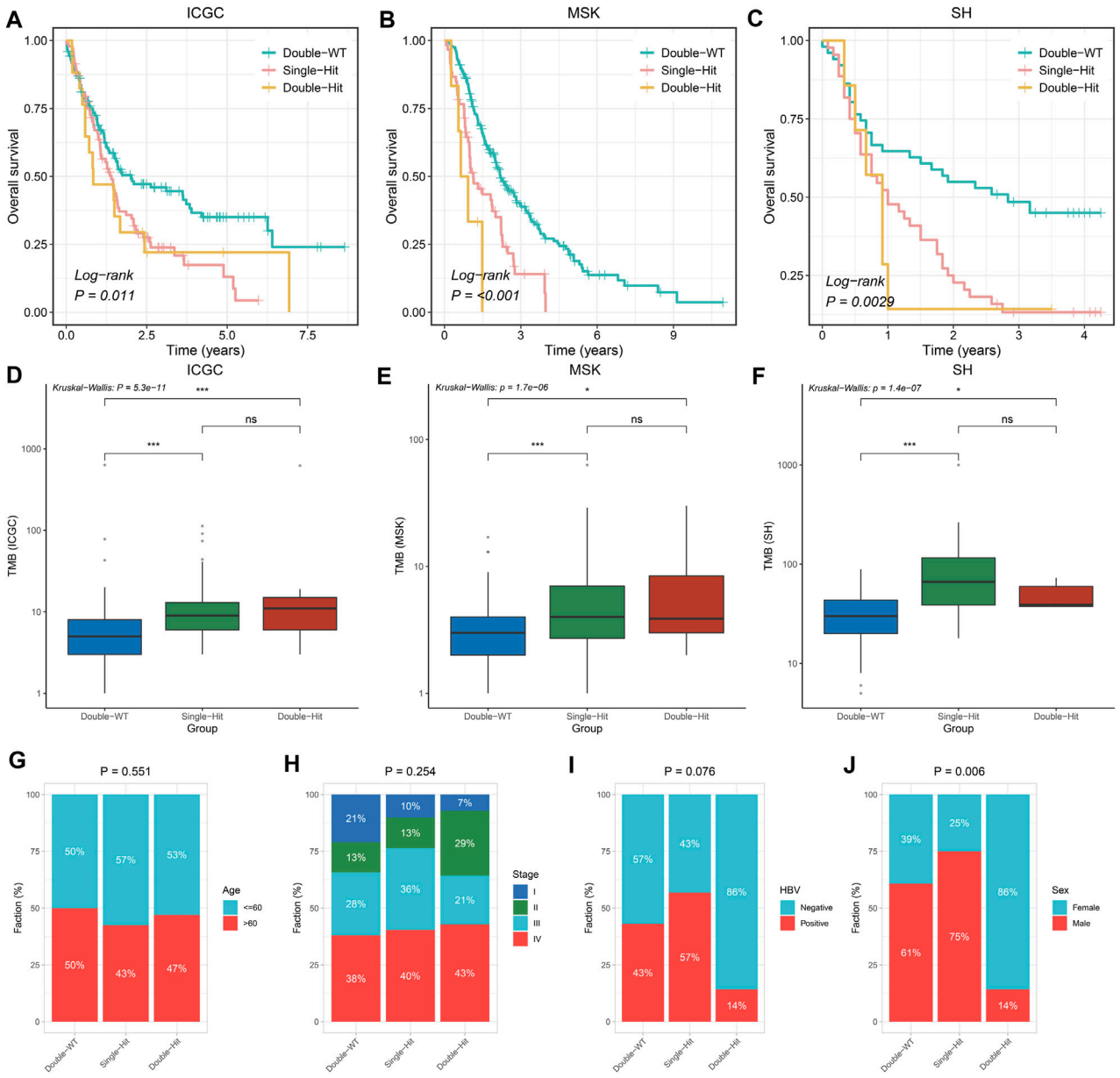


FIGURE 3 | Difference of clinical characteristics and prognosis among three TP53/KRAS mutant phenotypes in the three cohorts. **(A–C)** Kaplan–Meier survival analysis of the three phenotypes. **(D–F)** Boxplot of TMB for patients with three phenotypes. **(G–J)** Composition percentage of Age **(G)**, AJCC stage **(H)**, HBV **(I)**, and Sex **(J)** among the three phenotypes. ns $p > 0.05$; *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

Linsitinib, Trametinib, and VX-11e et al.), and only one drug was sensitive to "Single-Hit" (KIN001-270). Interestingly, the targets of sensitive drugs for "Double-Hit" phenotype patients mainly focused on the MAPK signaling pathway. Likewise, p53 signaling, VEGF signaling, and PI3K-AKT signaling were the targets of sensitive drugs for "Double-WT" patients. The drug sensitivity and target information may provide opportunities for targeted therapy in ICC patients with different phenotypes. Our study created conditions for chemotherapy for three mutation phenotypes.

DISCUSSION

In the current era of precision medicine, decoding the genetic information of tumors from the genetic levels is increasingly important for the treatment of ICC patients. In the present study, we comprehensively analyzed 899 clinical and genomics mutation data from ICGC, MSK, and Shanghai cohorts. TP53 and KRAS were common FMGs in ICC, and its mutation was associated with higher TMB and worse prognosis. Given the co-mutation phenomenon of TP53 and KRAS, three mutation

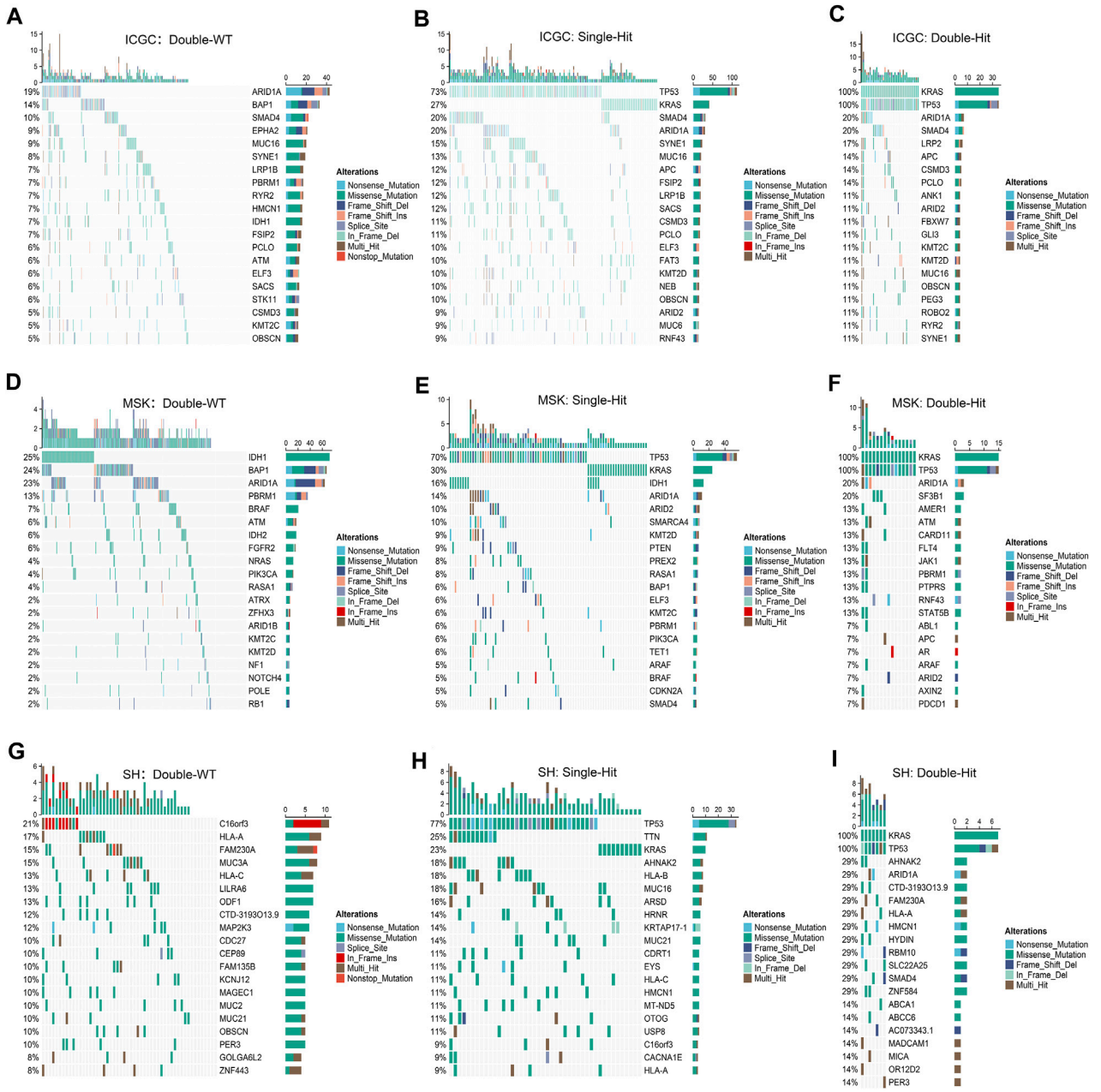


FIGURE 4 | Mutation landscapes of "Double-WT," "Single-Hit," and "Double-Hit" phenotypes in the three cohorts. **(A–C)** In the ICGC cohorts, the mutation landscapes of "Double-WT" **(A)**, "Single-Hit" **(B)**, and "Double-Hit" **(C)** phenotypes. **(D–F)** Oncoplot depicts the FMGs for the three phenotypes in the MSK cohort. **(G–I)** Oncoplot depicts the FMGs for the three phenotypes in the SH cohort.

phenotypes ("Double-WT," "Single-Hit," and "Double-Hit") were identified in ICC patients. With the cumulative mutation number in the three phenotypes, the prognosis of patients showed a tendency of dismalness. Noteworthily, we unearthed multiple potentially sensitive chemotherapeutic drugs of every phenotype, which provided a resource for precise chemotherapy of ICC patients in the clinic. In summary, our works presented a novel mutation classification and elucidated the importance of FMGs in guiding the treatment of ICC patients.

KRAS and TP53 mutations were known as major driver oncogenes in a variety of cancers, including pancreatic ductal carcinoma, non–small-cell lung cancer, and high-grade serous carcinoma (Bange et al., 2019; Sauriol et al., 2020; Tsutaho et al., 2020). Nevertheless, the clinical significance and molecular mechanism of this co-mutation phenomenon in ICC have not been elaborated. In our research, we found that TP53 and KRAS were the FMGs in cohorts from different countries. This suggests that the phenomenon of

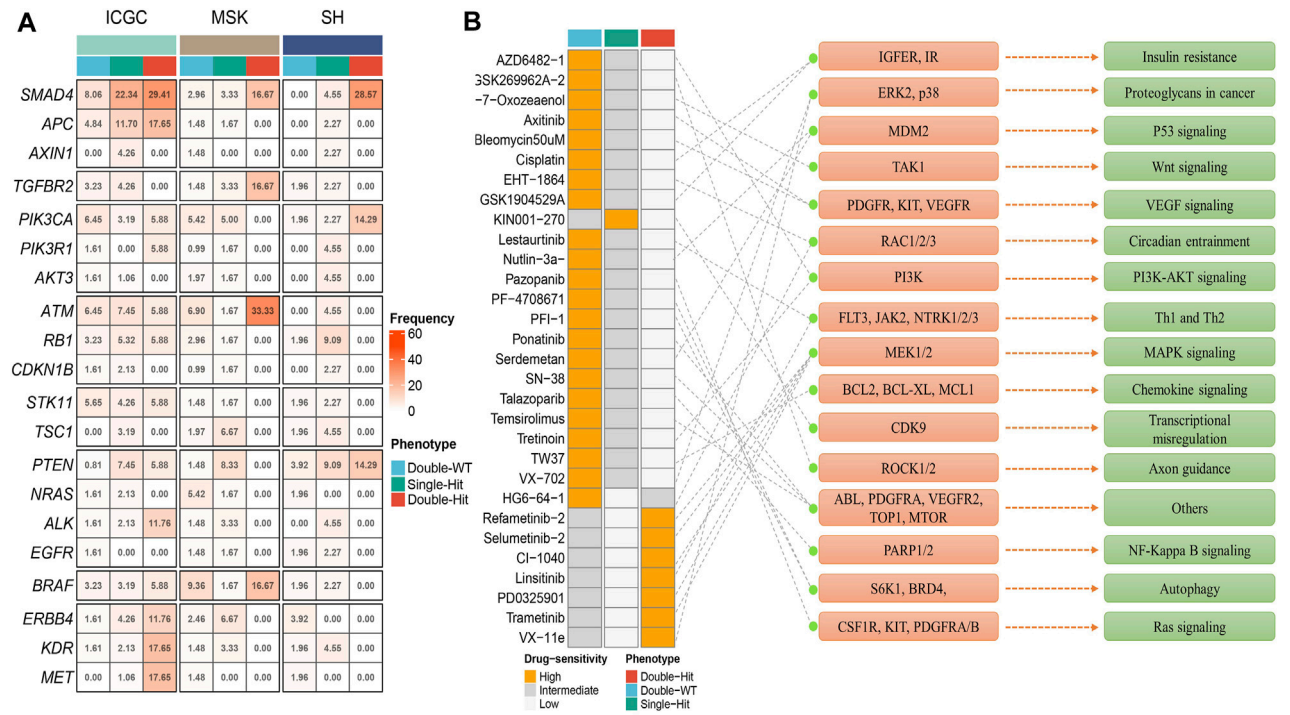


FIGURE 5 | Molecular and pharmacological landscape of three mutant phenotypes. **(A)** Mutation rate of driver genes among three mutation phenotypes in the three cohorts. **(B)** 30 potential chemotherapy drugs with specific sensitivity to each phenotype were identified in total. The left panel represents the drug names and the level of sensitivity in each phenotype, the middle panel represents the drug-targeted molecules, and the right panel represents the drug-targeted pathways.

TP53 and KRAS high-frequency mutations is not affected by race and sequencing platforms, which is important for the research of ICC. A previous study reported that mutation of TP53 would cause the download of p53, which is a tumor suppressor (Shi and Jiang, 2021). Dysfunction of p53 affects the T cell activation, which plays a key role in tumor immune escape. Similarly, KRAS mutation reduces tumor immunogenicity by inhibiting tumor neoantigen accumulation, thereby promoting tumor progression (Frost et al., 2021; Tran et al., 2021). Unsurprisingly, univariate and multivariate analysis displayed that TP53 and KRAS mutations were risk factors in multiple ICC cohorts. The prognosis of patients with the three phenotypes of "Double-WT," "Single-Hit," and "Double-Hit" was significantly indifferent, with "Double-Hit" having the worst prognosis and "Double-WT" having the best prognosis, which suggests an accumulative effect of the two mutations.

In addition, we found that we found that TMB tended to increase with the accumulation of TP53 and KRAS mutations in the ICGC and MSK cohort. However, due to the small number of patients in the "Double-Hit" group, the increase in TMB was not significant (ICGC cohort and MSK cohort) or even decreased (SH cohort) in the "Double-Hit" group compared with the "Single-Hit" and "Double-WT" groups. TMB quantifies the mutations found in the tumor and is correlated with quantity of neoantigens (Büttner et al., 2019; Grosser et al., 2019). Evidence indicated that patients with higher TMB also carry higher neoantigen loads (Büttner et al., 2019). This suggested that patients with "Double-Hit" (who tend to experience increase in TMB in the ICGC and MSK cohorts) are a potentially beneficial population for immunotherapy. In this study, we also found potentially sensitive chemotherapeutic agents for patients with different phenotypes. Patients with "Double-WT" phenotype were more sensitive to Axitinib, Cisplatin, and PFI-1. Likewise, patients with "Double-Hit" and "Single-Hit" phenotype also benefited from specific drugs, such as Trametinib and KIN001-270. Combining the benefits of immunotherapy and chemotherapy, our work provides guidance for the clinical management and individualized treatment of ICC patients with different phenotypes. However, this study has shortcomings, which are as follows: 1) further randomized clinical trials are necessary to validate the study findings and 2) some patients lacked clinical features, such as AJCC, tumor size, and lymph node metastasis. Although our results were derived from bioinformatics analysis rather than clinical experiments, we believe that comprehensive analysis based on the multicenter and larger sample can compensate for the shortcoming.

CONCLUSION

In conclusion, we defined a novel classification based on the common FMGs (TP53 and KRAS) in three large-scale cohorts. Patients with the three phenotypes showed significant differences

in mutation landscape, prognosis, and pharmacological sensitivity, which may provide new insights for individualized treatment and management of ICC 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 authors.

AUTHOR CONTRIBUTIONS

CG and ZL designed this work. CG, ZL, YY, YC, XH, and YG integrated and analyzed the data. CG, ZP, GC, ZH, and XH wrote this manuscript. CG and ZL edited and revised the manuscript. All authors approved this manuscript.

FUNDING

This study was supported by the National Natural Science Foundation of China (81873527).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.844800/full#supplementary-material

REFERENCES

Antwi, S. O., Mousa, O. Y., and Patel, T. (2018). Racial, Ethnic, and Age Disparities in Incidence and Survival of Intrahepatic Cholangiocarcinoma in the United States; 1995-2014. *Ann. Hepatol.* 17 (4), 604–614. doi:10.5604/01.3001.0012.0929

Bange, E., Marmarelis, M. E., Hwang, W. T., Yang, Y. X., Thompson, J. C., Rosenbaum, J., et al. (2019). Impact of KRAS and TP53 Co-mutations on Outcomes after First-Line Systemic Therapy Among Patients with STK11-Mutated Advanced Non-small-cell Lung Cancer. *JCO Precis Oncol.* 3, PO.18.00326. doi:10.1200/PO.18.00326

Büttner, R., Longshore, J. W., López-Ríos, F., Merkelbach-Bruse, S., Normanno, N., Rouleau, E., et al. (2019). Implementing TMB Measurement in Clinical Practice: Considerations on Assay Requirements. *ESMO Open* 4 (1), e000442. doi:10.1136/esmoopen-2018-000442

Chen, X., Wang, D., Liu, J., Qiu, J., Zhou, J., Ying, J., et al. (2021). Genomic Alterations in Biliary Tract Cancer Predict Prognosis and Immunotherapy Outcomes. *J. Immunother. Cancer* 9 (11), e003214. doi:10.1136/jitc-2021-003214

Cleary, J. M., Raghavan, S., Wu, Q., Li, Y. Y., Spurr, L. F., Gupta, H. V., et al. (2021). FGFR2 Extracellular Domain In-Frame Deletions Are Therapeutically Targetable Genomic Alterations that Function as Oncogenic Drivers in Cholangiocarcinoma. *Cancer Discov.* 11 (10), 2488–2505. doi:10.1158/2159-8290.cd-20-1669

Frost, N., Kollmeier, J., Vollbrecht, C., Grah, C., Matthes, B., Pultermann, D., et al. (2021). KRAS(G12C)/TP53 Co-mutations Identify Long-Term Responders to First Line Palliative Treatment with Pembrolizumab Monotherapy in PD-L1 High (>/=50%) Lung Adenocarcinoma. *Transl Lung Cancer Res.* 10 (2), 737–752. doi:10.21037/tlcr-20-958

Grosser, R., Cherkassky, L., Chintala, N., and Adusumilli, P. S. (2019). Combination Immunotherapy with CAR T Cells and Checkpoint Blockade for the Treatment of Solid Tumors. *Cancer Cell* 36 (5), 471–482. doi:10.1016/j.ccell.2019.09.006

Kelley, R. K., Bridgewater, J., Gores, G. J., and Zhu, A. X. (2020). Systemic Therapies for Intrahepatic Cholangiocarcinoma. *J. Hepatol.* 72 (2), 353–363. doi:10.1016/j.jhep.2019.10.009

Lamarca, A., Barriuso, J., McNamara, M. G., and Valle, J. W. (2020). Molecular Targeted Therapies: Ready for "prime Time" in Biliary Tract Cancer. *J. Hepatol.* 73 (1), 170–185. doi:10.1016/j.jhep.2020.03.007

Lee, C. H., Wang, H. E., Seo, S. Y., Kim, S. H., Kim, I. H., Kim, S. W., et al. (2016). Cancer Related Gene Alterations Can Be Detected with Next-Generation Sequencing Analysis of Bile in Diffusely Infiltrating Type Cholangiocarcinoma. *Exp. Mol. Pathol.* 101 (1), 150–156. doi:10.1016/j.yexmp.2016.07.010

Liu, Z., Guo, H., Zhu, Y., Xia, Y., Cui, J., Shi, K., et al. (2021). TP53 Alterations of Hormone-Naïve Prostate Cancer in the Chinese Population. *Prostate Cancer Prostatic Dis.* 24 (2), 482–491. doi:10.1038/s41391-020-00302-3

Liu, Z., Liu, L., Jiao, D., Guo, C., Wang, L., Li, Z., et al. (2021). Association of RYR2 Mutation with Tumor Mutation Burden, Prognosis, and Antitumor Immunity in Patients with Esophageal Adenocarcinoma. *Front. Genet.* 12, 669694. doi:10.3389/fgene.2021.669694

Liu, Z., Wang, L., Guo, C., Liu, L., Jiao, D., Sun, Z., et al. (2021). TTN/OBSCN 'Double-Hit' Predicts Favourable Prognosis, 'immune-hot' Subtype and Potentially Better Immunotherapeutic Efficacy in Colorectal Cancer. *J. Cel Mol Med* 25 (7), 3239–3251. doi:10.1111/jcmm.16393

Moeini, A., Sia, D., Bardeesy, N., Mazzaferro, V., and Llovet, J. M. (2016). Molecular Pathogenesis and Targeted Therapies for Intrahepatic Cholangiocarcinoma. *Clin. Cancer Res.* 22 (2), 291–300. doi:10.1158/1078-0432.ccr-14-3296

Rizvi, S., Khan, S. A., Hallemeier, C. L., Kelley, R. K., and Gores, G. J. (2018). Cholangiocarcinoma - Evolving Concepts and Therapeutic Strategies. *Nat. Rev. Clin. Oncol.* 15 (2), 95–111. doi:10.1038/nrclinonc.2017.157

Sauriol, A., Simeone, K., Portelance, L., Meunier, L., Leclerc-Desaulniers, K., de Ladurantaye, M., et al. (2020). Modeling the Diversity of Epithelial Ovarian Cancer through Ten Novel Well Characterized Cell Lines Covering Multiple Subtypes of the Disease. *Cancers (Basel)* 12 (8), 2222. doi:10.3390/cancers12082222

Shi, D., and Jiang, P. (2021). A Different Facet of P53 Function: Regulation of Immunity and Inflammation during Tumor Development. *Front. Cel Dev. Biol.* 9, 762651. doi:10.3389/fcell.2021.762651

Sirica, A. E., Gores, G. J., Groopman, J. D., Selaru, F. M., Strazzabosco, M., Wei Wang, X., et al. (2019). Intrahepatic Cholangiocarcinoma: Continuing Challenges and Translational Advances. *Hepatology* 69 (4), 1803–1815. doi:10.1002/hep.30289

Tannapfel, A., Sommerer, F., Benicke, M., Katalinic, A., Uhlmann, D., Witzigmann, H., et al. (2003). Mutations of the BRAF Gene in Cholangiocarcinoma but Not in Hepatocellular Carcinoma. *Gut* 52 (5), 706–712. doi:10.1136/gut.52.5.706

Tran, C. G., Goffredo, P., Mott, S. L., Hart, A., You, Y. N., Vauthey, J. N., et al. (2021). The Impact of KRAS Mutation, Microsatellite Instability, and Tumor Laterality on the Prognosis of Nonmetastatic colon Cancer. *Surgery* 171, 657. doi:10.1016/j.surg.2021.10.043

Tsutaho, A., Hashimoto, A., Hashimoto, S., Hata, S., Kachi, S., Hirano, S., et al. (2020). High Expression of AMAP1, an ARF6 Effector, Is Associated with Elevated Levels of PD-L1 and Fibrosis of Pancreatic Cancer. *Cell Commun Signal* 18 (1), 101. doi:10.1186/s12964-020-00608-8

Zhang, G. F., Qiu, L., Yang, S. L., Wu, J. C., and Liu, T. J. (2020). Wnt/β-catenin Signaling as an Emerging Potential Key Pharmacological Target in Cholangiocarcinoma. *Biosci. Rep.* 40 (3), BSR20193353. doi:10.1042/BSR20193353

Zhou, S. L., Luo, C. B., Song, C. L., Zhou, Z. J., Xin, H. Y., Hu, Z. Q., et al. (2021). Genomic Evolution and the Impact of SLIT2 Mutation in Relapsed Intrahepatic Cholangiocarcinoma. *Hepatology* 75, 831–846. doi:10.1002/hep.32164

Zou, S., Li, J., Zhou, H., Frech, C., Jiang, X., Chu, J. S. C., et al. (2014). Mutational Landscape of Intrahepatic Cholangiocarcinoma. *Nat. Commun.* 5, 5696. doi:10.1038/ncomms6696

Zou, W., Wang, Z., Zhang, X., Xu, S., Wang, F., Li, L., et al. (2021). PIWIL4 and SUPT5H Combine to Predict Prognosis and Immune Landscape in Intrahepatic Cholangiocarcinoma. *Cancer Cel Int* 21 (1), 657. doi:10.1186/s12935-021-02310-2

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 | Frontiers in Genetics

ORIGINAL RESEARCH
published: 04 May 2022
doi: 10.3389/fgene.2022.881345

Single-Cell RNA-Seq and Bulk RNA-Seq Reveal Intratumoral Heterogeneity and Tumor Microenvironment Characteristics in Diffuse Large B-Cell Lymphoma

*Yang Zhao, Hui Xu, Mingzhi Zhang and Ling Li**

Department of Oncology, First Affiliated Hospital of Zhengzhou University, Zhengzhou, China

Background: Diffuse large B-cell lymphoma (DLBCL) is the most common histologic subtype of non-Hodgkin's lymphoma (NHL) with highly heterogeneous genetic and phenotypic features. Therefore, a comprehensive understanding of cellular diversity and intratumoral heterogeneity is essential to elucidate the mechanisms driving DLBCL progression and to develop new therapeutic approaches.

Methods: We analyzed single-cell transcriptomic data from 2 reactive lymph node tissue samples and 2 DLBCL lymph node biopsy tissue samples to explore the transcriptomic landscape of DLBCL. In addition, we constructed a prognostic model based on the genes obtained from differential analysis.

Results: Based on gene expression profiles at the single cell level, we identified and characterized different subpopulations of malignant and immune cells. Malignant cells exhibited a high degree of inter-tumor heterogeneity. Tumor-infiltrating regulatory $CD4^+$ T cells showed highly immunosuppressive properties and exhausted cytotoxic $CD8^+$ T cells were highly expressed with markers of exhaustion. Cell communication analysis identified complex interactions between malignant cells and other cell subpopulations. In addition, the prognostic model we constructed allows for monitoring the prognosis of DLBCL patients.

Conclusion: This study provides an in-depth dissection of the transcriptional features of malignant B cells and tumor microenvironment (TME) in DLBCL and provides new insights into the tumor heterogeneity of DLBCL.

Keywords: diffuse large B-cell lymphoma, single-cell RNA sequencing, tumor microenvironment, tumor heterogeneity, prognosis

OPEN ACCESS

Edited by:
Rongshan Yu,
Xiamen University, China

Reviewed by:
Margarita Sánchez-Beato,
Hospital Universitario Puerta de Hierro Majadahonda, Spain
Eva Sahakian,
Moffitt Cancer Center, United States

****Correspondence:***
Ling Li
lingl510@126.com

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

Received: *22 February 2022*
Accepted: *22 April 2022*
Published: *04 May 2022*

Citation:
Zhao Y, Xu H, Zhang M and Li L (2022) Single-Cell RNA-Seq and Bulk RNA-Seq Reveal Intratumoral Heterogeneity and Tumor Microenvironment Characteristics in Diffuse Large B-Cell Lymphoma. Front. Genet. 13:881345. doi: 10.3389/fgene.2022.881345

INTRODUCTION

Diffuse large B-cell lymphoma (DLBCL) is the most common histologic subtype of non-Hodgkin's lymphoma (NHL) with highly heterogeneous genetic and phenotypic features. Gene expression profiling divides DLBCL into two distinct molecular subtypes, the activated B-cell-like and the germinal center B-cell-like subtypes (Scott et al., 2014; Reddy et al., 2017). Although the standard

first-line treatment regimen (R-CHOP) results in complete and durable remission in approximately 60% of cases, relapse occurs in 30–40% of patients and refractory disease in another 10% (Friedberg 2006; Friedberg 2011). Autologous stem cell transplantation (ASCT) after salvage chemotherapy is the standard second-line treatment for relapsed or refractory (R/R) DLBCL (Gisselbrecht et al., 2010). However, half of the patients are not eligible for transplantation due to ineffective salvage therapy, and the other half relapse after ASCT (Crump et al., 2017). The prognosis of this group of patients is extremely poor and the choice of treatment options is challenging.

The journal Science selected tumor immunotherapy as the most important scientific breakthrough of 2013 (Couzin-Frankel 2013). In 2017, the U.S. Food and Drug Administration approved two chimeric antigen receptor T-cells targeting CD19 for the treatment of R/R B-cell malignancies (Dwivedi et al., 2019). Tumor immunotherapy has become a more important treatment after the development of drug resistance in DLBCL patients. Studies have shown that the tumor immune microenvironment has a great impact on the efficacy of immunotherapy (Li et al., 2018). Thus, it has become a primary task to improve the current status of DLBCL treatment with important clinical significance to deeply explore the state of tumor microenvironment (TME) and drug resistance mechanism in DLBCL patients and find new therapeutic targets for DLBCL.

Tumor cells exist in a complex microenvironment composed of infiltrating immune cells and stromal cells. These immune cells and stromal cells, together with the cytokines and chemokines they secrete, as well as the intercellular stroma and microvasculature in the nearby area, constitute a complex network of TME (Hui and Chen 2015; Shen and Kang 2018). Tumor cells maintain their survival and proliferation by communicating with the TME network, which also allows tumor cells to develop immunosuppressive mechanisms to evade immune surveillance and promote disease progression (Coupland 2011; Ansell and Vonderheide 2013). The unique structure of the secondary lymphoid organs (including lymph nodes and spleen) in hematologic malignancies makes their microenvironment very different from that of solid tumors. In B-cell NHL, the TME is rich in immune cells, whereas in solid tumors, the number of infiltrating immune cells is relatively low (Ansell and Vonderheide 2013). Since the TME plays a crucial role in tumorigenesis, progression and recurrence, it is increasingly the focus of research on progression, metastasis and treatment resistance in solid and hematologic malignancies.

Here, we provide insight into the TME and tumor heterogeneity in DLBCL by analyzing single-cell transcriptomic data from 2 reactive lymph node tissue samples and 2 DLBCL lymph node biopsy tissue samples. We identified a high degree of inter-tumor heterogeneity in DLBCL samples and prominent immunosuppressive features in $CD4^+$ regulatory T cells ($CD4^+$ T_{REG}) and exhausted cytotoxic $CD8^+$ T cells ($CD8^+$ T_{EXH}). In addition, a prognostic model was constructed in a Bulk RNA-seq (Bulk-cell RNA sequencing) cohort containing 481 DLBCL samples based on the results of T cell subpopulation differential expression analysis, and the efficacy of the model in predicting prognosis and immunotherapy response was validated by the Gene Expression Omnibus (GEO) cohort and the Imvigor cohort.

MATERIALS AND METHODS

Acquisition and Processing of scRNA-Seq Data

Single cell transcriptome data containing 2 reactive lymph node tissue samples and 2 DLBCL lymph node biopsy tissue samples were obtained from the heiDATA database (https://heidata.uni-heidelberg.de) (**Supplementary Table S1**). Single cell samples were prepared and Single-cell RNA sequencing (scRNA-seq) as follows: single cell suspensions, synthetic complementary DNA and single cell libraries were prepared using Chromium Single Cell v2 3′ kits (10x Genomics) according to the manufacturer's instructions. Each was sequenced on a single NextSeq 550 lane (Illumina). The data were aligned to the hg38 reference genome with Cell Ranger (v2.1, 10x Genomics) using "mkfastq" and "count" commands and default parameters. The results of the Cell Ranger analysis contained the count values of unique molecular identifiers assigned to each gene in each of the cells for each individual sample using all mapped reads (**Supplementary Table S2**).

Filtering of scRNA-Seq Data

The R package Seurat (v4.0.2) (Butler et al., 2018) was used to perform quality control. Gene counts per cell, UMI counts per cell, and percentages of mitochondrial and ribosomal transcripts were calculated using the functions of the Seurat package. Genes expressed in three or fewer cells were excluded from downstream analysis. Before further analysis, libraries with >5% of mitochondrial transcripts, libraries with UMI numbers indicating an abnormal range of potential doublets, and libraries with less than 200 genes were screened out. After removing low-quality cells, we analyzed scRNA-seq profiles of 11,729 cells with an average sequencing depth of approximately 1,400 genes per cell.

Merging of Multisample Data With Correction for Batch Effects

The canonical correlation analysis (CCA) and mutual nearest neighbor (MNN) algorithms in the R package Seurat (v4.0.2) (Butler et al., 2018) were used for sample whole and correction of batch effects. After identifying the different cell types, the subsetdata function was used to split the dataset into subsets of different cell types.

Clustering and Dimensionality Reduction

We used Seurat (v4.0.2) (Butler et al., 2018) to perform clustering analysis of cells. Data was normalized to log scale using the "NormalizeData" function with a default scale parameter of 10,000. "FindVariableFeatures" function was used to identify highly variable genes with parameters for "selection.method = vst, nfeatures = 2000". We standardized the data with the "ScaleData"

function. These variable genes were used as input for PCA using the "RunPCA" function. The first 20 principal components (PCs) and a resolution of 0.5 were used for clustering using "FindClusters". Uniform manifold approximation and projection for dimension reduction (UMAP) was used for two-dimensional representation of first 20 PCs with "RunUMAP". We used the "FindAllMarkers" or "FindMarkers" function to determine the marker genes of each cluster relative to all other clusters or to a specific cluster. The selected parameters of marker genes were detected in at least 25% of the cells in the target cluster, under p value of Wilcoxon test <0.05 and the differential expression threshold of 0.25 log fold change. FeaturePlot, DotPlot, VlnPlot and DoHeatmap were used for visualization of gene expression levels. We labeled the obtained clusters as T cells, B cells, NK cells, Dendritic cells (DC) and monocytes by known classical markers (T cells: CD3D, CD3E, CD3G, TRAC; B cells: MS4A1, CD79A; NK cells: NKG7, GNLY; DC: IRF7, IRF8; monocytes: LYZ, CD68.).

Analysis of Intercellular Communication

Because DLBCL1 contains significantly more cells than DLBCL2, in order to perform a systematic analysis of intercellular communication, we re-clustered DLBCL1 for annotation and used the R package CellChat (v1.1.3) (Jin et al., 2021) to explore the expression of ligand-receptor pairs.

Cell Trajectory Analysis

Branching developmental trajectories of $CD8^+$ T cell subpopulations were calculated using the R package Monocle 2 (v2.16.0) (Qiu et al., 2017). Monocle introduces the strategy of ordering single cells in pseudo-time, by taking advantage of the asynchronous progression of individual cells in these processes and aligning them along trajectories corresponding to biological processes, such as cell differentiation.

Single-Cell Regulatory Network Inference and Clustering Analysis

After annotation of each cell type by characterization of cell type marker genes, we used the SCENIC package (v1.2.4) (Aibar et al., 2017) to analyze the enriched transcription factors in cell subpopulations. The input matrix is a normalized expression matrix, output by Seurat.

Gene Set Variation Analysis

Hallmark gene sets were downloaded from the MSigdb (Molecular Signatures Database) database and Gene Set Variation Analysis (GSVA) was performed using the R package GSVA to determine the molecular characteristics of different cell subpopulations. Gene-cell matrices are converted into gene set-cell matrices and GSVA scores are calculated for sets with at least 5 detected genes; all other parameters are default.

Prognostic Model Construction and Validation

RNA-seq data and clinical information of 481 DLBCL patients were downloaded from The Cancer Genome Atlas (TCGA) database (https://cancergenome.nih.gov/) for screening prognostic genes and developing prognostic models. RNA sequencing data and clinical information for 420 DLBCL patients from the external validation cohort GSE10846 dataset were obtained from the GEO database. Data for the IMvigor210 immunotherapy cohort were obtained from the website http://research-pub.gene. com/IMvigor210CoreBiology. Extracted $CD8^+$ T_{EXH} subpopulation-related genes obtained from differential gene expression analysis were used to construct prognostic models. In the TCGA cohort, univariate Cox regression analysis was performed using the R package Survival to screen prognosis-related genes ($p < 0.05$). Lasso regression analysis was performed using the R package glment to further screen prognosis-related genes, and finally six prognosis-related genes were obtained by multivariate Cox regression analysis for the construction of the prognostic risk model. The risk score of each patient was calculated as follows:

$$\text{Risk score} = \sum_{j=1}^{n} (\beta j \times \text{expG}j)$$

where β is the regression coefficient obtained by multivariate Cox regression analysis and expG is the prognostic gene expression level. Based on the median risk scores obtained from the prognostic model, the DLBCL samples were divided into high-risk and low-risk groups, and survival differences between the different risk subgroups were compared by Kaplan-Meier curves. We plotted time-dependent subject operating characteristic (ROC) curves with 1, 3 and 5 years as the defined points, calculated the corresponding area under the ROC curve to assess the predictive power of the risk model, and verified whether the risk score was an independent prognostic indicator for DLBCL by Cox regression analysis. The GSE10846 cohort was used as an independent external validation cohort to verify the efficacy of the prognostic model.

Tumor Microenvironment Score, Immune Cell Abundance and Immune Response Prediction

ESTIMATE is an algorithm that uses expression data to estimate stromal and immune cells in malignant tumor tissues, allowing estimation of stromal and immune scores for each DLBCL sample (Yoshihara et al., 2013). The deconvolution algorithm CIBERSORT is a method for characterizing cell composition from gene expression profiles of complex tissues, allowing inference of the relative content of immune cells from large amounts of tumor transcriptome data (Newman et al., 2015). Gene set enrichment analysis was performed using GSEA software (v4.1.0) to identify pathways that are predominantly enriched between high- and low-risk groups. Significantly enriched gene sets were screened with a threshold of $p < 0.05$. To validate the predictive power of prognostic models for immunotherapy response, the IMvigor210 immunotherapy cohort was used to assess differences in response to PD-L1 treatment in patients in different risk groups. Spearman correlation analysis was used to characterize the correlation between immune checkpoint genes and risk scores.

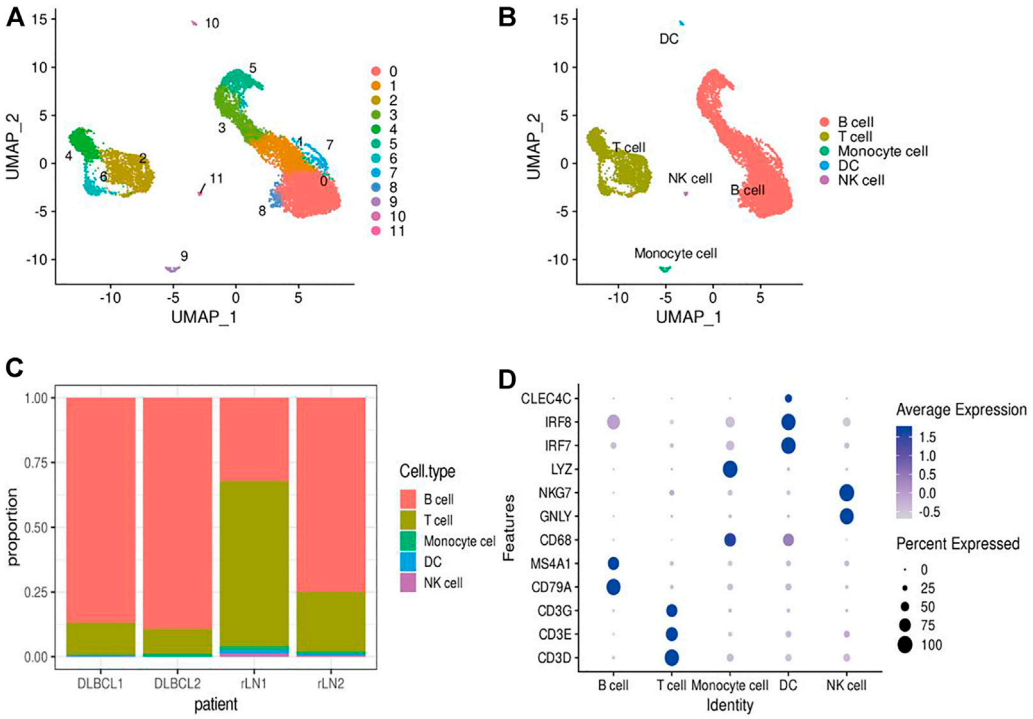


FIGURE 1 | Identification of cell types using scRNA-seq. **(A,B)** Cells from 4 samples were combined and visualized using UMAP association. Cells were colored according to their cluster **(A)** or type **(B)**. **(C)** Bar graph showing the proportion of cell types in each sample. **(D)** Typical marker genes for the immune cell types defined in **Figure 1B**. Coloration was based on expression level.

Statistical Analysis

All statistical analyses were performed in R (v4.0.5). Comparisons between groups were performed using the Wilcoxon test and *t*-test. Correlations were analyzed by using Spearmans correlation. Survival curves were compared using log-rank test. Statistical significance was accepted for $p < 0.05$. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

RESULTS

Single-Cell Transcriptomic Analysis Revealed the Complexity of Diffuse Large B-Cell Lymphoma

In this study, single-cell transcriptomic data obtained from 10x Genomics sequencing were used to investigate the cellular diversity and molecular features in DLBCL tissues. After data quality control and filtering, 11,729 cells were obtained for subsequent analysis. After normalization of gene expression data, descending and clustering were performed using principal component analysis and UMAP, respectively. Twelve cell subpopulations were obtained by dimensionality reduction and clustering (**Figure 1A**), and these cells were assigned to five different cell types using known marker genes (**Figures 1B,D**): B cells (marker genes: MS4A1 and CD79A), T cells (marker genes: CD3D, CD3E, CD3G and TRAC), NK cells (marker genes: GNLY and NKG7), DC cells (marker genes: IR7 and IR8), monocytes (marker genes: LYZ and CD68). Notably, B cells and T cells are the major cell subsets of DLBCL (**Figure 1C**).

Inter-Transcriptomic Heterogeneity of Malignant Cells in Diffuse Large B-Cell Lymphoma

To investigate the transcriptomic heterogeneity of malignant B cells in DLBCL tissues, we re-clustered the B cells and identified 13 cell subpopulations. (**Figure 2A**). To further distinguish malignant B cells from non-malignant B cells, we took advantage of the fact that the malignant B cell population expresses only one type of immunoglobulin light chain, i.e. κ or λ light chains. The ratio of light chains per B cell (κ/λ) was calculated based on the expression of the genes IGKC (encoding a constant portion of the κ light chain) and IGLC2 (λ light chain). Malignant lymph nodes contain malignant B cells that uniformly express κ light chains, whereas reactive lymph node samples contain only non-malignant B cells (**Figure 2B**). We then re-clustered the malignant B cells and obtained eight malignant B cell subpopulations (**Figure 2C**), which showed a high degree of heterogeneity. SCENIC analysis identified EGR1, FOS and STAT1 as potential transcription factors (**Figure 2D**). Gene differential expression analysis revealed different transcriptional profiles among malignant B cell

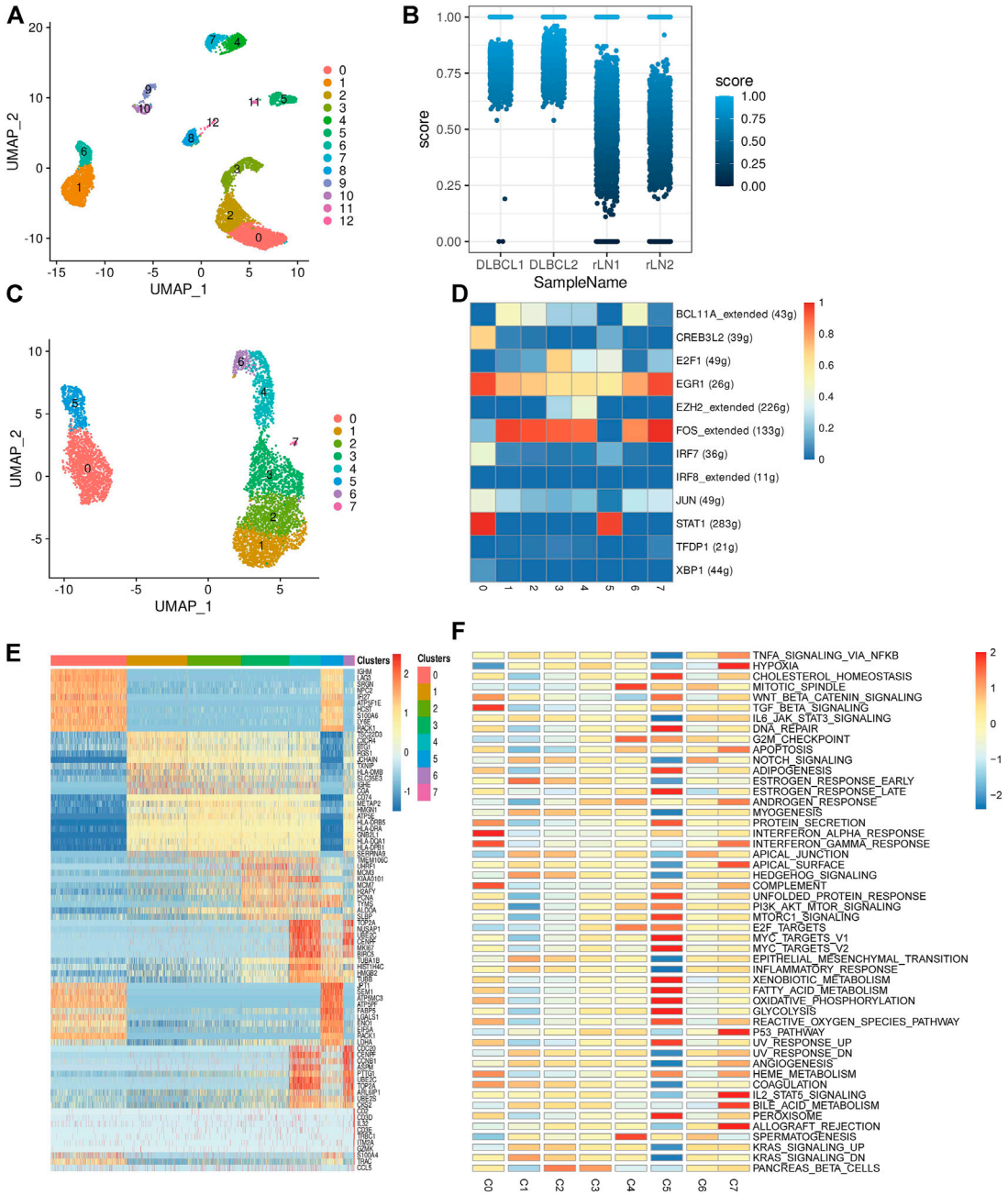


FIGURE 2 | Transcriptome heterogeneity in malignant cells. **(A)** B cells from 4 samples were combined and visualized using UMAP association. Cells were colored according to their clusters. **(B)** IGKC fraction, IGKC ÷ (IGKC + IGLC2), was calculated for each B cell. B cells were classified as κ+ if the fraction was >0.5 and as λ+ if the ratio was below 0.5. The percentage of B cells expressing κ or λ was calculated based on the transcriptionally distinct B cell clusters. Nonmalignant B cells contain approximately 50% κ and 50% λ-expressing B cells, whereas malignant B cells contain B cells that uniformly express the κ light chain. **(C)** The umap plot of malignant B cells. **(D)** Heat map of area under the curve scores for regulation of expression by transcription factors imputed with SCENIC. **(E)** Heat map showing the top 10 differential genes in the 8 malignant B cell subpopulations (Wilcoxon test). **(F)** Differential activity pathways in the 8 malignant B cell subpopulations (scored by GSVA for each cell).

subpopulations: subpopulation 0 showed high expression levels of the malignancy-promoting factors S100A6 and LY6E, subpopulation 1 showed high expression levels of the tumor suppressor BTG1 and TXNIP, subpopulation 2 showed high expression levels of the immune-related genes CD74 and HLA-DRA, subpopulation 3 and subpopulation 4 showed high expression levels of cell proliferation genes MCM3, H2AFY, PCN, MKI67, TK1, subpopulation 5 showed high expression levels of metabolism-related genes FABP5, LDHA, ENO1, and subpopulation 6 showed high expression levels of cell cycle-

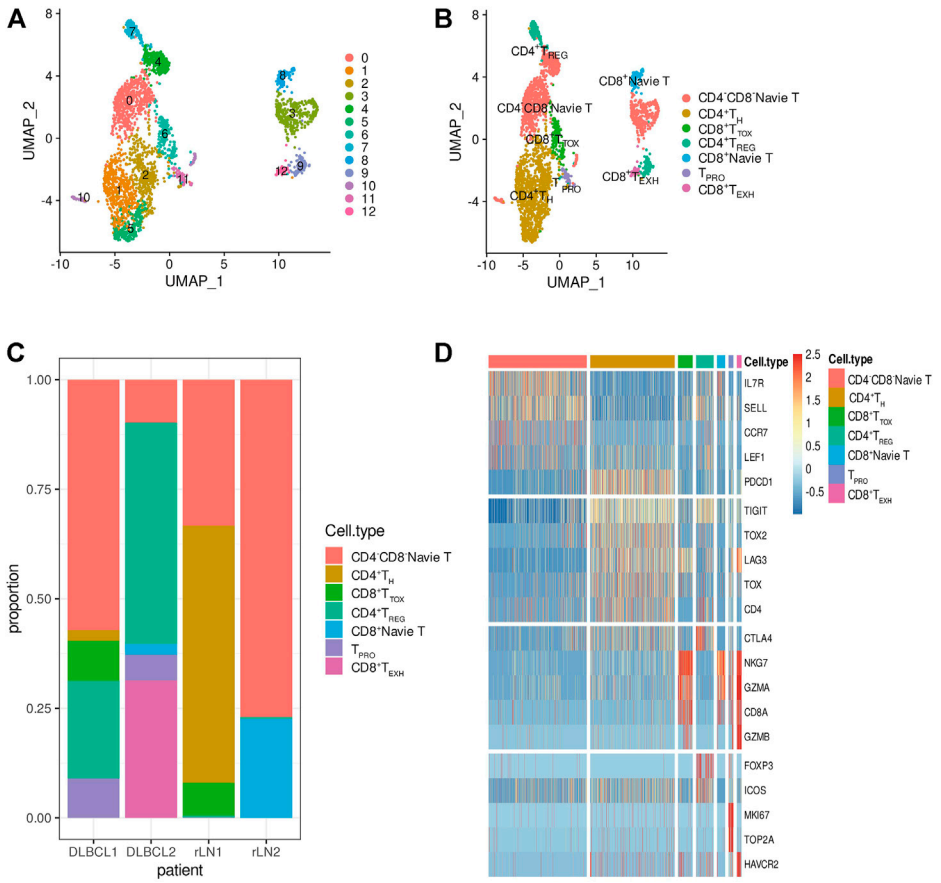


FIGURE 3 | Transcriptome heterogeneity in T cells. **(A,B)** T cells from 4 samples were combined and visualized using UMAP association. Cells were colored according to their cluster or subtype. **(C)** Bar graph showing the proportion of cell types in each sample. **(D)** Differentially expressed genes used to identify T cell subpopulations.

related genes CENPF, CCNB1, CDC20 (**Figure 2E**). GSVA analysis showed different molecular signatures among malignant B-cell subpopulations: interferon response-dominant signature (subpopulation 0), cell proliferation-dominant signature (subpopulation 3 and subpopulation 4), metabolism-dominant signature (subpopulation 5), and hypoxia-dominant signature (subpopulation 7) (**Figure 2F**). In conclusion, these results reveal a high degree of inter-tumor heterogeneity in DLBCL.

Enrichment of Immunosuppressive Tumor Infiltrating Regulatory T Cells in Diffuse Large B-Cell Lymphoma

Tumor-infiltrating immune cells are highly heterogeneous and play an important role in tumor cell immune evasion and response to immunotherapy. To investigate the transcriptomic heterogeneity of T cells in DLBCL tissues, we re-clustered T cells and identified 13 T cell subpopulations (**Figures 3A–D**). The T cell subpopulations were annotated by differentially expressed marker genes as: $CD4^{-}CD8^{-}$Navie T (IL7R,SELL, CCR7 and LEF1, subpopulations: 0, 3, 4 and 10), $CD4^{+}$ T_H (CD4 and TRAC, subpopulations: 1, 2 and 5), $CD8^{+}$ T_{TOX} (CD8A, GZMK and NKG7, subpopulation: 6), $CD4^{+}$ T_{REG} (FOXP3,TIGIT, ICOS and CTLA4, subpopulations: 7 and 9), $CD8^{+}$Navie T (CD8A, SELL and IL7R, subpopulation: 8), T_{PRO} (MKI67 and TOP2A, subpopulation: 11), $CD8^{+}$ T_{EXH} (CD8A, GZMA, NKG7, LAG3 and HAVCR2, subpopulation: 12). To understand the state transitions between $CD8^{+}$ T cell subtypes, we used Monocle2 to construct potential developmental trajectories of T cells. Developmental trajectories inferred from expression data or marker genes suggest (**Supplementary Figures S1A, B**) that $CD8^{+}$ T cells have two differentiation pathways: cytotoxic $CD8^{+}$ T cells ($CD8^{+}$ T_{TOX}) and exhausted $CD8^{+}$ T cells ($CD8^{+}$ T_{EXH}). GSVA analysis revealed different signaling pathway enrichment among subpopulations: WNT and TGF signaling ($CD4^{+}$ T_H), TGF and TNF signaling ($CD8^{+}$ T_{TOX}), IL6/STAT3, IL2/STAT5 and KRAS signaling ($CD4^{+}$ T_{REG}), and interferon response ($CD8^{+}$ T_{EXH}) (**Supplementary Figure S1C**). SCENIC analysis identified SREBF2, RAD21, IRF7 as potential transcription factors in different T cell subpopulations (**Supplementary Figure S1D**). Taken together, our single-cell analyses reveal that $CD4^{+}$ T_{REG} are

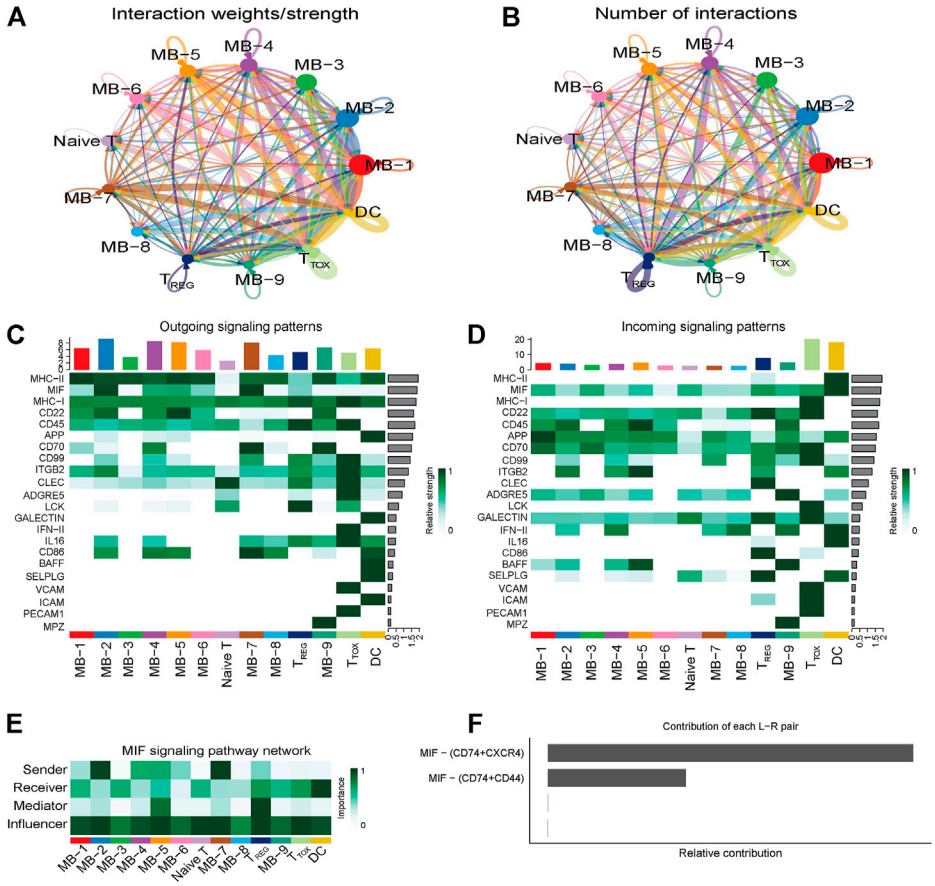


FIGURE 4 | Cellular communication in DLBCL. **(A,B)** Circle diagram showing the number of interactions or strength of interactions between any two groups of cells. **(C,D)** Heat map of the cell-cell communication network for incoming or outgoing signaling action analysis. **(E)** Heat map showing the relative importance of each cell group based on the four network centrality degrees of the calculated MIF signaling network. **(F)** Relative contribution of each ligand-receptor pair to the overall communication network of the MIF signaling pathway.

highly immunosuppressive and $CD8^+$ T_{EXH} highly express exhaustion markers such as LAG3, TIGIT and HAVCR2.

Cellular Communication in Diffuse Large B-Cell Lymphoma

To explore the interactions between cells in the DLBCL microenvironment, we used CellChat to infer and analyze intercellular communication networks. Dimension reduction, clustering and cell type annotation of sample DLBCL1 identified 13 cell subpopulations containing 9 malignant B cell subpopulations (MB1-9), 3 T cell subpopulations (T_{REG}, T_{TOX}, Naive T), and 1 DC cell subpopulation (DC). CellChat analysis revealed complex interactions between malignant B cell subpopulations and with other cell subpopulations, and 22 important pathways between 13 cell subpopulations were detected in DLBCL tissues, with the MIF signaling pathway being the prominent incoming and outgoing signaling mode (**Figures 4A–D**). Network centrality analysis of the inferred MIF signaling network showed that malignant B cell subpopulations (MB-2, MB-7) are the major senders and DCs are the major receivers of the MIF signaling pathway (**Figure 4E**). Notably, among all known ligand-receptor pairs, MIF signaling was predominantly dominated by the MIF ligand and its multimeric CD74/CXCR4 receptor (**Figure 4F**). CellChat uses a pattern recognition approach based on non-negative matrix decomposition to identify global communication patterns as well as key signals in different cell groups (i.e. pattern recognition modules). The output of this analysis is a set of the so-called communication patterns that connect cell groups with signaling pathways either in the context of outgoing signaling (i.e. treating cells as sources) or incoming signaling (i.e. treating cells as targets). The application of this pattern recognition module revealed three patterns of the outgoing signal and three patterns of the incoming signal (**Figures 5A,B**). The outgoing signaling of all malignant B cells is characterized by pattern #1, which includes the MHC-II, MIF, MHC-I, CD22, CD45 and other pathways, the outgoing signaling of T cells is characterized by pattern #2, which represents the ADGRE5, LCK, IFN-II, VCAM, PECAM1 and other pathways, and the outgoing

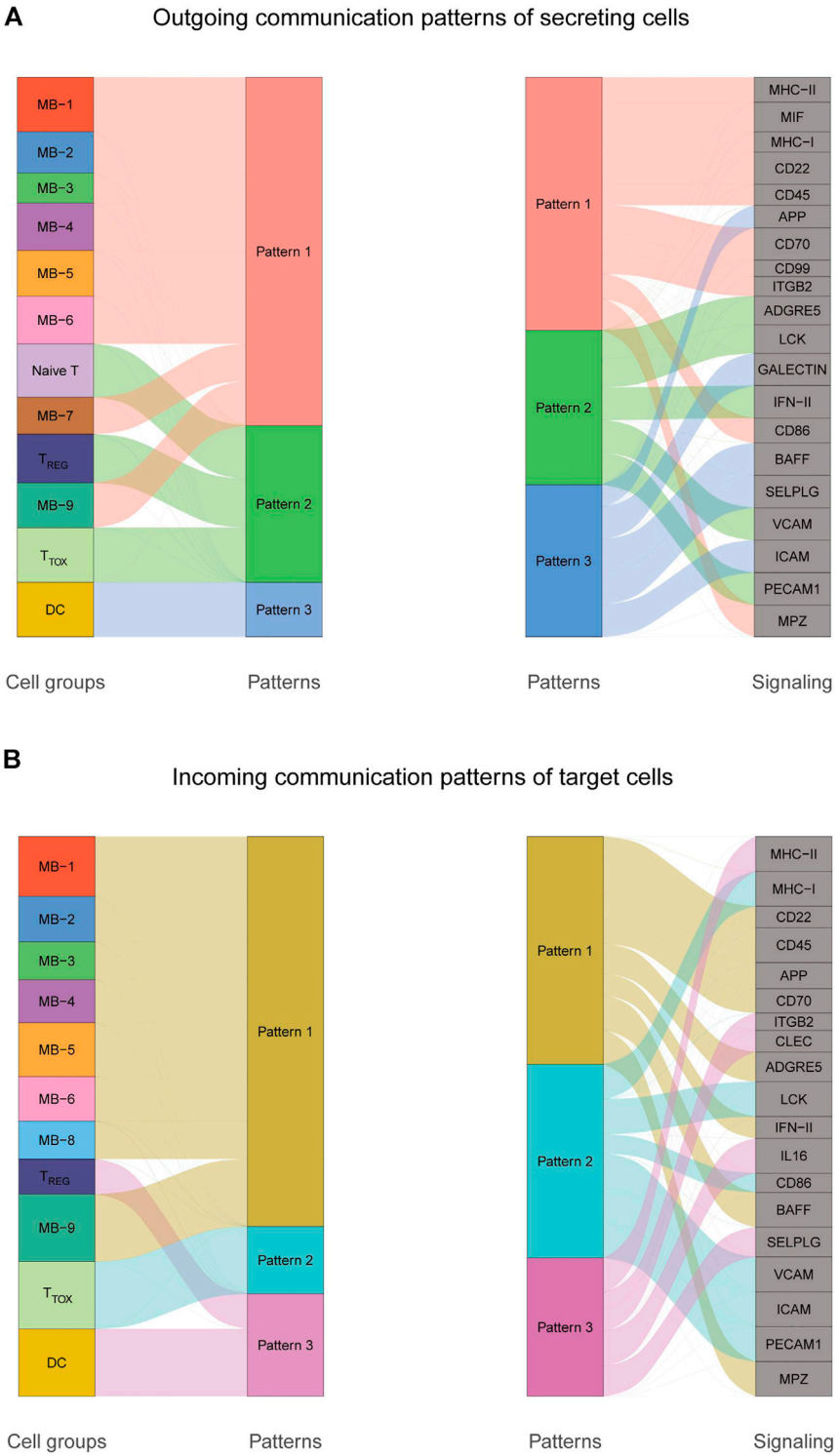


FIGURE 5 | Cellular communication patterns in DLBCL. **(A)** Visualization of outgoing communication patterns of secretory cells by alluvial plots showing the correspondence between inferred potential patterns and cell populations, as well as signaling pathways. The thickness of the flow indicates the contribution of the cell population or signaling pathway to each potential pattern. The height of each pattern is proportional to the number of cell populations or signaling pathways associated with it. Outgoing communication patterns reveal how sending cells coordinate with each other and how they coordinate with certain signaling pathways to drive communication. **(B)** Incoming communication patterns of target cells. Incoming communication patterns reveal how target cells coordinate with each other and how they coordinate with certain signaling pathways in response to incoming signals.

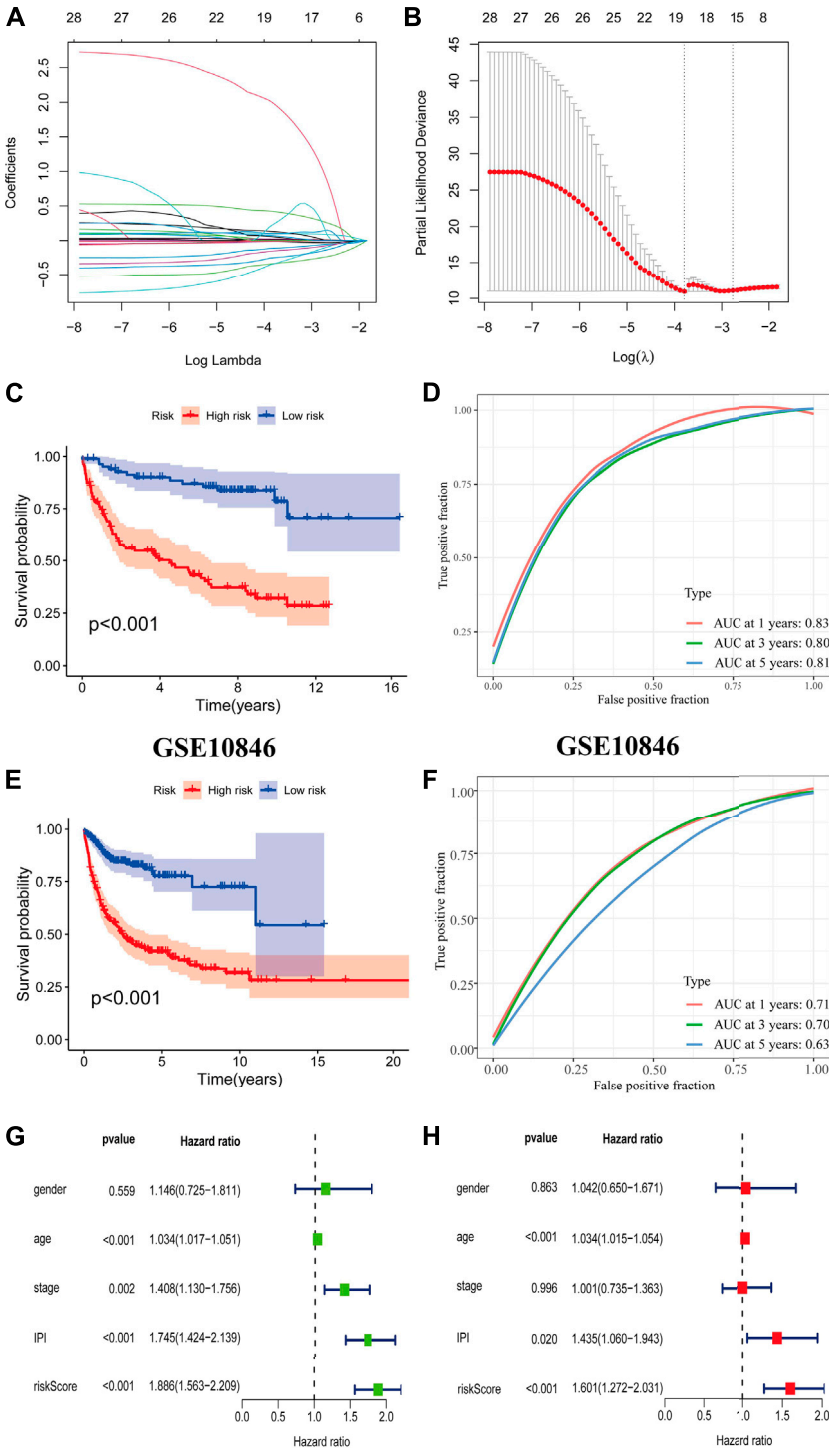


FIGURE 6 | Construction and validation of prognostic model. **(A,B)** Coefficients of selected characteristics are shown by the lambda parameter, the horizontal axis represents the value of the independent variable lambda and the vertical axis represents the coefficient of the independent variable; partial likelihood deviation is plotted against log(λ) using the lasso Cox regression model. **(C,E)** Survival analysis curves for high and low risk score groups. **(D,F)** ROC curves of the prognostic model. **(G)** Univariate Cox regression analysis of DLBCL risk factors. **(H)** Multivariate Cox regression analysis of DLBCL risk factors.

signaling of DC is characterized by pattern #3, which includes the APP, BAFF, ICAM and other pathways. On the other hand, the communication patterns of target cells show that incoming malignant B cell signaling is dominated by patterns #1, which includes signaling pathways such as CD22, CD45, CD70, BAFF, IFN-II, etc. Incoming T cell signaling is

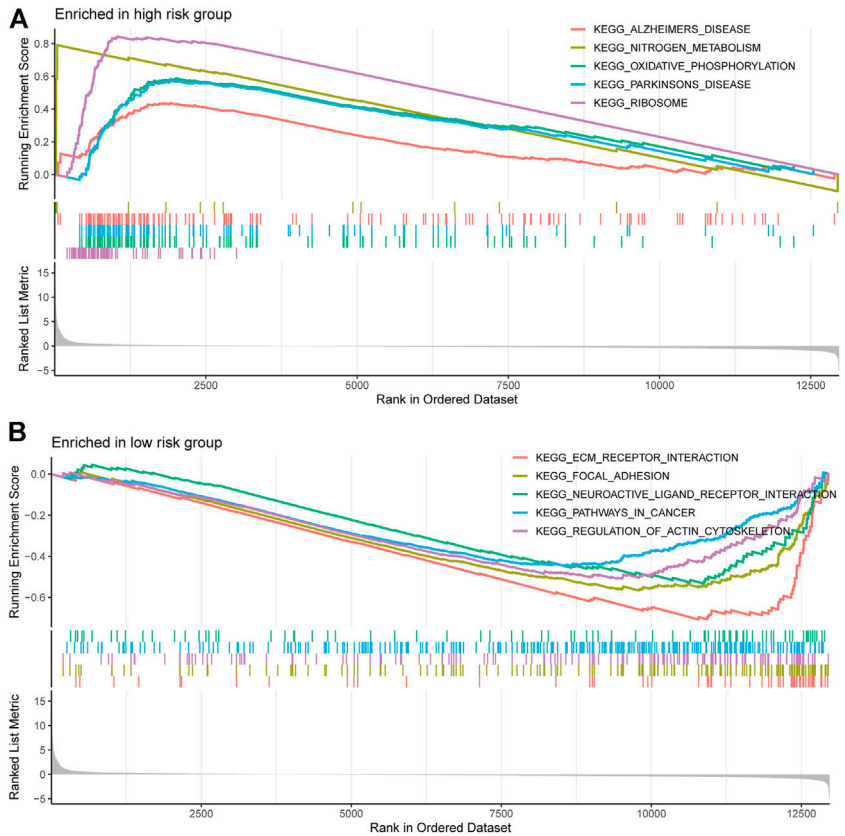


FIGURE 7 | Gene and enrichment analysis of different risk groups. **(A)** KEGG-enriched pathway in the high-risk group ($p < 0.05$ and fdr-adjusted q < 0.05). **(B)** KEGG-enriched pathway in the low-risk group ($p < 0.05$ and fdr-adjusted q < 0.05).

characterized by two patterns #2 and #3, driven by pathways such as MHC-I, LCK, VCAM, ICAM, etc., while incoming DC signaling is also characterized by patterns #3. These results suggest that different cell types in the same tissue have different signaling networks and the pattern of malignant B-cell communication is homogeneous.

Construction of a Prognostic Model Based on Exhausted CD8$^+$ T Cell-Associated Genes

The CD8$^+$ T_{EXH} subpopulation-related genes obtained from the differential analysis were extracted for the construction of the prognostic model. Nineteen genes were obtained by univariate Cox regression analysis and lasso regression analysis (**Figures 6A,B**), and finally six prognosis-related genes for model construction were obtained using multivariate Cox regression analysis (GABRA3, HOXC8, RTN4R,CRLF1, BIRC3, REXO5). Using the regression coefficients for each of the above 6 prognostic genes, we constructed a prognostic model for DLBCL patients and calculated the risk score according to the following formula: risk score = (2.201 × GABRA3 expression level) + (−0.719 × HOXC8 expression level) + (−0.765 × RTN4R expression level) + (0.545 × CRLF1 expression level) + (−0.013 × BIRC3 expression level) + (−0.226 × REXO5 expression level). Using the median value of the risk score as the threshold, we divided DLBCL patients into low-risk and high-risk groups. Survival analysis showed that patients in the high-risk group had a poorer prognosis ($p < 0.001$) (**Figure 6C**), with an area under the ROC curve of 0.83, 0.80 and 0.80 for 1-year, 3-years and 5-years OS, respectively (**Figure 6D**). In the external validation cohort, survival analysis also showed a poorer prognosis for patients in the high-risk group ($p < 0.001$) (**Figure 6E**), with an area under the ROC curve of 0.71, 0.70 and 0.63 for 1-year, 3-years and 5-years OS, respectively (**Figure 6F**). The results of univariate and multivariate Cox regression analyses indicated that risk score was an independent prognostic factor (**Figures 6G,H**).

Gene Set Enrichment Analysis for Different Risk Groups

We performed gene set enrichment analysis (GSEA) to identify potential biological processes between high- and low-risk groups. The results showed that pathways such as nitrogen metabolism, oxidative phosphorylation, ribosomes, Alzheimer's disease, and Parkinson's disease were enriched in

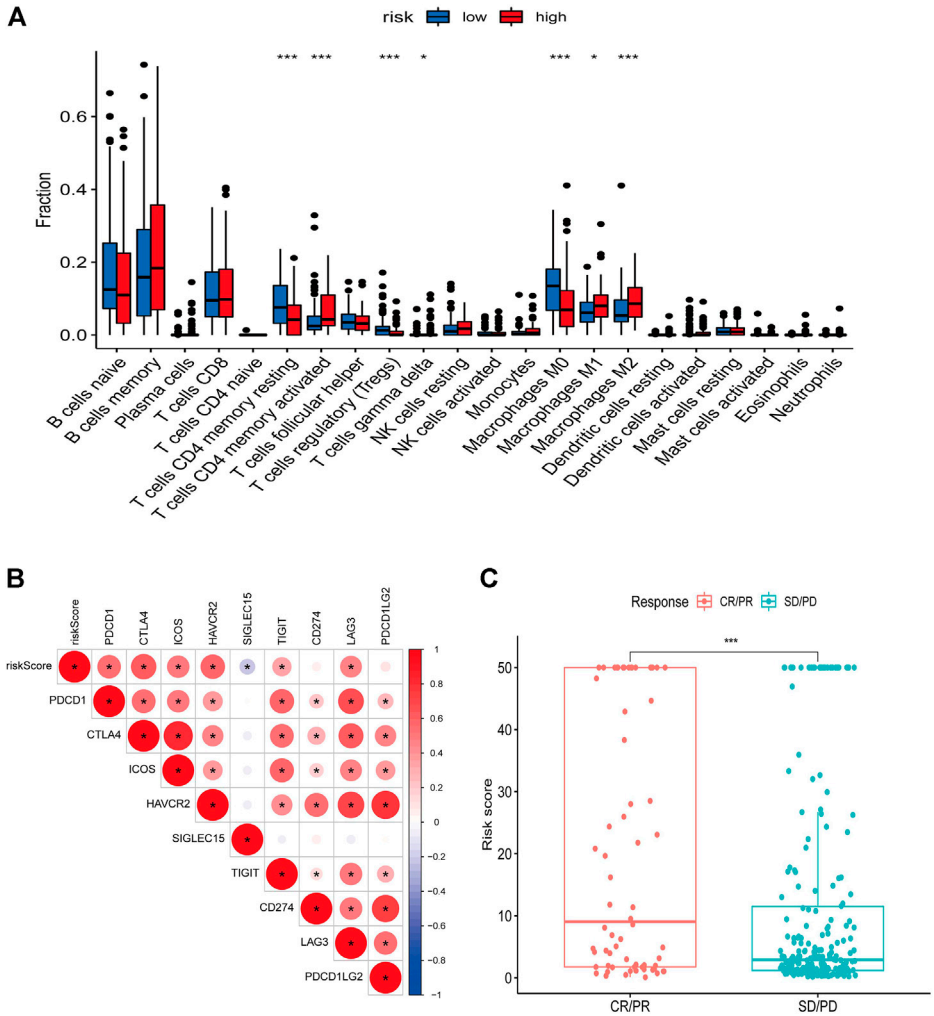


FIGURE 8 | Relationship between risk score and immune landscape. **(A)** Distribution of 22 immune cell types in high and low risk groups. **(B)** Correlation matrix heat map showing the correlation analysis of risk scores with immune checkpoint genes. **(C)** Boxplot showing the difference in the distribution of risk scores in different immunotherapy response groups.

the high-risk group, and pathways such as extracellular matrix receptor interactions, focal adhesion, gap linkage, pathways in cancer, and regulation of the actin cytoskeleton were enriched in the low-risk group (**Figures 7A,B**).

Immune Landscape and Response to Immunotherapy in Different Risk Groups

We used the ESTIMATE algorithm to assess the TME immune and stromal abundance in the different risk groups, and the results showed that the high-risk group had higher levels of immune and stromal component abundance (**Supplementary Figures S2A–C**). We also analyzed the proportion of 22 types of immune infiltrating cells among different risk groups in 481 DLBCL samples using the CIBERSORT algorithm (**Supplementary Figure S2D**), and the results showed that seven types of immune infiltrating cells were associated with risk scores: resting CD4 memory T cells, activated CD4 memory T cells, regulatory T cells, γδ T cells, and M0, M1, and M2 macrophages (**Figure 8A**). Correlation analysis of risk scores with immune checkpoint genes showed that most of the immune checkpoint gene expression levels were positively correlated with risk scores (**Figure 8B**). In addition, higher risk scores in the IMvigor210 immunotherapy cohort were associated with anti-PD-L1 treatment response (**Figure 8C**).

DISCUSSION

In this study, we combined scRNA-seq and bulk RNA-seq to investigate the tumor heterogeneity and TME characteristics of DLBCL. We showed the existence of malignant cell subpopulations with different transcriptional characteristics in DLBCL samples, such as a characteristic malignant cell subpopulation with predominantly cellular proliferation and a malignant cell subpopulation with predominantly metabolic characteristics. Roider T et al. investigated intra-tumor heterogeneity in B-NHL

at the level of drug response by scRNA-seq, with tumor subgroups in the same lymph node responding significantly differently to targeted and chemotherapeutic agents (Roider et al., 2020). This suggests that a rational combination of anticancer drugs is needed to target all tumor subgroups, especially those with proliferative and aggressive characteristics, to improve therapeutic response and avoid the development of tumor drug resistance.

Immunotherapy has become a major hot topic in oncology treatment research, and inhibitors targeting the PD1-PDL1 axis have been approved as second- or first-line therapies for an increasing number of types of malignancies, including melanoma, lymphoma, lung cancer, renal cell carcinoma, head and neck squamous cell carcinoma, bladder cancer, liver cancer, and gastroesophageal cancer. However, great progress has been made in clinical application, but most patients receiving immune checkpoint inhibitors (ICIs) have not benefited from them (Gong et al., 2018). ICIs have shown significant efficacy in relapsed/refractory classic Hodgkin's lymphoma (cHL), with an overall response rate (ORR) of 70–90% and have been approved for this indication (Ansell et al., 2015; Kasamon et al., 2017; Rossi et al., 2018). Unfortunately, ICIs are less effective in DLBCL, mainly due to its high biological heterogeneity. (Armand et al., 2013; Ansell et al., 2016; Lesokhin et al., 2016; Ansell et al., 2019; Frigault et al., 2020). By transcriptomic analysis of the microenvironment of multiple independent cohorts of DLBCL, Kotlov N et al. characterized four major lymphoma microenvironment (LME) categories associated with different biological abnormalities and clinical behaviors, namely GC-like, mesenchymal, inflammatory (IN), and depleted (DP) (Kotlov et al., 2021). Analysis of the correlation between LME category and response to chemoimmunotherapy showed that the number of responders was highest in GC-like patients and lowest in DP-LME patients. IN-LME is enriched in $CD8^+$ T cells and a subpopulation of $CD8^+$ T cells with high PD-1 expression and high expression of the immune checkpoint molecule PD-L1 and the tryptophanolytic enzyme IDO1, suggesting that this LME class may benefit from ICIs treatment. Steen CB et al. characterized clinically relevant DLBCL cell states and ecosystems with EcoTyper (a machine-learning framework integrating transcriptome deconvolution and single-cell RNA sequencing), identified 5 cell states of malignant B cells with different prognostic associations and differentiation status, and revealed nine multicellular ecosystems in DLBCL, known as lymphoma ecotypes (LE) (Steen et al., 2021). They found T-cell transcriptomic heterogeneity in DLBCL and that tumors high in LE4 are characterized by an immunoreactive T-cell state with widespread expression of co-inhibitory and stimulatory molecules, with potential implications for immunotherapeutic targeting. These studies suggest that exploring the heterogeneity of the DLBCL tumor microenvironment may better stratify patients to improve the efficacy of ICIs. Here, we identified seven different T cell subsets, $CD4^-CD8^-$Navie T, $CD4^+$ T_H, $CD8^+$ T_{TOX}, $CD4^+$ T_{REG}, $CD8^+$Navie T, T_{PRO}, and $CD8^+$ T_{EXH}. we found a significantly higher proportion of $CD4^+$ T_{REG} cells in DLBCL samples compared to reactive lymph node tissue. Recently, several studies have found that $CD4^+FOXP3^+$ T cells can be divided into three subpopulations: 1) effector Tregs (eTregs), which have a strong suppressive function; 2) naive Tregs, which have the potential to differentiate into eTregs upon antigen stimulation; and 3) non-Tregs, which are a non-suppressive subpopulation (Nishikawa and Sakaguchi 2014). Studies have shown that high infiltration of $FOXP3^+$ Tregs cells in DLBCL is associated with better prognosis, but these studies have targeted the entire FOXP3 population rather than the true Tregs cells (eTregs) that are essential for the impact of tumor immunity (Lee et al., 2008; Serag El-Dien et al., 2017). Nakayama S et al. found that high infiltration of FOXP3/CTLA-4 double-positive cells as eTregs was associated with a poorer prognosis (Nakayama et al., 2017). Recent animal studies with anti-CTLA-4 mAb using mice lacking antibody-dependent cytotoxic activity (by modulation of the Fc fraction or Fc receptor knockdown) showed that the anti-CTLA-4 mAb antitumor activity was attributed to depletion of $FOXP3^+CD4^+$ Treg cells from tumor tissue rather than direct activation of effector T cells (Bulliard et al., 2013; Selby et al., 2013; Simpson et al., 2013). Indeed, the reduction of $FOXP3^+CD4^+$ Treg cells in tumor tissue after anti-CTLA-4 mAb (Ipilimumab) treatment was strongly associated with clinical benefit (Hodi et al., 2008; Liakou et al., 2008). Furthermore, the critical role of CTLA-4 on $FOXP3^+CD4^+$ Treg cell function was revealed in animal studies, which showed that specific deletion of CTLA-4 in $FOXP3^+CD4^+$ Treg cells impairs their suppressive function and thus enhances antitumor immunity (Wing et al., 2008; Ise et al., 2010). Our single-cell analysis showed that the $CD4^+$ T_{REG} subpopulation (highly expressing FOXP3 and CTLA-4) in DLBCL showed highly immunosuppressive properties, attributed to the eTregs, suggesting that immunotherapy against eTregs could be an effective and novel treatment strategy for DLBCL patients with highly infiltrated FOXP3/CTLA-4 double-positive cells.

In addition to the classical immune checkpoint molecules PD-1 and CTLA-4, T cell immunoglobulin mucin receptor 3 (TIM3, or HAVCR2) and LAG-3 are also included in the field of tumor immunotherapy research. TIM-3 is a type I transmembrane protein that is expressed on T cells in a number of malignancies, including melanoma, lung cancer, hepatocellular carcinoma, and colon cancer. In these tumors, TIM-3 expression is usually associated with dysfunctional T cells and poorer prognosis in some tumor types (Anderson 2014). In hematologic malignancies, TIM-3 expression has been observed in adult T-cell leukemia/lymphoma and extranodal NK/T-cell lymphoma (Horlad et al., 2016; Feng et al., 2018). In addition, TIM-3 expression levels in DLBCL patients have been found to correlate with tumor stage and response to chemotherapy (Xiao et al., 2014; Zhang et al., 2015). LAG-3 is a member of the immunoglobulin superfamily and functions as a negative regulator of T cell homeostasis. LAG-3 has been shown to be expressed in tumor-infiltrating lymphocytes in a variety of tumor types, including breast, ovarian, and lung cancers, and is commonly associated with increased numbers of $PD\text{-}1^+$ T cells (Matsuzaki et al., 2010; Burugu et al., 2017; He et al., 2017). In follicular lymphoma, high expression of LAG-3 is associated with poorer patient prognosis and T-cell failure (Yang et al., 2017). Here, we characterized a population of $CD8^+$ T cells with high expression of LAG-3, TIM-3, TIGHT, i.e. exhausted cytotoxic $CD8^+$ T cells, which showed a molecular profile dominated by interferon response and retained the expression of GZMA, GZMB and NKG7. Furthermore, by SCENIC analysis, we revealed potential transcription factors, such

as STAT1 and IRF7, in the $CD8^+$ T_{EXH} cell subpopulation. Beltra JC et al. showed that in exhausted $CD8^+$ T cells are enriched with open chromatin regions that bind to STAT1 and IRF7 (Beltra et al., 2020), which is consistent with our findings.

We constructed prognostic models based on differential genes associated with $CD8^+$ T_{EXH} subpopulations obtained from previous differential gene expression analysis, and the efficacy of the prognostic models in predicting survival, and response to immunotherapy was validated by internal or external validation cohorts. This prognostic model could identify high-risk DLBCL patients and helped clinicians make better clinical decisions.

In conclusion, this study provides an in-depth dissection of the transcriptional features of malignant B cells and TME in DLBCL and provides new insights into the tumor heterogeneity of DLBCL. The data from our study can serve as a resource for subsequent in-depth studies to provide therapeutic targets and biomarkers for immunotherapy in DLBCL through deeper biological exploration. In addition, the prognostic model we developed can well predict the prognostic status and immunotherapeutic response of DLBCL patients with promising clinical applications.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

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

AUTHOR CONTRIBUTIONS

YZ designed this study. HX downloaded data. YZ analyzed data and wrote the initial version of the paper. MZ and LL revised manuscript.

ACKNOWLEDGMENTS

We thank all those who contributed to the article.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.881345/full#supplementary-material

REFERENCES

Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., et al. (2017). SCENIC: Single-Cell Regulatory Network Inference and Clustering. *Nat. Methods* 14 (11), 1083–1086. doi:10.1038/nmeth.4463

Anderson, A. C. (2014). Tim-3: an Emerging Target in the Cancer Immunotherapy Landscape. *Cancer Immunol. Res.* 2 (5), 393–398. doi:10.1158/2326-6066.CIR-14-0039

Ansell, S., Gutierrez, M. E., Shipp, M. A., Gladstone, D., Moskowitz, A., Borello, I., et al. (2016). A Phase 1 Study of Nivolumab in Combination with Ipilimumab for Relapsed or Refractory Hematologic Malignancies (Checkmate 039). *Blood* 128 (22), 183. doi:10.1182/blood.v128.22.183.183

Ansell, S. M., Lesokhin, A. M., Borrello, I., Halwani, A., Scott, E. C., Gutierrez, M., et al. (2015). PD-1 Blockade with Nivolumab in Relapsed or Refractory Hodgkin's Lymphoma. *N. Engl. J. Med.* 372 (4), 311–319. doi:10.1056/NEJMoa1411087

Ansell, S. M., Minnema, M. C., Johnson, P., Timmerman, J. M., Armand, P., Shipp, M. A., et al. (2019). Nivolumab for Relapsed/Refractory Diffuse Large B-Cell Lymphoma in Patients Ineligible for or Having Failed Autologous Transplantation: A Single-Arm, Phase II Study. *Jco* 37 (6), 481–489. doi:10.1200/JCO.18.00766

Ansell, S. M., and Vonderheide, R. H. (2013). Cellular Composition of the Tumor Microenvironment. *Am. Soc. Clin. Oncol. Educ. Book* 2013, 1. doi:10.14694/edbook_am.2013.33.e91

Armand, P., Nagler, A., Weller, E. A., Devine, S. M., Avigan, D. E., Chen, Y.-B., et al. (2013). Disabling Immune Tolerance by Programmed Death-1 Blockade with Pidilizumab after Autologous Hematopoietic Stem-Cell Transplantation for Diffuse Large B-Cell Lymphoma: Results of an International Phase II Trial. *Jco* 31 (33), 4199–4206. doi:10.1200/JCO.2012.48.3685

Beltra, J.-C., Manne, S., Abdel-Hakeem, M. S., Kurachi, M., Giles, J. R., Chen, Z., et al. (2020). Developmental Relationships of Four Exhausted CD8+ T Cell Subsets Reveals Underlying Transcriptional and Epigenetic Landscape Control Mechanisms. *Immunity* 52 (5), 825–841. e8. doi:10.1016/j.immuni.2020.04.014

Bulliard, Y., Jolicoeur, R., Windman, M., Rue, S. M., Ettenberg, S., Knee, D. A., et al. (2013). Activating Fc γ Receptors Contribute to the Antitumor Activities of Immunoregulatory Receptor-Targeting Antibodies. *J. Exp. Med.* 210 (9), 1685–1693. doi:10.1084/jem.20130573

Burugu, S., Asleh-Aburaya, K., and Nielsen, T. O. (2017). Immune Infiltrates in the Breast Cancer Microenvironment: Detection, Characterization and Clinical Implication. *Breast Cancer* 24 (1), 3–15. doi:10.1007/s12282-016-0698-z

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 (5), 411–420. doi:10.1038/nbt.4096

Coupland, S. E. (2011). The Challenge of the Microenvironment in B-Cell Lymphomas. *Histopathology* 58 (1), 69–80. doi:10.1111/j.1365-2559.2010.03706.x

Couzin-Frankel, J. (2013). Cancer Immunotherapy. *Science* 342 (6165), 1432–1433. doi:10.1126/science.342.6165.1432

Crump, M., Neelapu, S. S., Farooq, U., Van Den Neste, E., Kuruvilla, J., Westin, J., et al. (2017). Outcomes in Refractory Diffuse Large B-Cell Lymphoma: Results from the International SCHOLAR-1 Study. *Blood* 130 (16), 1800–1808. doi:10.1182/blood-2017-03-769620

Dwivedi, A., Karulkar, A., Ghosh, S., Rafiq, A., and Purwar, R. (2019). Lymphocytes in Cellular Therapy: Functional Regulation of CAR T Cells. *Front. Immunol.* 9, 3180. doi:10.3389/fimmu.2018.03180

Feng, Y., Zhong, M., Liu, Y., Wang, L., and Tang, Y. (2018). Expression of TIM-3 and LAG-3 in Extranodal NK/T Cell Lymphoma, Nasal Type. *Histol. Histopathol.* 33 (3), 307–315. doi:10.14670/HH-11-931

Friedberg, J. W. (2011). Relapsed/refractory Diffuse Large B-Cell Lymphoma. *Hematol. Am. Soc. Hematol. Educ. Program* 2011, 498–505. doi:10.1182/asheducation-2011.1.498

Friedberg, J. W. (2006). Rituximab for Early-Stage Diffuse Large-B-Cell Lymphoma. *Lancet Oncol.* 7 (5), 357–359. doi:10.1016/S1470-2045(06)70668-4

Frigault, M. J., Armand, P., Redd, R. A., Jeter, E., Merryman, R. W., Coleman, K. C., et al. (2020). PD-1 Blockade for Diffuse Large B-Cell Lymphoma after Autologous Stem Cell Transplantation. *Blood Adv.* 4 (1), 122–126. doi:10.1182/bloodadvances.2019000784

Gisselbrecht, C., Glass, B., Mounier, N., Singh Gill, D., Linch, D. C., Trneny, M., et al. (2010). Salvage Regimens with Autologous Transplantation for Relapsed Large B-Cell Lymphoma in the Rituximab Era. *Jco* 28 (27), 4184–4190. doi:10.1200/JCO.2010.28.1618

Gong, J., Chehrazi-Raffle, A., Reddi, S., and Salgia, R. (2018). Development of PD-1 and PD-L1 Inhibitors as a Form of Cancer Immunotherapy: a Comprehensive Review of Registration Trials and Future Considerations. *J. Immunother. cancer* 6 (1), 8. doi:10.1186/s40425-018-0316-z

He, Y., Yu, H., Rozeboom, L., Rivard, C. J., Ellison, K., Dziadziuszko, R., et al. (2017). LAG-3 Protein Expression in Non-small Cell Lung Cancer and its Relationship with PD-1/pd-L1 and Tumor-Infiltrating Lymphocytes. *J. Thorac. Oncol.* 12 (5), 814–823. doi:10.1016/j.jtho.2017.01.019

Hodi, F. S., Butler, M., Oble, D. A., Seiden, M. V., Haluska, F. G., Kruse, A., et al. (2008). Immunologic and Clinical Effects of Antibody Blockade of Cytotoxic T Lymphocyte-Associated Antigen 4 in Previously Vaccinated Cancer Patients. *Proc. Natl. Acad. Sci. U.S.A.* 105 (8), 3005–3010. doi:10.1073/pnas.0712237105

Horlad, H., Ohnishi, K., Ma, C., Fujiwara, Y., Niino, D., Ohshima, K., et al. (2016). TIM-3 Expression in Lymphoma Cells Predicts Chemoresistance in Patients with Adult T-Cell Leukemia/lymphoma. *Oncol. Lett.* 12 (2), 1519–1524. doi:10.3892/ol.2016.4774

Hui, L., and Chen, Y. (2015). Tumor Microenvironment: Sanctuary of the Devil. *Cancer Lett.* 368 (1), 7–13. doi:10.1016/j.canlet.2015.07.039

Ise, W., Kohyama, M., Nutsch, K. M., Lee, H. M., Suri, A., Unanue, E. R., et al. (2010). CTLA-4 Suppresses the Pathogenicity of Self Antigen-specific T Cells by Cell-Intrinsic and Cell-Extrinsic Mechanisms. *Nat. Immunol.* 11 (2), 129–135. doi:10.1038/ni.1835

Jin, S., Guerrero-Juarez, C. F., Zhang, L., Chang, I., Ramos, R., Kuan, C.-H., et al. (2021). Inference and Analysis of Cell-Cell Communication Using CellChat. *Nat. Commun.* 12 (1), 1088. doi:10.1038/s41467-021-21246-9

Kasamon, Y. L., de Claro, R. A., Wang, Y., Shen, Y. L., Farrell, A. T., and Pazdur, R. (2017). FDA Approval Summary: Nivolumab for the Treatment of Relapsed or Progressive Classical Hodgkin Lymphoma. *Oncol.* 22 (5), 585–591. doi:10.1634/theoncologist.2017-0004

Kotlov, N., Bagaev, A., Revuelta, M. V., Phillip, J. M., Cacciapuoti, M. T., Antysheva, Z., et al. (2021). Clinical and Biological Subtypes of B-Cell Lymphoma Revealed by Microenvironmental Signatures. *Cancer Discov.* 11 (6), 1468–1489. doi:10.1158/2159-8290.cd-20-0839

Lee, N.-R., Song, E.-K., Jang, K. Y., Choi, H. N., Moon, W. S., Kwon, K., et al. (2008). Prognostic Impact of Tumor Infiltrating FOXP3 Positive Regulatory T Cells in Diffuse Large B-Cell Lymphoma at Diagnosis. *Leukemia Lymphoma* 49 (2), 247–256. doi:10.1080/10428190701824536

Lesokhin, A. M., Ansell, S. M., Armand, P., Scott, E. C., Halwani, A., Gutierrez, M., et al. (2016). Nivolumab in Patients with Relapsed or Refractory Hematologic Malignancy: Preliminary Results of a Phase Ib Study. *Jco* 34 (23), 2698–2704. doi:10.1200/JCO.2015.65.9789

Li, J., Byrne, K. T., Yan, F., Yamazoe, T., Chen, Z., Baslan, T., et al. (2018). Tumor Cell-Intrinsic Factors Underlie Heterogeneity of Immune Cell Infiltration and Response to Immunotherapy. *Immunity* 49 (1), 178–193. e7. doi:10.1016/j.immuni.2018.06.006

Liakou, C. I., Kamat, A., Tang, D. N., Chen, H., Sun, J., Troncoso, P., et al. (2008). CTLA-4 Blockade Increases IFNγ-Producing CD4 + ICOS Hi Cells to Shift the Ratio of Effector to Regulatory T Cells in Cancer Patients. *Proc. Natl. Acad. Sci. U.S.A.* 105 (39), 14987–14992. doi:10.1073/pnas.0806075105

Matsuzaki, J., Gnjatic, S., Mhawech-Fauceglia, P., Beck, A., Miller, A., Tsuji, T., et al. (2010). Tumor-infiltrating NY-ESO-1-specific CD8 + T Cells Are Negatively Regulated by LAG-3 and PD-1 in Human Ovarian Cancer. *Proc. Natl. Acad. Sci. U.S.A.* 107 (17), 7875–7880. doi:10.1073/pnas.1003345107

Nakayama, S., Yokote, T., Akioka, T., Hiraoka, N., Nishiwaki, U., Miyoshi, T., et al. (2017). Infiltration of Effector Regulatory T Cells Predicts Poor Prognosis of Diffuse Large B-Cell Lymphoma, Not Otherwise Specified. *Blood Adv.* 1 (8), 486–493. doi:10.1182/bloodadvances.2016000885

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust Enumeration of Cell Subsets from Tissue Expression Profiles. *Nat. Methods* 12 (5), 453–457. doi:10.1038/nmeth.3337

Nishikawa, H., and Sakaguchi, S. (2014). Regulatory T Cells in Cancer Immunotherapy. *Curr. Opin. Immunol.* 27, 1–7. doi:10.1016/j.coi.2013.12.005

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A., et al. (2017). Reversed Graph Embedding Resolves Complex Single-Cell Trajectories. *Nat. Methods* 14 (10), 979–982. doi:10.1038/nmeth.4402

Reddy, A., Zhang, J., Davis, N. S., Moffitt, A. B., Love, C. L., Waldrop, A., et al. (2017). Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma. *Cell* 171 (2), 481–494. e15. doi:10.1016/j.cell.2017.09.027

Roider, T., Seufert, J., Uvarovskii, A., Frauhammer, F., Bordas, M., Abedpour, N., et al. (2020). Dissecting Intratumour Heterogeneity of Nodal B-Cell Lymphomas at the Transcriptional, Genetic and Drug-Response Levels. *Nat. Cell Biol.* 22 (7), 896–906. doi:10.1038/s41556-020-0532-x

Rossi, C., Gilhodes, J., Maerevoet, M., Herbaux, C., Morschhauser, F., Brice, P., et al. (2018). Efficacy of Chemotherapy or Chemo-Anti-PD-1 Combination after Failed Anti-PD-1 Therapy for Relapsed and Refractory Hodgkin Lymphoma: A Series from Lysa Centers. *Am. J. Hematol.* 2018, 1. doi:10.1002/ajh.25154

Scott, D. W., Wright, G. W., Williams, P. M., Lih, C.-J., Walsh, W., Jaffe, E. S., et al. (2014). Determining Cell-Of-Origin Subtypes of Diffuse Large B-Cell Lymphoma Using Gene Expression in Formalin-Fixed Paraffin-Embedded Tissue. *Blood* 123 (8), 1214–1217. doi:10.1182/blood-2013-11-536433

Selby, M. J., Engelhardt, J. J., Quigley, M., Henning, K. A., Chen, T., Srinivasan, M., et al. (2013). Anti-CTLA-4 Antibodies of IgG2a Isotype Enhance Antitumor Activity through Reduction of Intratumoral Regulatory T Cells. *Cancer Immunol. Res.* 1 (1), 32–42. doi:10.1158/2326-6066.cir-13-0013

Serag El-Dien, M. M., Abdou, A. G., Asaad, N. Y., Abd El-Wahed, M. M., and Kora, M. A. E.-H. M. (2017). Intratumoral FOXP3+ Regulatory T Cells in Diffuse Large B-Cell Lymphoma. *Appl. Immunohistochem. Mol. Morphol.* 25 (8), 534–542. doi:10.1097/PAI.0000000000000335

Shen, M., and Kang, Y. (2018). Complex Interplay between Tumor Microenvironment and Cancer Therapy. *Front. Med.* 12 (4), 426–439. doi:10.1007/s11684-018-0663-7

Simpson, T. R., Li, F., Montalvo-Ortiz, W., Sepulveda, M. A., Bergerhoff, K., Arce, F., et al. (2013). Fc-dependent Depletion of Tumor-Infiltrating Regulatory T Cells Co-defines the Efficacy of Anti-CTLA-4 Therapy against Melanoma. *J. Exp. Med.* 210 (9), 1695–1710. doi:10.1084/jem.20130579

Steen, C. B., Luca, B. A., Esfahani, M. S., Azizi, A., Sworder, B. J., Nabet, B. Y., et al. (2021). The Landscape of Tumor Cell States and Ecosystems in Diffuse Large B Cell Lymphoma. *Cancer cell* 39 (10), 1422–1437. e10. doi:10.1016/j.ccell.2021.08.011

Wing, K., Onishi, Y., Prieto-Martin, P., Yamaguchi, T., Miyara, M., Fehervari, Z., et al. (2008). CTLA-4 Control over Foxp3 + Regulatory T Cell Function. *Science* 322 (5899), 271–275. doi:10.1126/science.1160062

Xiao, T., Zhang, L., Chen, L., Liu, G., Feng, Z., and Gao, L. (2014). Tim-3 Expression Is Increased on Peripheral T Cells from Diffuse Large B Cell Lymphoma. *Tumor Biol.* 35 (8), 7951–7956. doi:10.1007/s13277-014-2080-0

Yang, Z.-Z., Kim, H. J., Villasboas, J. C., Chen, Y.-P., Price-Troska, T., Jalali, S., et al. (2017). Expression of LAG-3 Defines Exhaustion of Intratumoral PD-1+ T Cells and Correlates with Poor Outcome in Follicular Lymphoma. *Oncotarget* 8 (37), 61425–61439. doi:10.18632/oncotarget.18251

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

Zhang, L., Du, H., Xiao, T.-w., Liu, J.-z., Liu, G.-z., Wang, J.-x., et al. (2015). Prognostic Value of PD-1 and TIM-3 on CD3+ T Cells from Diffuse Large B-Cell Lymphoma. *Biomed. Pharmacother.* 75, 83–87. doi:10.1016/j.biopha.2015.08.037

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 | Frontiers in Genetics

TYPE Original Research
PUBLISHED 22 July 2022
DOI 10.3389/fgene.2022.938611

OPEN ACCESS

EDITED BY
Geng Chen,
GeneCast Biotechnology Co., Ltd., China

REVIEWED BY
Qingjia Chi,
Wuhan University of Technology, China
Ting Li,
National Center for Toxicological Research (FDA), United States

*CORRESPONDENCE
Jian Zhang,
zj301doctor@126.com

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

RECEIVED 07 May 2022
ACCEPTED 04 July 2022
PUBLISHED 22 July 2022

CITATION
Zhang S, Zhang W and Zhang J (2022), 8-Gene signature related to CD8+ T cell infiltration by integrating single-cell and bulk RNA-sequencing in head and neck squamous cell carcinoma.
Front. Genet. 13:938611.
doi: 10.3389/fgene.2022.938611

8-Gene signature related to CD8$^+$ T cell infiltration by integrating single-cell and bulk RNA-sequencing in head and neck squamous cell carcinoma

Shoujing Zhang[1], Wenyi Zhang[2] and Jian Zhang[1]*

[1]Department of Oral and Maxillofacial Surgery, Tianjin Medical University School and Hospital of Stomatology, Tianjin, China, [2]Department of Prosthodontics, Tianjin Medical University School and Hospital of Stomatology, Tianjin, China

Background: CD8$^+$ T cells, a critical component of the tumor immune microenvironment, have become a key target of cancer immunotherapy. Considering the deficiency of robust biomarkers for head and neck squamous cell carcinoma (HNSCC), this study aimed at establishing a molecular signature associated with CD8+T cells infiltration.

Methods: Single-cell RNA sequencing data retrieved from the Gene Expression Omnibus (GEO) database was analyzed to obtain the different cell types. Next, the cell proportions were investigated through deconvolution of RNA sequencing in the Cancer Genome Atlas (TCGA) database, and then the immune-related genes (IRGs) were identified by weighted gene co-expression network analysis (WGCNA). LASSO-Cox analysis was employed to establish a gene signature, followed by validation using a GEO dataset. Finally, the molecular and immunological properties, and drug responses between two subgroups were explored by applying "CIBERSORT", "ESTIMATE", and single sample gene set enrichment analysis (ssGSEA) methods.

Results: A total of 215 differentially expressed IRGs were identified, of which 45 were associated with the overall survival of HNSCC. A risk model was then established based on eight genes, including *DEFB1*, *AICDA*, *TYK2*, *CCR7*, *SCARB1*, *ULBP2*, *STC2*, and *LGR5*. The low-risk group presented higher infiltration of memory activated CD4$^+$ T cells, CD8$^+$ T cells, and plasma cells, as well as a higher immune score, suggesting that they could benefit more from immunotherapy. On the other hand, the high-risk group showed higher abundance of activated mast cells and M2 macrophages, as well as a lower immune score.

Conclusion: It was evident that the 8-gene signature could accurately predict HNSCC prognosis and thus it may serve as an index for clinical treatment.

KEYWORDS

CD8$^+$ T cells, head and neck squamous cell carcinoma, immunotherapy, prognosis, weighted gene co-expression network analysis

Introduction

Head and neck squamous cell carcinoma (HNSCC) is the seventh most common malignancy worldwide (Siegel et al., 2020). Despite the effective and aggressive treatment strategies involving surgery combined with radio- and chemotherapy, patients with advanced stage HNSCC only have a 50% five-year survival rate (Vigneswaran and Williams, 2014). In recent years, immunotherapy involving checkpoint inhibitors blocking programmed cell death protein 1 (*PD-1*) or programmed death ligand-1 (*PD-L1*) has been approved for clinical use, with preliminary results showing that the strategy significantly improves the overall survival of recurrent or metastatic HNSCC patients. However, several clinical trials have demonstrated that anti-*PD-1*/*PD-L1* therapy is only beneficial to a few patients (Ferris et al., 2016; Siu et al., 2019). Studies have suggested that $CD8^+$ T lymphocytes substantially express *PD-1* and may play an important role in the efficacy of immunotherapy (Jia et al., 2020). It is worth noting that high dense infiltration of $CD8^+$ T cells in HNSCC patients is generally associated with a good prognosis (Fridman et al., 2017). Moreover, PD-1+ $CD8^+$ T cells showed excellent anti-tumor effect in an anti-PD1-resistant murine HNSCC model (Xu et al., 2020). Therefore, there is an urgent need to explore the molecular mechanisms associated with $CD8^+$ T cells infiltration.

Single-cell RNA sequencing (scRNA-seq) has been the subject of rapid technological developments in the last decade, thereby resulting in significant improvements in describing and defining the tumor heterogeneity at a single-cell level (Qi et al., 2019). Besides, application of scRNA-seq to characterize the tumor microenvironment (TME) may provide valuable insights into immune landscapes and even effective immunotherapy strategies (Kurten et al., 2021). Similarly, the gene signature identified based on immune molecular characteristics might be a strong predictor of clinical outcome and immunotherapy response (Song et al., 2022). However, the predictive potential of the molecular mechanisms describing immunophenotypic features in HNSCC have not yet been elucidated.

This study explored the mechanism associated with infiltration of $CD8^+$ T cells through integrating bulk and scRNA sequencing. Specifically, a LASSO-Cox regression risk model was built and verified based on the hub immune-related genes (IRGs) identified by weighted gene co-expression network analysis (WGCNA) (Langfelder and Horvath, 2008). Next, we comprehensively represented the various immune features of an 8-gene signature using "ESTIMATE" (Yoshihara et al., 2013), "CIBERSORT" (Newman et al., 2015), single sample gene set enrichment analysis (ssGSEA) approaches, and immunophenoscore (IPS) data. It is expected that the identified risk score will not only be used as an efficient indicator for HNSCC prognosis, but also as a potential therapeutic target.

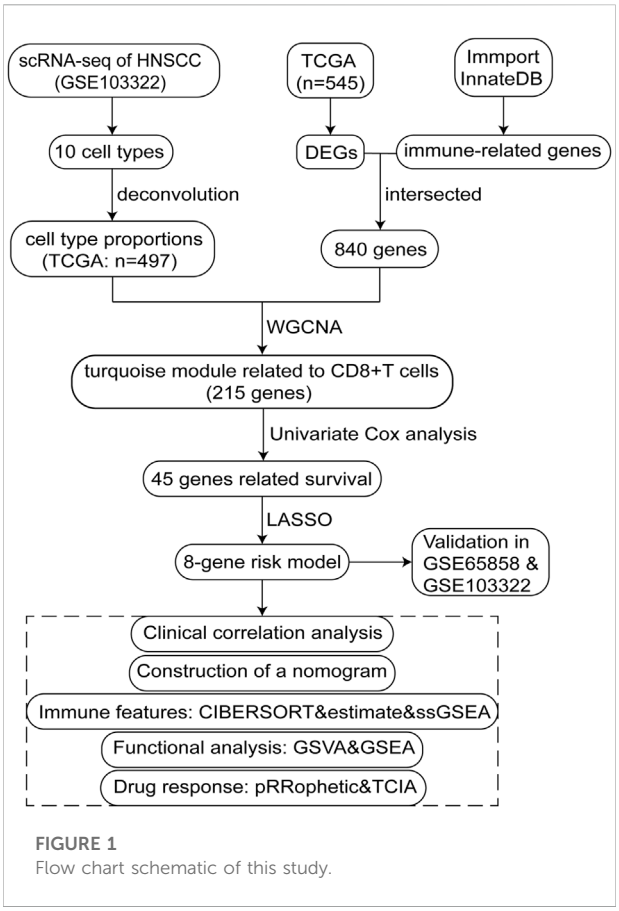


FIGURE 1
Flow chart schematic of this study.

Materials and methods

The study design is illustrated using a flow diagram (Figure 1).

Data acquisition

The single cell RNA-sequencing profile of GSE103322 dataset (Puram et al., 2017), comprising 5,902 single cells of 18 patients, was downloaded from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/) (accessed date 13 October 2021). HNSCC RNA-sequencing, clinical and mutation data were downloaded from The Cancer Genome Atlas (TCGA) database using the GDC Data Portal (https://portal.gdc.cancer.gov/ (accessed date 13 October 2021). The Fragments per Kilobase per Million (FPKM) values were first converted to transcripts per million kilobase

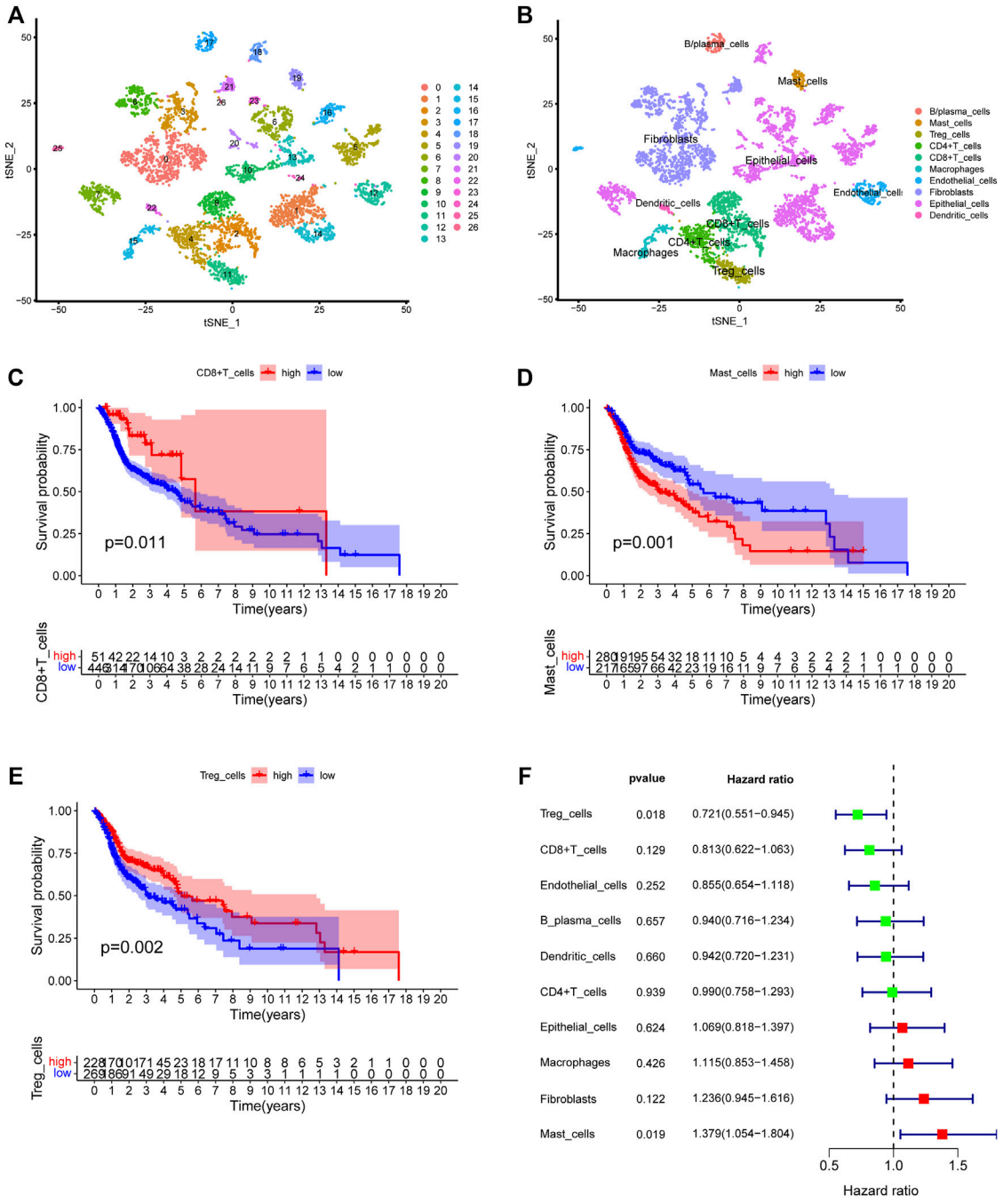


FIGURE 2
Identification of the HNSCC-associated cell subtypes. **(A)** t-SNE plot classified cell clusters based on scRNA sequencing data. **(B)** t-SNE plot identified the various cell subtypes. **(C–E)** Kaplan-Meier survival analysis of three cell subtypes using the deconvolved TCGA data. **(C)** $CD8^+$ T cells: $p = 0.011$, **(D)** Mast cells: $p = 0.001$, **(E)** Treg cells: $p = 0.002$. **(F)** Univariate analysis of ten cell subtypes.

(TPM) values. To validate the prognostic power of the model, the transcriptome and clinical files of the GSE65858 dataset, containing 270 HNSCC samples, were obtained from the GEO database (Wichmann et al., 2015). Notably, a total of 2,720 IRGs were obtained from the ImmPort (https://www.immport.org/home) and InnateDB (https://www.innatedb.com/) databases (accessed date 13 October 2021).

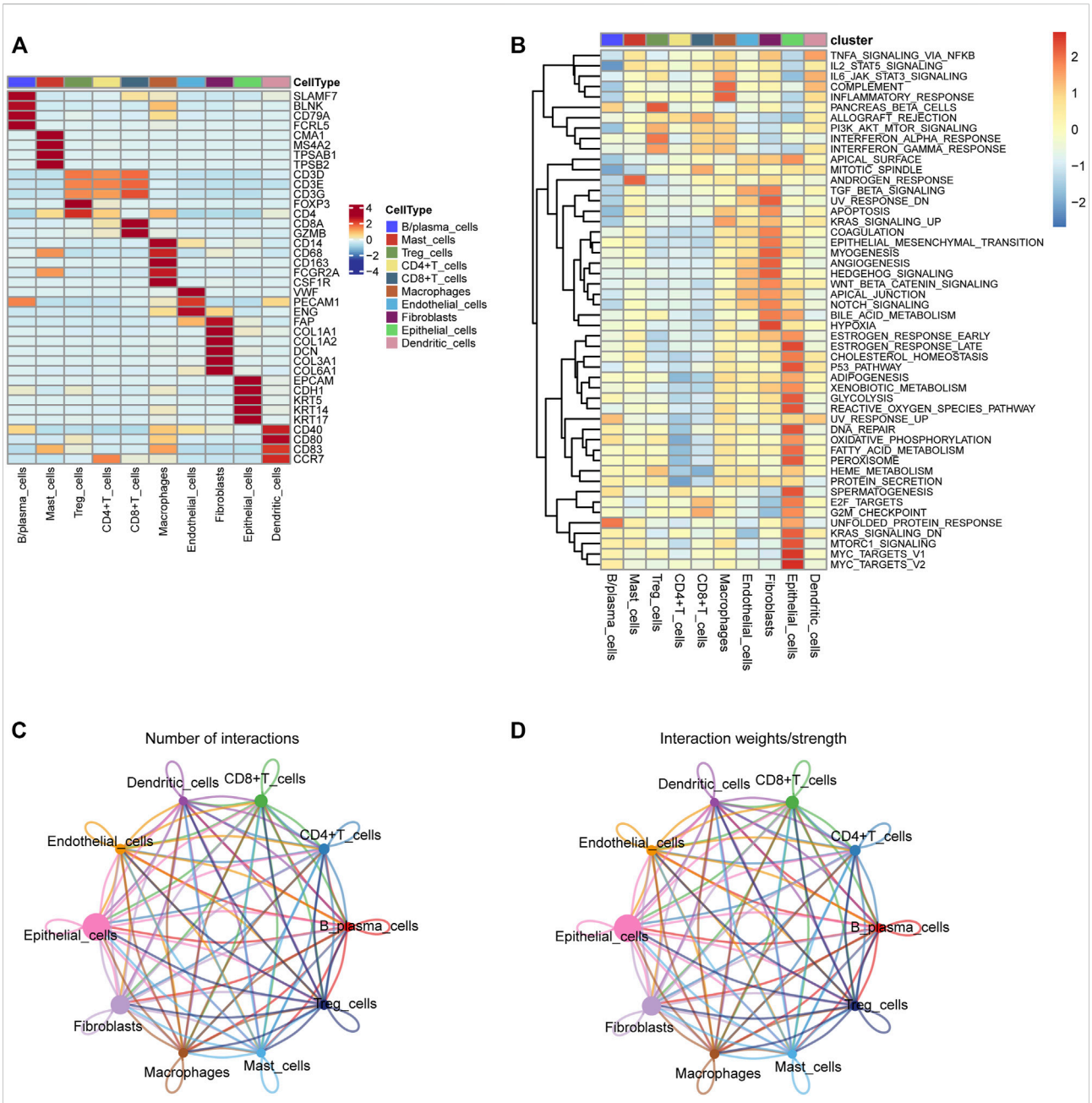


FIGURE 3
(A) The heatmap depicting marker genes associated with ten cell subtypes. **(B)** GSVA enrichment analysis of the cell subtypes. **(C,D)** Cell-cell communication network of ten cell subtypes.

Processing of single-cell and bulk RNA-seq files

The "Seurat" (version 4.1.1) package in R (version 4.1.2) was applied to group 5,902 cells into appropriate clusters, with the resolution set to 0.8. Results were presented by employing the T-distributed stochastic neighbor embedding (t-SNE) for dimension reduction. Next, diverse cell types, B/plasma cells, endothelial cells, regulatory T cells (Treg cells), mast cells, $CD8^+$ T cells, epithelial cells, dendritic cells, macrophages, fibroblasts, and $CD4^+$ T cells were identified based on their specific markers. The "Cellchat" (version 1.1.3) package was used to analyze the cell–cell communication, and then deconvolution was performed using the "BisqueRNA"

(version 1.0.5) method (Jew et al., 2020) to calculate the cells fractions of TCGA bulk profiles. Based on the TCGA RNA-seq profiles, differentially expressed genes (DEGs) were identified with FDR < 0.05 and $|\log_2 FC| > 1$ set as the cutoff values.

Determination of immune-related candidate genes

The differential IRGs were determined by overlapping DEGs and IRGs, and then used to screen the hub genes by WGCNA (version 1.7.0). First, Pearson correlation coefficient was determined for every gene, and a suitable soft threshold β was automatically selected through the pick Soft Threshold function. Next, gene expression similarity matrix was transformed into an adjacency matrix using a network type of signed and soft powers $\beta = 3$, followed by employing TOM (topological overlap measure) to cluster genes into network modules. The 1-TOM (dissimilarity TOM) was then applied as the input for hierarchical clustering and the "DynamicTreeCut" algorithm was employed to detect modules (clusters of highly interconnected genes) as branches of the dendrogram. Finally, we identified and selected a module (215 genes) that significantly correlated with $CD8^+$ T cells content. Kaplan–Meier (KM) survival and univariate Cox analysis were utilized to determine the hub genes associated with survival at a threshold of $p < 0.05$.

Development of a prognostic signature in TCGA (n = 498)

LASSO-Cox analysis was performed using "glmnet" package to determine the optimal prognostic gene set. The risk score of each HNSCC patient was determined as the sum of normalized gene expression values weighted by their LASSO-Cox coefficients in accordance with the following formula:

$$\text{risk score} = \sum_{i=1}^{n} Coef_i^* Exp_i$$

Where $Coef_i$ indicates the calculated regression coefficient of each gene in the LASSO-Cox model and Exp_i represents the mRNA expression value. Kaplan-Meier (KM) analysis, receiver operating characteristic (ROC) curves, and univariate and multivariate Cox regression analyses were employed to validate the independent prognostic factors in TCGA-HNSC and GSE65858 datasets. For better clinical prediction of HNSCC patient survival probabilities, a nomogram was constructed using the "rms" R package based on multivariate Cox analysis results. The concordance index (C-index) of the nomogram was calculated to assess the discriminative ability.

Immune features and therapy prediction in distinct risk groups

"CIBERSORT" (version 1.03) and "ESTIMATE" (version 1.0.13) analyses were applied to determine the abundance of 22 immune cells and immune infiltration scores. The ssGSEA approach was employed via the "GSVA" (version 1.42.0) package to compute the enrichment scores of 29 immune features (Hänzelmann et al., 2013). To predict the susceptibility of eight common chemotherapeutic drugs (5-Fluorouracil, bleomycin, cetuximab, cisplatin, docetaxel, methotrexate, rapamycin, and sunitinib) for HNSCC, the "pRRophetic" (version 0.5) method was performed to evaluate the half-maximal inhibitory concentration (IC50) of patients in distinct groups (Geeleher et al., 2014). The immunophenoscore (IPS) of HNSCC patients, which is a scoring scheme that characterizes the determinants of tumor immunogenicity (Charoentong et al., 2017), were downloaded from The Cancer Imaging Archive (TCIA) database (https://tcia.at/home, accessed date 15 November 2021). To predict the anti-CTLA4 and anti-PD1 responses, patients with different IPS were further compared between the two risk groups. Finally, the "Maftools" (version 2.10.05) (Mayakonda et al., 2018) package was used to determine the tumor mutational burden (TMB) and identify the driver genes.

Enrichment analysis

The reference gene sets of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway (c2. cp.kegg.v7.5.1. symbols.gmt) were obtained from the MSigDB database (https://www.gsea-msigdb.org/gsea/msigdb, accessed date 15 November 2021). GSEA software (version 4.2.3) and Gene Set Variation Analysis (GSVA) were conducted to determine the KEGG pathways with FDR < 0.05.

Results

Cell typing in head and neck squamous cell carcinoma scRNA-seq and deconvolution in the Cancer Genome Atlas-HNSC

We first collected the Smart-seq2 profile data of 5,902 cells in the GSE103322 dataset. Principal component analysis (PCA) and t-SNE analysis identified 27 cell clusters (Figure 2A). According to expressions of marker genes, 10 distinct cell clusters were identified, including $CD8^+$ T cells, macrophages, $CD4^+$ T cells, fibroblasts, endothelial cells, B/plasma cells, mast cells, Treg cells, epithelial cells, and dendritic cells (Figures 2B, 3A). GSVA results showed that "MYC_TARGETS_V2" and "MYC_TARGETS_V1"

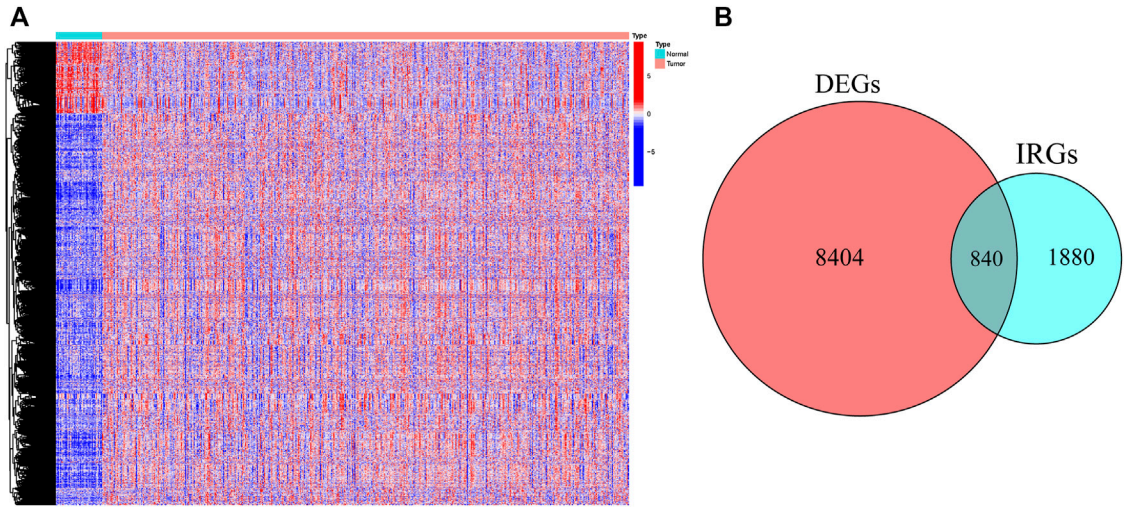


FIGURE 4
The heatmap **(A)** and Venn diagram **(B)** identified the differentially expressed genes (DEGs) and immune-related DEGs between tumor and normal samples in TCGA.

were activated in epithelial cells, whereas "HYPOXIA" was abundant in fibroblasts (Figure 3B). Results obtained after applying the "CellChat" method showed that there was a strong connectivity between different cell types (Figures 3C,D). Next, the BisqueRNA approach was performed to calculate proportions of the 10 cell types by deconvoluting the TCGA bulk profiles. Supplementary Table S1 shows proportion of the 10 cell types in 497 samples. Survival analysis demonstrated that mast cells (p = 0.001), $CD8^+$ T cells (p = 0.011), and Treg cells (p = 0.002) were significantly associated with HNSCC outcome (Figures 2C–E). Moreover, univariate Cox analysis indicated that Treg cells were associated with good outcome (p = 0.018), whereas mast cells were intimately linked to poor prognosis (p = 0.019) (Figure 2F).

Construction and validation of a gene risk signature associated with $CD8^+$ T cells

First, 9,244 DEGs were obtained from the TCGA-HNSC dataset comprising 501 tumor and 44 normal samples (Figure 4A). Subsequently, 2,720 IRGs from ImmPort and InnateDB databases were matched with DEGs, from which 840 differentially expressed IRGs were obtained for further analysis (Figure 4B). Based on the 840 IRGs and proportions of the 10 cell types in TCGA, the weighted gene co-expression network was generated using the soft-thresholding power β = 3, which resulted in identification of 10 modules (Figures 5A,B). To further explore the features of $CD8^+$ T cells infiltration, we selected the turquoise module (215 genes) which had the strongest correlation with $CD8^+$ T cells (r = 0.86, p = 1e-17). Univariate Cox analysis demonstrated that 45 of the 215 hub genes were closely associated with HNSCC survival (Figure 5C). Therefore, the 45 genes were subjected to LASSO regression analysis to identify the optimal penalty coefficient (Figures 5D,E). The survival analysis identified eight genes, including *DEFB1*, *AICDA*, *TYK2*, *CCR7*, *SCARB1*, *ULBP2*, *STC2*, and *LGR5*, which were significantly associated with HNSCC prognosis (Figures 6A–H). The eight risk regression coefficients were then employed to compute individual risk score of HNSCC patients according to the following formula:

$$\begin{aligned}\text{Risk score} = {} & (-0.097)^{*}\text{DEFB1} + (-0.444)^{*}\text{AICDA} \\ & + (-0.175)^{*}\text{TYK2} + (-0.071)^{*}\text{CCR7} \\ & + 0.020^{*}\text{SCARB1} + 0.079^{*}\text{ULBP2} + 0.161^{*}\text{STC2} \\ & + (-0.128)^{*}\text{LGR5}\end{aligned}$$

Next, the 498 HNSCC patients were stratified into high- and low-risk groups based on the median risk score. KM survival analysis results indicated that the high-risk group patients showed poorer outcomes compared to the low-risk group (p < 0.001, Figure 6I). Consistently, similar results were observed in the GSE65858 dataset (p = 0.005, Figure 6J).

Validation in the Cancer Genome Atlas-HNSC and GSE65858 cohorts, and scRNA-seq data

The risk score, survival status distributions of HNSCC patients, and correlation analysis are displayed in Figure 7. Results demonstrated that survival reduced with rising risk score, and there was a significant correlation between risk

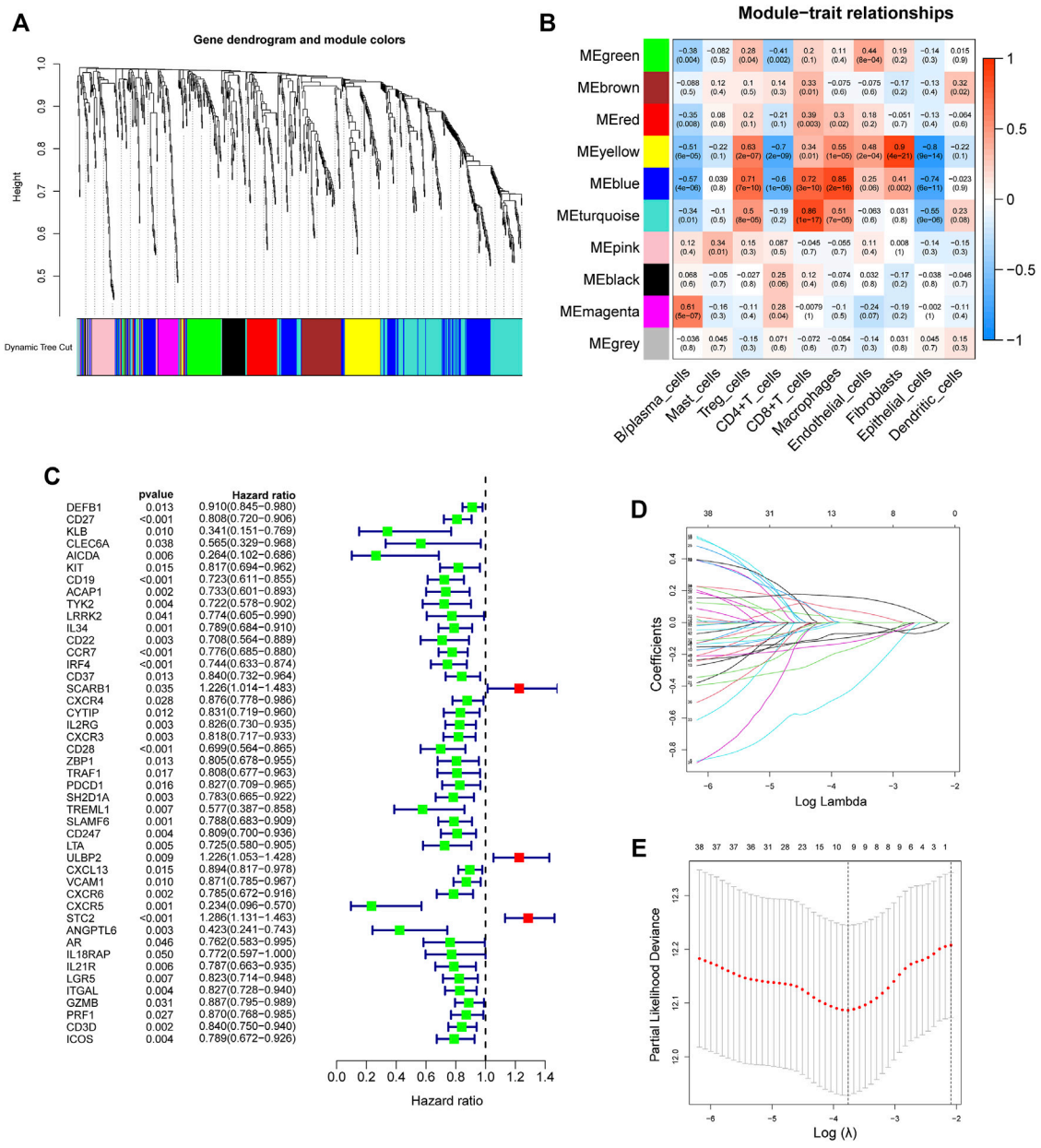


FIGURE 5
Development of an 8-gene signature. **(A)** The Cluster dendrogram of co-expression network modules obtained by WGCNA. **(B)** Correlation heatmap among ten co-expression modules and the levels of cell subtypes. The turquoise module had the greatest correlation with $CD8^+$ T cells (r = 0.86, p = 1e-17). **(C)** Univariate analysis of 45 immune-related hub genes. **(D)** LASSO coefficient profiles of 45 immune-related genes. **(E)** Tuning parameter selection in the LASSO model using ten-time cross-validation.

score and survival in TCGA cohort (r = -0.2, p = 6.4e-06). Time-dependent ROC and calibration curves at one-, three-, and five-years were then constructed (Figures 8A,B). In the TCGA cohort, the areas under the ROC curves (AUCs) were 0.679, 0.703, and 0.644 for 1-, 3-, and 5-years survival, respectively. In both the TCGA and GSE65858 cohorts, univariate and multivariate Cox analyses demonstrated that the risk score was an independent predictor for prognosis (Figures 8C–F). To determine the cells that these eight genes were enriched, the distribution plots for expressions of the eight genes in the 10 cell types identified in the GSE103322 dataset were generated and are shown in Figures 9A–I. Results showed that the expression levels of *DEFB1* and *ULBP2* were higher in epithelial cells, whereas *TYK2* and *CCR7* levels were abundant in dendritic cells. In addition, the endothelial cells had higher expressions of *SCARB1* and *STC2*, and *LGR5* was highly expressed in both dendritic cells and

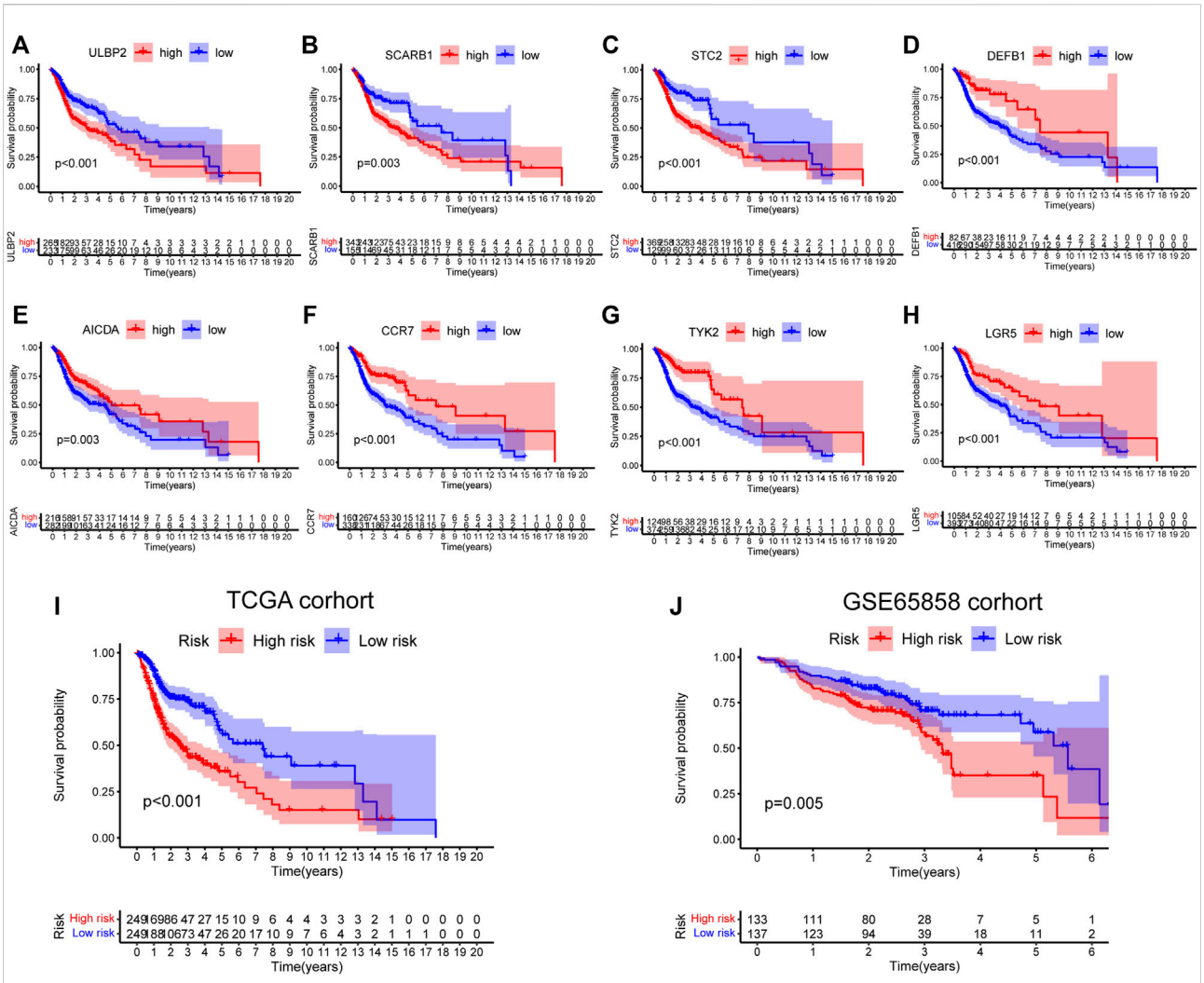


FIGURE 6
Survival analysis of eight genes in risk signature. **(A)** ULBP2: $p < 0.001$. **(B)** SCARB1: $p = 0.003$. **(C)** STC2: $p < 0.001$. **(D)** DEFB1: $p < 0.001$. **(E)** AICDA: $p = 0.003$. **(F)** CCR7: $p < 0.001$. **(G)** TYK2: $p < 0.001$. **(H)** LGR5: $p < 0.001$. **(I,J)** Survival analysis of the 8-gene signature in TCGA ($p < 0.001$) and GSE65858 cohorts ($p = 0.005$).

fibroblasts. Based on proportions of the 10 cell types obtained after deconvolution, correlation analysis was performed to evaluate the association among proportion of $CD8^+$ T cells and risk score. Obtained results revealed that fractions of $CD8^+$ T cells declined as the risk score increased ($r = -0.41$, $p < 2.2$ e-16, Figures 9J,K).

Construction of a nomogram for clinical practice

A heatmap was generated to depict the changes in expression of the eight genes between different clinical subgroups (Figure 10A). The performance of the risk score was then explored in different clinicopathological subgroups, including clinical stage (stage I-III and stage IV), age (<=60 and >60), grade (G1-2 and G3-4), T stage (T0-2 and T3-4), N stage (N0-1 and N2-3), and gender (female and male). According to the survival analysis results, HNSCC patients with high-risk scores consistently had a poorer outcome in all subgroups (Figures 10B–G). Next, the three remarkable variables in the multivariate analysis, including age, N stage, and risk score, were selected and used to build a nomogram (C-index: 0.676) for estimating the 1-, 3-, and 5-year survival rate (Figure 11A). By drawing a vertical line to the axis points, we could estimate patient survival based on total points. Overall, the calibration curves and the AUC's (1-, 3-, and 5-year: 0.733, 0.749, and 0.691, respectively) suggested that the risk model could accurately predict the HNSCC survival rate (Figures 11B,C).

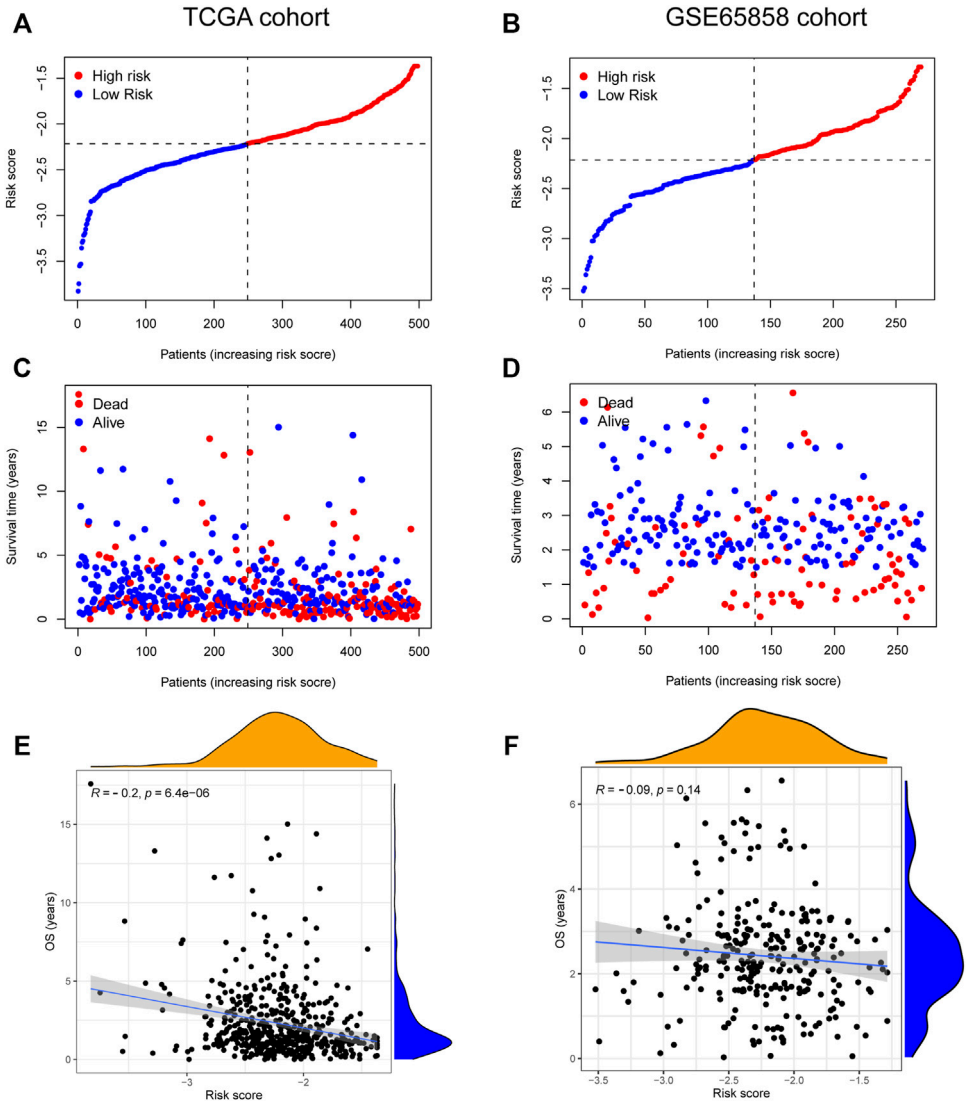


FIGURE 7
The relationship between risk score and HNSCC survival. **(A–D)** Distribution of risk score and survival status of 8-gene signature in TCGA **(A,C)** and GSE65858 **(B,D)** cohorts. **(E,F)** The correlation analysis between overall survival (OS) and risk score in TCGA **(E)** and GSE65858 **(F)** cohorts.

The immune landscape of the two risk groups

To elucidate the biological characteristics activated in distinct risk groups, KEGG pathway enrichment analysis was performed using GSVA and GSEA methods. By setting the adjusted *p* value (FDR) < 0.05, a total of 51 and 16 pathways were obtained in GSVA and GSEA, respectively (Figures 12A,B). Several overlapping immunoregulatory processes were enhanced in the low-risk group, including "hematopoietic cell lineage", "T cell receptor signaling pathway", "antigen processing and presentation" and "natural killer cell-mediated cytotoxicity". To describe the patterns of immune infiltrations, CIBERSORT and ESTIMATE methods were implemented for calculating the cell fractions and immune-related scores of HNSCC samples (Figures 13A,B). The low-risk group showed more significant infiltrations of $CD8^+$ T cells, M1 macrophages, follicular helper T cells, plasma cells, regulatory T cells, and memory activated $CD4^+$ T cells, as well as a higher immune score. With regard to the high-risk group, abundant infiltrations of activated mast cells, M2 macrophages, resting NK cells, and low immune score were observed. The ssGSEA approach was then applied to estimate the scores of specific immune functions and cells. Results revealed significant differences of most immune cells and functions between high- and low-risk groups (Figure 13C). Besides, 15 immune checkpoint molecules (IFNG, GZMB, HAVCR2, CD274, CD8A, PDCD1,

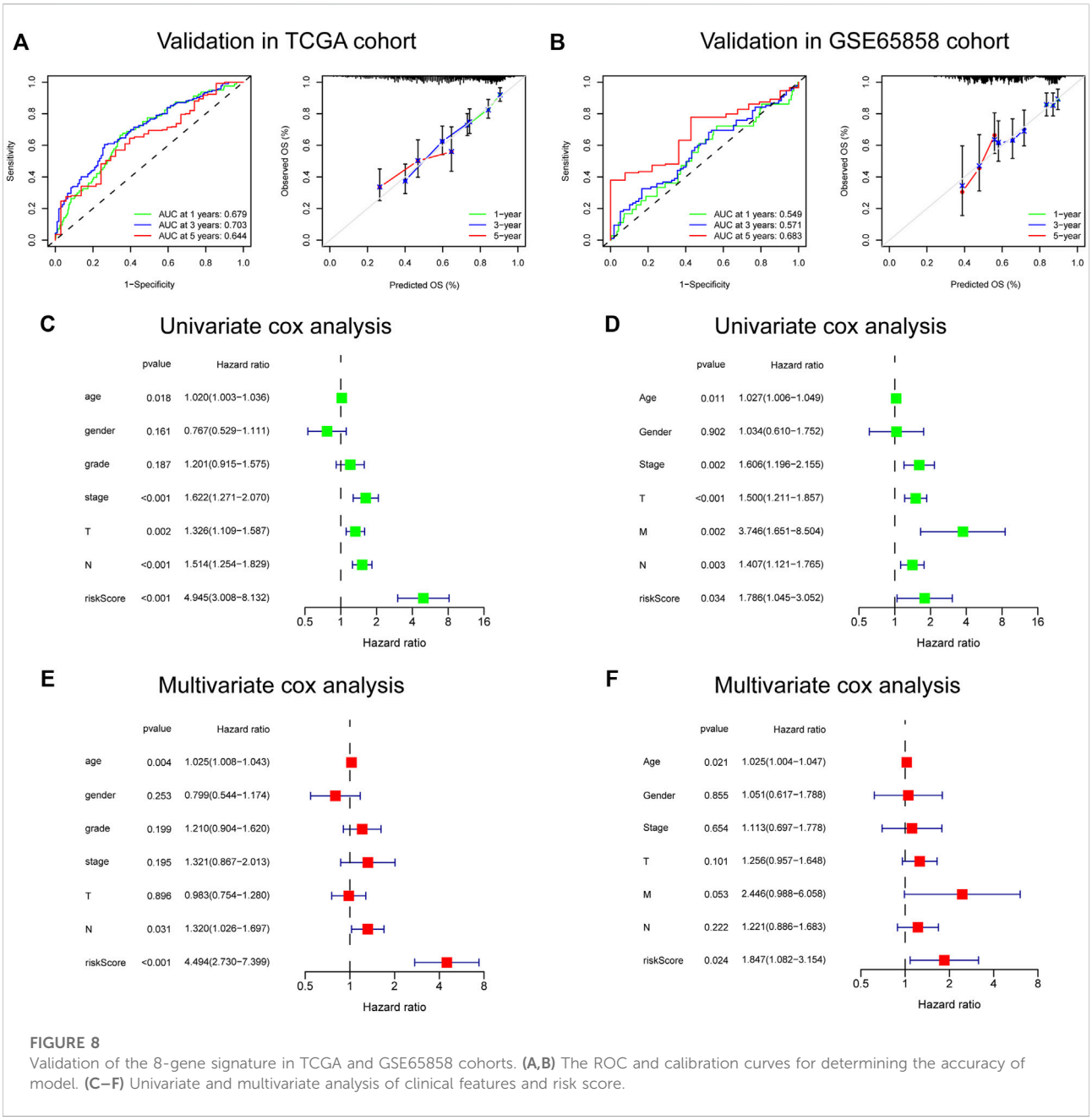


FIGURE 8
Validation of the 8-gene signature in TCGA and GSE65858 cohorts. **(A,B)** The ROC and calibration curves for determining the accuracy of model. **(C–F)** Univariate and multivariate analysis of clinical features and risk score.

TBX2, IDO1, GZMA, LAG3, CXCL10, CTLA4, PRF1, CXCL9, and TNF) were selected and their expressions were compared between the two risk groups (Figure 13D). Based on the correlation analysis results, it was evident that the expressions of CD274 and CTLA4 in the two groups were significantly different (CD274: p = 0.0006; CTLA4: p = 2.5e-14), and decreased as the risk score rose (CD274: r = −0.18, p = 3.6 e-05; CTLA4: r = −0.43, p < 2.2 e-16) (Figures 13E–H). Next, the pRRophetic algorithm was applied to predict the IC50 of eight common chemotherapeutic drugs between the two groups. Patients with a high-risk score showed an increased susceptibility to bleomycin (p = 0.00014), cisplatin (p = 3.2e-05), and methotrexate (p = 0.039). On the other hand, low-risk group patients showed increased sensitivity to rapamycin (p = 5.6e-06) (Figures 14A–H). To forecast the response to anti-PD1 and anti-CTLA4 immunotherapy, the IPS scores of HNSCC patients were used to compare the two risk groups (Figure 14I–L). Results indicated that patients in the low-risk group exhibited higher IPS scores and showed greater response to anti-PD1 therapy and anti-PD1 plus anti-CTLA4 therapy (ips_ctla4_neg_pd1_pos: p = 0.0054, ips_ctla4_pos_pd1_pos: p = 1.6e -05) relative to patients in the high-risk group. Given the important role of TMB in prognosis, the intrinsic connection

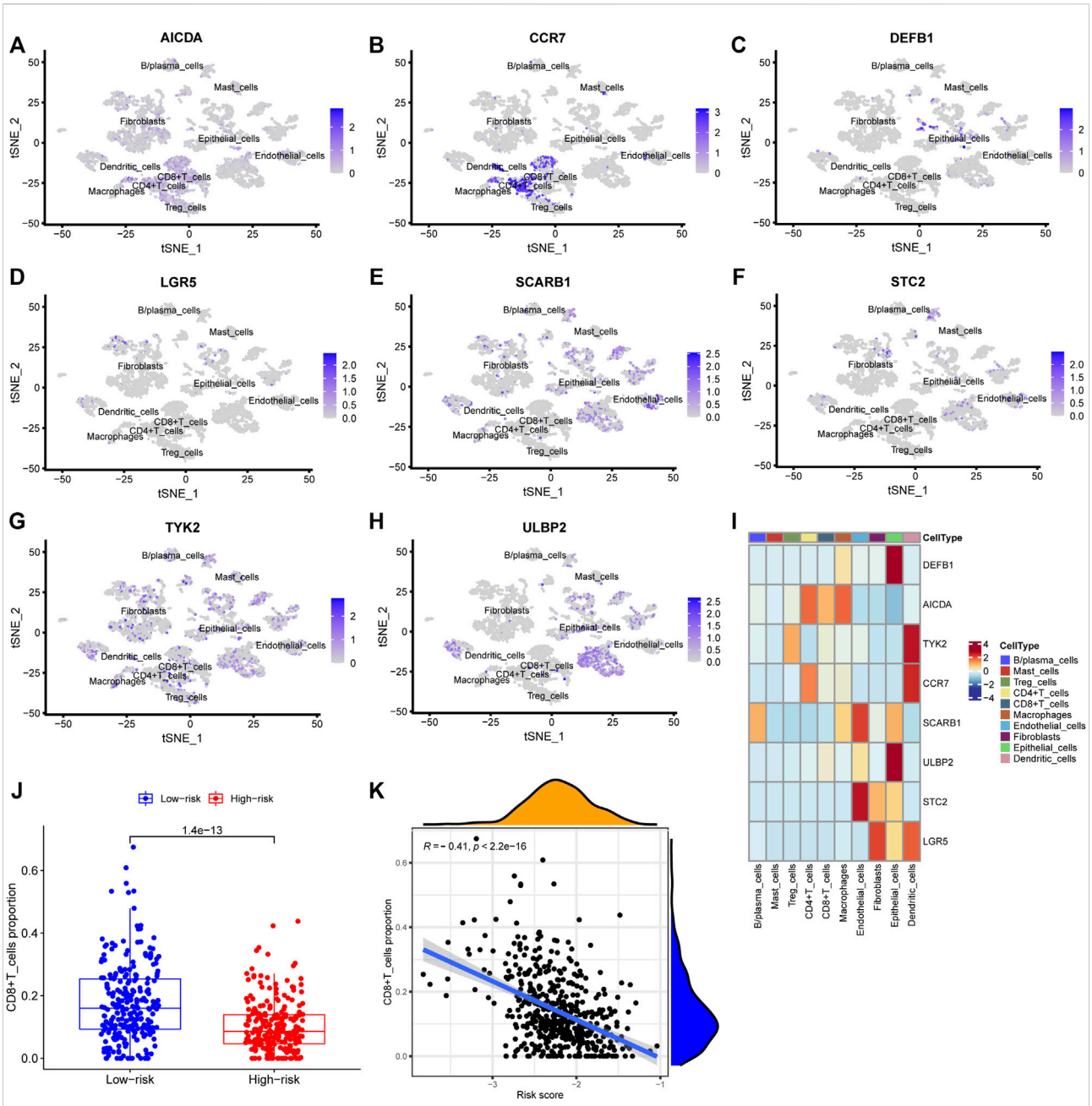


FIGURE 9
Verification using single-cell sequencing data. **(A–H)** Colors indicating the localization of the expression of eight genes: AICDA, CCR7, DEFB1, LGR5, SCARB1, STC2, TYK2, and ULBP2. **(I)** Heatmap depicting expressions of the eight genes in the cell subtypes. **(J)** The levels of $CD8^+$ cells in TCGA deconvoluted data between low- and high-risk groups (p = 1.4 e-13). **(K)** Correlation analysis between $CD8^+$ T cells levels and risk score (r = −0.41, p < 2.2 e-16).

between TMB and risk score was explored to assess genetic signature. It was found that the high-risk group exhibited higher TMB (Figure 15A). A significant correlation was observed between TMB and risk score (r = 0.22, p = 1.3e-06, Figure 15B). Survival curve suggested that a low TMB/low risk group showed a great outcome compared with the other groups (p < 0.001, Figure 15C). The top 20 driver genes with the highest alteration frequency were analyzed (Figures 15D,E) and four genes (*TP53*, *PKHD1L1*, *DNAH9*, *FAT1*) were significantly different between high- and low-risk groups (Supplementary Table S2).

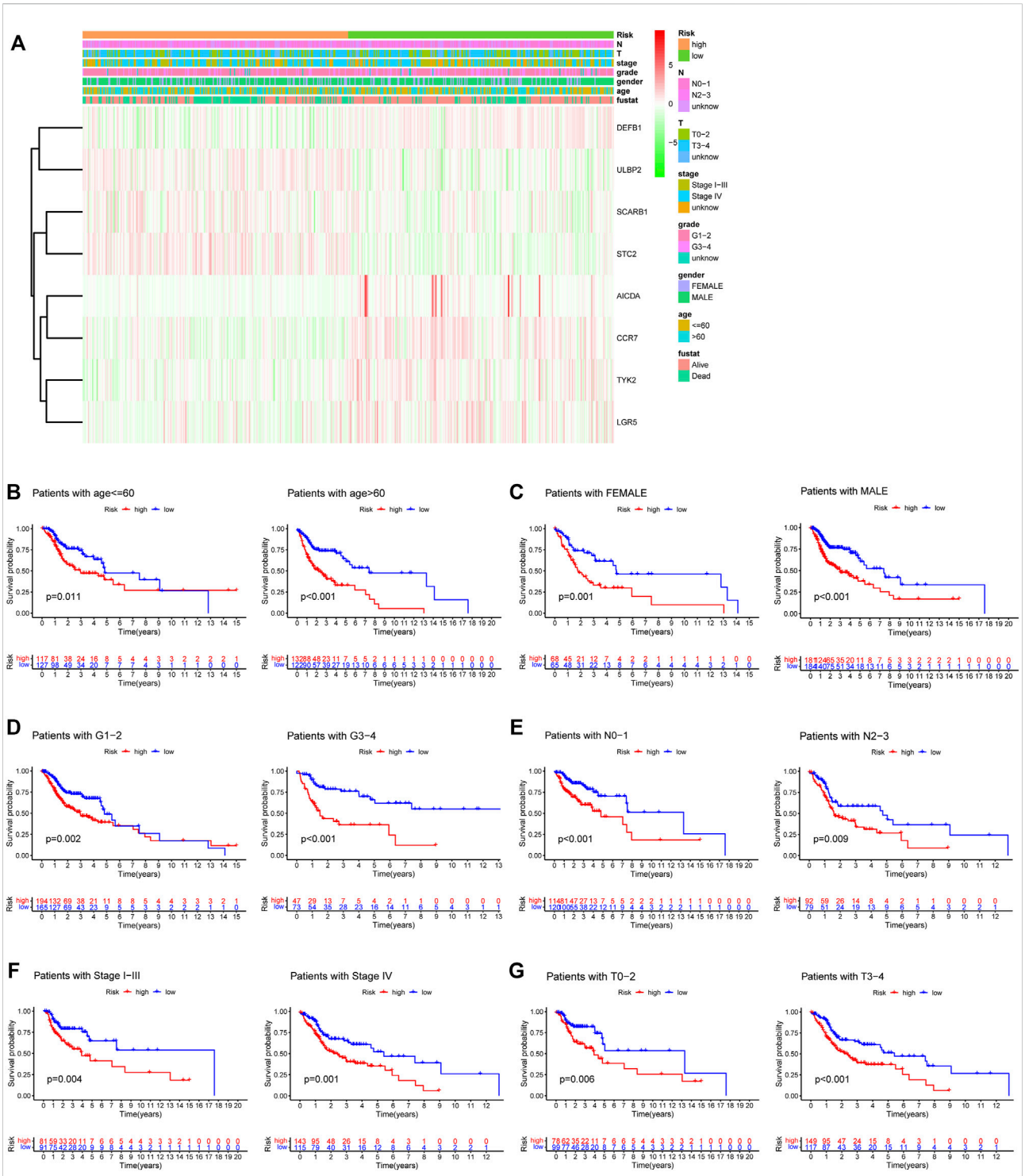


FIGURE 10
The relationship between risk signature and the clinical characteristics. **(A)** The heatmap depicting eight gene expressions among distinct clinical patterns. **(B–G)** Kaplan-Meier survival analysis according to the 8-gene signature stratified by clinicopathological factors. **(B)** age<=60: $p = 0.011$, age>60: $p < 0.001$. **(C)** Female: $p = 0.001$, Male: $p < 0.001$. **(D)** G1-2: $p = 0.002$, G3-4: $p < 0.001$. **(E)** N0-1: $p < 0.001$, N2-3: $p = 0.009$. **(F)** Stage I-III: $p = 0.004$, Stage IV: $p = 0.001$. **(G)** T0-2: $p = 0.006$, T3-4: $p < 0.001$.

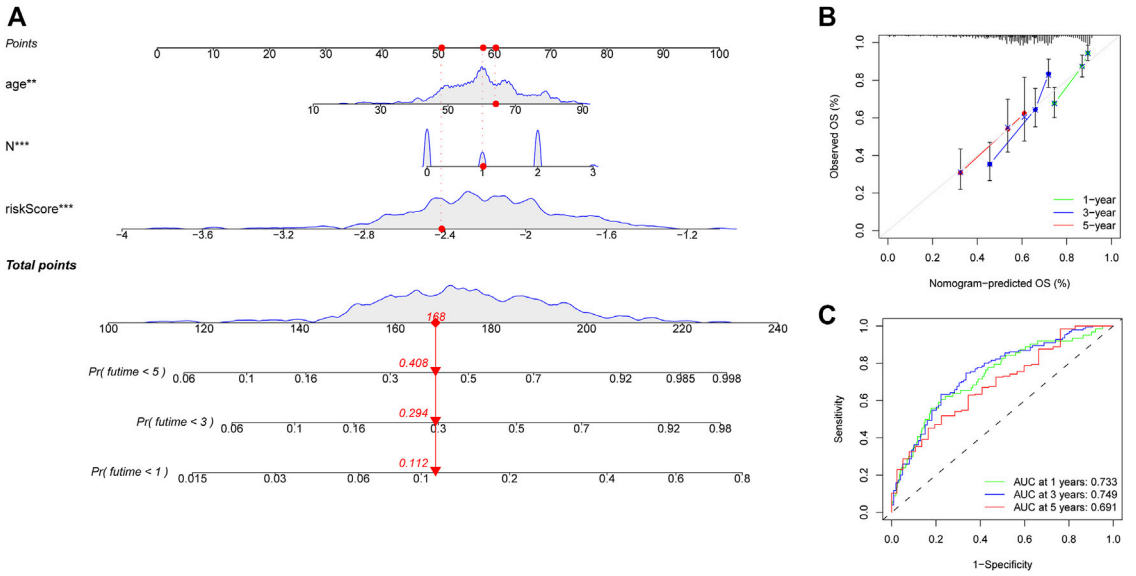


FIGURE 11
Construction of a nomogram for predicting survival of HNSCC patients. **(A)** Nomogram using two clinical traits (N stage and age) and the risk score. **(B,C)** The calibration and ROC curves for determining the reliability of the nomogram to predict one-, three-, and five-year survival rates.

Discussion

Immunotherapy has been successful used to treat cancer patients in the advanced tumor stage. Nevertheless, clinical application of the strategy is hampered by several limitations, including low response rates, development of serious side effects, and drug resistance (Sacco et al., 2021). One of the key reasons for these limitations is the paucity of potential predictive markers. In the present study, we calculated the proportion of $CD8^+$ T cells, and selected IRGs-related to $CD8^+$ T cells infiltration by integrating scRNA and bulk sequencing profiles. As a result, 215 differential IRGs were identified by WGCNA, of which 45 genes were significantly associated with HNSCC survival. Subsequently, we developed and validated an 8-gene risk model which may be useful for predicting prognosis and immunotherapeutic effect.

The eight critical genes, including *DEFB1*, *AICDA*, *TYK2*, *CCR7*, *SCARB1*, *ULBP2*, *STC2*, and *LGR5*, play essential roles in tumor progression and immune-modulatory effects. For example, DEFB1, the human antimicrobial peptide defensin β 1, is considered as a potential tumor suppressor gene and has been shown to mediate PI3K/mTOR signaling, thereby leading to death of tumor cells (Sun et al., 2006; Lee et al., 2015). *DEFB1* was also found to be theoretically useful as a prognostic biomarker for HNSCC (Han et al., 2014). Moreover, *DEFB1* was commonly detected in epithelial cells, which is consistent with our results. UL16-binding protein 2 (*ULBP2*), a ligand of the activating NK cell receptor *NKG2D*, was found to be engaged in target recognition by NK cells (Textor et al., 2011). A previous study confirmed that the soluble *ULBP2* secreted by cancer cells contributed to the immune escape (Waldhauer and Steinle, 2006). Herein, we observed that *ULBP2* was upregulated in epithelial cells. Meanwhile, *ULBP2* has been shown to be a prognosis indicator for several cancers, such as lung cancer and pancreatic cancer (Chang et al., 2011; Yamaguchi et al., 2012). The activation-induced cytidine deaminase (*AICDA*) is an essential enzyme of the adaptive immune system. A recent study found that elevated expression of *AICDA* regulates the function of B cells in regional lymph nodes and significantly improves prognosis of HNSCC patients (Pylaeva et al., 2021). Tyrosine kinase 2 (*TYK2*), a member of the Janus kinase (JAK) family, has emerged as both a promising biomarker and a target for anti-cancer therapies (Borcherding et al., 2021). It has been reported that high expression of *TYK2* is associated with better prognosis of HNSCC (Fang et al., 2021). A recent review concluded that CC motif chemokine receptor (*CCR7*) is correlated with good outcomes of HNSCC patients (Korbecki et al., 2020). However, if located on cancer cells, *CCR7* and its ligands (*CCL19*/*CCL21*) is a vital axis for carcinogenic properties, such as epithelial-mesenchymal transition (EMT) tumor invasion and migration (Chen et al., 2020; Korbecki et al., 2020). Notably, the present study found that *CCR7* was predominantly expressed in dendritic cells. *SCARB1* has been demonstrated to be involved in cholesterol metabolism, thereby facilitating cancer progression (Gutierrez-Pajares et al., 2016). In addition, stanniocalcin-2 (*STC2*) exerted a significant role in a wide variety of signaling pathways in HNSCC apoptosis and autophagy (Li et al., 2020). Studies have revealed that

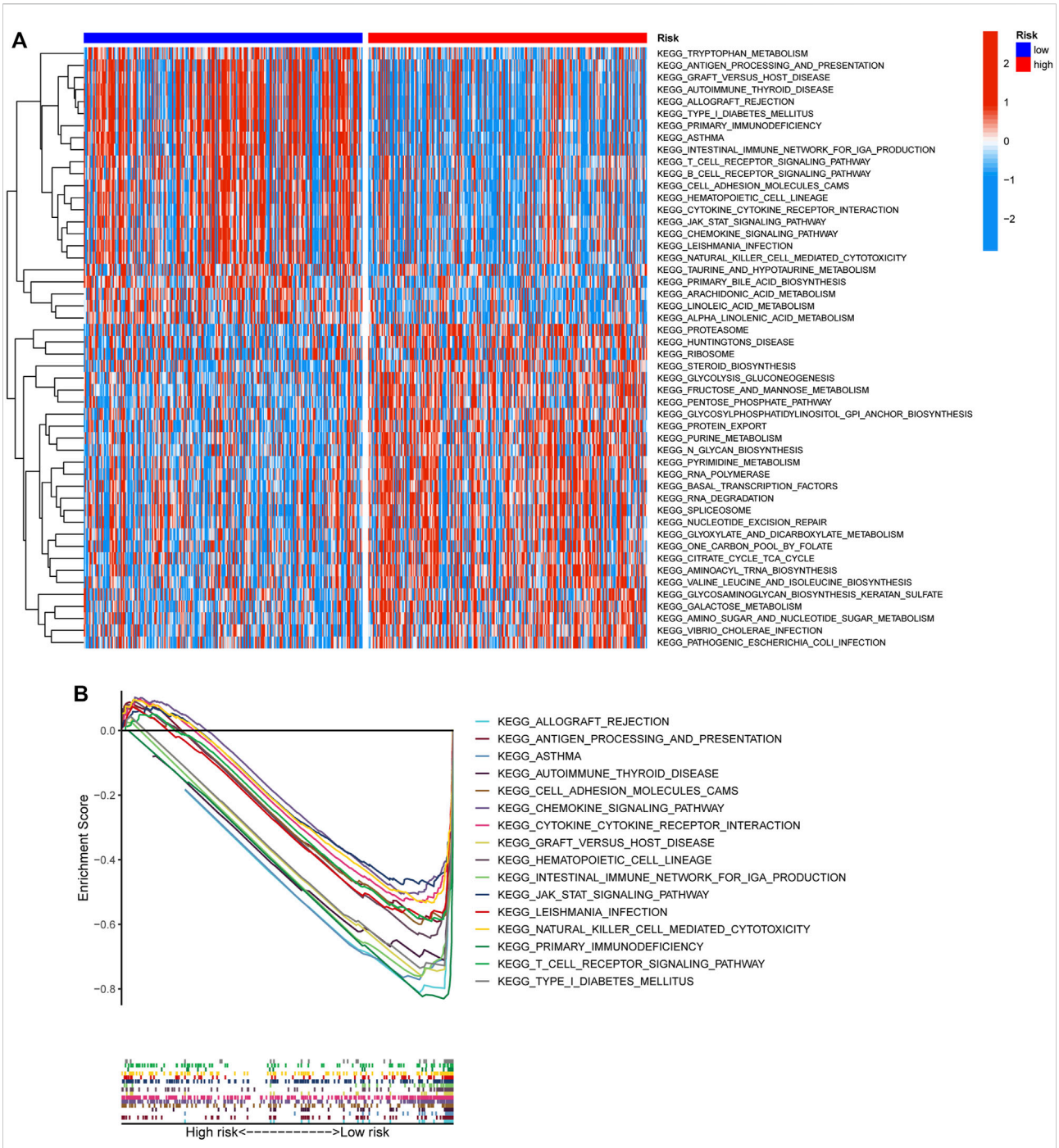


FIGURE 12
Functional enrichment characteristics of the risk signature. **(A)** Different activities of KEGG pathway scored by GSVA between high- and low-risk groups. **(B)** GSEA analysis showing the sixteen KEGG functional pathways enriched in low-risk group.

downregulated expression of *STC2* can suppress growth of HNSCC cells (Li et al., 2019; Li et al., 2020). Moreover, the leucine-rich repeat-containing G protein-coupled receptor *LGR5* participated in Wnt signaling and was intimately linked to the severity of HNSCC (Dalley et al., 2015).

Given the important role of immune cell infiltrations in the diagnosis and treatment of diseases, we further explored the immune landscape in different HNSCC groups. Based on the degree of immune cell infiltrations, particularly $CD8^+$ T cells, the tumor phenotypes can be defined as two major patterns,

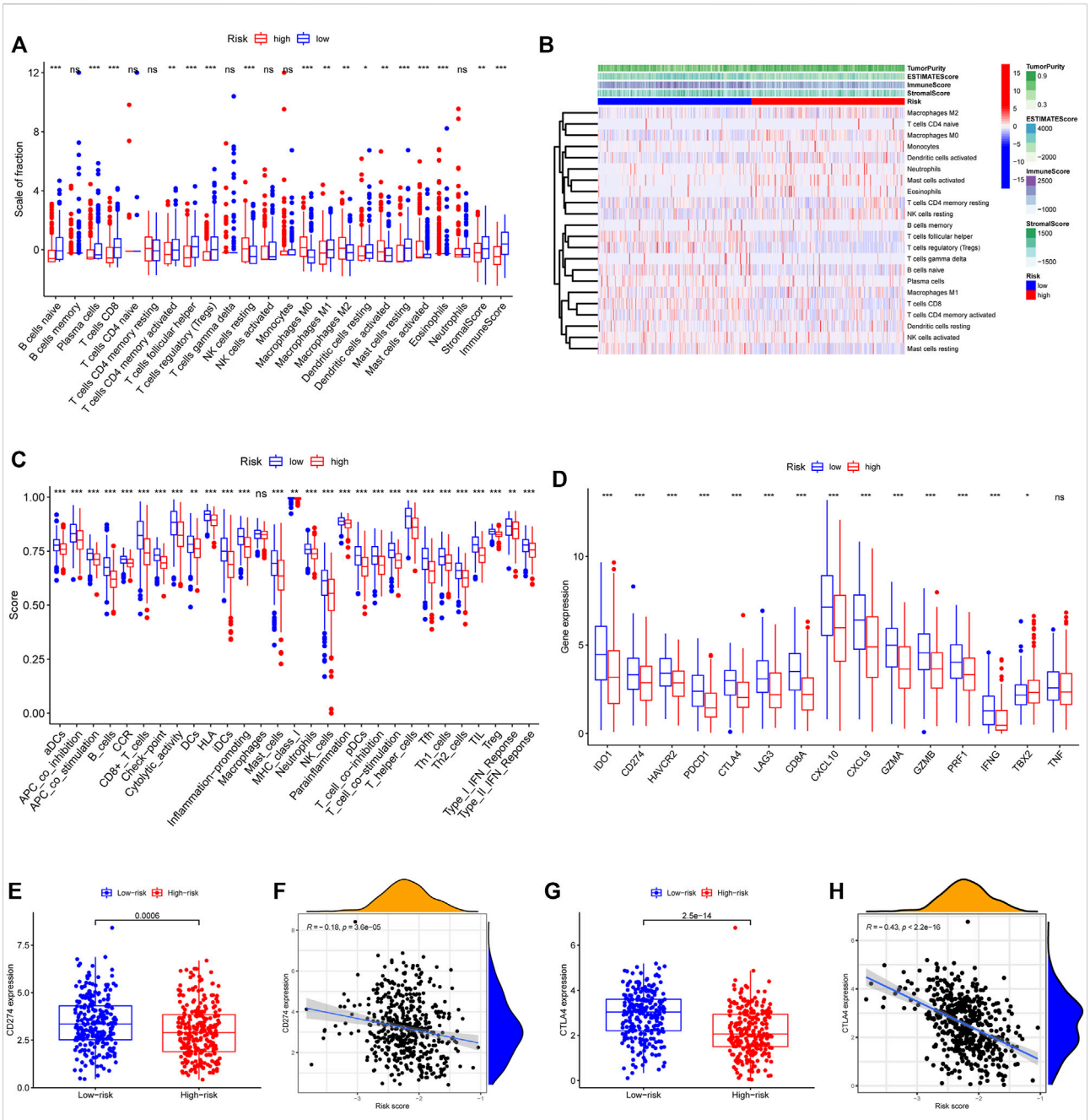


FIGURE 13
Patterns of immune cells infiltration in two risk groups. **(A)** The box plot showing the fractions of 22 infiltrating immune cells and immune-related scores based on CIBERSORT and ESTIMATE algorithms. **(B)** A heatmap presenting the 22 immune cells in the two risk score subgroups with different immune-related scores. **(C)** The ssGSEA scores for 29 immune gene sets. **(D)** Differential expressions of the 15 immune checkpoint-related genes. **(E)** CD274 expression difference among the high- and low-risk groups (p = 0.0006). **(F)** The spearman correlation plot between CD274 expression and risk score (r = −0.18, p = 3.6 e-05). **(G)** CTLA4 expression difference among the high- and low-risk groups (p = 2.5 e-14). **(H)** The spearman correlation plot between CTLA4 expression and risk score (r = −0.43, p < 2.2 e-16). $*p < 0.05$; $**p < 0.01$; $***p < 0.001$; ns: no significance.

"hot" and "cold", which are associated with good and poor antitumor immune responses, respectively (Galon and Bruni, 2019). This study explored the abundance of immune cells and functions using CIBERSORT, ESTIMATE, and ssGSEA methods. According to the obtained results, the low-risk group exhibited more infiltration of $CD8^+$ T cells, memory activated $CD4^+$ T cells, and plasma cells, as well as higher immune score, and thus can be categorized as "hot" tumor phenotype. On the other hand, the high-

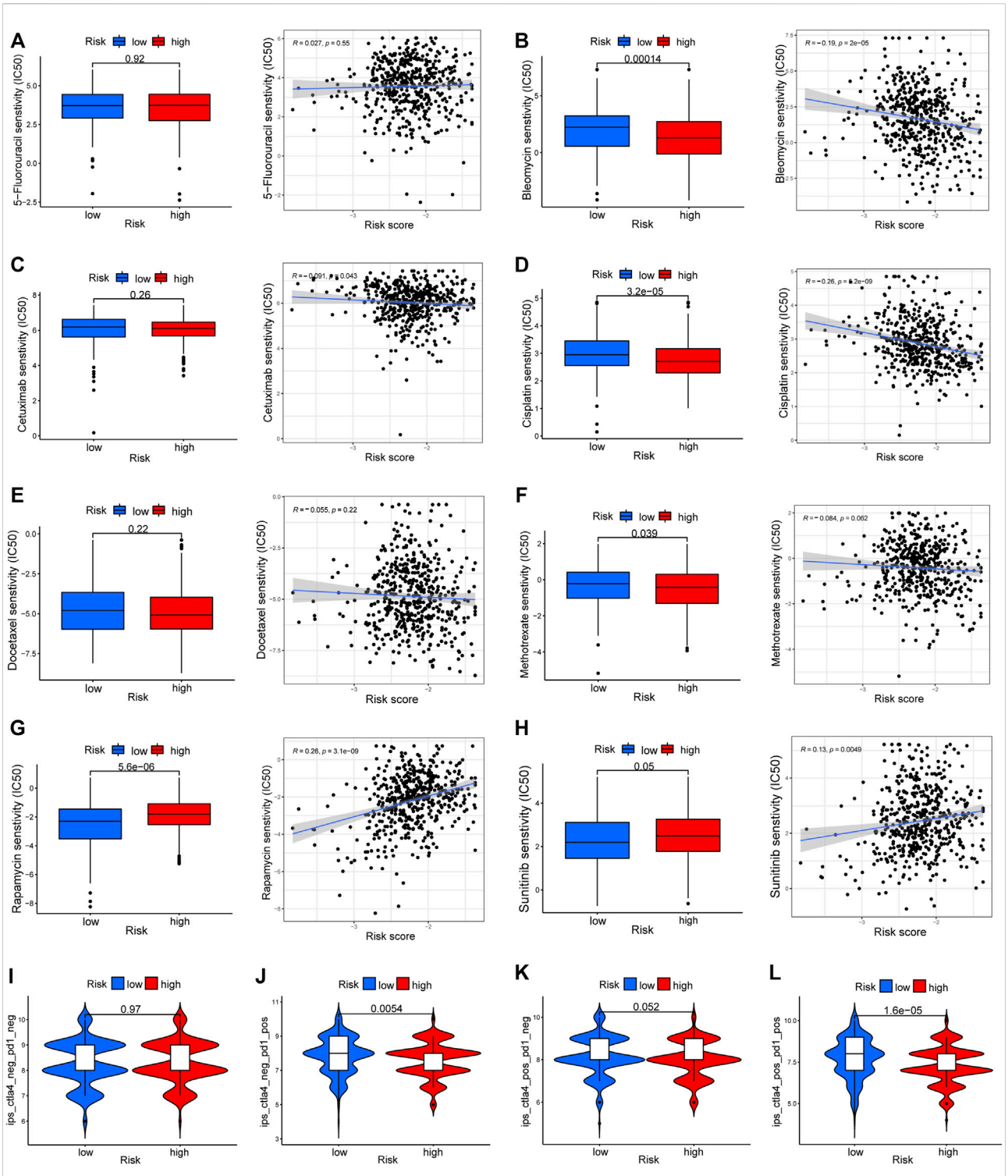


FIGURE 14
Drug response prediction between the two risk groups. **(A–H)** The IC50 of eight common chemotherapeutic agents (5-Fluorouracil, Bleomycin, Cetuximab, Cisplatin, Docetaxel, Methotrexate, Rapamycin, and Sunitinib) and correlation analysis with risk score. **(I–L)** The difference of immunophenoscore (IPS) scores among high- and low-risk groups.

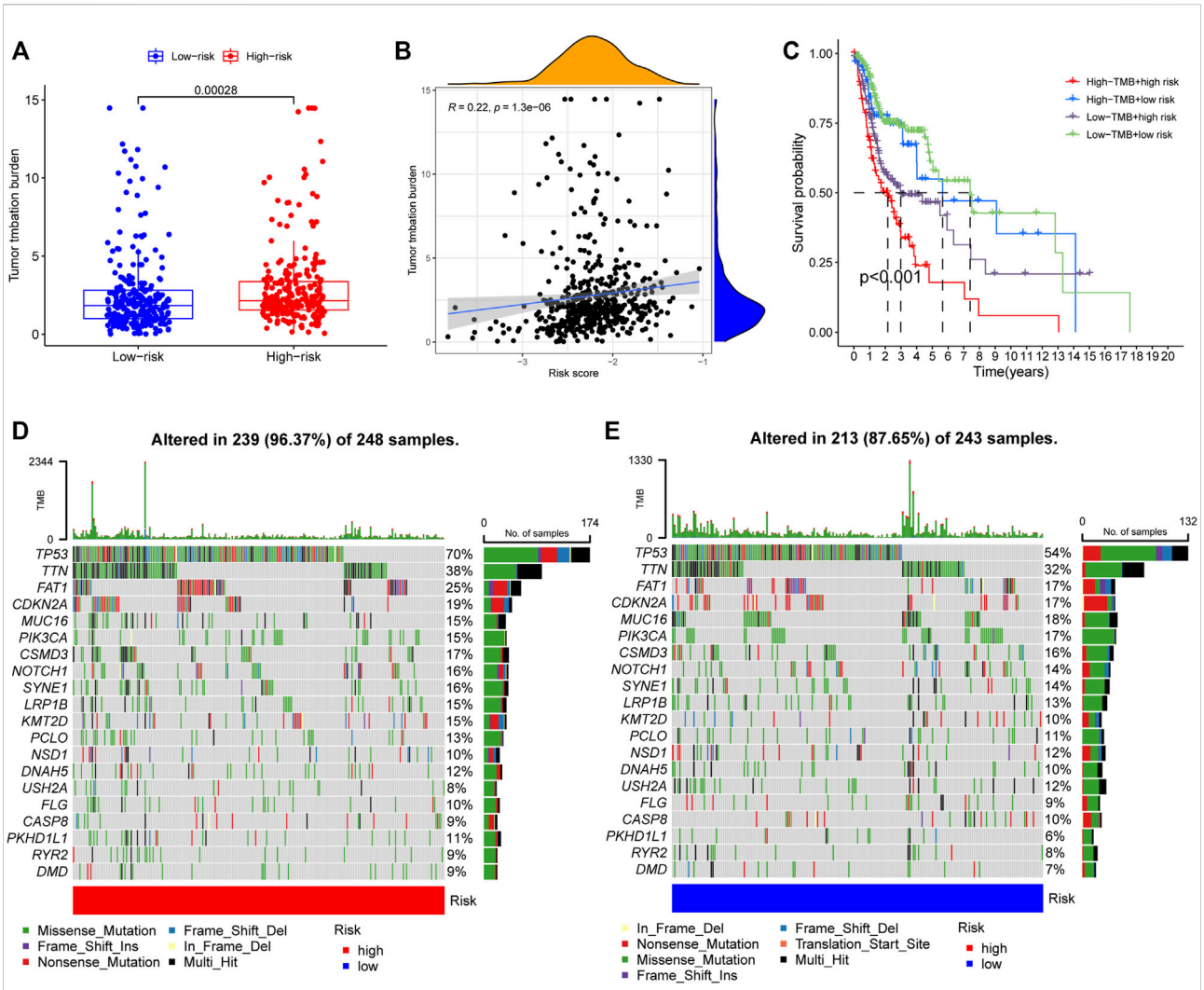


FIGURE 15
(A) TMB difference among the high and low risk groups. p = 0.00028. **(B)** The Spearman correlation analysis between risk score and TMB. r = 0.22, p = 1.3e-06. **(C)** K-M survival analysis stratified by both TMB and risk scores. $p < 0.001$. **(D,E)** Distribution of the top 20 variant mutated genes among high **(D)** and low **(E)** risk groups. The waterfall plot showing the genetic alterations types.

risk group showed greater abundance of activated mast cells, resting NK cells, and M2 macrophages, and lower immune score, suggesting the "cold" tumor phenotype. Furthermore, the immune checkpoint-related genes exhibited relatively high expressions in the low-risk group, including *IFNG*, *PRF1*, *GZMA*, *GZMB*, *CXCL10*, *CXCL9*, *CD8A*, *CD274* (*PD-L1*), *HAVCR2*, *IDO1*, *LAG3*, *CTLA4*, and *PDCD1*. Studies have confirmed that infiltration of M2 macrophages is associated with tumorigenic chronic inflammation with secretion of protumorigenic factors, such as *IL-6*, *VEGF*, and *TGFβ* (Ruffell and Coussens, 2015). Accumulating evidence suggests that preexisting $CD8^+$ T cells and *PD-L1* expression are generally correlated with improved efficacy of immunotherapy (Farhood et al., 2019; Gavrielatou et al., 2020). Consistently, our results suggested that patients with low-risk score, as a consequence of higher IPS scores, had more vigorous immune responses to anti-PD1 therapy and anti-PD1 plus anti-CTLA4 therapy. Moreover, patients in the two groups exhibited varying sensitivity to four common chemotherapeutic drugs, including bleomycin, cisplatin, methotrexate, and rapamycin (Cramer et al., 2019). Notably, previous studies have verified the therapeutic safety and effectiveness of chemotherapy in combination with *PD-L1* blockade (Burtness et al., 2019; Cohen et al., 2019). Nevertheless, different sensitivities to 5-Fluorouracil, cetuximab, docetaxel, and sunitinib were not observed in this study. TMB level was considered to be an indicator of immunotherapy response (Rizvi et al., 2015). We then examined the relationship between TMB and the risk score. The alteration frequency of *TP53*, *PKHD1L1*, *DNAH9* and

FAT1 was significantly different between high- and low-risk groups. *TP53* is one of the most frequently mutated genes in HNSCC and *TP53* mutations play a critical role in tumorigenesis and progression (Nathan et al., 2022). Understanding the *DNAH9* and *FAT1* mutations may contribute to cancer surveillance and treatment (Huang et al., 2021; Yang et al., 2022). Investigation of the mutational signatures may allow for an improved selection of immunotherapies in individual patients.

However, this study was limited by the fact that it lacked experimental and clinical pathology studies to validate the function of the eight genes. Therefore, further clinical trials are needed to confirm the predictive potential of the risk signature.

Conclusion

In conclusion, by comprehensively analyzing the single-cell and bulk RNA sequencing of HNSCC, this study developed and externally validated a novel and robust model based on eight $CD8^+$ T cells-related genes. It is expected that the 8-gene signature will facilitate understanding of HNSCC immune characteristics, predict prognosis of HNSCC patients, and guide the clinical use of immunotherapy.

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

JZ designed the study and collected the data. SZ and WZ carried out the data analyses, prepared all figures and tables and wrote the manuscript. All authors participated in improving the writing of the manuscript and approved the submitted version.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.938611/full#supplementary-material

References

Borcherding, D. C., He, K., Amin, N. V., and Hirbe, A. C. (2021). TYK2 in cancer metastases: Genomic and proteomic discovery. *Cancers (Basel)* 13 (16), 4171. doi:10.3390/cancers13164171

Burtness, B., Harrington, K. J., Greil, R., Soulières, D., Tahara, M., de Castro, G., et al. (2019). Pembrolizumab alone or with chemotherapy versus cetuximab with chemotherapy for recurrent or metastatic squamous cell carcinoma of the head and neck (KEYNOTE-048): A randomised, open-label, phase 3 study. *Lancet* 394 (10212), 1915–1928. doi:10.1016/s0140-6736(19)32591-7

Chang, Y. T., Wu, C. C., Shyr, Y. M., Chen, T. C., Hwang, T. L., Yeh, T. S., et al. (2011). Secretome-based identification of ULBP2 as a novel serum marker for pancreatic cancer detection. *PLoS One* 6 (5), e20029. doi:10.1371/journal.pone.0020029

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 (1), 248–262. doi:10.1016/j.celrep.2016.12.019

Chen, Y., Shao, Z., Jiang, E., Zhou, X., Wang, L., Wang, H., et al. (2020). CCL21/CCR7 interaction promotes EMT and enhances the stemness of OSCC via a JAK2/STAT3 signaling pathway. *J. Cell. Physiol.* 235 (9), 5995–6009. doi:10.1002/jcp.29525

Cohen, E. E. W., Soulières, D., Le Tourneau, C., Dinis, J., Licitra, L., Ahn, M.-J., et al. (2019). Pembrolizumab versus methotrexate, docetaxel, or cetuximab for recurrent or metastatic head-and-neck squamous cell carcinoma (KEYNOTE-040): A randomised, open-label, phase 3 study. *Lancet* 393 (10167), 156–167. doi:10.1016/s0140-6736(18)31999-8

Cramer, J. D., Burtness, B., Le, Q. T., and Ferris, R. L. (2019). The changing therapeutic landscape of head and neck cancer. *Nat. Rev. Clin. Oncol.* 16 (11), 669–683. doi:10.1038/s41571-019-0227-z

Dalley, A. J., Abdul Majeed, A. A., Pitty, L. P., Major, A. G., and Farah, C. S. (2015). LGR5 expression in oral epithelial dysplasia and oral squamous cell carcinoma. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.* 119 (4), 436–440. e431. doi:10.1016/j.oooo.2014.11.014

Fang, L., Wang, W., Shi, L., Chen, Q., and Rao, X. (2021). Prognostic values and clinical relationship of TYK2 in laryngeal squamous cell cancer. *Med. Baltim.* 100 (34), e27062. doi:10.1097/MD.0000000000027062

Farhood, B., Najafi, M., and Mortezaee, K. (2019). CD8(+) cytotoxic T lymphocytes in cancer immunotherapy: A review. *J. Cell. Physiol.* 234 (6), 8509–8521. doi:10.1002/jcp.27782

Ferris, R. L., Blumenschein, G., Jr., Fayette, J., Guigay, J., Colevas, A. D., Licitra, L., et al. (2016). Nivolumab for recurrent squamous-cell carcinoma of the head and neck. *N. Engl. J. Med.* 375 (19), 1856–1867. doi:10.1056/NEJMoa1602252

Fridman, W. H., Zitvogel, L., Sautes-Fridman, C., and Kroemer, G. (2017). The immune contexture in cancer prognosis and treatment. *Nat. Rev. Clin. Oncol.* 14 (12), 717–734. doi:10.1038/nrclinonc.2017.101

Galon, J., and Bruni, D. (2019). Approaches to treat immune hot, altered and cold tumours with combination immunotherapies. *Nat. Rev. Drug Discov.* 18 (3), 197–218. doi:10.1038/s41573-018-0007-y

Gavrielatou, N., Doumas, S., Economopoulou, P., Foukas, P. G., and Psyrri, A. (2020). Biomarkers for immunotherapy response in head and neck cancer. *Cancer Treat. Rev.* 84, 101977. doi:10.1016/j.ctrv.2020.101977

Geeleher, P., Cox, N., and Huang, R. S. (2014). pRRophetic: an R package for prediction of clinical chemotherapeutic response from tumor gene expression levels. *PLoS One* 9 (9), e107468. doi:10.1371/journal.pone.0107468

Gutierrez-Pajares, J. L., Ben Hassen, C., Chevalier, S., and Frank, P. G. (2016). SR-BI: Linking cholesterol and lipoprotein metabolism with breast and prostate cancer. *Front. Pharmacol.* 7, 338. doi:10.3389/fphar.2016.00338

Han, Q., Wang, R., Sun, C., Jin, X., Liu, D., Zhao, X., et al. (2014). Human beta-defensin-1 suppresses tumor migration and invasion and is an independent predictor for survival of oral squamous cell carcinoma patients. *PLoS One* 9 (3), e91867. doi:10.1371/journal.pone.0091867

Hänzelmann, S., Castelo, R., and Guinney, J. (2013). Gsva: Gene set variation analysis for microarray and RNA-seq data. *BMC Bioinforma.* 14, 7. doi:10.1186/1471-2105-14-7

Huang, C., Chen, L., Savage, S. R., Eguez, R. V., Dou, Y., Li, Y., et al. (2021). Proteogenomic insights into the biology and treatment of HPV-negative head and neck squamous cell carcinoma. *Cancer Cell* 39 (3), 361–379. e316. doi:10.1016/j.ccell.2020.12.007

Jew, B., Alvarez, M., Rahmani, E., Miao, Z., Ko, A., Garske, K. M., et al. (2020). Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nat. Commun.* 11 (1), 1971. doi:10.1038/s41467-020-15816-6

Jia, L., Zhang, W., and Wang, C. Y. (2020). BMI1 inhibition eliminates residual cancer stem cells after PD1 blockade and activates antitumor immunity to prevent metastasis and relapse. *Cell Stem Cell* 27 (2), 238–253. e236. doi:10.1016/j.stem.2020.06.022

Korbecki, J., Grochans, S., Gutowska, I., Barczak, K., and Baranowska-Bosiacka, I. (2020). CC chemokines in a tumor: A review of pro-cancer and anti-cancer properties of receptors CCR5, CCR6, CCR7, CCR8, CCR9, and CCR10 ligands. *Int. J. Mol. Sci.* 21 (20), E7619. doi:10.3390/ijms21207619

Kurten, C. H. L., Kulkarni, A., Cillo, A. R., Santos, P. M., Roble, A. K., Onkar, S., et al. (2021). Investigating immune and non-immune cell interactions in head and neck tumors by single-cell RNA sequencing. *Nat. Commun.* 12 (1), 7338. doi:10.1038/s41467-021-27619-4

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

Lee, M., Wiedemann, T., Gross, C., Leinhauser, I., Roncaroli, F., Braren, R., et al. (2015). Targeting PI3K/mTOR signaling displays potent antitumor efficacy against nonfunctioning pituitary adenomas. *Clin. Cancer Res.* 21 (14), 3204–3215. doi:10.1158/1078-0432.CCR-15-0288

Li, T., Feng, Z., Wang, Y., Zhang, H., Li, Q., Schiferle, E., et al. (2020). Antioncogenic effect of MicroRNA-206 on neck squamous cell carcinoma through inhibition of proliferation and promotion of apoptosis and autophagy. *Hum. Gene Ther.* 31 (23-24), 1260–1273. doi:10.1089/hum.2020.090

Li, T., Qin, Y., Zhen, Z., Shen, H., Cong, T., Schiferle, E., et al. (2019). Long non-coding RNA HOTAIR/microRNA-206 sponge regulates STC2 and further influences cell biological functions in head and neck squamous cell carcinoma. *Cell Prolif.* 52 (5), e12651. doi:10.1111/cpr.12651

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

Nathan, C. A., Khandelwal, A. R., Wolf, G. T., Rodrigo, J. P., Mäkitie, A. A., Saba, N. F., et al. (2022). TP53 mutations in head and neck cancer. *Mol. Carcinog.* 61 (4), 385–391. doi:10.1002/mc.23385

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* 12 (5), 453–457. doi:10.1038/nmeth.3337

Puram, S. V., Tirosh, I., Parikh, A. S., Patel, A. P., Yizhak, K., Gillespie, S., et al. (2017). Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* 171 (7), 1611–1624. doi:10.1016/j.cell.2017.10.044

Pylaeva, E., Ozel, I., Squire, A., Spyra, I., Wallner, C., Korek, M., et al. (2021). B-helper neutrophils in regional lymph nodes correlate with improved prognosis in patients with head and neck cancer. *Cancers (Basel)* 13 (12), 3092. doi:10.3390/cancers13123092

Qi, Z., Barrett, T., Parikh, A. S., Tirosh, I., and Puram, S. V. (2019). Single-cell sequencing and its applications in head and neck cancer. *Oral Oncol.* 99, 104441. doi:10.1016/j.oraloncology.2019.104441

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

Ruffell, B., and Coussens, L. M. (2015). Macrophages and therapeutic resistance in cancer. *Cancer Cell* 27 (4), 462–472. doi:10.1016/j.ccell.2015.02.015

Sacco, A. G., Chen, R., Worden, F. P., Wong, D. J. L., Adkins, D., Swiecicki, P., et al. (2021). Pembrolizumab plus cetuximab in patients with recurrent or metastatic head and neck squamous cell carcinoma: An open-label, multi-arm, non-randomised, multicentre, phase 2 trial. *Lancet. Oncol.* 22 (6), 883–892. doi:10.1016/s1470-2045(21)00136-4

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

Siu, L. L., Even, C., Mesia, R., Remenar, E., Daste, A., Delord, J. P., et al. (2019). Safety and efficacy of durvalumab with or without tremelimumab in patients with PD-L1-low/negative recurrent or metastatic HNSCC: The phase 2 CONDOR randomized clinical trial. *JAMA Oncol.* 5 (2), 195–203. doi:10.1001/jamaoncol.2018.4628

Song, P., Li, W., Wu, X., Qian, Z., Ying, J., Gao, S., et al. (2022). Integrated analysis of single-cell and bulk RNA-sequencing identifies a signature based on B cell marker genes to predict prognosis and immunotherapy response in lung adenocarcinoma. *Cancer Immunol. Immunother.* 1, 1. doi:10.1007/s00262-022-03143-2

Sun, C. Q., Arnold, R., Fernandez-Golarz, C., Parrish, A. B., Almekinder, T., He, J., et al. (2006). Human beta-defensin-1, a potential chromosome 8p tumor suppressor: Control of transcription and induction of apoptosis in renal cell carcinoma. *Cancer Res.* 66 (17), 8542–8549. doi:10.1158/0008-5472.CAN-06-0294

Textor, S., Fiegler, N., Arnold, A., Porgador, A., Hofmann, T. G., Cerwenka, A., et al. (2011). Human NK cells are alerted to induction of p53 in cancer cells by upregulation of the NKG2D ligands ULBP1 and ULBP2. *Cancer Res.* 71 (18), 5998–6009. doi:10.1158/0008-5472.CAN-10-3211

Vigneswaran, N., and Williams, M. D. (2014). Epidemiologic trends in head and neck cancer and aids in diagnosis. *Oral Maxillofac. Surg. Clin. North Am.* 26 (2), 123–141. doi:10.1016/j.coms.2014.01.001

Waldhauer, I., and Steinle, A. (2006). Proteolytic release of soluble UL16-binding protein 2 from tumor cells. *Cancer Res.* 66 (5), 2520–2526. doi:10.1158/0008-5472.CAN-05-2520

Wichmann, G., Rosolowski, M., Krohn, K., Kreuz, M., Boehm, A., Reiche, A., et al. (2015). The role of HPV RNA transcription, immune response-related gene expression and disruptive TP53 mutations in diagnostic and prognostic profiling of head and neck cancer. *Int. J. Cancer* 137 (12), 2846–2857. doi:10.1002/ijc.29649

Xu, K., Fu, Y., Han, Y., Xia, R., Xu, S., Duan, S., et al. (2020). Fewer tumour-specific PD-1(+)CD8(+) TILs in high-risk "Infiltrating" HPV(-) HNSCC. *Br. J. Cancer* 123 (6), 932–941. doi:10.1038/s41416-020-0966-8

Yamaguchi, K., Chikumi, H., Shimizu, A., Takata, M., Kinoshita, N., Hashimoto, K., et al. (2012). Diagnostic and prognostic impact of serum-soluble UL16-binding protein 2 in lung cancer patients. *Cancer Sci.* 103 (8), 1405–1413. doi:10.1111/j.1349-7006.2012.02330.x

Yang, F., Long, N., Anekpuritanang, T., Bottomly, D., Savage, J. C., Lee, T., et al. (2022). Identification and prioritization of myeloid malignancy germline variants in a large cohort of adult patients with AML. *Blood* 139 (8), 1208–1221. doi:10.1182/blood.2021011354

Yoshihara, K., Shahmoradgoli, M., Martinez, E., Vegesna, R., Kim, H., Torres-Garcia, W., et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* 4, 2612. doi:10.1038/ncomms3612

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 17 August 2022
DOI 10.3389/fgene.2022.916400

OPEN ACCESS

EDITED BY
Geng Chen,
Genecast Biotechnology Co., Ltd., China

REVIEWED BY
Ting Li,
National Center for Toxicological Research (FDA), United States
Michal Linial,
Hebrew University of Jerusalem, Israel

*CORRESPONDENCE
Hong-Jiu Wang,
wanghongjiu@hainmc.edu.cn
Zhen-Zhen Wang,
wangzhenzhen@hainmc.edu.cn

†These authors have contributed equally to this work

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

RECEIVED 09 April 2022
ACCEPTED 11 July 2022
PUBLISHED 17 August 2022

CITATION
Wang N, He D-N, Wu Z-Y, Zhu X, Wen X-L, Li X-H, Guo Y, Wang H-J and Wang Z-Z (2022), Oncogenic signaling pathway dysregulation landscape reveals the role of pathways at multiple omics levels in pan-cancer. *Front. Genet.* 13:916400. doi: 10.3389/fgene.2022.916400

Oncogenic signaling pathway dysregulation landscape reveals the role of pathways at multiple omics levels in pan-cancer

Na Wang[1,2,3†], Dan-Ni He[2†], Zhe-Yu Wu[2†], Xu Zhu[1,3], Xiao-Ling Wen[1,3], Xu-Hua Li[1,3], Yu Guo[1,3], Hong-Jiu Wang[1,2,3*] and Zhen-Zhen Wang[1,2,3*]

[1]Key Laboratory of Tropical Translational Medicine of Ministry of Education, College of Biomedical Information and Engineering, Hainan Medical University, Haikou, China, [2]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China, [3]College of Biomedical Information and Engineering, Hainan Medical University, Haikou, China

Dysregulation of signaling pathways plays an essential role in cancer. However, there is not a comprehensive understanding on how oncogenic signaling pathways affect the occurrence and development with a common molecular mechanism of pan-cancer. Here, we investigated the oncogenic signaling pathway dysregulation by using multi-omics data on patients from TCGA from a pan-cancer perspective to identify commonalities across different cancer types. First, the pathway dysregulation profile was constructed by integrating typical oncogenic signaling pathways and the gene expression of TCGA samples, and four molecular subtypes with significant phenotypic and clinical differences induced by different oncogenic signaling pathways were identified: TGF-β+ subtype; cell cycle, MYC, and NF2− subtype; cell cycle and TP53+ subtype; and TGF-β and TP53− subtype. Patients in the TGF-β+ subtype have the best prognosis; meanwhile, the TGF-β+ subtype is associated with hypomethylation. Moreover, there is a higher level of immune cell infiltration but a slightly worse survival prognosis in the cell cycle, MYC, and NF2− subtype patients due to the effect of T-cell dysfunction. Then, the prognosis and subtype classifiers constructed by differential genes on a multi-omics level show great performance, indicating that these genes can be considered as biomarkers with potential therapeutic and prognostic significance for cancers. In summary, our study identified four oncogenic signaling pathway–driven patterns presented as molecular subtypes and their related potential prognostic biomarkers by integrating multiple omics data. Our discovery provides a perspective for understanding the role of oncogenic signaling pathways in pan-cancer.

KEYWORDS

signaling pathways, dysregulation landscape, molecular subtypes, multi-omics, biomarkers, pan-cancer

Introduction

A large number of studies have shown that the oncogenic signaling pathways play important roles in cancers, and multi-omics changes that occurred in these signaling pathways are identified as the common biomarkers in cancers. Therefore, the identification of oncogenic signaling pathways has become a key step in cancer drug screening and cancer treatment. Although the roles of individual pathways in the development of single cancer have been successively discovered and demonstrated, it is interesting to study how these signaling pathways affect cancer development and progression from a pan-cancer perspective.

There are many studies on oncogenic signaling pathways and the genes involved (Joerger and Fersht, 2016; Taciak et al., 2018; Calses et al., 2019). It has been reported that the RTK-RAS pathway, PI3K/Akt signaling pathway, TP53 signaling pathway, APC, and other signaling pathways often undergo genetic changes in cancer. Then, the molecular mechanism of these pathways and the role of each gene in these pathways and the relationship between these pathways and the occurrence and development of cancer were integrated (Vogelstein and Kinzler, 2004). Francisco used multi-omics data to analyze the mechanisms and patterns of 10 pathways, including cell cycle, Hippo, MYC, NOTCH, Nrf2, PI3Ki-Akt, RTK-RAS, TGF-β, p53, and β-catenin/WNT, and identified the interaction of pathways (Sanchez-Vega et al., 2018). The study has proven that the main functions of the Hippo pathway include restriction of tissue growth and regulation of cell proliferation, differentiation, and migration in developing organs. In addition, the dysregulation of the Hippo pathway can also lead to abnormal cell growth and the occurrence of tumors (Meng et al., 2016). Giachino et al. (2015) explored the role of the NOTCH signaling pathway in promoting and suppressing cancer and analyzed the molecular mechanisms of the NOTCH signaling pathway in hematological cancers and solid tumors, which have also been linked to therapeutic strategies targeting the NOTCH pathway in human cancer treatment.

In recent years, the research on subtype analysis of single cancer based on pathways has been continuously developed (Bild et al., 2006; Liu et al., 2015; Kaunitz et al., 2017; Thanki et al., 2017). Bidkhori et al. (2018) classified hepatocellular carcinoma (HCC) patients into three subtypes with significant differences based on graph and control theory concepts to the topology of genome-scale metabolic networks and identified drug targets for effective treatment of HCC patients.

Gong et al. (2021) discovered three subtypes of triple-negative breast cancer (TNBC) with significant prognosis, molecular subtype distribution, and genomic alterations by investigating metabolic pathways, which demonstrated the metabolic heterogeneity of TNBC and made it possible to develop personalized treatments for unique tumor metabolism characteristics. Park et al. (2019) identified glioblastoma multiforme (GBM) subtypes with prognostic core genes, prognostic chromosomal aberrations, and mutations. The aim was to verify that the failure of targeted therapy in patients with glioblastoma is associated with high heterogeneity and activation of multiple oncogenic pathways. It is believed that subtype-specific alterations can be used as new prognostic biomarkers and therapeutic targets for GBM. Moreover, although the pan-cancer analysis can open the doors to identification of the commonalities in cancer and offer insights that could expand further discoveries and cancer treatments, there are few studies focused on the dysregulated patterns of multiple signaling pathways systematically in pan-cancer, and the cooperative mode of oncogenic signaling pathways is not clear.

Here, we proposed a method to identify different roles of oncogenic signaling pathways from the perspective of pan-cancer. The four molecular subtypes named by different signaling pathways were identified based on the gene expression of TCGA data, which shows distinct phenotypic and clinical features. In addition, combining multi-omics data, we studied the differences in differentially expressed genes, copy number variations, chromatin accessibility, DNA methylation levels, and tumor microenvironment of the four subtypes, and identified differential genes of each omics which were used to construct the prognostic models with significant results, such as WNT7A, CNTN6, and CDR1. These differential signatures were characterized as biomarkers with potential therapeutic and prognostic significance for cancer. In conclusion, the research helps to further understand the role of oncogenic signaling pathways in pan-cancer.

Results

Four pathway-driven subtypes were identified based on oncogenic signaling pathways

In order to investigate the mechanism of 10 pathways in cancers (Ciriello et al., 2013; Imperial et al., 2019; Paczkowska et al., 2020), we collected 333 genes of 10 canonical oncogenic signaling pathways confirmed in the previous research. Based on those gene expression levels for 7,518 patients (TCGA training set, Supplementary Table S1), we first characterized the oncogenic signaling pathway dysregulation landscape by calculating the enrichment scores of 10 pathways for each patient with the GSVA package in R (Supplementary Figure S1D) (Hanzelmann et al., 2013), and then using the consensus cluster analysis (Wilkerson and Hayes, 2010; Gan et al., 2018), we identified distinct clusters with the oncogenic signaling pathway dysregulation landscape. To get the more robust clustering results, the consistency of the clustering results was evaluated between different cluster methods and measurements. There were about 81% of the clustering results whose consistency

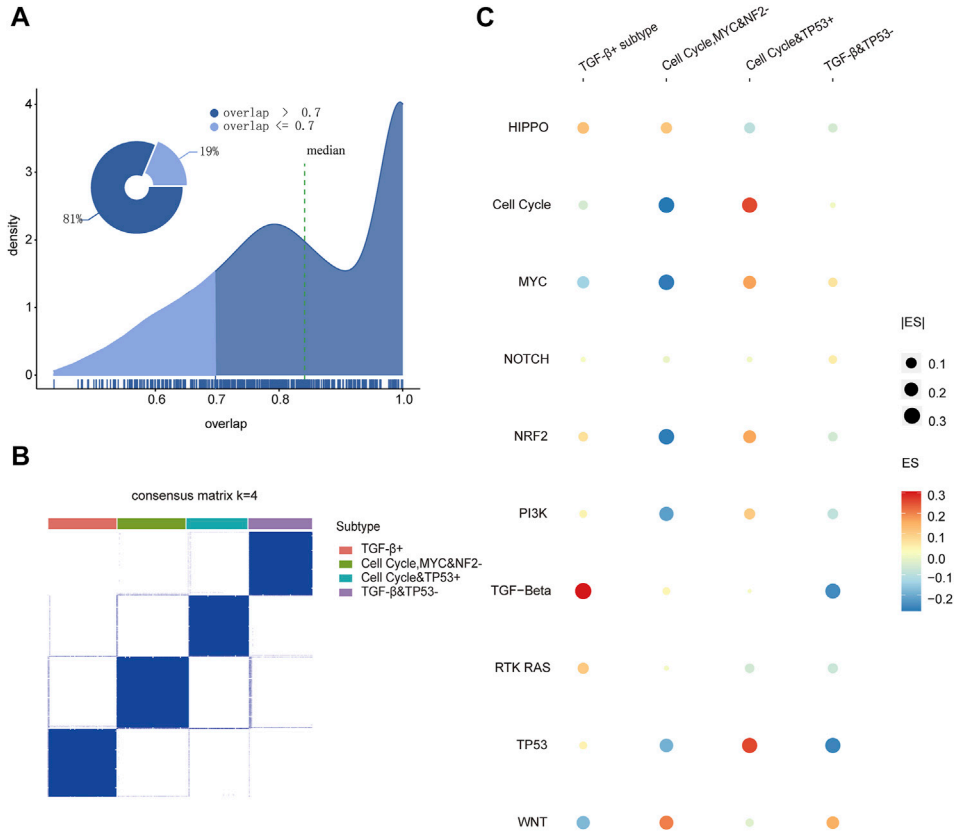


FIGURE 1
(A) Coincidence rate of the clustering results under different clustering methods and measures. **(B)** Consistent clustering result by the kmdist-Spearman method to cluster the TCGA training set into four classes. **(C)** Dot plot of the enrichment scores for 10 pathways in four subtypes.

rate reached 0.7 in all the cluster results. It showed that the clustering results were consistent under different clustering methods and measurements, which suggests that there are significant different subtype patient groups in pan-cancer (Figure 1A). Then, the consensus clustering results when k = 2–8 were discussed (Supplementary Figure S1). The variation trend of the area under cumulative density function curve (CDF) is shown in Supplementary Figure S1B, and the result at k = 4 was the inflection point in all outcomes. Under k = 4, we observed the clarity of classification of clustering results among 112 clustering results, and then, the result with the measurement of the kmdist-Spearman method was considered as the final result of clustering (Figure 1B), which indicates that four robust consensus molecular subtypes driven by specific oncogenic signaling pathways were identified. The heat map shows the enrichment score profile of 10 pathways for four molecular subtypes in Figure 1C; it exhibits that the TGF-β pathway is upregulated in subtype 1, then cell cycle, MYC and NF2 pathways are downregulated in subtype 2, while subtype 3 is basically opposite to subtype 2, and the TGF-β and TP53 pathways are downregulated in subtype 4. Therefore, we named the four molecular subtypes based on the characteristics of being driven by the pathways as the TGF-β+ subtype (subtype 1); cell cycle, MYC, and NF2− subtype (subtype 2); cell cycle and TP53 + subtype (subtype 3); and TGF-β and TP53− subtype (subtype 4), respectively.

Four subtypes based on oncogenic signaling pathways show phenotypic and clinical heterogeneity

To explore if there is the phenotypic and clinical heterogeneity among those oncogenic signaling pathway–driven molecular subtypes, we first continued to compare the survival differences among patients in various molecular subtypes using the Kaplan–Meier curve and the log-rank test (Xie and Liu, 2005). It represents significant differences in overall survival and disease-free survival time among the patients of the four subtypes. The TGF-β+ subtype had significantly better overall survival (OS) and disease-free survival (DFS) (OS: Log rank, $p < 0.0001$, Figure 2A; DFS: Log

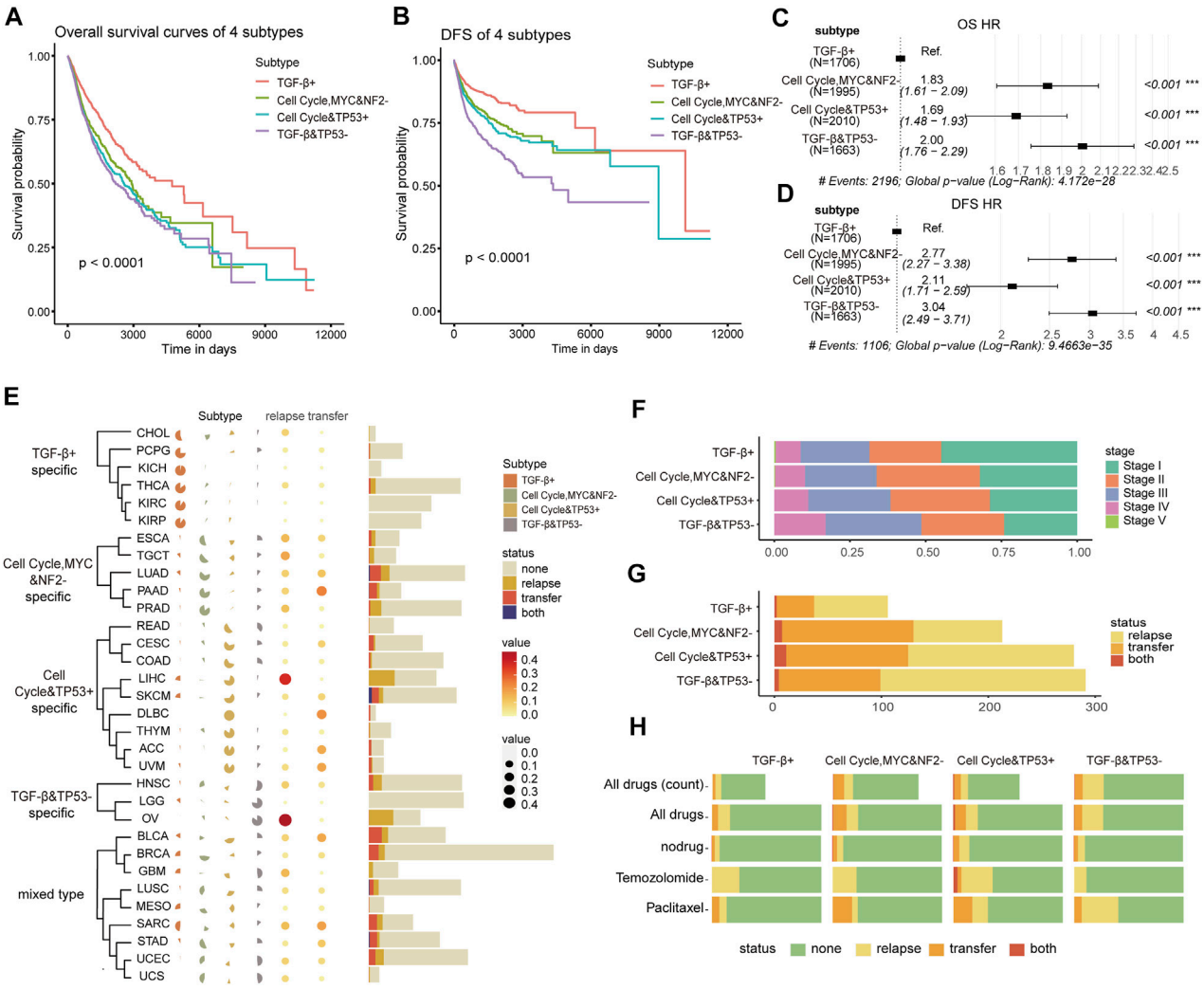


FIGURE 2
(A,B) Kaplan–Meier curves of the overall survival (OS) and disease-free survival (DFS) among the four subtypes in TCGA training cohort. **(C,D)** Forest plot of single Cox regression analysis on subtypes for OS and DFS. The hazard ratios are shown with 95% confidence intervals. **(E)** Percentage heat map shows the distribution of 32 cancers in four subtypes (left), and the dotted heat map and histogram show the distribution of 32 cancers in the recurrence and metastasis state (right). **(F)** Distribution of the four subtypes in the pathological stage. **(G)** Distribution of the four subtypes in the recurrence and metastasis state. **(H)** Sensitivity to drugs of four subtypes' patients. (None: sensitive; relapse, transfer, both: insensitive).

rank, $p < 0.0001$, Figure 2B) than other subtypes. To investigate whether these results hold for a specific cancer type or were only valid to "pan-cancer", we ran an analysis of the differences in survival curves across the four subtypes within each cancer type. It showed that the results of the survival curves remained similar when compiling everything into pan-cancer; there were differences only in CHOL, COAD, and THCA, but it was considered due to the small sample size of the subtype (Supplementary Figure S3).

A Cox hazard regression analysis was used to compare the hazard ratio of OS and DFS among the four subtypes. Using the TGF-β+ subtype as the reference group, we found that the other three subtypes were significantly at a high risk for both OS and DFS, suggesting that there was a relationship between poor prognosis and the molecular subtypes driven by oncogenic signaling pathways (Figures 2C,D). The results showed that the hazard ratio of the TGF-β+ subtype was different from the other three subtypes, indicating that the subtype characteristics were independent predictors of patient survival. The multiple Cox regression analysis also revealed that the pathological stage was a risk factor for poor prognosis (Supplementary Figures S2A,C). Then, we investigated if age and sex contributed to the different hazard ratios among these subtypes and found out that age > 60 was an important high risk

factor for survival both in OS and DFS, but the sex information contributed to the hazard ratio only in OS (Supplementary Figures S2B,D).

Next, we analyzed the distribution of cancer types among patients to find out whether a cancer type is specifically enriched in these subtypes. Cancers in the kidney with relatively better prognosis are mainly enriched in the TGF-β+ subtype, intestinal cancers are predominant in cell cycle and TP53 + subtype-specific, and head and neck cancers are enriched in TGF-β and TP53− subtype. This demonstrated that the distribution of cancer types in the molecular subtypes may be tendentious, so we categorized cancer types by molecular subtypes to understand whether cancer type specific to the same subtype tend to be driven by the same pathways, leading to similar mechanisms of cancer pathogenesis. The TGF-β+ subtype was significantly enriched in CHOL, PCPG, KICH, THCA, KIRC, and KIRP; THE cell cycle, MYC, and NF2− subtype was significantly enriched in ESCA, TGCT, LUAD, PAAD and PRAD, and the cell cycle the TP53 + subtype was significantly enriched in READ, CESC, COAD, LIHC, SKCM, DLBC, THYM, ACC, and UVM; the TGF-β and TP53− subtype was significantly enriched in HNSC, LGG, and OV. Nonetheless, the other nine mixed cancer types of BLCA, BRCA, GBM, LUSC, MESO, SARC, STAD, UCEC, and UCS were classified as mixed carcinomas, and there was no significant difference enrichment among those subtypes (Figure 2E). Furthermore, we continued to check if the patients of cancer types enriched in the subtypes with poor prognosis tend to metastasis or recurrence. The proportion of recurrence and metastasis of the patients in CHOL, PCPG, KICH, THCA, KIRC, and KIRP enriched in the TGF-β+ subtype were significantly lower than those of other cancers, and the recurrence rate of the cell cycle, MYC and NF2 subtype-specific patients was significantly higher than that of metastasis. Most patients with cell cycle and TP53+ subtype-specific cancers were more likely to develop metastases than local recurrence. There was no significant difference in recurrence and metastasis of mixed carcinomas. In other words, four oncogenic pathway–related subtypes have tissue specificity and are closely related to the recurrence and metastasis.

We further explored the reasons for differences in patient survival and analyzed the pathological stage distribution of patients among the four subtypes. From the TGF-β+ subtype to TGF-β and TP53− subtype, the proportion of patients in the early stage gradually decreased and the proportion in the late stage gradually increased, which was consistent with the survival analysis, indicating that the four subtypes' patients have significant differences in pathological stages (Figure 2F). At the same time, the patients of the four subtypes also showed differences in recurrence and metastasis rates. The patients of the TGF-β+ subtype with the best prognosis owned the lowest rate of recurrence and metastasis, while the patients of the TGF-β and TP53− subtype with the worst prognosis owned a lower metastasis rate than the patients of the cell cycle, MYC, and NF2− subtype and cell cycle and TP53 + subtype, but it had a significantly higher recurrence rate (Figure 2G), indicating that those subtypes' patients owned specific pathogenic molecular mechanisms which determined the postoperative pathological stage of the patient. Then, we used Fisher's test to analyze the status of recurrence and metastasis of patients after drug treatment in four subtypes. First, we screened out the drugs which were used by more than 50 patients for analysis (Supplementary Figure S4A). The patients of the TGF-β+ subtype showed the smallest proportion of recurrence or metastasis after drug treatment. Temozolomide was significantly less sensitive in cell cycle and TP53 + subtype patients ($p < 0.05$), and paclitaxel was significantly less responsive in TGF-β and TP53− subtype patients (Figure 2H; Supplementary Figure S4B). It means that temozolomide may be related to the upregulation of the activity of the cell cycle and the TP53 pathway and is also effective for the diseases caused by the dysregulation of these two pathways.

Collectively, the patients of four subtypes based on oncogenic signaling pathways had significant differences in clinical phenotypes, such as survival time, tissue specificity, tumor stage, recurrence and metastasis rates, and drug response. The patients with the upregulated TGF-β pathway had the best prognosis, while patients with downregulated TGF-β and TP53 pathways had the worst prognosis. These data imply that the pathogenesis of cancer is strongly correlated with the molecular mechanisms of oncogenic signaling pathways, and the dysregulation of pathways might be the driving factor for cancer development.

Novel subtype and prognostic classifiers were constructed based on the genes related to prognosis among subtypes

To figure out whether transcriptional changes among subtypes are related to the dysregulation of specific signaling pathways, we estimated the gene expression difference in these pathways in TCGA training cohort. According to 333 cancer-related pathway gene expressions, 65 differentially expressed genes between each subtype and other subtypes were identified ($p < 0.05$ and fold change $|\log2FC| > 1$, Supplementary Table S2) (Robinson et al., 2010). Most genes showed high expression in the cell cycle and TP53 + subtype and the TGF-β and TP53− subtype, and just a few genes showed high expression in the TGF-β+ subtype and the cell cycle, MYC, and NF2− subtype (Figure 3A). Thereafter, by mapping differentially expressed genes into these oncogenic signaling pathways, some subtype-specific key sub-pathways with consistent transcriptional change were identified (Figure 3B; Supplementary Figures S5–S7). For example, NF2 and WWC1 were highly expressed in the TGF-β+ subtype, which promotes the high expressions of LATS1, SAV1, and other genes

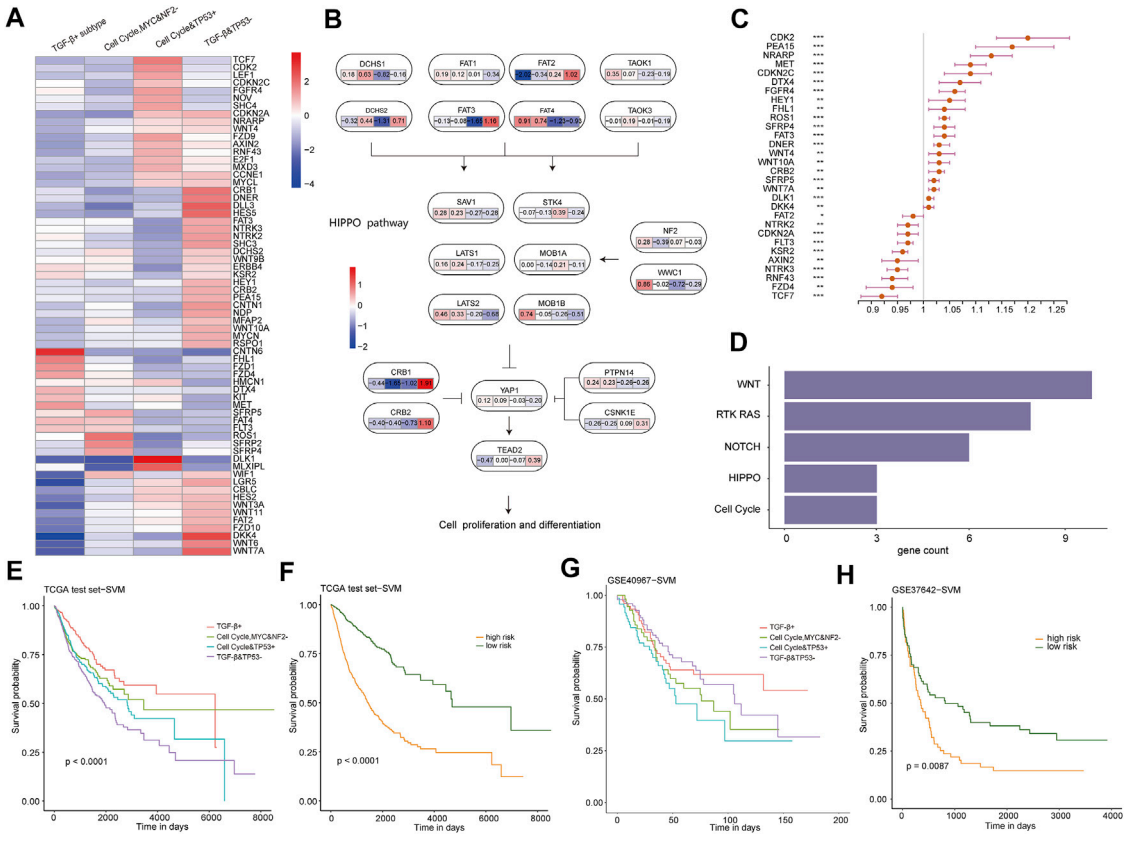


FIGURE 3
(A) Heat map of the log2FC value of differentially expressed genes in four subtypes. (FC, fold change, the ratio of the average mRNA expression for each cancer pathway–related subtype to the average mRNA expression for samples not of the aforementione subtype. Red, upregulated; blue, downregulated.). **(B)** Interaction of genes in the HIPPO pathway and the FC value of the four subtypes of the gene. **(C)** Forest plot of multivariate Cox regression analysis for 30 genes related to prognosis. The hazard ratios are shown with 95% confidence intervals (***$p < 0.001$; **$p < 0.01$; *$p < 0.05$; and $p < 0.1$). **(D)** Pathways in which 30 differentially expressed genes are enriched. **(E)** KM survival curves of the subtype classifier constructed using samples from TCGA test cohort by the SVM method. **(F)** KM survival curves of the prognosis classifier constructed using samples from TCGA test cohort by the SVM method. **(G)** KM survival curves of the subtype classifier constructed using GSE40967 by the SVM method. **(H)** KM survival curves of the prognosis classifier constructed using GSE37642 by the SVM method.

in the TGF-β+ subtype, whereas CRB1 and CRB2 were highly expressed in the TGF-β and TP53− subtype, inhibiting the *YAP1* gene, making it lowly expressed in the TGF-β and TP53− subtype in the HIPPO pathway. This result showed that the different driver genes might lead to the different pathway changes in the TGF-β+ subtype and the TGF-β and TP53− subtype, which suggests that the oncogenic signaling pathways own subtype-specific driving sub-pathways, resulting in different states of dysregulation of downstream pathways.

Furthermore, we explored whether these 65 differentially expressed genes would predict a worse prognosis in pan-cancer. A total of 56 prognostic-related genes were identified by using a single Cox regression analysis, and multivariate Cox proportional hazard models revealed 30 genes which can predict worse prognosis (Figure 3C). These 30 genes were enriched into WNT, RTK-RAS, NOTCH, HIPPO, and cell cycle pathways (Figure 3D). In particular, there are 10 differentially expressed genes associated with prognosis enriched in the WNT pathway, which might be part of the reason for the upregulation of the WNT pathway activity in the cell cycle, MYC, and NF2− subtype and TGF-β and TP53− subtype. Collectively, our results demonstrate that there are strong relationships between pathway dysregulation and the subtypes. We further explored to construct a subtype and prognostic classifiers, based on the expression profiles of these 30 genes, by using the support vector machine (SVM) method. Then, TCGA test cohort was used to verify these 30 genes as biomarkers for predicting subtype and prognosis, and the classifiers' results in survival were also very significant (Figures 3E,F, $p < 0.0001$). At the same time, GSE40967 and GSE37642 data on the GPL570 platform from the Gene Expression Omnibus (GEO) database were downloaded as validating (external verification) data sets. Then, SVM was used to build the classifiers, and the classification result had significant survival differences between our prognostic

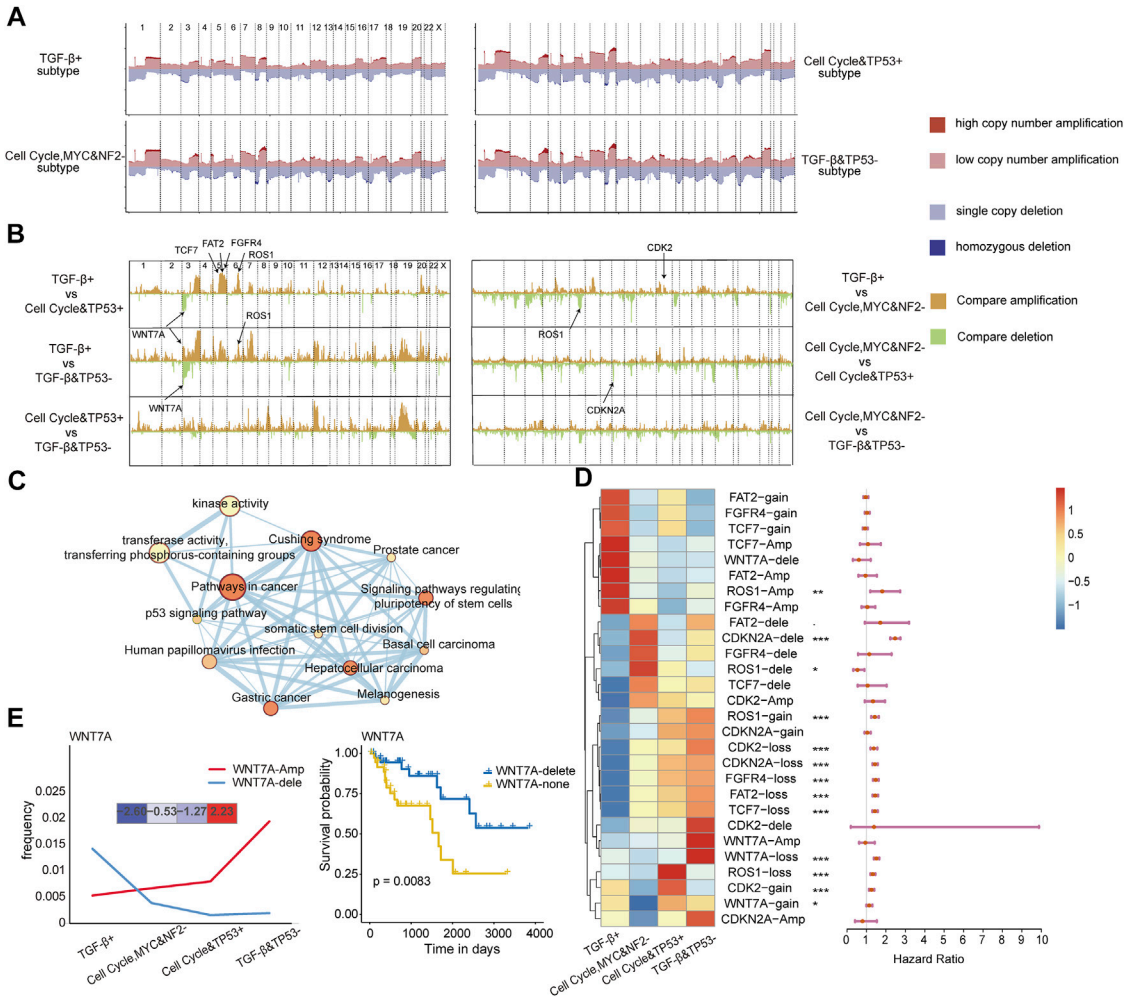


FIGURE 4
(A) Somatic CNA frequency of individual genes in each subtype plotted along the chromosomes. **(B)** Comparisons of somatic CNA between subtypes with −log10 FDR plotted along the chromosomes (Fisher's exact test). **(C)** Interaction of the enriched pathways. The size represents the number of genes, and the color represents the *p*-value. **(D)** Differences in copy number variation across the four subtypes of the four copy number variation states of the seven genes and their relationship with the prognosis. **(E)** Changes in the number of amplified and deleted samples of WNT7A in the four subtypes; the expression of WNT7A in the four subtypes (left) and the difference in survival between the two categories (right).

subgroups (Figure 3H). The Kaplan–Meier curve of GSE40967 data also revealed distinct prognostic outcomes among the predicted subtypes, although the difference was not statistically significant ($p = 0.08$, Figure 3G), possibly due to the single cancer type included in the data.

These results suggest that these 30 differentially expressed genes associated with prognosis among subtypes could be recognized as key genes in oncogenic signaling pathways and biomarkers for identifying molecular subtypes and risk groups, and their expression changes can also affect the expression of upstream and downstream genes through the relationship of promotion or inhibition between genes, leading to dysregulation of oncogenic signaling pathways.

Oncogenic signaling pathway–based subtypes show distinct genomic alteration features

Genomic alterations can drive oncogenic signaling pathway reprogramming in cancers. We further explored to compare genomic alterations among the four subtypes with the copy number variation data on 22,445 genes obtained from UCSC Xena for TCGA pan-cancer patients. Genome-wide copy number variation revealed that the TGF-β and TP53– subtype had a significantly higher copy number variation, especially on chromosomes 3, 4, and 19, as shown in Figure 4A. We further examined the detailed characterization of copy number variation

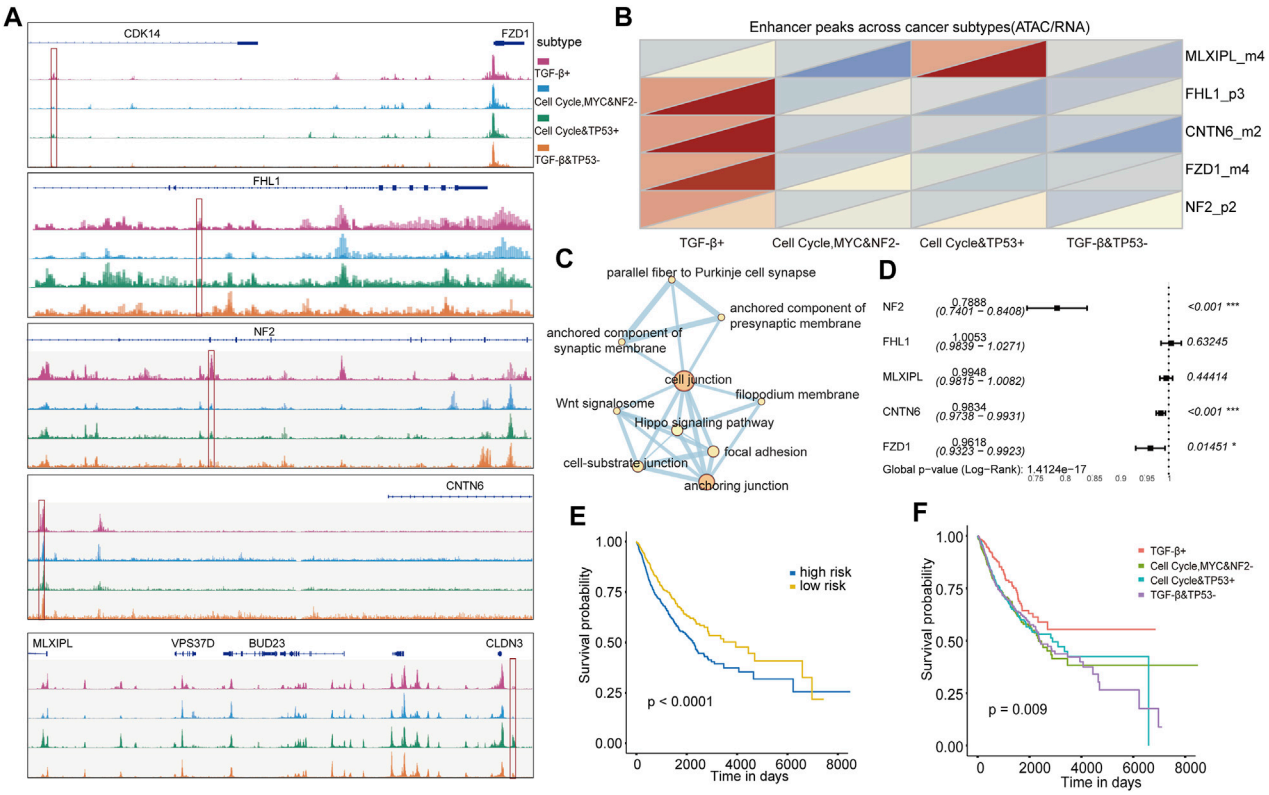


FIGURE 5
(A) Five enhancers visualized using IGV. **(B)** ATAC accessibility (upper triangle) and expression (lower triangle) heat map of five enhancer-related genes in four subtypes. **(C)** Pathways in which five enhancers are enriched. **(D)** Multi-Cox risk regression model of five enhancers. **(E)** KM survival curve for classifying the high and low risks of TCGA validation set samples was constructed by five enhancers as features. **(F)** KM survival curve of TCGA validation set samples divided into four subtypes by the classifier was constructed with five enhancers as features.

across the subtypes. Between any two subtypes, the differences of all genes in copy number amplification and deletion (−log10 FDR value) were calculated using Fisher's exact test (Figure 4B). There were significant difference peaks on chromosomes 3, 5, and 6 between the TGF-β+ subtype and the cell cycle and TP53+ subtype; on chromosomes 3, 5, 7, and 19 chromosomes between the TGF-β+ subtype and the TGF-β and TP53− subtype; and on chromosomes 12 and 19 between the cell cycle and TP53 + subtype and the TGF-β and TP53− subtype. Combined with pathway-related genes, especially in chromosome 3, we found that there were seven genes, namely, *FAT2*, *CDK2*, *CDKN2A*, *WNT7A*, *TCF7*, *FGFR4*, and *ROS1*(−log10 FDR>2), which had significant differences in the copy number between subtypes.

To further explore the biological functions of these seven genes, we performed a pathway enrichment analysis for these genes. In addition to affecting oncogenic pathways, we further examined the biological functions of these genes to see if they affect cancer development from other perspectives. The results showed that the seven genes were also enriched in the pathways, including Cushing syndrome, and the pathways of cancer and kinase activity (Figure 4C). These genes were indeed involved in cancer development as a multifunctional model, and this result suggests that the genomic alterations of these genes may drive the dysregulation of oncogenic signaling pathways. Then, the copy number variation states of the seven genes which had different copy number changes between subtypes were disassembled to analyze. We found that FAT2-amp, FGFR4-amp, TCF7-amp, and WNT7A-delete showed upregulation in the TGF-β+ subtype, and most other genes showed upregulation in the other three subtypes. Multivariate Cox proportional hazard models also revealed the prognosis-related states in non-diploid normal copy states of the seven genes (Figure 4D). The amplification frequency of WNT7A gradually increased from the TGF-β+ subtype to the TGF-β and TP53− subtype, and the frequency of WNT7A deletion gradually decreased from the TGF-β+ subtype to the TGF-β and TP53− subtype. The most deleted changes and the least amplification changes of WNT7A in copy number variation were observed in the TGF-β+ subtype, which was similar to the WNT7A gene expression trend among the four subtypes. It shows that the copy number variation change of WNT7A affects its expression on the transcriptome and thereby affects the function of the WNT pathway, and this

result suggests that the copy number variation of WNT7A could be a driver factor for WNT pathway dysregulation. Then, we continued to select the four copy number variation states of WNT7A as biomarkers for diagnosis. Notably, the survival of patients with homozygous deletion of WNT7A was significantly better than that of patients with normal diploid copies of WNT7A (p = 0.0083, Figure 4E), which validates the efficacy of WNT7A as a prognostic marker.

Five subtype-specific enhancers were identified by a chromatin accessibility analysis

The integration of transcriptome data and ATAC-seq could determine a great deal of putative distal enhancers (Corces et al., 2018). We continued to identify subtype-specific transcriptional regulators that influence patterns of oncogenic pathway dysregulation at the level of chromatin accessibility by integrating ATAC-seq data with RNA-seq data for pan-cancer cases in TCGA. A total of 2,579 differential ATAC peaks between any subtypes were identified, and we found that there were five enhancers showing subtype-specific activity in the oncogenic signaling pathways such as CNTN6 in the NOTCH pathway and MLXIPL in the MYC pathway. Furthermore, a location analysis of these peaks showed that these subtype-specific enhancers' chromosome locations were distinct. For example, FZD1_m4 and CNTN6_m2 were located in the distal upstream of the related genes, FHL1_p3 and NF2_p2 were located in the inner gene, and MLXIPL_m4 was located in the distal downstream of related genes (Figure 5A). We further investigated whether these enhancers located in different chromosomal regions could lead to its expression change. Subsequently, expression of these subtype-specific enhancer-related genes was analyzed, and it was found that the changes in chromosome accessibility and gene expression showed a consistent trend (Figure 5B). For example, MLXIPL displayed high chromatin accessibility and gene expression level in the cell cycle and TP53 + subtype, whereas it showed the opposite trend in the cell cycle, MYC, and NF2− subtype. This result suggests that these subtypes own their specific transcriptional regulators, which drive oncogenic signaling pathway dysregulation by distinct molecular mechanisms.

To understand the molecular function of the enhancers, we performed the functional enrichment analysis and found that the five genes were enriched in several other pathways, including cell junction and anchoring junction pathways (Figure 5C). The junctions of these pathways might affect cancer cell adhesion and further affect the possibility of metastasis. We further explored whether these subtype-specific enhancers could be used to predict a worse OS and construct subtype classifiers. NF2, CNTN6, and FZD1 presented very low risk. It suggests that the genes of *NF2*, *CNTN6*, and *FZD1* might be low risk factors for poor prognosis (Figure 5D). Using these five genes as features, we constructed subtype and prognostic classifiers using the random forest method. The patients were divided into high- and low-risk groups with significant survival differences according to the prognostic model risk score ($p < 0.0001$, Figure 5E). Also, the survival differences were also significant for subtype classifiers (p = 0.009, Figure 5F). Overall, our analysis revealed that these five genes can serve as key biomarkers for identifying patient prognostic risk and subtypes based on oncogenic signaling pathways.

Pathway-driven subtype-associated methylation sites were identified

In tumor cells, proto-oncogenes are in a state of hypomethylation and activated, while tumor suppressor genes are in a state of hypermethylation and inhibited (Kulis and Esteller, 2010; Gyorffy et al., 2016; Chen et al., 2021). Next, we explored whether some methylated CPG sites had DNA methylation abnormalities due to subtypes driven by the oncogenic signaling pathway. We further performed a methylated CPG site analysis, and 11,122 differential methylated sites were identified. According to the methylation sites' position on the gene, the differential methylation site of each subtype was classified (Figures 6A,B). There were the least differential methylated sites in 3′UTR, and most of the differential methylated sites were located on CpG islands. There were a few differential methylated sites in the TGF-β+ subtype, but much more in the cell cycle and TP53 + subtype and the TGF-β and TP53− subtype.

To compare methylation sites' difference across subtypes, we further used the weighted gene co-expression network (WGCNA) (Langfelder and Horvath, 2008) to explore the subtype-specific driving methylation sites from the 11,122 methylation sites. After screening, the soft thresholding power of the WGCNA was 12 (Figure 6C). The network was constructed to classify all methylation sites into four modules (gray, brown, turquoise, and blue, Figure 6D). The correlations between the four modules and the subtype characteristics were obtained by using the phenotypic data on the patients (Figure 6E). It can be seen that methylation sites in the turquoise module are not only related to the turquoise module but also to its corresponding traits (Figure 6F), which further indicates that these sites are worthy of in-depth exploration.

We continued to identify subtypes related to the methylation sites, whose correlation with the turquoise module was greater than 0.8 and the subtype with a correlation greater than 0.25, and it revealed the strong correlations among these methylation sites. A total of 15 genes mapped by the DMSs (differentially methylated sites with a degree greater than 300) were identified (Supplementary Table S3). Furthermore, we

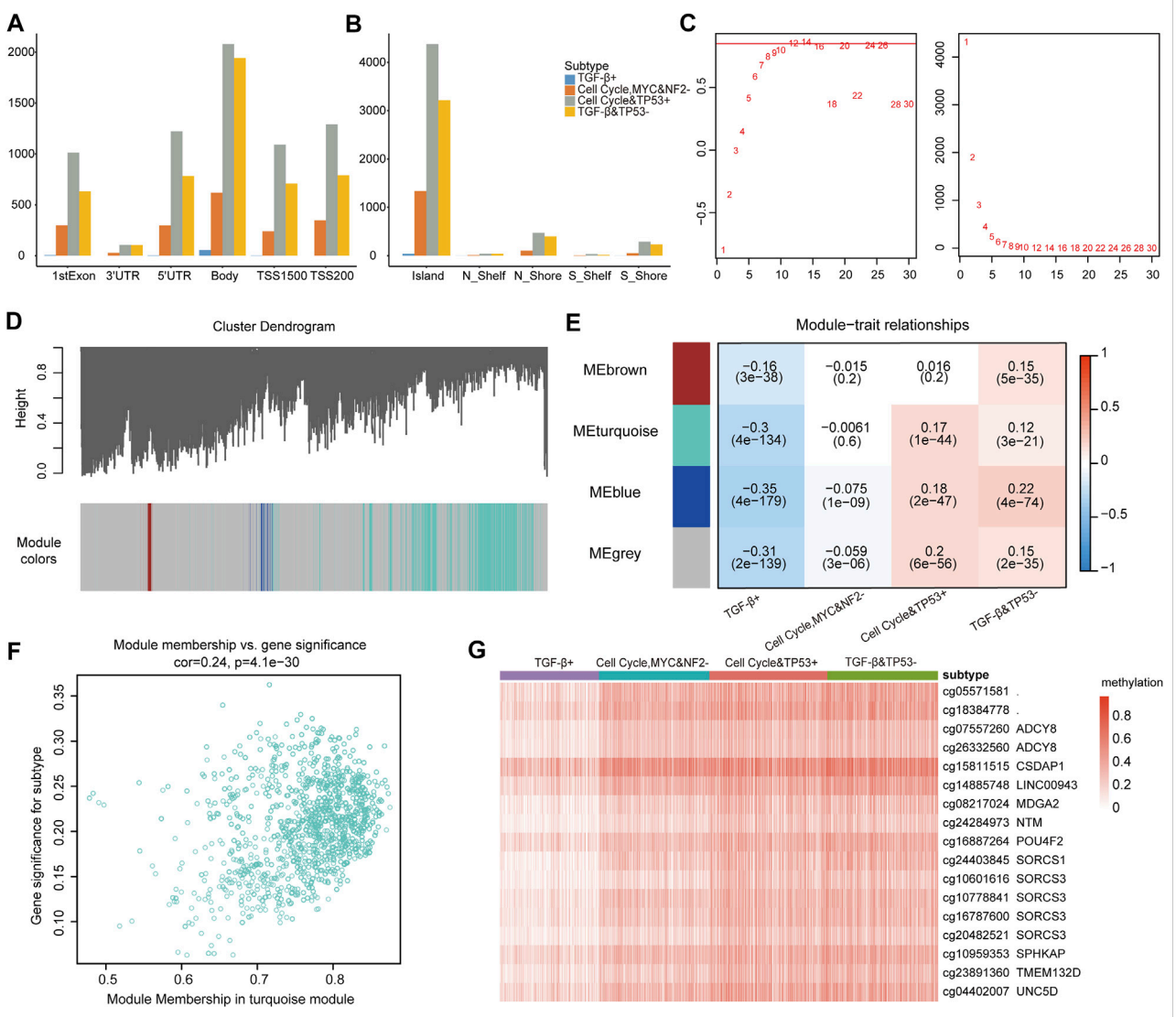


FIGURE 6
(A) Location of methylation sites on genes. **(B)** Location of methylation sites on CPG islands. **(C)** Analysis of network topology for different soft thresholding powers. **(D)** Module colors and gene dendrogram. **(E)** Correlations between the four modules and the subtype characteristics of samples. **(F)** Weighted interaction gene network of the CpG sites in the turquoise module. **(G)** Methylation profiles of the 15 CpG sites.

estimated the methylation level of genes across subtypes. The methylation level of patients from the TGF-β+ subtype was significantly lower than that of patients from other subtypes, and the patients from the TGF-β+ subtype also had better prognosis than patients from other subtypes (Figure 6G). Then, these genes were mainly enriched in WNT, NOTCH, and RTK-RAS pathways. The results indicate that oncogenic signaling pathway–based subtypes are closely related to the methylation status, and the genes annotated at these 15 CpG sites are closely related to the dysregulation of oncogenic signaling pathways; also, hypomethylation is associated with a better prognosis for patients.

Identification of tumor microenvironment–associated immune biomarkers across subtypes

The tumor microenvironment (TME), the environment for tumor cells to survive, could facilitate tumor cell growth, metastasis, and immune escape. We estimated whether these oncogenic signaling pathway–based subtypes would show distinct tumor microenvironment characteristics. We first analyzed the infiltration level of immune cells estimated by TIMER and MCP of the four subtypes' patients and found these subtype patients with specific tumor microenvironments.

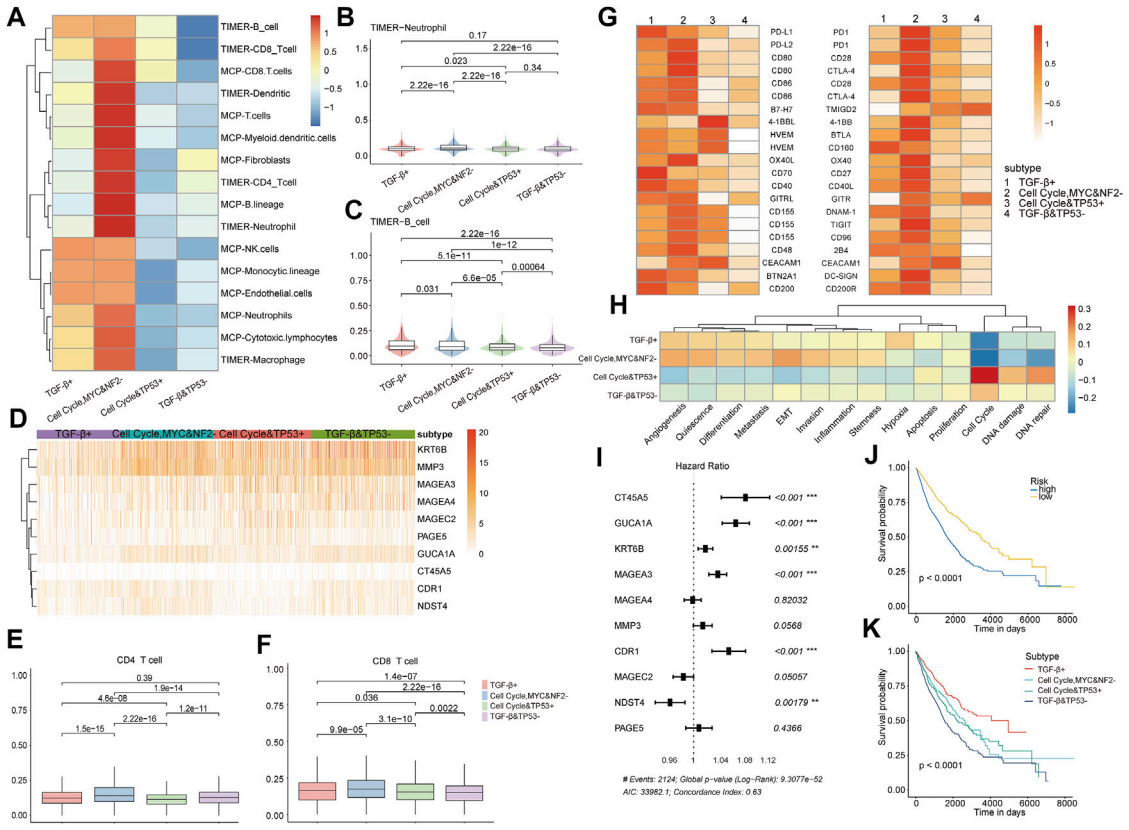


FIGURE 7
(A) Normalized infiltration of immune cells from TIMER and MCP in the four subtypes. **(B,C)** Violin chart of TIMER neutrophils and B-cell infiltration. **(D)** Gene expression heat map of subtype-specific T-cell dysfunction genes. **(E,F)** Boxplot of TIMER CD4 T-cell and CD8 T-cell infiltration. **(G)** Immune checkpoint tumor cells (left) and immune checkpoint T cells' (right) corresponding gene normalized expression in the four subtypes. **(H)** Cell state enrichment scores' heat map. **(I)** HR values of 10 subtype-specific T-cell dysfunction genes. **(J,K)** KM survival curves of the prognostic classifier (left) and subtype classifier (right) constructed with 10 subtype-specific T-cell dysfunction genes as features.

It was mainly reflected in the fact that the infiltration of most immune cells in the cell cycle, MYC, and NF2− subtype was significantly higher than that of other subtypes (Figure 7A, Supplementary Figure S8), especially neutrophils (Figure 7B) and B cells (Figure 7C), whereas we found that patients of the cell cycle, MYC, and NF2− subtype had a higher level of immune cell infiltration but a poor prognosis. Therefore, we continued to analyze this issue from the perspective of immune cell function such as T-cell dysfunction (Jiang et al., 2018; Zhao et al., 2020) and immune checkpoints. A total of 10 differentially expressed T-cell dysfunction–related genes were identified across subtypes, and these genes all showed significantly high expression in the cell cycle, MYC, and NF2− subtype (Figures 7D–F), which suggested that most patients in the cell cycle, MYC, and NF2− subtype exhibited a state of T-cell dysfunction. We continued to check the immune checkpoint genes' expression level across subtypes and found that immune checkpoint genes also tended to be highly expressed in the cell cycle, MYC, and NF2− subtype (Figure 7G). Immune checkpoint genes were overexpressed, which can lead to suppressed immune function and cause low body immune capacity. In general, our analysis suggests that high gene expression of T-cell dysfunction and immune checkpoint genes might be responsible for the patients owning a higher level of immune cell infiltration but a lower prognosis in the cell cycle, MYC, and NF2− subtype. Next, we analyzed 14 cell states of the four subtypes' patients based on the gene set variation analysis (GSVA) (Yuan et al., 2019). Most cell states except the cell cycle showed upregulation in the TGF-β+ subtype and the cell cycle, MYC, and NF2− subtype, and cell cycle, DNA damage, and DNA repair showed upregulation in the cell cycle and TP53 + subtype (Figure 7H). Overall, the aforementioned results reveal significant differences in immune cell infiltration, T-cell function, and cell state across subtypes.

Next, we continued to analyze whether the aforementioned 10 T-cell dysfunction gene expression models could predict patient prognosis and subtype. There were five genes with significantly high risk, and only one gene showed significantly

low risk (Figure 7I). These genes were identified as key prognostic factors and then used as features to construct prognosis and subtype classifiers; both classifiers showed great performance (KM survival curve, log rank: $p < 0.0001$, Figures 7J,K).

Materials and methods

TCGA data sets

The gene expression data on 32 cancers including 9,398 samples were downloaded from UCSC Xena (https://xenabrowser.net/), and the data types were mRNA count-UQ and mRNA FPKM-UQ. We divided all TCGA patients into the training data set (80%) and the test data set (20%).

Then, the copy number variation data on 22,445 genes were obtained from UCSC Xena. The copy number variation data on TCGA samples included the four non-diploid normal copy states of homozygous deletion (−2), single copy deletion (−1), low copy number amplification (1), and high copy number amplification (2).

The clinical data on TCGA samples including gender, age, tumor weight, TNM stage, and survival time were downloaded by the GDC tool (https://portal.gdc.cancer.gov/).

Gene expression omnibus data sets

We downloaded GSE40967 and GSE37642 data on the GPL570 platform from the Gene Expression Omnibus (GEO) database as an external validation data set. (https://www.ncbi.nlm.nih.gov/geo/). GSE40967 contained two sets of data, GSE39582 had 585 tumor samples including 566 CC samples and 19 non-tumor samples. GSE40966 had 566 tumor samples. The data contained clinical information including sex, age, TNM stage, treatment strategy, survival time, and mutation information. GSE37642 contained the expression data on 562 samples of adult acute myeloid leukemia (AML) patients. The clinical information included age and survival status.

ATAC-seq data set

The genome-wide chromatin accessibility profiles (Corces et al., 2018) of 410 tumor samples spanning 23 cancer types from TCGA were downloaded by the GDC tool (https://portal.gdc.cancer.gov/).

Immune cells

The tumor purity of the six immune cells, namely, B cells, CD4+ T cells, CD8+ T cells, neutrophils, macrophages, and dendritic cells of TCGA cancer patients, were available from TIMER (version 1.0) (Li et al., 2017) (http://cistrome.dfci.harvard.edu/TIMER/).

DNA methylation data

The DNA methylation 450 k data on 31 cancers were downloaded from UCSC Xena. The data recorded the DNA methylation value (β value) of each array probe in each sample. The DNA methylation value is a continuous variable between 0 and 1, which represents the degree of methylation. A higher β value represents hypermethylation, and a lower β value represents hypomethylation.

We used the Xena probeMap derived from GEO GPL13534 to map the microarray probes to the coordinates of the human genome, displaying the annotation information of all methylation sites, including base changes, chromosomes, CPGs, and gene positions.

Gene set variation analysis to calculate the enrichment score of each pathway

Gene set variation analysis (GSVA) (Hanzelmann et al., 2013) is a non-parametric, unsupervised method that estimates the enrichment score of each gene set based on the gene expression level. We used the R package "GSVA" (version 1.38.2) to calculate the enrichment scores of 10 oncogenic signaling pathways for each sample and built a pathway dysregulation profile. In the profile, the enrichment score greater than 0 means that the pathway activity is upregulated, while an enrichment score less than 0 indicates that the pathway activity is downregulated. The enrichment score is close to 0, which means that there is little difference in the pathway activity (http://www.bioconductor.org/packages/release/bioc/html/GSVA.html).

Consensus cluster on training samples

We used the ConsensusClusterPlus package (version 1.54.0) in R (Wilkerson and Hayes, 2010) to perform consistent clustering on the pathway dysregulation profile obtained by the GSVA method. The optimal number of clusters is determined by the cumulative density function (CDF), which plots the corresponding empirical cumulative distribution defined in the range between 0 and 1, and the optimal cluster is determined by calculating the proportional increase in the area under the CDF curve number. When any further increase in the number of clusters (K) does not result in a corresponding significant increase in the area of the CDF, the number of clusters is determined.

Our consistent clustering methods included pam, kmdist, and hc, and clustering measures included Pearson, Spearman, maximum, Minkowski, Manhattan, binary, Canberra, and Euclidean methods. Using each method and each measurement to cluster cancer samples, the number of categories ranged from 2 to 8, reps = 50, pItem = 0.8, and pFeature = 1, and a total of 112 clustering results were obtained.

Then, under the same clustering number, we compared the overlapping rate among these clustering results using the Wilcoxon rank-sum test.

Kaplan–Meier and log-rank tests

We used the R packages "survival" (version 3.2–7) and "survminer" (version 0.4.9) to calculate the survival difference among subtypes; log rank $p < 0.05$ represents a significant difference.

Identification of differentially expressed genes

Subtype-specific differentially expressed genes were identified (Wilcoxon text $p < 0.05$; $|\log2FC| > 1$) by using the R packages "edgeR" (version 3.32.1) (Robinson et al., 2010) and "limma" (3.46.0).

Cox proportional hazards regression model

We performed a univariate Cox regression analysis on 65 differentially expressed genes among subtypes ($p < 0.01$), and then, 56 genes that correlated with the prognosis were identified ($p < 0.01$). Then, the multivariate Cox proportional analysis was performed, and 30 genes were regarded as candidate prognostic genes. To identify independent predictors that significantly contributed to OS or RFS, we constructed a risk model based on these 30 genes and calculated the risk score of each patient using the predict() function in the survival package.

$$\text{RiskScore} = \sum \beta i \times Xi,$$

where βi represents the risk regression coefficient of the multiple Cox analysis corresponding to each gene, and Xi represents the gene expression value. The samples were divided into high- and low-risk groups based on the median value of the risk score for subsequent analysis.

Random forest and support vector machine to construct the subtype and prognosis classifiers

We used random forest and support vector machine (SVM) methods to construct the subtype and prognostic classifiers by using the R packages "randomForest" (version 4.6–14) and "e1071" (version 1.7–6) in the training data set and then used the test data set to test the performance of the classifiers. In the random forest method, we set the cutoff to 0.5 so that every tree "votes". Next, we used the importance function to calculate the accuracy of the model variables and the gini coefficient to judge the importance of the variables. The mean value of the gini index change was used as a measure of the importance of the variables, and all features were sorted according to their importance.

Fisher's exact test

We used Fisher's exact test (Blevins and McDonald, 1985) to calculate the difference in copy number amplification and deletion between each two subtypes ($p < 0.01$; FDR > 2).

Integrative genomics viewer to visualize ATAC-seq data

IGV (Integrative Genomics Viewer) (Thorvaldsdottir et al., 2013) is a tool that can visualize sequencing data on a local computer. For the ATAC-seq bw file of each sample, IGV (version 2.7.0) was used to visualize the chromatin accessibility at the genome position of each subtype.

Weighted gene co-expression network to identify the methylation sites

The R package WGCNA (version 1.70–3) (Langfelder and Horvath, 2008) was used to build a weighted gene co-expression network. First, the soft threshold β was screened to ensure that the constructed network was more in line with the characteristics of the scale-free network. Next, the one-step method was used to construct the network, and gene clustering was performed based on TOM. Then, we used the hierarchical clustering tree to display each module and obtained the correlation between the modules. The correlations between characteristic methylation sites and clinical phenotypes were assessed by Pearson's correlation analysis, and the correlation coefficients between modules and clinical phenotypes were used to select modules for a downstream analysis.

MCP to calculate the cell infiltration fraction

We used the R package MCPcounter (version 1.2.0) (https://github.com/ebecht/MCPcounter) to calculate the infiltration fraction of T cells, CD8 T cells, cytotoxic lymphocytes, NK cells, B cells, monocytes, bone marrow dendrites, neutrophils, endothelial cells, and fibroblasts based on gene expression data in GDC.

Discussion

Cancer subtypes have broad prospects in understanding cancer and personalized treatment (Cao et al., 2018; Guo et al., 2019). However, many studies so far have been based on single cancer. Analyzing from a pan-cancer perspective can identify the differences and commonalities across different cancer types. Signaling pathways change in different combinations among cancers, and there are complex interactions between pathways (Jackstadt et al., 2019; Li et al., 2020). But the extent, mechanism, and co-occurrence of these pathway changes varied across tumors and tumor types.

We divided patients of TCGA 32 cancer types into four molecular subtypes; although our project covered most tissues and organ systems, some tumor types including most hematologic cancers were not included. Also, we did not combine the known molecular subtypes of certain cancer types for our analysis. Then, the biomarkers among subtypes were identified at the multi-omics levels. A multi-omics analysis is of great significance for revealing cancer development, treatment resistance, and recurrence risk, and it is the key to advancing precision medicine in clinical practice. However, we did not conduct further and deeper mining of multi-omics biomarkers we found. In addition, drug sensitivity requires clinical evaluation; then well-designed clinical trials are expected to test the possibility of translating our results to clinical practice in the future.

In conclusion, our study provided a new perspective to understand the relationship of the dysregulation of oncogenic signaling pathways and cancers and identified potential prognostic biomarkers from multiple omics data, and it further might have implications for clinical applications in the future.

Conclusion

Here, based on gene set variation analysis (GSVA), we constructed a pathway dysregulation landscape and identified four subtypes based on oncogenic signaling pathways in pan-caner, which may provide an increased understanding of the common molecular mechanisms driven by oncogenic signaling pathways underlying the pathogenesis of the malignancy. These four subtypes showed distinct patient prognosis, cancer type distributions, transcriptional changes, chromatin accessibility, genomic alterations, methylation degree, and tumor microenvironment characteristics. Several signature sets were identified by integrating multi-omics profiles, which were used to construct a subtype classifier and a prognosis prediction model. Overall, our analysis demonstrates that the molecular heterogeneity of oncogenic signaling pathways, improves the understanding of the mechanisms of oncogenic signaling pathways driving tumor progression, and enables the development of personalized therapies targeting unique tumor oncogenic signaling pathway dysregulation profiles.

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

Conceptualization: Z-ZW and H-JW; methodology: NW; validation: NW, D-NH, and Z-YW; formal analysis: NW; resources: XZ, X-LW, YG, and X-HL; writing—original draft preparation: NW; writing—review and editing: H-JW and Z-ZW; visualization: NW; and supervision: Z-ZW.

Funding

This research was funded by the Natural Science Foundation of Hainan Province [Nos. 821MS045, 822MS074, 821MS0777, and 621MS041], the National Natural Science Foundation of China [No. 31701159 and 32160179], the Major Science and Technology Program of Hainan Province [No. ZDKJ202003], and the Key R&D Projects of Hainan Province [No. ZDYF2022SHFZ055].

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.916400/full#supplementary-material

SUPPLEMENTARY FIGURE S1
(A) CDF curve with k = 2–8 using the kmdist–Spearman method. **(B)** Area under the CDF curve for the kmdist–Spearman method, k = 2–8. **(C)** Sample cluster distribution using the kmdist–Spearman method, k = 2–8. **(D)** Signaling pathway–based clustering results of TCGA training cohort (*n* = 7518). The heat map shows normalized enrichment scores of the four subtypes.

SUPPLEMENTARY FIGURE S2
(A,C) Forest plot of the multiple Cox regression analysis on the pathological stage for OS **(A)** and DFS**(C)**. **(B,D)** Forest plot of the multiple Cox regression analysis on age and sex for OS **(B)** and DFS **(D)**. The hazard ratios are shown with 95% confidence intervals.

SUPPLEMENTARY FIGURE S3
Kaplan–Meier survival curves among the four subtypes of 22 cancers.

SUPPLEMENTARY FIGURE S4
(A) Drugs used by more than 50 patients (red). **(B)** Sensitivity of four subtypes of patients to different situations with treatment. (None: sensitive; relapse, transfer, both: insensitive).

SUPPLEMENTARY FIGURE S5
(A) Interactions of genes in the cell cycle pathway and fold change values of four subtypes of genes. **(B)** Interaction of genes in the MYC pathway and the fold change values of the four subtypes of the genes. **(C)** Interactions of genes in the NOTCH pathway and fold change values of four subtypes of genes.

SUPPLEMENTARY FIGURE S6
Interactions of genes in the RTK/RAS pathway and fold change values of four subtypes of genes.

SUPPLEMENTARY FIGURE S7
Interactions of genes in the WNT pathway and fold change values of four subtypes of genes.

SUPPLEMENTARY FIGURE S8
Infiltration of immune cells in MCP and TIMER in four subtypes.

SUPPLEMENTARY TABLE S1
TCGA dataset.

SUPPLEMENTARY TABLE S2
Differentially expressed genes across the four subtypes.

SUPPLEMENTARY TABLE S3
Differentially methylated sites.

References

Bidkhori, G., Benfeitas, R., Klevstig, M., Zhang, C., Nielsen, J., Uhlen, M., et al. (2018). Metabolic network-based stratification of hepatocellular carcinoma reveals three distinct tumor subtypes. *Proc. Natl. Acad. Sci. U. S. A.* 115 (50), E11874–E11883. doi:10.1073/pnas.1807305115

Bild, A. H., Yao, G., Chang, J. T., Wang, Q., Potti, A., Chasse, D., et al. (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 439 (7074), 353–357. doi:10.1038/nature04296

Blevins, L., and McDonald, C. J. (1985). Fisher's exact test: An easy-to-use statistical test for comparing outcomes. *Md. Comput.* 2 (1), 15–19. doi:10.1016/j.juro.2007.05.156

Calses, P. C., Crawford, J. J., Lill, J. R., and Dey, A. (2019). Hippo pathway in cancer: Aberrant regulation and therapeutic opportunities. *Trends Cancer* 5 (5), 297–307. doi:10.1016/j.trecan.2019.04.001

Cao, B., Wang, Q., Zhang, H., Zhu, G., and Lang, J. (2018). Two immune-enhanced molecular subtypes differ in inflammation, checkpoint signaling and outcome of advanced head and neck squamous cell carcinoma. *Oncoimmunology* 7 (2), e1392427. doi:10.1080/2162402X.2017.1392427

Chen, H., Qin, Q., Xu, Z., Chen, T., Yao, X., Xu, B., et al. (2021). DNA methylation data-based prognosis-subtype distinctions in patients with esophageal carcinoma by bioinformatic studies. *J. Cell. Physiol.* 236 (3), 2126–2138. doi:10.1002/jcp.29999

Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., Sander, C., et al. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* 45 (10), 1127–1133. doi:10.1038/ng.2762

Corces, M. R., Granja, J. M., Shams, S., Louie, B. H., Seoane, J. A., Zhou, W., et al. (2018). The chromatin accessibility landscape of primary human cancers. *Science* 362 (6413), eaav1898. doi:10.1126/science.aav1898

Gan, Y., Li, N., Zou, G., Xin, Y., and Guan, J. (2018). Identification of cancer subtypes from single-cell RNA-seq data using a consensus clustering method. *BMC Med. Genomics* 11 (6), 117. doi:10.1186/s12920-018-0433-z

Giachino, C., Boulay, J. L., Ivanek, R., Alvarado, A., Tostado, C., Lugert, S., et al. (2015). A tumor suppressor function for notch signaling in forebrain tumor subtypes. *Cancer Cell* 28 (6), 730–742. doi:10.1016/j.ccell.2015.10.008

Gong, Y., Ji, P., Yang, Y. S., Xie, S., Yu, T. J., Xiao, Y., et al. (2021). Metabolic-pathway-based subtyping of triple-negative breast cancer reveals potential therapeutic targets. *Cell Metab.* 33 (1), 51–64. doi:10.1016/j.cmet.2020.10.012

Guo, L., Chen, G., Zhang, W., Zhou, L., Xiao, T., Di, X., et al. (2019). A high-risk luminal A dominant breast cancer subtype with increased mobility. *Breast Cancer Res. Treat.* 175 (2), 459–472. doi:10.1007/s10549-019-05135-w

Gyorffy, B., Bottai, G., Fleischer, T., Munkacsy, G., Budczies, J., Paladini, L., et al. (2016). Aberrant DNA methylation impacts gene expression and prognosis in breast cancer subtypes. *Int. J. Cancer* 138 (1), 87–97. doi:10.1002/ijc.29684

Hanzelmann, S., Castelo, R., and Guinney, J. (2013). Gsva: Gene set variation analysis for microarray and RNA-seq data. *BMC Bioinforma.* 14, 7. doi:10.1186/1471-2105-14-7

Imperial, R., Toor, O. M., Hussain, A., Subramanian, J., and Masood, A. (2019). Comprehensive pancancer genomic analysis reveals (RTK)-RAS-RAF-MEK as a key dysregulated pathway in cancer: Its clinical implications. *Semin. Cancer Biol.* 54, 14–28. doi:10.1016/j.semcancer.2017.11.016

Jackstadt, R., van Hooff, S. R., Leach, J. D., Cortes-Lavaud, X., Lohuis, J. O., Ridgway, R. A., et al. (2019). Epithelial NOTCH signaling rewires the tumor microenvironment of colorectal cancer to drive poor-prognosis subtypes and metastasis. *Cancer Cell* 36 (3), 319–336. e317. doi:10.1016/j.ccell.2019.08.003

Jiang, P., Gu, S., Pan, D., Fu, J., Sahu, A., Hu, X., et al. (2018). Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nat. Med.* 24 (10), 1550–1558. doi:10.1038/s41591-018-0136-1

Joerger, A. C., and Fersht, A. R. (2016). The p53 pathway: Origins, inactivation in cancer, and emerging therapeutic approaches. *Annu. Rev. Biochem.* 85, 375–404. doi:10.1146/annurev-biochem-060815-014710

Kaunitz, G. J., Cottrell, T. R., Lilo, M., Muthappan, V., Esandrio, J., Berry, S., et al. (2017). Melanoma subtypes demonstrate distinct PD-L1 expression profiles. *Lab. Invest.* 97 (9), 1063–1071. doi:10.1038/labinvest.2017.64

Kulis, M., and Esteller, M. (2010). DNA methylation and cancer. *Adv. Genet.* 70, 27–56. doi:10.1016/B978-0-12-380866-0.60002-2

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

Li, F., Wu, T., Xu, Y., Dong, Q., Xiao, J., Xu, Y., et al. (2020). A comprehensive overview of oncogenic pathways in human cancer. *Brief. Bioinform.* 21 (3), 957–969. doi:10.1093/bib/bbz046

Li, T., Fan, J., Wang, B., Traugh, N., Chen, Q., Liu, J. S., et al. (2017). Timer: A web server for comprehensive analysis of tumor-infiltrating immune cells. *Cancer Res.* 77 (21), e108–e110. doi:10.1158/0008-5472.CAN-17-0307

Liu, Y., Broaddus, R. R., and Zhang, W. (2015). Identifying aggressive forms of endometrioid-type endometrial cancer: New insights into molecular subtyping. *Expert Rev. Anticancer Ther.* 15 (1), 1–3. doi:10.1586/14737140.2015.992420

Meng, Z., Moroishi, T., and Guan, K. L. (2016). Mechanisms of Hippo pathway regulation. *Genes Dev.* 30 (1), 1–17. doi:10.1101/gad.274027.115

Paczkowska, M., Barenboim, J., Sintupisut, N., Fox, N. S., Zhu, H., Abd-Rabbo, D., et al. (2020). Integrative pathway enrichment analysis of multivariate omics data. *Nat. Commun.* 11 (1), 735. doi:10.1038/s41467-019-13983-9

Park, A. K., Kim, P., Ballester, L. Y., Esquenazi, Y., and Zhao, Z. (2019). Subtype-specific signaling pathways and genomic aberrations associated with prognosis of glioblastoma. *Neuro. Oncol.* 21 (1), 59–70. doi:10.1093/neuonc/noy120

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26 (1), 139–140. doi:10.1093/bioinformatics/btp616

Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W. K., Luna, A., La, K. C., et al. (2018). Oncogenic signaling pathways in the cancer genome atlas. *Cell* 173 (2), 321–337. doi:10.1016/j.cell.2018.03.035

Taciak, B., Pruszynska, I., Kiraga, L., Bialasek, M., and Krol, M. (2018). Wnt signaling pathway in development and cancer. *J. Physiol. Pharmacol.* 69 (2), 185–196. doi:10.26402/jpp.2018.2.07

Thanki, K., Nicholls, M. E., Gajjar, A., Senagore, A. J., Qiu, S., Szabo, C., et al. (2017). Consensus molecular subtypes of colorectal cancer and their clinical implications. *Int. Biol. Biomed. J.* 3 (3), 105–111.

Thorvaldsdottir, H., Robinson, J. T., and Mesirov, J. P. (2013). Integrative genomics viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.* 14 (2), 178–192. doi:10.1093/bib/bbs017

Vogelstein, B., and Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nat. Med.* 10 (8), 789–799. doi:10.1038/nm1087

Wilkerson, M. D., and Hayes, D. N. (2010). ConsensusClusterPlus: A class discovery tool with confidence assessments and item tracking. *Bioinformatics* 26 (12), 1572–1573. doi:10.1093/bioinformatics/btq170

Xie, J., and Liu, C. (2005). Adjusted Kaplan-Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Stat. Med.* 24 (20), 3089–3110. doi:10.1002/sim.2174

Yuan, H., Yan, M., Zhang, G., Liu, W., Deng, C., Liao, G., et al. (2019). CancerSEA: A cancer single-cell state atlas. *Nucleic Acids Res.* 47 (D1), D900–D908. doi:10.1093/nar/gky939

Zhao, Y., Shao, Q., and Peng, G. (2020). Exhaustion and senescence: Two crucial dysfunctional states of T cells in the tumor microenvironment. *Cell. Mol. Immunol.* 17 (1), 27–35. doi:10.1038/s41423-019-0344-8

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 02 September 2022
DOI 10.3389/fgene.2022.925564

OPEN ACCESS

EDITED BY
Rongshan Yu,
Xiamen University, China

REVIEWED BY
Yi Qin,
Fudan University, China
Lei Li,
East China Normal University, China
Jianru Xiao,
Shanghai Jiao Tong University, China
Elizabeth Thomas Bartom,
Northwestern University, United States

*CORRESPONDENCE
Yangbai Sun,
drsunyb@fudan.edu.cn
Kun Li,
kunli12345@163.com
Wangjun Yan,
yanwj@fudan.edu.cn

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

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

RECEIVED 07 June 2022
ACCEPTED 27 July 2022
PUBLISHED 02 September 2022

CITATION
Sun Z, Yin M, Ding Y, Zhu Z, Sun Y, Li K and Yan W (2022), Integrative analysis of synovial sarcoma transcriptome reveals different types of transcriptomic changes.
Front. Genet. 13:925564.
doi: 10.3389/fgene.2022.925564

Integrative analysis of synovial sarcoma transcriptome reveals different types of transcriptomic changes

Zhengwang Sun[1†], Mengchen Yin[2†], Yi Ding[3†], Zixu Zhu[4], Yangbai Sun[1*], Kun Li[1*] and Wangjun Yan[1*]

[1]Department of Musculoskeletal Oncology, Fudan University Shanghai Cancer Center, Shanghai, China, [2]Longhua Hospital, Shanghai University of Traditional Chinese Medicine, Shanghai, China, [3]Shanghai Key Laboratory of Regulatory Biology and School of Life Sciences, East China Normal University, Shanghai, China, [4]No.2 High School of East China Normal University, Shanghai,200000, China

Background: Synovial sarcoma (SS) is a rare and aggressive cancer that can come from distinct soft tissue types including muscle and ligaments. However, the transcriptomic landscape of SS is still poorly understood. This study aimed to systematically dissect the changes in SS transcriptome from different perspectives.

Methods: We performed deep total RNA sequencing on ten paired Synovial sarcoma and tumor-adjacent tissues to systematically dissect the transcriptomic profile of SS in terms of gene expression, alternative splicing, gene fusion, and circular RNAs.

Results: A total of 2,309 upregulated and 1,977 downregulated genes were identified between SS and tumor-adjacent tissues. Those upregulated genes could lead to the upregulation of the cell cycle, ribosome, and DNA replication pathways, while the downregulated genes may result in the downregulation of a set of metabolic biological processes and signaling pathways. Moreover, 2,511 genes (including 21 splicing factors) were differentially alternative spliced, indicating that the deregulation of alternative splicing could be one important factor that contributes to tumorigenesis. Additionally, we identified the known gene fusions of SS18-SSX1/SSX2 as well as 11 potentially novel gene fusions. Interestingly, 49 circular RNAs were differentially expressed and their parental genes could function in muscle contraction and muscle system processes.

Conclusions: Collectively, our comprehensive dissection of the transcriptomic changes of SS from both transcriptional and post-transcriptional levels provides novel insights into the biology and underlying molecular mechanism of SS.

KEYWORDS

synovial sarcoma, gene expression, alternative splicing, gene fusion, circular RNA

Introduction

Synovial sarcoma (SS) is a rare and aggressive soft tissue cancer, which tends to occur near large joints, particularly in the extremities of the arms or legs, in young adults (Ladanyi et al., 2002). At present, surgery is still the main treatment strategy for SS. Cytogenetically, a significant portion of SS cases involve nonrandom translocations between SS18 and SSX (Przybyl et al., 2012). Although a range of studies has investigated the genetic profile of SS from different cascades, a comprehensive transcriptomic profile of SS from different aspects is still lacking (Cancer Genome Atlas Research Network and Electronic address, 2017; He et al., 2017). RNA sequencing (RNA-Seq) technologies provide unprecedented opportunities to gain insights into the transcriptome from various aspects, including expression level, alternative splicing (AS), gene fusions, and circular RNAs. These analyses are essential to systematically reveal and better understand the abnormally transcriptomic changes of SS; however, a comprehensive exploration of the SS transcriptome from these aspects is still currently lacking.

AS is a crucial mechanism of post-transcriptional modification responsible for increasing both transcriptome and proteome diversity of a cell in eukaryotes (Wang et al., 2008; Keren et al., 2010). Since AS play important role in a variety of physiological processes (e.g. developmental programming), the misregulation of AS can result in splicing defects which may have a pathogenic function to cause severe diseases, including cancers (Wang and Cooper, 2007; Zhang and Manley, 2013). However, the genome-wide AS profile of SS is rarely studied to date. Furthermore, besides the common gene fusions formed by the translocation between chromosome X and 18 in SS, other gene fusions could also contribute to tumorigenesis or progression (Edwards, 2010). In addition, many circular RNAs (circRNAs) have been demonstrated to be functional as miRNA sponges and modulators of transcription (Chen, 2016; Li et al., 2018), which could be vital for different aspects of malignant phenotypes, such as cell cycle, apoptosis, and invasion (Qu et al., 2015; Greene et al., 2017). Moreover, some circRNAs are potentially important biomarkers for certain cancers (Abu and Jamal, 2016; Dong et al., 2017; Greene et al., 2017). But little is known about the expression profile of circRNAs in SS and almost no study has investigated this in SS. Thus, systematic dissection of the SS transcriptome from both transcriptional and post-transcriptional layers is necessary to better understand the underlying mechanisms of SS development.

Here we performed Ribo-Zero RNA-seq on ten pairs of Chinese SS and corresponding tumor-adjacent tissues to comprehensively explore the transcriptome profile of SS from various aspects. We first carried out differential expression calling and detected a number of upregulated and downregulated genes. Then the AS deregulation of a multitude of genes and a set of tumor-specific gene fusion events were identified. We also investigated the expression changes of circRNAs between SS and tumor-adjacent tissues. Moreover, we constructed an interaction network among circRNAs, miRNAs, and their target genes, which enabled us to further gain insights into the potential function of circRNAs in SS.

Materials and methods

RNA extraction

Total RNA was extracted from the 10 mg synovial sarcoma and tumor-adjacent tissues after grinding by Homogenizer (Scientz) using TRIzol® Reagent (Invitrogen) and RNeasy MinElute spin column (Qiagen) according to the manufacturer's instructions. Then the integrity of the total RNA was determined by 2100 Bioanalyser (Agilent) and quantified using the NanoDrop (Thermo Scientific). About 1 ug high-quality or media-quality RNA sample (OD260/280 = 1.9–2.0, RIN≥4) was used to construct the sequencing library.

Library preparation and RNA sequencing

RNA purification, reverse transcription, library construction, and sequencing were performed at WuXi NextCODE in Shanghai according to the manufacturer's instructions (Illumina). The rRNA-depleted sequencing libraries from total RNA were prepared using Illumina TruSeq® Stranded Total RNA Gold preparation Kit. About 1 ug total RNA was used as input material, and then the Ribo-Zero Gold kit was used to remove both cytoplasmic and mitochondrial rRNA. After purification of the remaining RNA without rRNA, the RNA was fragmented into small pieces using divalent cations under elevated temperature. The cleaved RNA fragments are copied into the first-strand cDNA using reverse transcriptase and random primers, followed by second-strand cDNA synthesis. These cDNA fragments then were subjected to end-repair, phosphorylation, and 'A' base addition according to Illumina's library construction protocol. The products were purified and enriched with PCR, and the AMPure XP Beads (Beckmen) were used to clean up the amplified target fragments to create the final cDNA library. After library construction, Qubit 2.0 fluorometer dsDNA HS Assay (Thermo Fisher Scientific) was used to quantify the concentration of the resulting sequencing libraries, while the size distribution was analyzed using Agilent BioAnalyzer 2100 (Agilent). Sequencing was performed using the Illumina system following Illumina-provided protocols for 2 x150 paired-end sequencing in WuXi NextCODE at Shanghai, China.

Short-read mapping and gene expression quantification

The RNA-seq reads of each sample for 10 pairs of SS and tumor-adjacent tissues were separately aligned to the human reference genome GRCh38 using HISAT2 (version 2.1.0) (Kim et al., 2015). Then we quantified the gene expression of each sample by employing StringTie (version 1.3.6) (Pertea et al., 2015). The human gene annotation file in the GTF format of version 95 from the Ensembl database (http://www.ensembl.org) was used. The mapped read count and expression value in transcript per million (TPM) for each gene were obtained from StringTie and used for downstream analysis.

Differential gene expression calling

For differential expression analysis, the read count mapped to each gene was used as input. The gene expression changes between SS and tumor-adjacent tissues were examined using DESeq2 (version 1.24.0) (Love et al., 2014). We defined the differentially expressed genes (DEGs) with the threshold of |fold change| >2 and adjusted p-value < 0.01.

Detection of alternative splicing events

We investigated the alternative splicing (AS) profile of genes between SS and tumor-adjacent tissues by employing rMATS (version 4.0.2) (Shen et al., 2014). The bam files outputted by HISAT2 after read mapping were used as the input. Five common AS types of exon skipping (ES), alternative 3′ acceptor sites (A3AS), alternative 5′ donor sites (A5DS), intron retention (IE), and mutually exclusive exons (ME) were investigated. The differential alternative splicing events were identified with the cutoff of FDR <0.05.

Identification of gene fusions

In order to explore the genetic alterations, we employed TopHat-Fusion (version 2.1.0) with default parameters to identify the gene fusion events in all tumor and normal samples (Kim and Salzberg, 2011). Only the fusions with at least 3 supporting reads and 2 supporting pairs were considered. Finally, 14 and 11 gene fusion pairs were detected in SS and tumor-adjacent tissues, respectively. We only kept the 14 gene fusions that are unique to SS and discarded the fusions detected in tumor-adjacent tissues.

Circular RNA detection and differential expression analysis

We investigated the expression profiles of circRNAs in SS and tumor-adjacent tissues using CIRI (version 0.1.0) (Gao et al., 2015). Then differential expression analysis was conducted by employing DESeq2 (version 1.24.0) based on the expression count of circRNAs identified by CIRI. Only the circRNAs with expression changes of |fold change| >2 and adjusted p-value < 0.01 were considered as differentially expressed. The official IDs of circRNAs were obtained by coordinate mapping using the circBase database (Glazar et al., 2014).

Construction of interaction network among circRNAs, miRNAs, and target genes

To gain insights into the function of circRNAs, we built an interaction network among the circRNAs, miRNAs, and the target genes of miRNAs. The PPI interactions were downloaded from the STRING database (version 11.0) (Szklarczyk et al., 2019). The regulatory relationship between miRNAs and target genes, as well as the known miRNA-circRNAs interactions, were obtained from the starBase database (version 3.0) (Li et al., 2014). We only used the circRNA-miRNA pairs supported by > 5 CLIP-seq experiments and the miRNA-target gene pairs supported by > 2 CLIP-seq experiments and >2 degradome-seq experiments in the StarBase2 database. Then we incorporated these interactions to construct the interaction network among circRNAs, miRNAs, and the genes targeted by miRNAs using Cytoscape (version 3.7.2) (Shannon et al., 2003). Only the parental genes of differentially expressed circRNAs, DEGs, DASGs, and fusion genes were considered in the interaction network construction.

Gene functional enrichment analysis

We conducted gene ontology (GO) and KEGG pathway enrichment analyses using GSEA (version 4.0.1) for the upregulated and downregulated DEGs between SS and tumor-adjacent tissues (Subramanian et al., 2005). The functional enrichment analysis of biological processes and pathways for the differentially alternative spliced genes, fusion genes, and the parental genes of circular RNAs were carried out with cluster Profiler (version 3.12.0) (Yu et al., 2012). The enriched items with adjusted p-value < 0.05 were defined as significant.

TABLE 1 Detailed information of 10 synovial sarcoma patients.

Patient ID	Age	Sex	Tumor Location	Tumor Size (cm)	Tumor status	Outcome
1	26	Female	Thigh	12*10.5*10	Primary	Alive
2	18	Male	Foot	6.0*3.5*2.0	Local recurrence	Died
3	37	Male	Groin	9.5*6*6	Local recurrence	Alive
4	29	Female	Lung	2.0*2.0*1.3	Primary	Alive
5	59	Female	Iliac Bone	5.5*4.5*3.5	Local recurrence	Alive
6	28	Female	Foot	2*1.7*0.7	Primary	Alive
7	20	Female	Neck	6*5*4	Primary	Alive
8	27	Male	Shank	7*6*2	Primary	Alive
9	41	Male	Shank	8*6*4	Local recurrence	Alive
10	71	Female	Thigh	9*6*3	Primary	Alive

Results

An abundance of important genes is differentially expressed between SS and tumor-adjacent tissues

To gain insights into the transcriptomic changes of SS patients, we deeply sequenced the tumor and tumor-adjacent tissues of 10 SS patients with total RNA sequencing (including both poly (A+) and poly (A-) RNAs) (Table 1). We first aligned the RNA-seq reads of each sample to the human reference genome GRCh38 using HISAT2 (Kim et al., 2015) and then conducted differential expression calling by employing DEseq2 (Love et al., 2014). A total of 4,286 differentially expressed genes (DEGs) were detected using the threshold of |fold change| >2 and adjusted p-value < 0.01, of which 2,309 (including 432 lncRNA genes) and 1,977 (including 333 lncRNA genes) genes were separately upregulated and downregulated in SS compared to tumor-adjacent tissues (Figure 1A, Supplementary Table S1). Interestingly, we found that 340, 185, 124, and 7 of those DEGs are oncogenes, tumor suppressor genes (TSGs), transcription factors (TFs), and splicing factors (see Supplementary Figure S1 for differentially expressed splicing factors), respectively (Figure 1B). Specifically, 52 TFs (such as AES and BCL6) were down-regulated and 72 TFs (e.g. ARID3A and BRCA2) were up-regulated, suggesting that the expression changes of these TFs could influence the expression of their downstream target genes including related oncogenes and TSGs. Since oncogenes and TSGs are closely correlated with cancer, their expression changes may play an important role in the development of SS. Specifically, in consideration of the crucial function of splicing factors in AS regulation (Lee and Rio, 2015), we further conducted a qPCR experiment to validate the expression profiles of the seven splicing factors (ELAVL2, HNRNPA1, HNRNPH2, MBNL1, PCBP1, QKI, and TIA1) in DEGs (Supplementary Figure S2). As expected, the experimental results were consistent with the RNA-seq data, indicating the robustness of our analysis. Therefore, the differential expression of these splicing factors could result in the AS deregulation of corresponding genes in SS.

Gene ontology (GO) and KEGG pathway enrichment analyses showed that these upregulated and downregulated DEGs were mainly involved in the fundamental and tumor-related biological processes and pathways (Figures 1C–E FDR <0.05). For example, the up-regulated DEGs were primarily enriched in the cell-cycle-related biological processes (e.g. chromosome organization, chromatin organization, and DNA conformation change) and pathways of systemic lupus erythematosus, cell cycle, DNA replication, and P53 signaling (Figure 1C). Several previous studies also identified the cell-cycle-related genes in sarcoma as a major category of up-regulated genes (Chibon et al., 2010; Yen et al., 2012), which was in line with our findings. By contrast, the down-regulated DEGs were mainly involved in the metabolic-related biological processes (such as energy derivation by oxidation of organic compounds, muscle system process, and glucan metabolic process) and the pathways of oxidative phosphorylation, insulin signaling pathway, and vascular smooth muscle contraction (Figure 1D). Thus, the result suggests that a multitude of genes prominently altered their expression levels in SS, which could be one of the main factors accounting for tumorigenesis through up-regulating and down-regulating corresponding pathways.

Deregulation of alternative splicing could contribute to the tumorigenesis of SS

Considering that the misregulation of AS can lead to the production of aberrant proteins that contribute to tumorigenesis (Zhang and Manley, 2013), we further compared the AS profile between SS and tumor-adjacent tissues by employing rMATS (Shen et al., 2014). Five classical splicing categories of exon

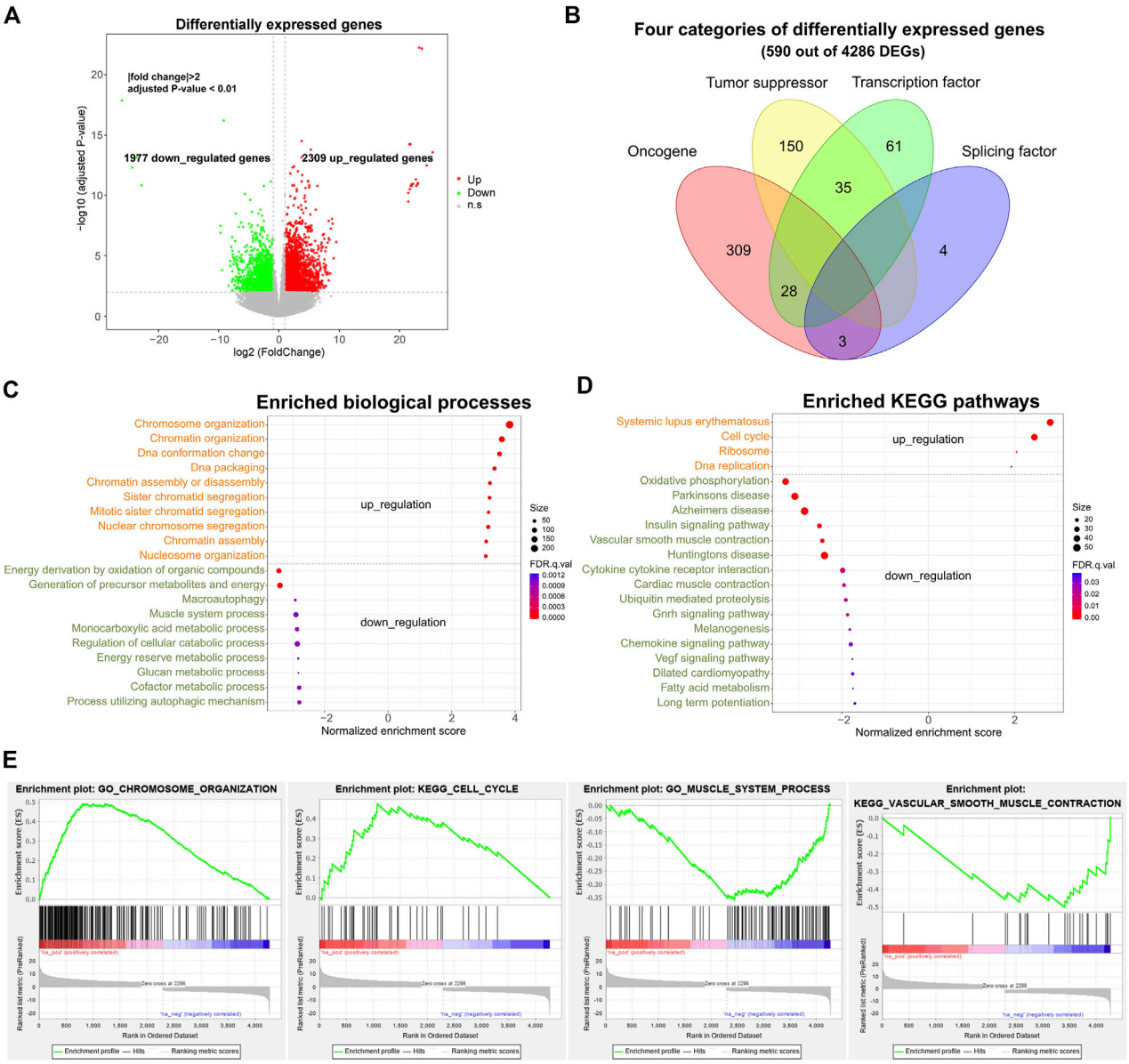


FIGURE 1
Differential expression profile and corresponding gene functions of SS. **(A)** Volcano plot displaying differentially expressed genes (DEGs) between ten pairs of SS and tumor-adjacent tissues. |fold change| >2 and adjusted *p*-value < 0.01. **(B)** Different categories of DEGs. **(C)** Top enriched up-regulated and down-regulated biological processes of DEGs. **(D)** Top enriched up-regulated and down-regulated KEGG pathways for DEGs. **(E)** Examples of enriched biological processes and pathways. Adjusted *p*-value < 0.05. DEGs: differentially expressed genes.

skipping (ES), alternative 5′ donor sites (A5DS), alternative 3′ acceptor sites (A3AS), mutually exclusive exons (ME), and intron retention (IR) were analyzed. In total, we identified 2511 (including 41 lncRNA genes) significantly differential AS genes (DASGs), of which 2018, 223, 242, 486, and 159 belong to the splicing mode changes of ES, A5DS, A3AS, ME, and IR, respectively (Figure 2A, FDR <0.05, Supplementary Table S2). As expected, ES was the most common differential splicing mode (80.37%, 2018 out of 2511 DASGs), whereas IR was the least (6.33%, 159 out of 2511 DASGs). Notably, the majority of those DASGs among the five classical splicing categories were largely different, only a small portion of them simultaneously exhibited three or more distinct splicing types (Figure 2A).

Gene functional enrichment analysis indicated that those 2511 DASGs were mainly involved in the RNA splicing and cancer-related biological processes and KEGG pathways (Figures 2B,C, adjusted *p*-value < 0.05), which was highly correlated with the AS process. For instance, the top enriched biological

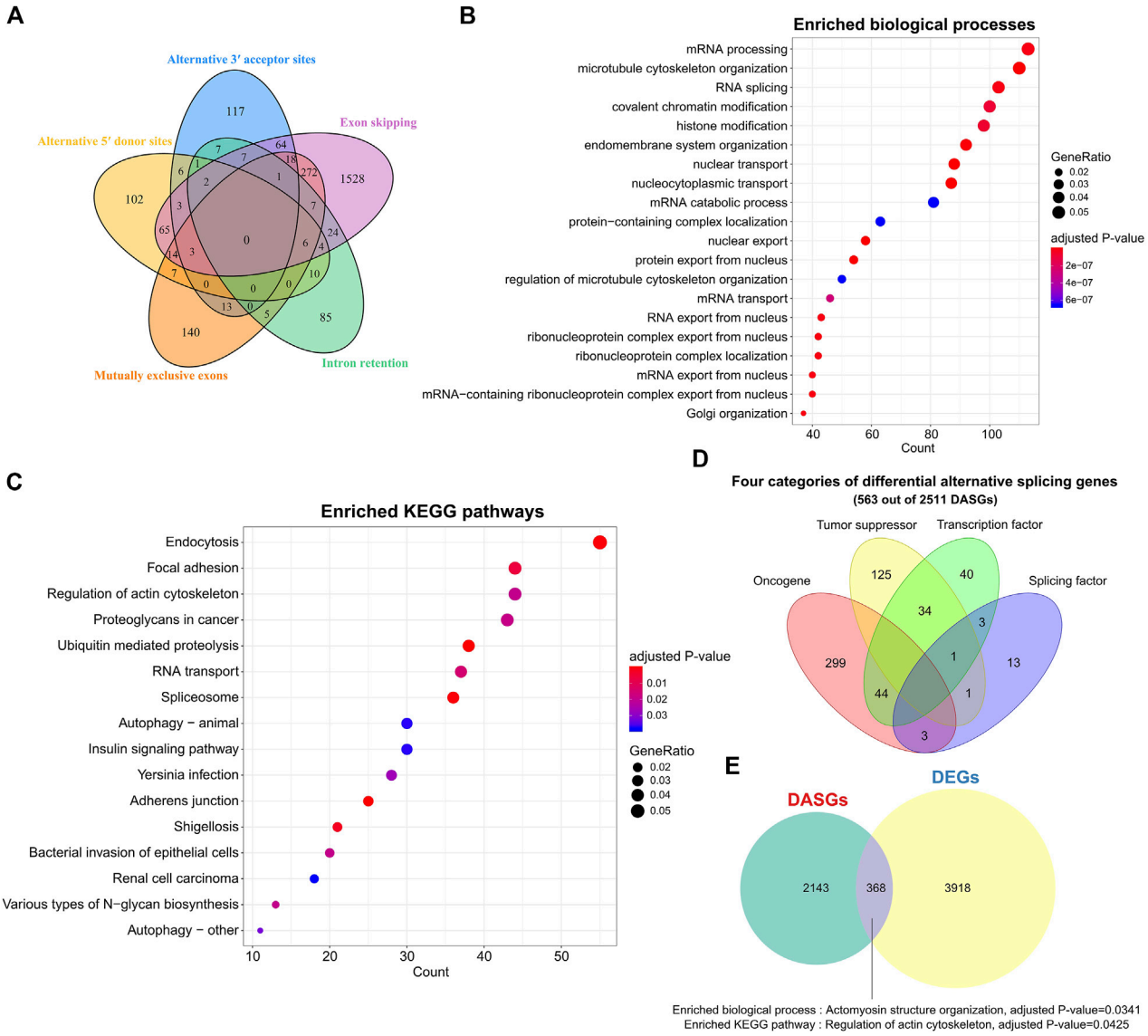


FIGURE 2
Alternative splicing patterns and related gene functions of SS. **(A)** Venn plot of the DASGs for five common splicing modes. **(B)** Top 20 enriched biological processes of DASGs. FDR <0.05. **(C)** Top 20 enriched pathways for DASGs. Adjusted *p*-value < 0.05. **(D)** Distinct categories of DASGs. **(E)** Comparison between DEGs and DASGs. DASGs: differentially alternative spliced genes; DEGs: differentially expressed genes.

processes of those DASGs were mRNA processing, microtubule cytoskeleton organization, and RNA splicing, while the enriched pathways are endocytosis, RNA transport, proteoglycans in cancer, and spliceosome. Moreover, we observed that 21 splicing factor genes showed significantly differential AS between SS and tumor-adjacent tissues, such as HNRNPA1, PTBP2, QKI, RBFOX2, and TRA2A. It is well known that the splicing factors are crucial for AS regulation (Lee and Rio, 2015), the deregulation of those splicing factors could drastically disrupt the splicing process of many corresponding genes and contribute to the tumorigenesis of SS (Dvinge et al., 2016). Furthermore, we found that 346, 204, and 122 oncogenes, TSGs, and TFs were also differentially spliced (Figure 2D). The abnormal splicing of these TFs could influence the expression of their downstream target genes and contribute to the development and progression of SS. Only 368 genes shared between DASGs and DEGs, leaving most of them were distinct (Figure 2E). These common 368 genes were enriched in the biological process of actomyosin structure organization and pathway of regulation of actin cytoskeleton (Figure 2E, adjusted *p*-value < 0.05). Thus, the genes that showed differential expression were quite distinct from those that exhibited differential splicing, suggesting that AS is

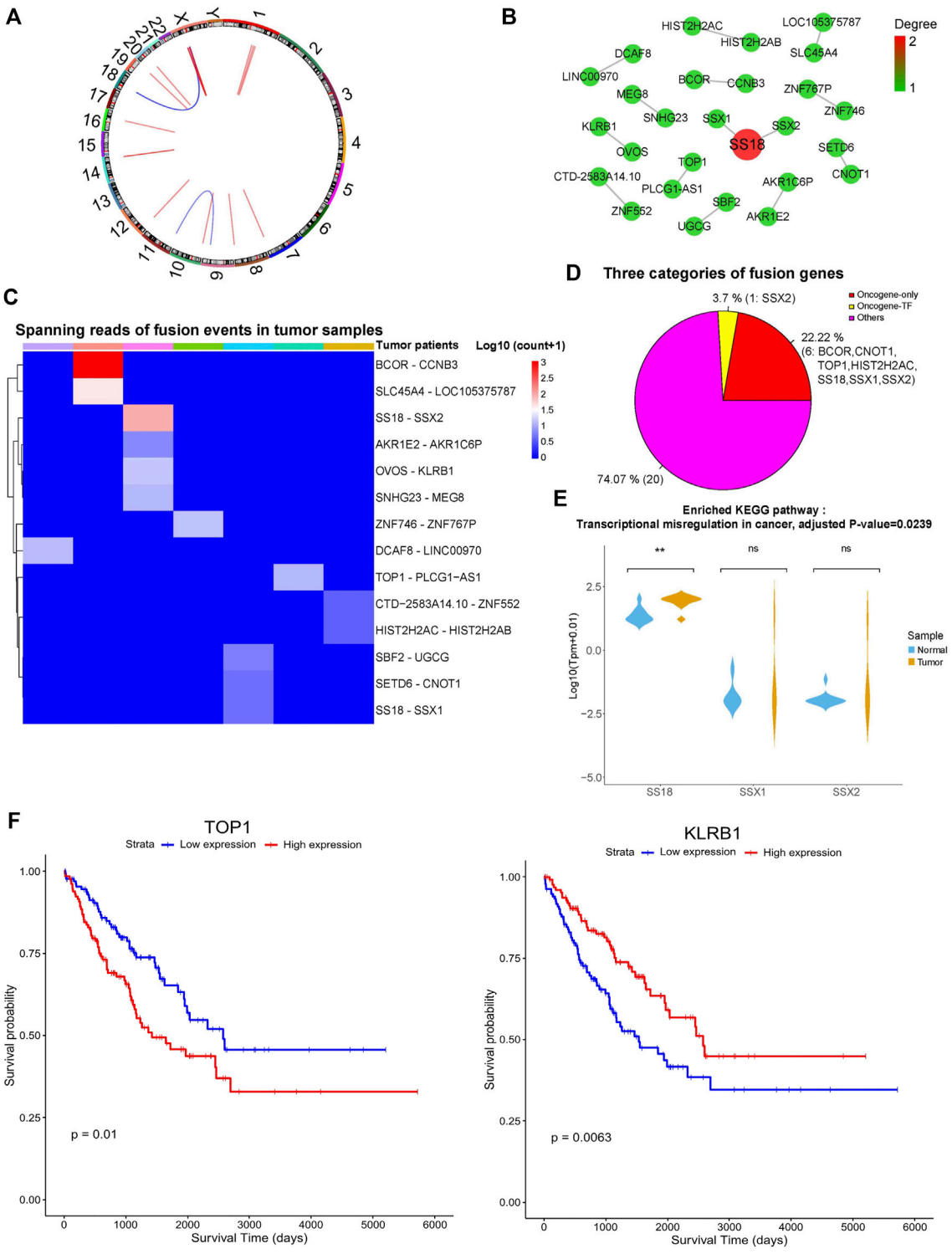


FIGURE 3
Gene fusion landscape of SS. **(A)** Circos plot showing the 14 tumor-specific fusion pairs in SS. **(B)** Network of tumor-specific fusion genes. The size and color of each circle correspond to the degree of fusion edges. **(C)** Heatmap displaying the supporting junction reads for tumor-specific gene fusions. **(D)** Different categories of the fusion genes. **(E)** Expression profile of the fusion genes involved in significantly enriched biological processes. Adjusted *p*-value < 0.05. **(F)** The fusion genes significantly correlated with the survival of TCGA sarcoma patients. *p*-value < 0.05.

complementary with expression level in revealing the transcriptomic changes. These results indicate that the abnormal AS changes of genes could be another important factor responsible for the tumorigenesis of SS.

Dissection of the gene fusions in SS

We further explored the gene fusion events in SS patients using TopHat-Fusion (Kim and Salzberg, 2011). A total of 14 and 11 unique gene fusion pairs were separately identified in SS and tumor-adjacent tissues, and no fusion was shared between them. The 14 tumor-specific gene fusion pairs were from seven SS patients, most of which (11 out of 14) resulted from the rearrangements within the same chromosome, while 3 of them were generated by breaking and rejoining two disparate chromosomes (Figure 3A). In total, 27 genes were involved in these tumor-specific gene fusions. SS18 was fused with SSX1 and SSX2, which was in line with previous studies (Edwards, 2010). In contrast, other genes were mainly fused with one partner (Figure 3B).

As shown in Figure 3C, the maximum number of gene fusion pairs detected in individual patients was four and the gene fusion events in those SS patients were quite distinct. Intriguingly, these tumor-specific gene fusion events contain one TF of SSX2 and seven oncogenes of SS18, SSX1, SSX2, BCOR, CNOT1, HIST2H2AC, and TOP1 (Figure 3D). Oncogene SS18 was fused with the TF and oncogene of SSX2 as well as the oncogene SSX1, which is consistent with the known findings (Kawai et al., 1998). Besides, other oncogenes of BCOR, CNOT1, HIST2H2AC, and TOP1 formed the fusion events of BCOR-CCNB3, CNOT1-SETD6, HIST2H2AC-HIST2H2AB, and TOP1-PLCG1-AS1, respectively. Previous studies have shown that BCOR-CCNB3 fusion tends to occur in the undifferentiated small round-cell sarcomas like Ewing sarcoma and has the potential to drive sarcoma progression (Pierron et al., 2012; Li et al., 2016; Kao et al., 2018). Other gene fusions could be novel for SS, and the involved genes could be functionally important. For example, CNOT1 encodes the CCR4-NOT transcription complex subunit 1, which mainly participates in deadenylating mRNAs (Pavanello et al., 2018). HIST2H2AC and HIST2H2AB can generate the replication-dependent histones that are basic nuclear proteins responsible for the nucleosome structure of the chromosomal fiber. TOP1 encodes the enzyme of DNA topoisomerase for controlling and altering the topologic states of DNA during transcription (Baranello et al., 2016). Since TF could regulate the expression of many downstream target genes and oncogenes are closely associated with cancer, the fusion events of those TFs and oncogenes may contribute to the tumorigenesis/progression of SS. Interestingly, lncRNA genes of LINC00970, LOC105375787, and PLCG1-AS1 were also involved in the gene fusion events, but their functions were still unknown. Gene functional enrichment analysis showed that those fusion genes were significantly enriched in the KEGG pathway of transcriptional misregulation in cancer (Figure 3E, adjusted p-value < 0.05).

Moreover, we further explored the expression profile of these fusion genes using synovial sarcoma data from The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013). As expected, these fusion genes showed similar expression patterns between the synovial sarcoma samples of us and TCGA (Supplementary Figure S3A). Additionally, we also found that these fusion genes exhibited slightly different expression profiles across distinct types of TCGA sarcomas (Supplementary Figure S3B). Considering that the number of synovial sarcoma samples is limited, we used all the TCGA sarcoma samples to do the survival analysis based on our identified fusion genes. Interestingly, the expression levels of KLRB1 and TOP1 were significantly associated with the survival of sarcoma patients (Figure 3F, p < 0.05), indicating that they could be potential prognostic markers.

Circular RNAs may play a role in SS formation

Emerging evidence shows that circRNAs can involve in various aspects of tumor biology (Dong et al., 2017; Zhang and Xin, 2018), thus we further investigated the expression profile of circRNAs in SS and tumor-adjacent tissues. We detected 49 differentially expressed circular RNAs by employing CIRI with the threshold of |fold change| >2 and adjusted p-value < 0.01. As shown in Figure 4A, 21 of them were significantly up-regulated in SS, whereas the other 28 were down-regulated. Furthermore, we found that the great majority (46 out of 49, 93.88%) of those differentially expressed circRNAs were formed by the circulation of exons of their parental genes, only two circRNAs of 10:24380869|24384423 (parental gene: KIAA1217) and 17:35168061|35168685 (parental gene: UNC45B) were produced from the intronic region and another one (5:137757867|137759020) was generated in the intergenic region (Figure 4B).

Intriguingly, 7, 5, and 3 of the parental genes for those differentially expressed circRNAs were oncogenes, TSGs, and TFs (Figure 4C). Previous studies have shown that circRNAs can form posttranscriptional regulators to regulate the expression of their parental genes (Memczak et al., 2013; Zhang et al., 2013). Thus, these circRNAs have the potential to affect the expression of their parental oncogenes, TSGs, and TFs. Gene functional enrichment analysis showed that the parental genes of those differentially expressed circRNAs were mainly enriched in the muscle system process (such as MAP2K4, HDAC4, TMEM38A, MYH1, MYH2, CAMK2G, TRDN, and SULF2) and muscle contraction (e.g. HDAC4, TMEM38A, MYH1, MYH2, TRDN, and SULF2) (Figure 4D, adjusted p-value < 0.05).

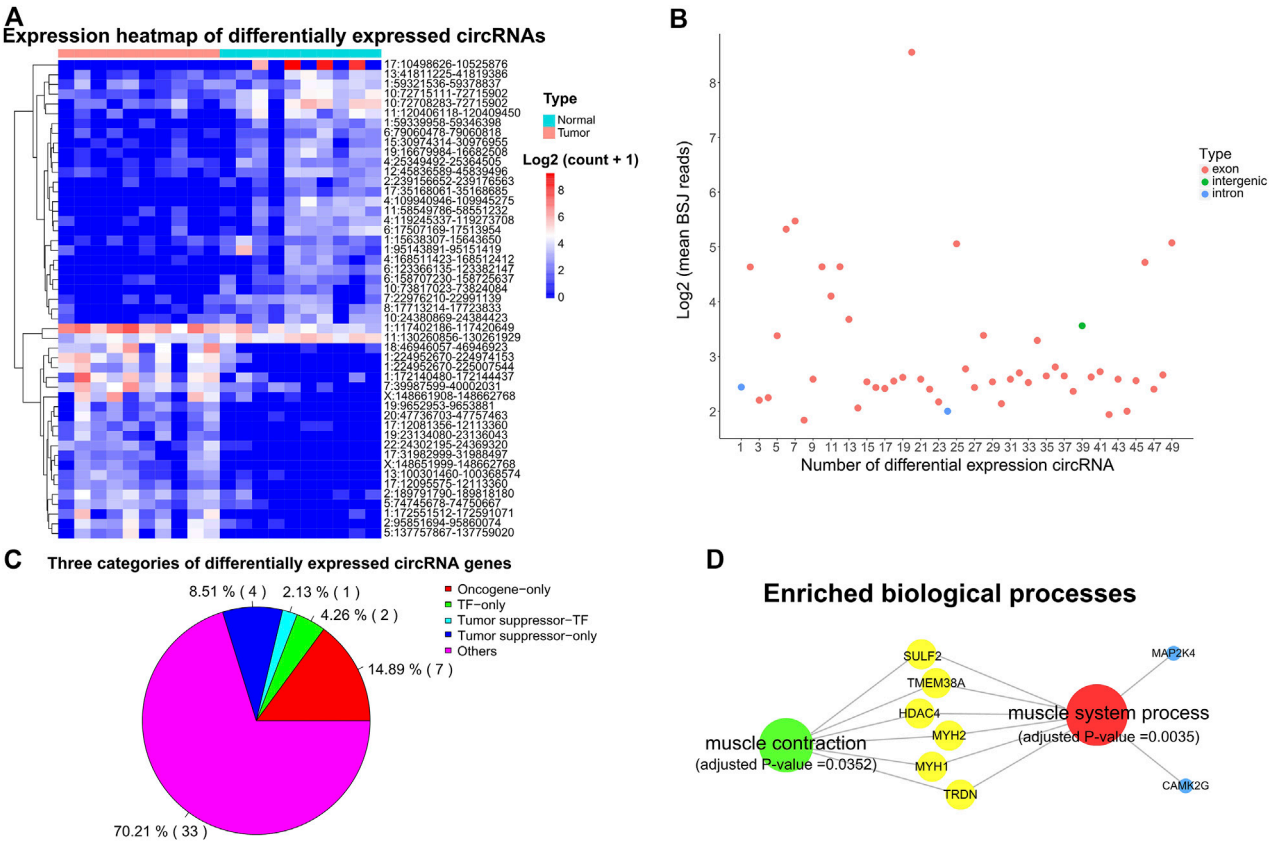


FIGURE 4
Expression profile and characteristics of circRNAs. **(A)** Expression heatmap showing the differentially expressed circRNAs. Supporting reads of circRNAs were used in the heatmap. **(B)** Genomic region of the differentially expressed circRNAs. BSJ: back-spliced junction. **(C)** Distinct types of the parental genes of differentially expressed circRNAs. **(D)** Significantly enriched biological processes for the parental genes of differentially expressed circRNAs. Adjusted *p*-value < 0.05.

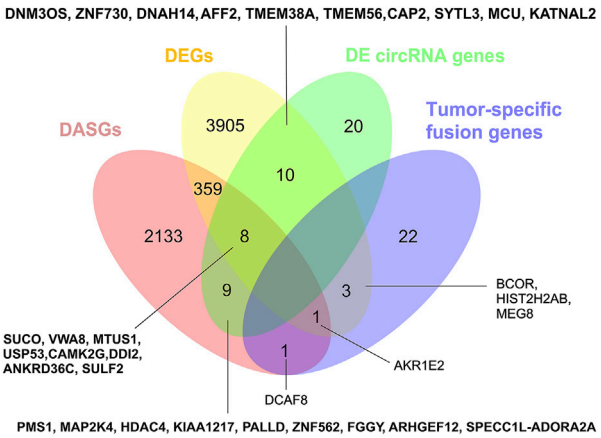


FIGURE 5
Comparison of DEGs, DASGs, fusion genes, and the parental genes of differentially expressed circRNAs. The overlapped genes for each part were shown in the figure. DEGs: differentially expressed genes; DASGs: differentially alternative spliced genes.

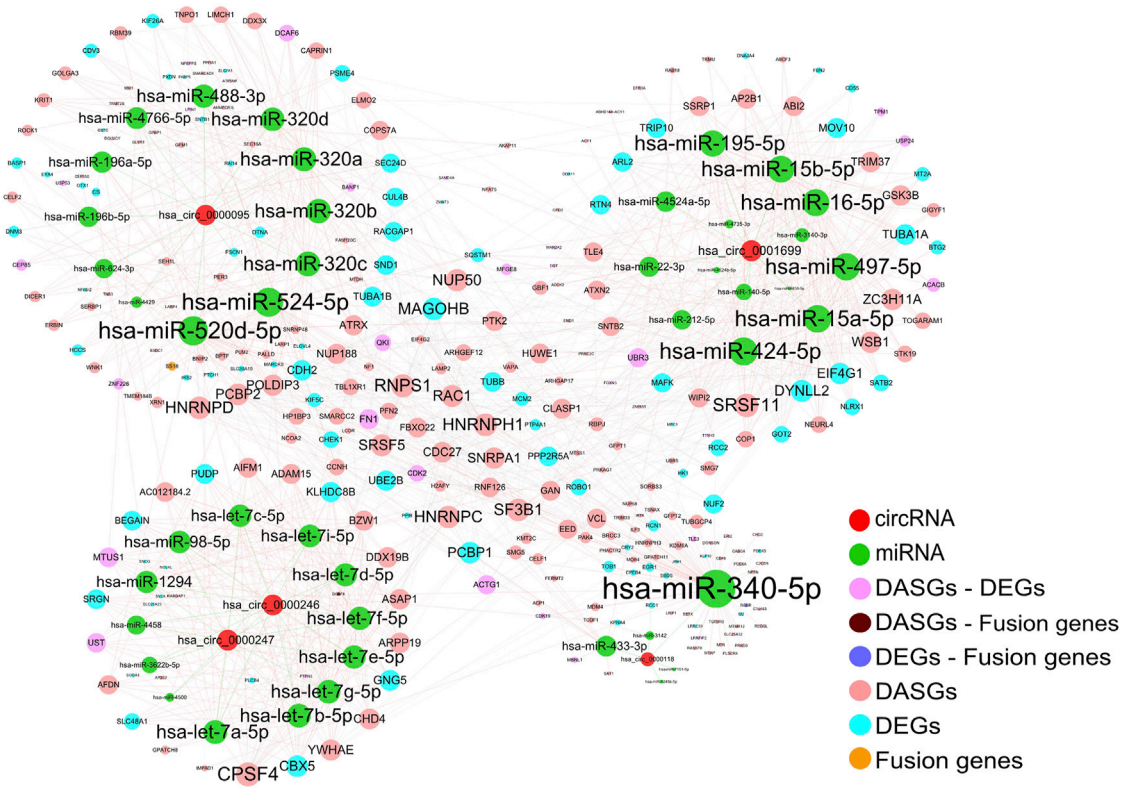


FIGURE 6
Interaction network among circRNAs, miRNAs, and related target genes. The edges represent the potential interactions between different types of genes and the size of each node is in proportion to the degree of edges. Only the target genes of DEGs, DASGs, fusion genes, and the parental genes of differentially expressed circRNAs were considered for corresponding miRNAs. The protein-protein interactions were obtained from the STRING database, while the known circRNA -miRNA interactions were downloaded from starBase. The resulting interaction network contains 5 circRNAs, 44 miRNAs, and 293 protein-coding genes. DEGs: differentially expressed genes; DASGs: differentially alternative spliced genes.

The genes involved in different types of transcriptomic changes are largely distinct

We further compared the four gene types of DEGs, DASGs, the fusion genes, and the parental genes of differentially expressed circRNAs. As shown in Figure 5, the genes in one type were largely distinct from that of other types, and no genes were common among the four categories. Only a fraction of them was involved in two or three types of changes (Figure 5). Intriguingly, the DEGs of BCOR, HIST2H2AB, and MEG8, and the DASGs of AKR1E2 and DCAF8 overlapped with the fusion genes, suggesting that the fusion events may influence the expression and/or AS profile of these genes. BCOR is an oncogene, while MEG8 is an imprinted gene. Moreover, 18 DEGs (e.g. DNM3OS, ZNF730, DNAH14, and AFF2) shared with the parental genes of differentially expressed circRNAs, implying that expression changes of these genes could affect the expression of circRNAs as well. In addition, 17 DASGs (such as SUCO, VWA8, MTUS1, and USP53) were common to the parental genes of differentially expressed circRNAs. Since circRNAs are mainly formed by AS of pre-mRNAs through backsplicing (Barrett and Salzman, 2016), the AS changes of these DASGs have the potential to influence the expression of corresponding circRNAs. Collectively, our results show that all the four transcriptomic aspects of expression changes, AS, gene fusions, and circRNAs could be closely correlated with the tumorigenesis/progression of SS.

CircRNAs could potentially regulate the expression of a multitude of genes by acting as miRNA sponges

An increasing number of studies suggested that endogenous circRNAs can act as miRNA sponges to regulate corresponding gene expression (Kulcheski et al., 2016; Panda, 2018). We further constructed the interaction network among differentially expressed circRNAs, miRNAs, and the miRNA target genes of DEGs, DASGs, and fusion genes to elucidate the functional roles of those differentially expressed circRNAs. Based on the known

miRNA-circRNA regulations, and the miRNA-targets relationships in the starBase database (Li et al., 2014) as well as the protein-protein interactions (PPIs) in the String database (Szklarczyk et al., 2019), the resulting interaction network involved in 5 circRNAs (hsa_circ_0001699, hsa_circ_0000247, hsa_circ_0000246, hsa_circ_0000095, and hsa_circ_0000118), 44 miRNAs, 293 protein-coding genes, containing 57 miRNA-circRNA interactions, 789 miRNA-mRNA interactions and 350 PPIs (Figure 6).

It is well known that circRNAs can regulate gene expression by influencing transcription, mRNA turnover, and translation by sponging RNA-binding proteins (RBPs) and miRNAs (Panda, 2018). Our resulting network showed that circRNAs hsa_circ_0001699, hsa_circ_0000247, hsa_circ_0000246, hsa_circ_0000095, and hsa_circ_0000118 could act as the sponges of 14, 13, 13, 12, and 5 miRNAs, respectively. Moreover, these miRNAs have the potential to regulate the expression of 119, 202, and 3 genes of DEGs, DASGs, and/or fusion genes. Based on the findings in previous studies (Kulcheski et al., 2016; Panda, 2018). The expression of these miRNA target genes could be indirectly influenced by corresponding circRNAs through competing for the interaction with miRNAs. Consequently, our result suggests that circRNAs could potentially function as miRNA sponges to regulate the expression of an abundance of corresponding genes.

Discussion

In this study, we systematically explored the transcriptome alterations of SS in terms of gene expression and AS, as well as gene fusions and circRNAs. A total of 4286 genes (including 765 lncRNA genes) were differentially expressed between SS and paired tumor-adjacent tissues, which were mainly involved in fundamental biological processes and cancer-related pathways. Moreover, we experimentally validated the differential expression of seven splicing factors using qPCR. We also detected 2511 genes (including 41 lncRNA genes) that showed differential AS, where the most common AS mode was ES (80.37% of these DASGs), followed by ME, A3AS, A5DS, and IR. Gene functional enrichment analysis also showed that these DASGs were enriched in splicing-related biological processes and pathways. Surprisingly, those DEGs and DASGs were largely distinct, only a small portion of them were the same, suggesting that AS is complementary with expression level for investigating transcriptomic changes. Notably, a fraction of those DEGs and DAGs were oncogenes, tumor suppressors, and TFs, indicating that they could be closely associated with the tumorigenesis of SS. Moreover, we identified 14 tumor-specific gene fusion pairs in SS, which not only included the known gene fusions of SS18-SSX but also contained novel fusion events involving both protein-coding and lncRNA genes. Additionally, we observed that 49 circRNAs markedly changed expression in SS compared to tumor-adjacent tissues, and their parental genes were enriched in the muscle system process.

To the best of our knowledge, we are the first to study the SS transcriptome from a comprehensive view covering both transcriptional and post-transcriptional levels. Specifically, the deregulation of AS and the role of circRNAs were rarely explored in SS previously. An increasing number of studies have shown that imbalances in the AS process can affect the development of various human diseases, especially the oncogenesis, progression, and metastasis of a range of cancers (Scotti and Swanson, 2016). We identified 122 differentially spliced TFs and 124 differentially expressed TFs, suggesting that these TFs could be responsible for the expression level changes of an abundance of their target genes (Vaquerizas et al., 2009; Lambert et al., 2018). Moreover, we observed that 7 and 21 splicing factors were dramatically changed in expression level or AS profile. Since splicing factors are essential in regulating the AS of genes, these abnormally changed splicing factors may significantly contribute to the AS changes of many related genes (Anczukow and Krainer, 2016). On the other hand, circRNAs have critical regulatory functions and play key roles in the initiation and progression of diverse diseases including cancers (Zhang et al., 2018; Haddad and Lorenzen, 2019). The differentially expressed circRNAs identified by us were mainly generated from the genes correlated with the muscle system process and contraction. We also constructed the interaction network among circRNAs, miRNAs, and downstream target genes to elucidate their potential regulatory mechanism. The resulting network indicated that those differentially expressed circRNAs have the potential to act as the sponge for dozens of miRNAs to indirectly regulate the expression of hundreds of DEGs and DASGs.

Conclusion

Collectively, we systematically dissected the transcriptomic profile of SS and identified a number of DEGs, DASGs, fusion genes, and circRNAs that could be closely associated with the tumorigenesis of SS. Our study not only gained novel insights into SS transcription and post-transcription but also shed light on the underlying molecular mechanisms.

Data availability statement

The data presented in the study are deposited in the Gene Expression Omnibus (GEO) repository, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE144190.190, accession number GSE144.

Ethics statement

Clinical samples were collected from Fudan University Shanghai Cancer Center of China. This study was approved by the Ethics Committee of Fudan University Shanghai Cancer

Center. All patients in this study provided written informed consent for sample collection and data analyses.

Author contributions

ZS, MY, and YD are co-first authors. SYB, LK and YWJ are co-response authors.

Funding

The study was supported by Shanghai Municipal Health Commission (202140393).

Conflict of interest

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

The reviewers (YQ, LL) declared a shared affiliation with the author(s) (ZS, YD, YS, YC, KL, and WY) to the handling editor at the time of review

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.925564/full#supplementary-material

References

Abu, N., and Jamal, R. (2016). Circular RNAs as promising biomarkers: A mini-review. *Front. Physiol.* 7, 355. doi:10.3389/fphys.2016.00355

Anczukow, O., and Krainer, A. R. (2016). Splicing-factor alterations in cancers. *Rna* 22 (9), 1285–1301. doi:10.1261/rna.057919.116

Baranello, L., Wojtowicz, D., Cui, K. R., Devaiah, B. N., Chung, H. J., Chan-Salis, K. Y., et al. (2016). RNA polymerase II regulates topoisomerase 1 activity to favor efficient transcription. *Cell* 165 (2), 357–371. doi:10.1016/j.cell.2016.02.036

Barrett, S. P., and Salzman, J. (2016). Circular RNAs: analysis, expression and potential functions. *Development* 143 (11), 1838–1847. doi:10.1242/dev.128074

Cancer Genome Atlas Research NetworkElectronic Address (2017). Comprehensive and integrated genomic characterization of adult soft tissue sarcomas. *Cell* 171 (4), 950–965. doi:10.1016/j.cell.2017.10.014

Chen, L. L. (2016). The biogenesis and emerging roles of circular RNAs. *Nat. Rev. Mol. Cell Biol.* 17 (4), 205–211. doi:10.1038/nrm.2015.32

Chibon, F., Lagarde, P., Salas, S., Perot, G., Brouste, V., Tirode, F., et al. (2010). Validated prediction of clinical outcome in sarcomas and multiple types of cancer on the basis of a gene expression signature related to genome complexity. *Nat. Med.* 16 (7), 781–787. doi:10.1038/nm.2174

Dong, Y. P., He, D., Peng, Z. Z., Peng, W., Shi, W. W., Wang, J., et al. (2017). Circular RNAs in cancer: an emerging key player. *J. Hematol. Oncol.* 10, 2. doi:10.1186/s13045-016-0370-2

Dvinge, H., Kim, E., Abdel-Wahab, O., and Bradley, R. K. (2016). RNA splicing factors as oncoproteins and tumour suppressors. *Nat. Rev. Cancer* 16 (7), 413–430. doi:10.1038/nrc.2016.51

Edwards, P. A. (2010). Fusion genes and chromosome translocations in the common epithelial cancers. *J. Pathol.* 220 (2), 244–254. doi:10.1002/path.2632

Gao, Y., Wang, J. F., and Zhao, F. Q. (2015). CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. *Genome Biol.* 16, 4. doi:10.1186/s13059-014-0571-3

Glazar, P., Papavasileiou, P., and Rajewsky, N. (2014). circBase: a database for circular RNAs. *Rna* 20 (11), 1666–1670. doi:10.1261/rna.043687.113

Greene, J., Baird, A. M., Brady, L., Lim, M., Gray, S. G., Mcdermott, R., et al. (2017). Circular RNAs: Biogenesis, function and role in human diseases. *Front. Mol. Biosci.* 4, 38. doi:10.3389/fmolb.2017.00038

Haddad, G., and Lorenzen, J. M. (2019). Biogenesis and function of circular RNAs in Health and in disease. *Front. Pharmacol.* 10, 428. doi:10.3389/fphar.2019.00428

He, R. Q., Wei, Q. J., Tang, R. X., Chen, W. J., Yang, X., Peng, Z. G., et al. (2017). Prediction of clinical outcome and survival in soft-tissue sarcoma using a ten-lncRNA signature. *Oncotarget* 8 (46), 80336–80347. doi:10.18632/oncotarget.18165

Kao, Y. C., Owosho, A. A., Sung, Y. S., Zhang, L., Fujisawa, Y., Lee, J. C., et al. (2018). BCOR-CCNB3 fusion positive sarcomas A clinicopathologic and molecular analysis of 36 cases with comparison to morphologic spectrum and clinical behavior of other round cell sarcomas. *Am. J. Surg. Pathol.* 42 (5), 604–615. doi:10.1097/PAS.0000000000000965

Kawai, A., Woodruff, J., Healey, J. H., Brennan, M. F., Antonescu, C. R., and Ladanyi, M. (1998). SYT-SSX gene fusion as a determinant of morphology and prognosis in synovial sarcoma. *N. Engl. J. Med.* 338 (3), 153–160. doi:10.1056/NEJM199801153380303

Keren, H., Lev-Maor, G., and Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. *Nat. Rev. Genet.* 11 (5), 345–355. doi:10.1038/nrg2776

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12 (4), 357–360. doi:10.1038/nmeth.3317

Kim, D., and Salzberg, S. L. (2011). TopHat-fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.* 12 (8), R72. doi:10.1186/gb-2011-12-8-r72

Kulcheski, F. R., Christoff, A. P., and Margis, R. (2016). Circular RNAs are miRNA sponges and can be used as a new class of biomarker. *J. Biotechnol.* 238, 42–51. doi:10.1016/j.jbiotec.2016.09.011

Ladanyi, M., Antonescu, C. R., Leung, D. H., Woodruff, J. M., Kawai, A., Healey, J. H., et al. (2002). Impact of SYT-SSX fusion type on the clinical behavior of synovial sarcoma: a multi-institutional retrospective study of 243 patients. *Cancer Res.* 62 (1), 135–140. doi:10.1146/annurev-biochem-060614-034316

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., et al. (2018). The human transcription factors. *Cell* 172 (4), 650–665. doi:10.1016/j.cell.2018.01.029

Lee, Y., and Rio, D. C. (2015). Mechanisms and regulation of alternative pre-mRNA splicing. *Annu. Rev. Biochem.* 84, 291–323. doi:10.1146/annurev-biochem-060614-034316

Li, J. H., Liu, S., Zhou, H., Qu, L. H., and Yang, J. H. (2014). starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* 42 (D1), D92–D97. doi:10.1093/nar/gkt1248

Li, W. S., Liao, I. C., Wen, M. C., Lan, H. H. C., Yu, S. C., and Huang, H. Y. (2016). BCOR-CCNB3-positive soft tissue sarcoma with round-cell and spindle-cell histology: a series of four cases highlighting the pitfall of mimicking poorly differentiated synovial sarcoma. *Histopathology* 69 (5), 792–801. doi:10.1111/his.13001

Li, X., Yang, L., and Chen, L. L. (2018). The biogenesis, functions, and challenges of circular RNAs. *Mol. Cell* 71 (3), 428–442. doi:10.1016/j.molcel.2018.06.034

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15 (12), 550. doi:10.1186/s13059-014-0550-8

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., et al. (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 495 (7441), 333–338. doi:10.1038/nature11928

Panda, A. C. (2018). Circular RNAs act as miRNA sponges. *Adv. Exp. Med. Biol.* 1087, 67–79. doi:10.1007/978-981-13-1426-1_6

Pavanello, L., Hall, B., Airhihen, B., and Winkler, G. S. (2018). The central region of CNOT1 and CNOT9 stimulates deadenylation by the Ccr4-Not nuclease module. *Biochem. J.* 475, 3437–3450. doi:10.1042/BCJ20180456

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33 (3), 290–295. doi:10.1038/nbt.3122

Pierron, G., Tirode, F., Lucchesi, C., Reynaud, S., Ballet, S., Cohen-Gogo, S., et al. (2012). A new subtype of bone sarcoma defined by BCOR-CCNB3 gene fusion. *Nat. Genet.* 44 (4), 461–466. doi:10.1038/ng.1107

Przybyl, J., Sciot, R., Rutkowski, P., Siedlecki, J. A., Vanspauwen, V., Samson, I., et al. (2012). Recurrent and novel SS18-SSX fusion transcripts in synovial sarcoma: description of three new cases. *Tumour Biol.* 33 (6), 2245–2253. doi:10.1007/s13277-012-0486-0

Qu, S., Yang, X., Li, X., Wang, J., Gao, Y., Shang, R., et al. (2015). Circular RNA: a new star of noncoding RNAs. *Cancer Lett.* 365 (2), 141–148. doi:10.1016/j.canlet.2015.06.003

Scotti, M. M., and Swanson, M. S. (2016). RNA mis-splicing in disease. *Nat. Rev. Genet.* 17 (1), 19–32. doi:10.1038/nrg.2015.3

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13 (11), 2498–2504. doi:10.1101/gr.1239303

Shen, S., Park, J. W., Lu, Z. X., Lin, L., Henry, M. D., Wu, Y. N., et al. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. U. S. A.* 111 (51), E5593–E5601. doi:10.1073/pnas.1419161111

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 (43), 15545–15550. doi:10.1073/pnas.0506580102

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47 (D1), D607–D613. doi:10.1093/nar/gky1131

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., and Luscombe, N. M. (2009). A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* 10 (4), 252–263. doi:10.1038/nrg2538

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., et al. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* 456 (7221), 470–476. doi:10.1038/nature07509

Wang, G. S., and Cooper, T. A. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet.* 8 (10), 749–761. doi:10.1038/nrg2164

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., et al. (2013). The cancer genome Atlas pan-cancer analysis project. *Nat. Genet.* 45 (10), 1113–1120. doi:10.1038/ng.2764

Yen, C. C., Yeh, C. N., Cheng, C. T., Jung, S. M., Huang, S. C., Chang, T. W., et al. (2012). Integrating bioinformatics and clinicopathological research of gastrointestinal stromal tumors: Identification of aurora kinase A as a poor risk marker. *Ann. Surg. Oncol.* 19 (11), 3491–3499. doi:10.1245/s10434-012-2389-0

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

Zhang, J., and Manley, J. L. (2013). Misregulation of pre-mRNA alternative splicing in cancer. *Cancer Discov.* 3 (11), 1228–1237. doi:10.1158/2159-8290.CD-13-0253

Zhang, M. C., and Xin, Y. (2018). Circular RNAs: a new frontier for cancer diagnosis and therapy. *J. Hematol. Oncol.* 11, 21. doi:10.1186/s13045-018-0569-5

Zhang, Y., Zhang, X. O., Chen, T., Xiang, J. F., Yin, Q. F., Xing, Y. H., et al. (2013). Circular intronic long noncoding RNAs. *Mol. Cell* 51 (6), 792–806. doi:10.1016/j.molcel.2013.08.017

Zhang, Z. R., Yang, T. T., and Xiao, J. J. (2018). Circular RNAs: Promising biomarkers for human diseases. *Ebiomedicine* 34, 267–274. doi:10.1016/j.ebiom.2018.07.036

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 19 September 2022
DOI 10.3389/fgene.2022.989985

OPEN ACCESS

EDITED BY
Geng Chen,
Stemirna Therapeutics Co., Ltd., China

REVIEWED BY
Milind B. Ratnaparkhe,
ICAR Indian Institute of Soybean Research, India
Rasime Kalkan,
Cyprus Health and Social Sciences University, Cyprus

*CORRESPONDENCE
Dongguo Li,
ldg213@ccmu.edu.cn

†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 29 July 2022
ACCEPTED 30 August 2022
PUBLISHED 19 September 2022

CITATION
Wang R, Zhao L, Wang S, Zhao X, Liang C, Wang P and Li D (2022), Regulatory pattern of abnormal promoter CpG island methylation in the glioblastoma multiforme classification. *Front. Genet.* 13:989985. doi: 10.3389/fgene.2022.989985

Regulatory pattern of abnormal promoter CpG island methylation in the glioblastoma multiforme classification

Rendong Wang[1,2†], Lei Zhao[3†], Shijia Wang[1,2], Xiaoxiao Zhao[1,2], Chuanyu Liang[3], Pei Wang[3] and Dongguo Li[1,2*]

[1]School of Biomedical Engineering, Capital Medical University, Beijing, China, [2]Beijing Key Laboratory of Fundamental Research on Biomechanics in Clinical, Capital Medical University, Beijing, China, [3]Department of Anesthesiology, Xuanwu Hospital, Capital Medical University, Beijing, China

Glioblastoma (GBM) is characterized by extensive genetic and phenotypic heterogeneity. However, it remains unexplored primarily how CpG island methylation abnormalities in promoter mediate glioblastoma typing. First, we presented a multi-omics scale map between glioblastoma sample clusters constructed based on promoter CpG island (PCGI) methylation-driven genes, using datasets including methylation profiles, expression profiles, and single-cell sequencing data from multiple highly annotated public clinical cohorts. Second, we identified differences in the tumor microenvironment between the two glioblastoma sample clusters and resolved key signaling pathways between cell clusters at the single-cell level based on comprehensive comparative analyses to investigate the reasons for survival differences between two of these clusters. Finally, we developed a diagnostic map and a prediction model for glioblastoma, and compared theoretical differences of drug sensitivity between two glioblastoma sample clusters. In summary, this study established a classification system for dissecting promoter CpG island methylation heterogeneity in glioblastoma and provides a new perspective for the diagnosis and treatment of glioblastoma.

KEYWORDS

glioblastoma 1, subtype classification 2, CpG island 3, DNA methylation 4, tumor microenviroment 5, single-cell RNA sequencing 6, intercellular communication 7

Introduction

Glioblastoma (GBM) is a malignant primary brain cancer characterized by high infiltration into the parenchyma and wide phenotypic heterogeneity (Hua et al., 2015). Despite advances in surgical techniques and clinical regimens, the standard therapies, including surgical resection, chemotherapy, are predominantly ineffective for GBMs due to therapeutic resistance, rapid recurrence, and the patient outcomes remain between

Abbreviations: PCGI, promoter CpG island; MMPs, matrix metalloproteinases.

12 and 15 months survival rate, 5-year survival rates at only 10% (Tao et al., 2020). In light of the molecular complexity and histopathological grading of GBM (Vitucci et al., 2017), there is a critical need to complement the inaccurate prediction of disease progression and the deviation of therapy with genomic information.

The significant factors contributing to the pathogenesis of GBM were epigenetic molecular mechanisms (Kosti et al., 2020). DNA methylation, the most common epigenetic event in cancer, contributes to carcinogenesis and frequently occurs in the promoter region of genes (Agundez et al., 2011; Wang et al., 2022). With the help of multi-omics datasets, profiles of GBM at the transcriptome and methylation levels have been increasingly reported to investigate the extensive heterogeneity in the tumor and single-cell level regarding transcriptomic expression (Oh et al., 2020). Several extensive cohort studies indicate an important association between DNA methylation of the promoter region and phenotypic of GBM (Guo et al., 2015). For instance, the discordance of promoter methylation with O-6-methylguanine-DNA-methyltransferase (MGMT) expression in GBM has been a plausible strategy for sensitizing temozolomide (TMZ) therapy and provides a strong rationale for the development of new drugs (Yi et al., 2019). Furthermore, numerous potential prognostic biomarkers, including long non-coding RNA (lncRNA) and mRNA, were identified with aberrant methylation (Han et al., 2020). The characterization of the epigenome by DNA methylation assay has been progressively used to stratify and integrate molecular and phenotypic features. Nevertheless, with advances in genomics, the single-gene methylation status has limited its clinical utility.

During cancer development, aberrant DNA methylation occurs within the gene promoter, CpG island, and their shores (Hardy et al., 2017). However, CpG island has received little individual attention. CpG sites methylation patterns are believed to differ considerably between GBM patients (Etcheverry et al., 2010). In particular, some cancers show an apparent CpG island methylator phenotype (CIMP), of which a critical milestone highlighting the clinical importance of the epigenetic profile of gliomas was the discovery of the glioma CpG island methylation phenotype (G-CIMP) (Northcott et al., 2017; Ogino et al., 2018). Specifically, patients carrying G-CIMP have a better prognosis than patients who do not carry this phenotype. The clusters identified by separating Isocitrate dehydrogenase (IDH) mutation status showed overall concordance with G-CIMP, which exemplifies the particularity of CpG island in the molecular diagnosis of GBM (Geisenberger et al., 2015; Park et al., 2019). Recent studies also suggest that the tumor microenvironment (TME) plays an essential role in clinical outcomes and response to therapy (Gangoso et al., 2021). The tumor microenvironment of GBM contains a large number of infiltrating macrophages (Chen et al., 2019). However, few studies have assessed the epigenetic alterations and the TME simultaneously, especially at the single-cell level. Here we explored a comprehensive genomic and transcriptomic analysis. We resolve the comprehensive characterization of GBM subgroups by integrating CpG island methylation, expression profiling, and single-cell sequencing data. Finally, we constructed a planetary diagnostic view and performed a drug sensitivity analysis to illustrate the clinical contribution of the results.

Materials and methods

Data sources

The HM450k DNA methylation data were downloaded from The Cancer Genome Atlas (TCGA, https://portal.gdc.cancer.gov/) database and GSE41826 in Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/), which includes 155 tumor samples and 56 normal samples. The methylation level of each probe was represented by the β-value (from 0 to 1). β-Value = $I_{meth}/I_{meth} + I_{unmeth}$, I_{meth} is the intensity of methylation, and I_{unmeth} is the intensity of unmethylation. CpG methylation probes were annotated with the platform annotations in GEO (GPL13534). Clinical information and expression data were downloaded from the TCGA database, and the expression level was quantified as fragments per kilobase of transcription per million mapped reads (FPKM) values. Besides, we downloaded gene expression data from the Chinese Glioma Genome Atlas (CGGA, http://www.cgga.org.cn/) database as a supplementary dataset, which includes 282 GBM patients who possessed complete clinical information (Zhao et al., 2021a). The annotation file for mRNAs and promoter region was derived from the GENCODE database (https://www.gencodegenes.org/) (Di Risi et al., 2021). The single-cell sequencing data were obtained from GSE162631 in the GEO database, and cells derived from the tumor cores of three GBM patients in the dataset were selected (Xie et al., 2021a). Expression profile data of human cancer cell lines (CCLs) were obtained from the Broad Institute Cancer Cell Line Encyclopedia (CCLE) project (https://portals.broadinstitute.org/ccle/). Drug sensitivity data of CCLs were achieved from the Cancer Therapeutics Response Portal (CTRPv.2.0, https://portals.broadinstitute.org/ctrp), which contains the sensitivity data for 481 compounds over 835 CCLs (Lauria et al., 2020; Bagaev et al., 2021). The dataset provides the area under the dose-response curve (area under the curve AUC) values as a measure of drug sensitivity, and lower AUC values indicate increased sensitivity to treatment.

Gene regulation patterns and GBM molecular cluster classification

DEGs were identified with the Limma R package (version 3.48.3), and adjusted p-value<0.05 and

$|\log_2$ Fold Change (FC)|>1 were considered to have a significant difference (Liu et al., 2021). We used the MethylMix R package (version 2.22.0) with the $|\log_2 FC|>0.5$, Cor < -0.3, p-value<0.05 to extract the PCGI methylation-driven genes (Xu et al., 2019). Based on the expression of genes, GBM samples were clustered into K (2–9) groups using the ConsensusClusterPlus package (version 1.56.0) in R software (Wilkerson and Hayes, 2010). The optimal K value was determined to obtain a stable cluster, of which correlation coefficients were computed by spearman, and partitioning around medoids was selected as a clustering algorithm.

Single sample gene set enrichment analysis

The corresponding enrichment score was computed with the GSVA R package (version 1.40.1) (Lauria et al., 2020), which estimated the biological similarity of immune cells by multi-dimensional scaling and a Gaussian fitting model to represent the relative abundance of each immune cell type in gene set enrichment analysis (ssGSEA). Specifically, the tumor microenvironment was assessed by immunohistochemistry for markers of immune cell types (Supplementary Table S1). Further, the ssGSEA score was normalized to unity distribution for each immune cell type, and the estimate scores, including purity, stromal and immune values, were calculated with the estimate R package (version 1.0.13) (Yoshihara et al., 2013).

Single-cell analysis

We collected three separate tissue samples originating from the tumor core in GBM patients from GSE162631 (21). The raw count data were loaded into the Seurat package (version 4.0.5) for quality control (QC), data filtering, normalization, Principal Component Analysis (PCA), Uniform Manifold Approximation, and Projection (UMAP) visualization, clustering. The single-cell sequencing data from three patients were integrated by the Harmony R package (version 0.1.0) and the cells with mitochondrial genes greater than 10% or fewer than 300 detected genes were filtered out. A scale factor of 10,000 was used to normalize all the remaining cells (Xie et al., 2021a). We used the FindAllMarkers function in Seurat to determine the genes enriched in each cluster and set a logFC threshold of 0.25. It applies a Wilcoxon Rank Sum test and performs multiple test corrections using the Bonferroni method. We used Cellchat R package (version 1.1.3) with the cellchatDB.Human database, which includes supporting evidence for each signaling interaction and considers the structural composition of ligand-receptor interactions and cofactor molecules to identify and visualize cell-cell interactions (Jin et al., 2021).

Co-expression network

We calculated the Spearman correlation between ligand-receptor genes with PCGI methylation-driven genes. The regulation pairs with Cor>0.4 and p-value<0.05 were used to construct the co-expression network, which visualized in Cytoscape (version 3.9.0). We used cytoHubba plug-ins built into Cytoscape to calculate key genes in the network.

Statistical analysis

All statistical tests were performed in R software (v4.0.3). For the comparisons of the normally distributed groups, statistical analysis was performed by t-tests, and for non-normally distributed variables, statistical analysis was analyzed by Wilcoxon rank-sum tests. The Chi-square test is used to compare clinical, pathological parameters, and other categorical variables. Correlation between two continuous variables was measured by either Pearson's correlation or Spearman's correlation. For survival analysis, the differences in prognosis between clusters were assessed via Kaplan-Meier OS analysis, and log-rank tests were utilized to judge the differences between clusters. The cluster prediction model was constructed with LASSO regression in the glmnet R package (version 4.1.2) (Huang et al., 2021). The pROC package (version 1.18.0) in R was utilized to calculate the ROC curves and AUC values. For all statistical analyses, a two-tailed $p < 0.05$ was considered significant. Significance values correspond to p-value as follows: ns > 0.05, *<0.05, **<0.01, ***<0.001, ****<0.0001.

Drug sensitivity

We used the Ridge regression analysis in the pRRophetic R package (version 0.5) to predict differences in drug sensitivity between the two clusters of GBM cancer samples using default settings (Yang et al., 2021). K nearest neighbor imputation was applied to impute the missing AUC values. We used the normalization method to modify the drug sensitivity data matrix of CCLs (Roy et al., 2019). The drugs with $|\log_2 FC|>0.1$ were considered to have differential sensitivity in different clusters (Yang et al., 2021).

Results

Identification of glioblastoma clusters based on promoter CpG island methylation-driven genes

To investigate the DNA methylation of promoter CpG island (PCGI) associated with GBM disease progression, we established

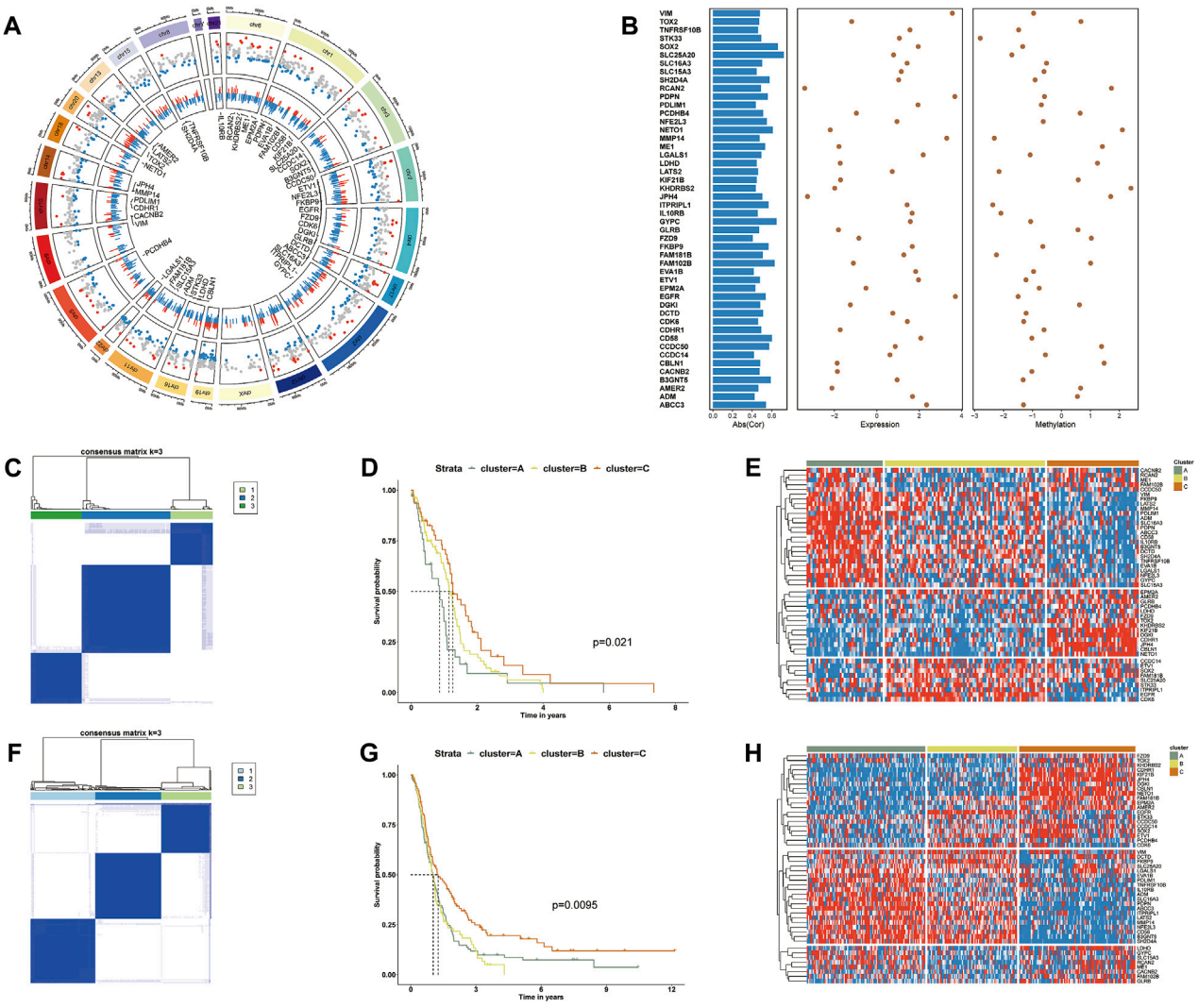


FIGURE 1
(A) The Screening process of PCGI methylation-driven genes. The first circle shows the chromosomal location and ordering information. The second circle represents the distribution of differential methylation genes. Red points represent hypermethylation, and blue ones represent hypomethylation relative to normal samples. The third circle shows the relationship between methylation and gene expression. Blue and red bars represent values with negative and positive correlations, respectively. **(B)** Overview of 48 PCGI methylation-driven genes. The first column shows the absolute values of the correlation coefficients between DNA methylation and expression level. The second and third columns show the fold change of the expression and methylation level. **(C)** Consensus cluster for GBM patients of TCGA based on PCGI methylation-driven genes. **(D)** Kaplan–Meier survival analysis for TCGA sample clusters. **(E)** Hierarchically clustered heatmap for the expression of PCGI methylation-driven genes across clusters in TCGA. **(F)** Consensus cluster for GBM patients of CGGA based on PCGI methylation-driven genes. **(G)** Kaplan–Meier survival analysis for CGGA sample clusters. **(H)** Hierarchically clustered heatmap for the expression of PCGI methylation-driven genes across clusters in CGGA.

a richly computational strategy that maps the Infinium HumanMethylation450K microarray to gene PCGI methylation profiles and summarizes DNA methylation patterns at the gene level (Zhao et al., 2021b). First, based on the gene annotation derived from the GENCODE database and GPL13534 platform file containing the methylation probes information (Li et al., 2020; Di Risi et al., 2021), we defined the promoter region as 2 kb located upstream of the transcription start site (Hollstein et al., 2019). We extracted the relevant probes on the PCGI from the annotation file for subsequent analysis (Carro et al., 2010). The mean value for probes was calculated as the methylation level of genes (Liu et al., 2020). In total, 46,072 probes in the DNA methylation microarray were annotated to 15,067 genes, of which we selected 10,895 coding genes according to the gene annotation file. The DNA methylation profiles exhibit the distribution of DNA

methylation across the CpG island with a typical DNA hypomethylation tendency in GBM (Supplementary Figure S1A).

Overall, studies on DNA methylation are thought to be associated with an opposite gene expression pattern. Thus, we identified the differentially expressed genes (DEGs) with Limma R package (p-value<0.05; $|\log_2 FC|>1$; Supplementary Figure S1B) and calculated the methylation differences and the correlation between expression and methylation with MethylMix R package ($|\log_2 FC|>0.5$; Cor ≤ -0.3; p-value<0.05). Ultimately, we identified 48 PCGI methylation-driven genes (Figures 1A,B and Supplementary Table S2) (Xu et al., 2019).

To clarify the heterogeneity of PCGI methylation-driven genes in TCGA-GBM tumor samples, we performed the consensus cluster method to cluster the samples based on the similarity of PCGI methylation-driven genes expression signature (Datta et al., 2021). It is worth noting that all samples were likely categorized into three clusters named ClusterA, ClusterB, and ClusterC because the interference between clusters can be minimized when K = 3 was selected (Figure 1C and Supplementary Figures S1C–L) (Gong et al., 2020). The epigenomic analysis demonstrates that GBM patients exhibit different levels of abnormal methylation in promoter CpG island, reflecting the heterogeneity of GBM. Particular clustering results for each sample are listed in Supplementary Table S3. The prognostic characteristics of clusters were further appraised by survival analysis, indicating that PCGI methylation is a significant prognostic factor in GBM patients (Figure 1D). The heatmap showed significant disparities in PCGI methylation-driven genes between clusters (Figure 1E). We further collected 283 GBM samples from the Chinese Glioma Genome Atlas (CGGA) RNA-seq database with clinical information data available and performed the analogous analysis to verify the rationality of results obtained from TCGA(19): we determined the clustering results for CGGA patients based on similarity in gene expression and calculated the survival probabilities between different clusters (Figures 1F,G). Of particular interest, the expression pattern of PCGI methylation-driven genes in the CGGA database is similar to that of the TCGA database, with samples divided into three clusters based on gene expression (Figure 1H). As could be expected, we observed high concordance between the clustering results and prognostic features of CGGA and TCGA.

Tumor microenvironment heterogeneity between glioblastoma clusters

As substantial changes in the tumor microenvironment with infiltrating immune cells and gene regulation machinery can influence tumor progression (Tian et al., 2021), we put the DNA methylation data into a broader GBM context to identify the effect of PCGI methylation in-depth on the tumor microenvironment. We first identified the DEGs between clusters with significant survival differences (clusters A and C) and performed the single sample gene set enrichment analysis (ssGSEA) analysis based on the immune cell signature gene set to investigate the differences between clusters (Bastola et al., 2020; Krug et al., 2020). The results showed that specific PCGI methylation-driven genes were substantially different between clusterA and clusterC within the top differentially expressed genes, including *KIF21B*, *JPH4*, *NET O 1*, *FAM181B*, *AMER2* being up-regulated in clusterC, while *ABCC3*, *MMP14*, *LGALS1*, *SLC16A3*, *PDPN*, *ADM* exhibiting up-regulated in clusterA (Figures 2A,B). Additionally, we got the same results in the CGGA database (Figure 2C). Remarkably, we observed significant differences in the immune cell infiltration between the two clusters. The abundance score of immune cells calculated with ssGSEA was lower in clusterC and higher in clusterA, as shown in Figure 2D. Collectively, it is worth investigating this apparent inconsistency between clusterA and clusterC in the tumor microenvironment as a possible reason for the difference in clinical survival of patients (Gangoso et al., 2021). We further utilized the ESTIMATE R package on the expression profiles of TCGA samples to infer immune and stromal scores for estimating Tumor Purity, Stromal, and Immune Scores (Figure 2E) (Riaz et al., 2017; Stewart et al., 2017; Krug et al., 2020). Studies exist demonstrating that the mesenchymal subtype has many immune cells, in concordance with our work, this subtype showed lower cell density and large necrotic areas in histopathology (Klughammer et al., 2018). We observed high levels of macrophages in clusterA and low levels in clusterC (Figure 2D). We have reviewed the available studies that higher with increased macrophages is associated with lower overall survival (Chen et al., 2019). We also found that the matrix metalloproteinases (MMPs), which might influence the expression of multiple proteins in the extracellular matrix, were differentially expressed between two clusters (Supplementary Table S4) (Theodoris et al., 2015). We speculated that the differences in immune cells could be responsible for the survival status between clusters A and C. Similarly, we obtained practically consistent results by validating with the CGGA database, which proved that our analysis was reliable (Figures 2F,G).

Linking single cell analysis and communication patterns to glioblastoma clusters

To accurately assess the tumor microenvironment between clusters A and C, we analyzed the single-cell data from the core tumor region of three GBM patients (GSE162631). Specifically, the cells were analyzed with the Seurat package in R and annotated according to the expression of canonical cell class markers and the SingleR R package (Xie et al., 2021a; Lu et al.,

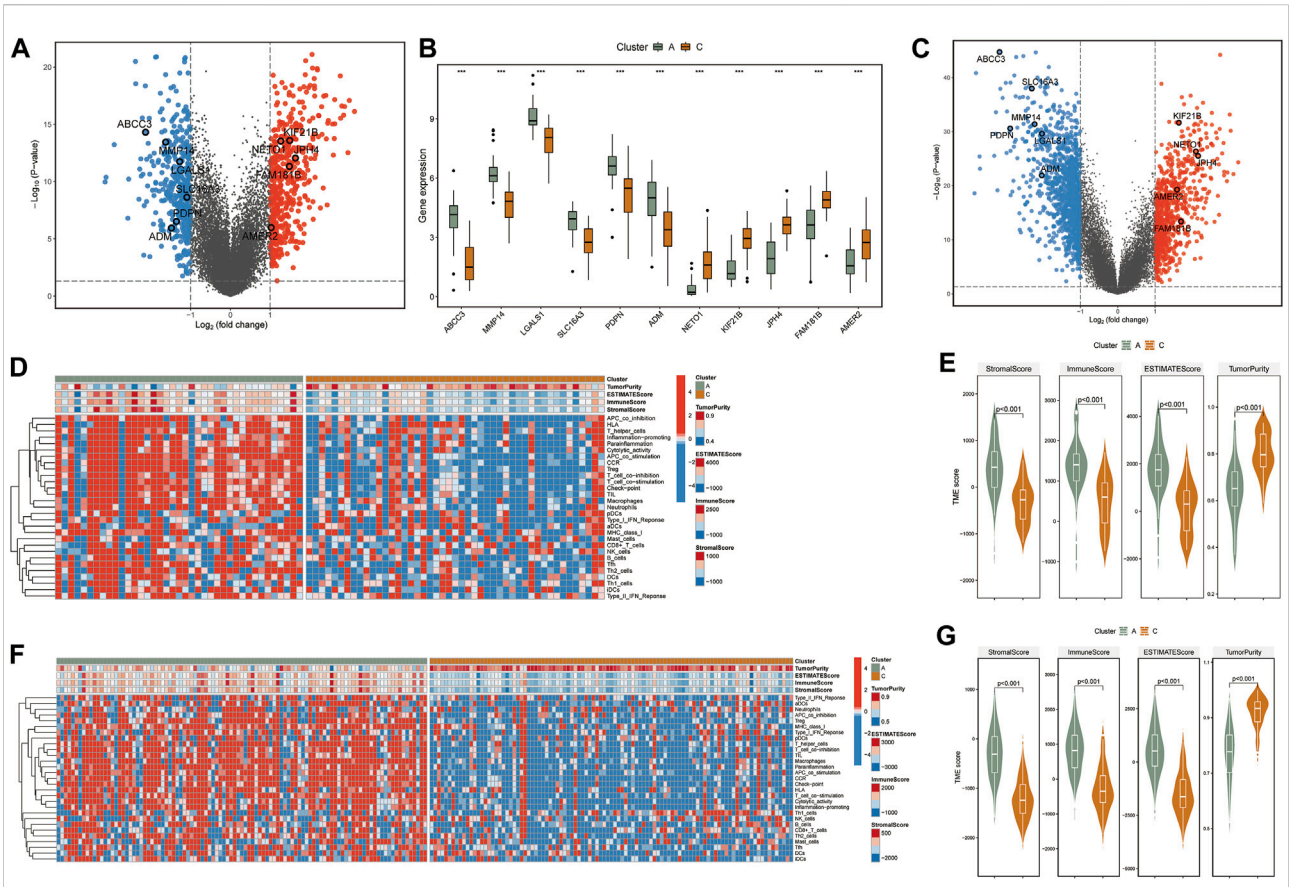


FIGURE 2
(A) The volcano plot shows fold changes for genes differentially expressed between TCGA-clusters A and C. The differentially expressed PCGI methylation-driven genes are highlighted in the figure. **(B)** Boxplots of differentially expressed PCGI methylation-driven genes between clusters A and C. **(C)** The volcano plot shows fold changes for genes differentially expressed between CGGA-clusters A and C. The differentially expressed PCGI methylation-driven genes are highlighted in the figure. **(D)** Heatmap and hierarchical clustering of normalized immune cell infiltration score of TCGA samples. **(E)** Violin plots for the distributions of StromalScore, immuneScore, ESTIMATEScore, and TumorRurity of TCGA samples. **(F)** Heatmap and hierarchical clustering of normalized immune cell infiltration score of CGGA samples. **(G)** Violin plots for the distributions of StromalScore, immuneScore, ESTIMATEScore, and TumorRurity of CGGA samples.

2021). After the data preprocessing pipeline, the dataset contains 14,926 cells, which cluster into 10 cell groups. The clusters included Macrophages (*APOC1*, *CD163*, *F13A1*), Microglia (*CX3CR1*, *P2RY12*, *P2RY13*), Neutrophils (*IL1R2*, *CXCR2*, *FPR2*), T cells (*CD3D*, *CD3E*, *GZMK*), B cells (*IGHG1*, *IGHG3*, *CD79A*), Dendritic cells (*HLA-DQA1*, *HLA-DPB1*), Glial/Neuronal cells (*FABP7*, *PTPRZ1*), Endothelial cells (*CD34*, *VWF*, *CLDN5*) and Mural cells (*RGS5*, *PDGFRB*, *NOTCH3*) were identified in this data set (Figures 3A,B) (Xie et al., 2021a). We observed a high content of macrophages, monocytes, and microglia in the single-cell sequencing data of GBM samples. Figure 3B illustrates the overlap in gene expression between these 3 cell groups. Consistent with previous studies, the gene expression patterns of these 3 cell groups are similar, and it has always been a challenge to accurately distinguish them in the GBM microenvironment (Ryan et al., 2017; Yao et al., 2020).

Based on published research, we recognized that microglia and tumor-associated macrophages, which accumulate in the tumor region secreting MMPs to promote tumor invasion and secrete tumor cell proliferation promoting factors are distinct subpopulations derived from mononuclear phagocytes (Fan et al., 2020; Ma et al., 2020). The available gene markers do not reliably discriminate between microglia and macrophages. In contrast, the B-cell content was shallow in the GBM microenvironment (Figure 3A). In the central nervous system, B cells are responsible for the antigenic presentation of tumor antigens and participate in anti-tumor immunity (Galstyan et al., 2019).

To predict cell signaling and inferred the precise connections between identified cell clusters to uncover coordinated responses among different cell types. We assessed not only the cell types in the tumor microenvironment but also the interactions between cells within the GBM tumor microenvironment, which constitute

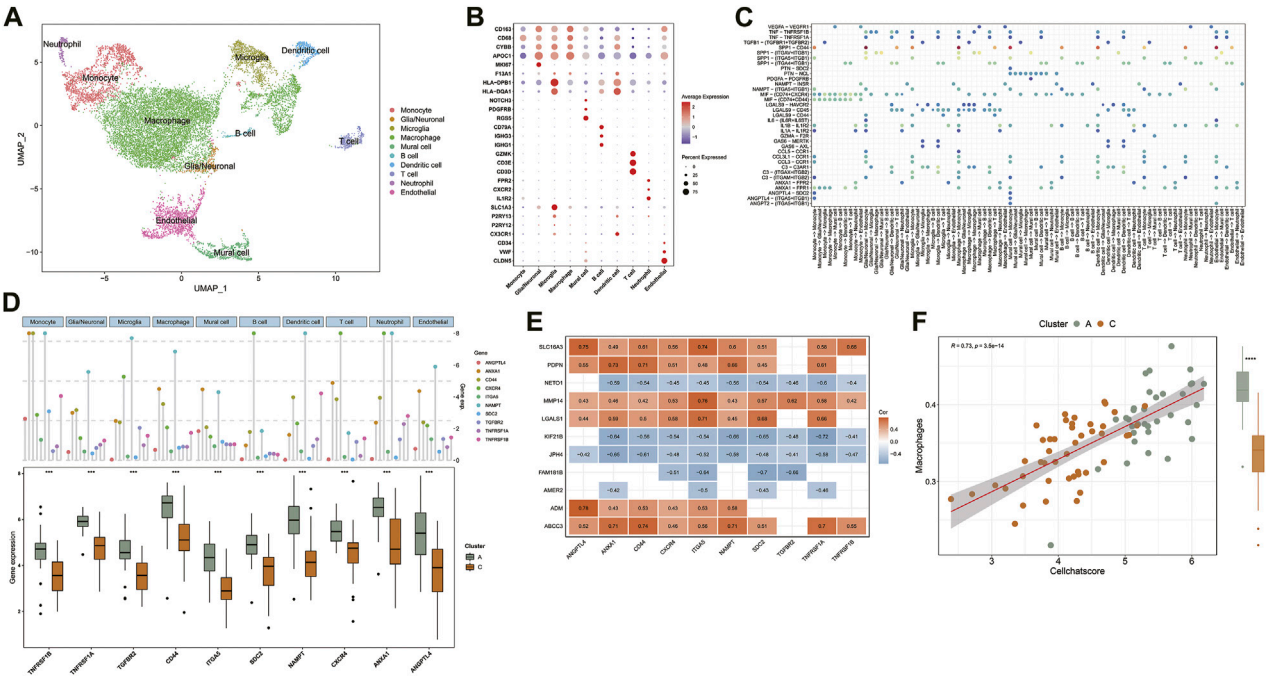


FIGURE 3
(A) UMAP of Single-cell sequencing data, colored for the 10 cell clusters. **(B)** Dot plot heatmap of the marker genes in individual cell clusters. **(C)** The dot plot shows the significant signaling patterns and ligand-receptor pairs. Dot color reflects communication probabilities and dot size represents computed *p*-values. The highlighted signals are pathways in which ligand-receptor genes are differentially expressed between clusters. **(D)** Expression distribution of differentially expressed ligand-receptor genes in 10 cell clusters and comparison of expression between TCGA sample clusters A and C. **(E)** Heatmap shows Spearman's correlations between PCGI methylation-driven genes and ligand-receptor genes. **(F)** The correlation between Cellchat score and macrophage infiltration score.

an additional layer of information for the integration of DNA methylation data (Lennon et al., 2016).

Normalized single-cell data was then loaded into the Cellchat R package, which integrates cell gene expression and prior knowledge of the interactions between signaling ligands, receptors, and their cofactors to model ligand-receptor mediated signaling interactions (Jin et al., 2021; Leimkuhler et al., 2021). Lastly, we calculated the probability of intercellular communication through Cellchat's standard process. We detected 35 significant ligand-receptor pairs categorized into 18 signaling pathways, including SPP1, MIF, COMPLEMENT, IL1, ANNEXIN, VISFATIN, GALECTIN, CCL, TNF, PTN, VEGF, GAS, ANGPT, ANGPTL, TGFβ, PARs, IL6, PDGF (Figure 3C). Signaling contribution analysis of cell populations revealed that monocytes were the most important source of SPP1 pathway receptors and the most important source of ANGPTL pathway ligands. Additionally, the communication patterns of multiple cell populations are clustered in the GALECTIN pathway which provided compelling evidence that different cells may depend on the same signals (Supplementary Figures S2D–F).

We further intersected the identified ligand-receptor genes with the list of differentially expressed genes between clusters A and C in the TCGA database simultaneously. The results indicate that ten ligand-receptor genes were differentially expressed between clusters A and C. Notably, all ten ligand-receptor genes were up-regulated in cluster A and down-regulated in cluster C, showing a consistent pattern of differential expression in general (Figure 3D). Compared to the ssGSEA results, these significant differences in the expression distribution trends of the ten ligand-receptor genes in the GBM sample clusters are comparable to the differences in immune cell abundance between clusters A and C (Figures 2D,F). Specifically, multiple signaling pathways may be activated in the tumor microenvironment of subtype A, including TNF, SPP1, MIF, ANGPL, and ANGPTL (Figure 3C). For example, TNF receptor superfamily members might participate in the progression of GBM through responses to TNF signaling pathway and are associated with poor prognosis (Xie et al., 2021b). This cross-referencing of single-cell sequencing data with epigenetic analysis models provides rapid insight into the mechanisms underlying the analysis of the GBM tumor microenvironment.

The correlation between PCGI methylation-driven genes and ligand-receptor genes was further evaluated to explore the effect of PCGI methylation-driven genes on patients' tumor microenvironment. The correlation heatmap shows that 78.2% of the correlation coefficient matrices had absolute values greater than

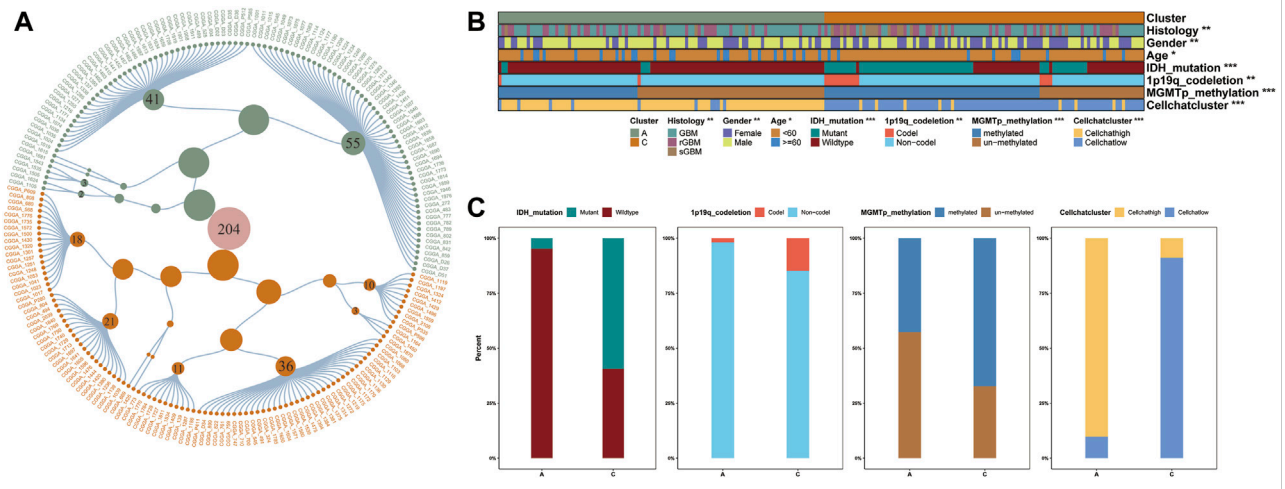


FIGURE 4
(A) Clinical diagnostic map based on GBM samples in CGGA database. Constellation map of GBM sample distribution based on the results of this study. The green points represent clusterA, and orange ones represent clusterC, which is consistent with the color annotation of other figures in this study. **(B–C)** The distribution of CGGA clinical characteristics and Cellchat score we defined.

0.4, embodying a critical regulatory relationship between PCGI methylation-driven genes and ligand-receptor genes (Figure 3E). Next, we applied weighted co-expression network analysis to the correlation coefficient matrices and explored the critical nodes in the network. Ranked by the degree method, we found that ITGA5 may play an essential role in the network as a key node (Supplementary Figure S2B and Supplementary Table S5). Our results show that the major signaling pathways of ITGA5 include SPP1, ANGPTL, ANGPT, which are characterized by monocytes in the incoming interaction environment, but the communication patterns of outcoming interaction are dominated by macrophages (Supplementary Figure S2C). Then, we defined the mean value of ligand-receptor genes expression in each sample of the TCGA database as a Cellchat score, which quantified the strength of cell communication. We observed that Cellchat score played a significant positive correlation with the abundance of macrophages (Cor = 0.73; $p < 0.001$) (Figure 3F). Given the crucial role of macrophages in the GBM tumor microenvironment, significant heterogeneity in the expression profile of ligand-receptor genes could help us differentiate the infiltration of macrophages in GBM clusters regulated by PCGI methylation-driven genes, eliminating the dependence of epigenetic typing on high-quality methylation data (Klughammer et al., 2018).

Building diagnostic map and cluster prediction model for the clinical improvement of GBM clusters

Precise molecular clustering and clinical features may become key components in prognostic index models (Giacopelli et al., 2019). Therefore, to evaluate the contribution of the GBM clustering results to diagnosis and prognosis in this study, we constructed a diagnostic map containing all clinical features in the database, based on the samples from the CGGA database (Figure 4A and Supplementary Figure S3A) (Bagaev et al., 2021). Additionally, to characterize the clinicopathological relevance of our results, we not only compared the clinical characteristics of the samples in the CGGA database but also calculated the Cellchat score of each sample (Figure 4B). We compared the distribution of multiple clinical features and showed that the majority of patients in clusterA were IDH wild-type. At the same time, we divided the patients into high and low subgroups by the median value of the Cellchat score. The distribution between Cellchat groups and GBM clusters was assessed in the CGGA cohort. The samples in Cellchat high score largely overlapped with clusterA and the Cellchat low score overlapped with clusterC, which showed a comparable outcome to TCGA database analysis (Figure 4C). The reliability of our results was verified in the CGGA database, synthetically validating our conclusions.

The clinical utility of our GBM clustering results provided new insights into GBM progression compared to a single clinical feature. For example, we observed that patients with MGMT promoter unmethylated, IDH wild-type, 1p/19q non-codel in clusterC were more likely to achieve a survival advantage at a later stage of GBM progression than patients in clusterA (Figures 5A–C). These findings may suggest that the abnormal methylation profile of promoter CpG islands does not necessarily reflect initial risk factors in GBM progression but is a late result of complex gene-tumor microenvironment interactions throughout GBM progression.

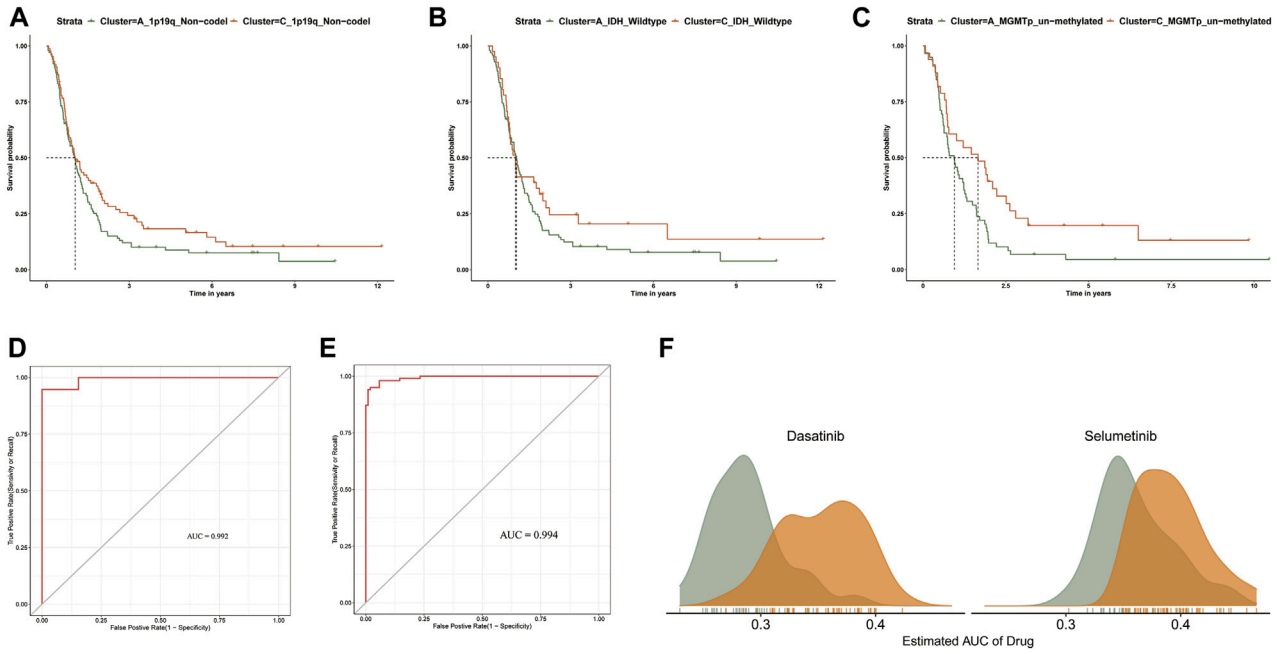


FIGURE 5
(A–C) Kaplan-Meier survival curve analysis between clusters A and C in 1p/19q_Non-codel, MGMTp_un-methylated, and IDH_wildtype). **(D)** The ROC curves of TCGA sample prediction results. **(E)** The ROC curves of CGGA sample prediction results. **(F)** Comparison of estimated Dasatinib and Selumetinib sensitivity.

We developed an accurate performance model which can explore a prompt diagnosis (Pyonteck et al., 2013). Specifically, the PCGI methylation-driven genes differentially expressed in clusters A and C were used to construct the prediction model. We randomly divided the samples of TCGA into a training dataset (70%) and test dataset (30%) and brought the CGGA samples as an independent test dataset to verify the repeatability of the cluster prediction model. Then we fit the LASSO logistic regression with the best lambda value to get a stable set of selected features (Supplementary Figures S3B,C). Lastly, the Area Under the Curve (AUC) area was used to quantify response prediction, which exhibited reasonable prediction accuracy in GBM patients with an AUC of 0.975 in the TCGA database and 0.969 in the CGGA database (Figures 5D,E).

For refining the diagnostic map, we concentrated on predicting drug response between GBM clusters based on the CTRP dataset, which contains the gene expression profiles and drug sensitivity profiles of cancer cell lines (CCLs) (Basu et al., 2013; Yang et al., 2021). We excluded the compounds containing NAs in more than 20% of the samples and excluded the CCLs derived from hematopoietic and lymphoid tissue. After pre-processing the data, we used 658 CCLs containing 266 compounds in CTRP and expression profile data from GBM patients to predict patient response to drugs between clusters A and C, based on pRRpphetic with a built-in ridge regression model (Yang et al., 2021). The difference of estimated AUC values of compounds between two clusters was compared with the Wilcoxon rank-sum test, and the results indicated that patients in clusterA showed significantly lower estimated AUC values of Dasatinib and Selumetinib than clusterC ($p < 0.001$) (Figure 5F and Supplementary Figures S3D,E). Previous studies have shown that the combination of Crizotinib and Dasatinib induced an anti-proliferative effect in GBM cell lines, exerting a potent effect on different GBM cell lines when investigating different tyrosine kinase inhibitors (Nehoff et al., 2015; Wang et al., 2020). Additionally, Selumetinib, a kinase inhibitor affecting actionable kinase targets associated with intracranial tumor growth rate, has been selected for single and combination therapy to develop a miniaturized system for drug testing (Gilbert et al., 2018). The difference in estimated AUC values suggest that patients in clusterA may be more sensitive to these two drugs in clinical treatment (Yang et al., 2021). Overall, we believe that our results provide new insights into improving clinical outcomes for GBM patients and the basis for new treatment options for GBM.

Discussion

Multi-omics data analysis has significantly propelled the understanding of GBM biology, enabling scientists to provide new insights into the GBM precision medicine (Bock et al., 2016).

Although the importance of aberrant DNA methylation is well established in various cancers, comprehensive analyses of genomic and single-cell sequencing data based on tumor typing of CpG island within promoter regions remain deficiency. Collectively, elucidating the complexity of the epigenome in GBM typing and therapeutic response specificity may reveal potential mechanisms of targeted therapy and immunotherapy resistance (Manuyakorn et al., 2010). Hence, we performed a consensus clustering analysis with PCGI methylation-driven genes expression profiles and identified three clusters in patients of TCGA and CGGA database, which helped frame the development of GBM precision diagnosis. The identification of PCGI methylation-driven genes comprehensively reflects the influence of methylation information layer on genes and avoids the noise of miscellaneous methylation probe data. Many PCGI methylation-driven genes have been proven extremely valuable in diverse GBM research. For example, the up-regulation of PDPN by cancer cells has recently been linked to an increased risk for venous thromboembolism in GBM (Tawil et al., 2021). Moreover, Hernando et al. found that forced expression of reprogramming transcription factor SOX2, which is highly expressed in GBM, reduces expression of TET2 and 5hmC, thus contributing to the hyper-methylated phenotype of GSCs (Lopez-Bertoni et al., 2022). In terms of other clinical features and diagnostics, our results can complement existing molecular typing while identifying new clinical differences in the integration process.

Because of the particular proliferation form and development process of the tumor, TME exhibits significant differences compared to the normal tissue environment, leading to exclusive characteristics of the tumor [58]. In this study, clusters A and C we identified differed in the degree of immune infiltration in GBM. Combined with the results of single-cell sequencing, differences in the extent of macrophage infiltration in the TME may account for the significant differences in survival between clusters. Macrophages and microglia are significantly abundant in the GBM microenvironment and provide 10%–34% of the tumor mass, which is supported by previous observations (Jacobs et al., 2012). In studies on GBM typing, macrophages and microglia are more increased in recurrent mesenchymal GBM than in primary non-mesenchymal GBM (Wang et al., 2017). Classifying GBM samples based on the TME has predictive power, so efforts to characterize PCGI methylation-driven genes will prove invaluable for identifying the immunosuppressed patients. Additionally, matrix metalloproteinases (MMPs), a key factor degrading almost all proteins in the extracellular matrix, were found substantially distinct between clusters. MMPs can degrade a variety of proteins in the extracellular matrix, and their increased expression levels are positively correlated with the malignancy of GBM. For example, MMP14 was reported to be up-regulated in some types of cancer and to promote cancer cell invasion (Theodoris et al., 2015).

Single-cell heterogeneity, essential for the precise application of biomarkers and selecting appropriate drugs for clinical use, plays an important role in tumor therapy and diagnostic [63]. The signaling pathways identified by the Cellchat R package help us measure the dynamic interactions between tumor cells and their microenvironment. For instance, multiple studies have shown that macrophages maintain GBM cells and stimulate angiogenesis through the SPP1 pathway, which correlates positively with a higher macrophage density in GBM patients. The maintenance of macrophage infiltration and its immunosuppressive phenotype in GBM requires the SPP1 pathway, which induces a positive feedback loop for macrophage production of SPP1 [17]. Previous studies have shown that ITGA5 was increased in GBM tissues and promoted tumor cell proliferation and invasiveness, which is consistent with our results (Figure 3D). Further experiments revealed that NEAT1 promoted ITGA5 expression through competitive binding with miR-128–3p, which might offer a potential strategy for the treatment of GBM (Chen et al., 2021; Shaim et al., 2021). Although many methodological issues need further discussion, the ligand-receptor genes differently expressed between clusters validate the reasonableness of the typing results from different perspectives, indicating the combination of gene methylation and TME may be a beneficial strategy for GBM patients.

Lastly, the diagnostic map refined the former classification and proposed new points for molecular typing [63]. As new criteria and classification methods provide a more detailed understanding of GBM, relying exclusively on a single molecular marker could not satisfy an accurate diagnosis. The observed GBM sample clusters based on PCGI methylation-driven genes in this study improve homogeneous tumor diagnosis and provide insights into the prognosis of GBM patients at later stages of progression (Brennan et al., 2013; Geisenberger et al., 2015). Strikingly, the Cellchat score we defined distinguished the GBM subtypes with clear separation in the CGGA and TCGA databases. This comprehensive DNA methylation- and tumor microenvironment-based classification of biomarker arrays improves molecular understanding of pathway signaling among GBM cell clusters. Here, our results also show that sample classification of GBM can further stratify patient response to different drugs, which could ultimately compensate for personalized therapies in groups of GBM patients.

In conclusion, the results of our analysis adequately discuss the heterogeneous profile of promoter CpG island methylation in GBM. The GBM typing constructed by integrating PCGI methylation-driven genes and the GBM tumor microenvironment in our study contributes to improving the understanding of homogeneous intra-tumor diagnostics.

Data availability statement

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

Author contributions

DL, RW, LZ and SW conceived and designed the study. RW and ZL wrote the manuscript. RW, LZ, SW, XZ, CL and PW made substantial contributions to the acquisition of data. RW and LZ analyzed and interpreted the data. All authors reviewed and revised the manuscript. DL edited and supervised this study. All authors read and approved the final manuscript.

Funding

This work was funded by the Beijing-Tianjin-Hebei Cooperation Special Project (grant numbers: J200002).

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.989985/full#supplementary-material

References

Agundez, M., Grau, L., Palou, J., Algaba, F., Villavicencio, H., and Sanchez-Carbayo, M. (2011). Evaluation of the methylation status of tumour suppressor genes for predicting bacillus Calmette-Guerin response in patients with T1G3 high-risk bladder tumours. *Eur. Urol.* 60 (1), 131–140. doi:10.1016/j.eururo.2011.04.020

Bagaev, A., Kotlov, N., Nomie, K., Svekolkin, V., Gafurov, A., Isaeva, O., et al. (2021). Conserved pan-cancer microenvironment subtypes predict response to immunotherapy. *Cancer Cell.* 39 (6), 845–865.e7. doi:10.1016/j.ccell.2021.04.014

Bastola, S., Pavlyukov, M. S., Yamashita, D., Ghosh, S., Cho, H., Kagaya, N., et al. (2020). Glioma-initiating cells at tumor edge gain signals from tumor core cells to promote their malignancy. *Nat. Commun.* 11 (1), 4660. doi:10.1038/s41467-020-18189-y

Basu, A., Bodycombe, N. E., Cheah, J. H., Price, E. V., Liu, K., Schaefer, G. I., et al. (2013). An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell.* 154 (5), 1151–1161. doi:10.1016/j.cell.2013.08.003

Bock, C., Farlik, M., and Sheffield, N. C. (2016). Multi-omics of single cells: Strategies and applications. *Trends Biotechnol.* 34 (8), 605–608. doi:10.1016/j.tibtech.2016.04.004

Brennan, C. W., Verhaak, R. G., McKenna, A., Campos, B., Noushmehr, H., Salama, S. R., et al. (2013). The somatic genomic landscape of glioblastoma. *Cell.* 155 (2), 462–477. doi:10.1016/j.cell.2013.09.034

Carro, M. S., Lim, W. K., Alvarez, M. J., Bollo, R. J., Zhao, X., Snyder, E. Y., et al. (2010). The transcriptional network for mesenchymal transformation of brain tumours. *Nature* 463 (7279), 318–325. doi:10.1038/nature08712

Chen, J., Wang, H., Wang, J., Niu, W., Deng, C., and Zhou, M. (2021). LncRNA NEAT1 enhances glioma progression via regulating the miR-128-3p/ITGA5 Axis. *Mol. Neurobiol.* 58 (10), 5163–5177. doi:10.1007/s12035-021-02474-y

Chen, P., Zhao, D., Li, J., Liang, X., Li, J., Chang, A., et al. (2019). Symbiotic macrophage-glioma cell interactions reveal synthetic lethality in PTEN-null glioma. *Cancer Cell.* 35 (6), 868–884. doi:10.1016/j.ccell.2019.05.003

Datta, I., Noushmehr, H., Brodie, C., and Poisson, L. M. (2021). Expression and regulatory roles of lncRNAs in G-CIMP-low vs G-CIMP-high glioma: An *in-silico* analysis. *J. Transl. Med.* 19 (1), 182. doi:10.1186/s12967-021-02844-z

Di Risi, T., Vinciguerra, R., Cuomo, M., Della Monica, R., Riccio, E., Cocozza, S., et al. (2021). DNA methylation impact on Fabry disease. *Clin. Epigenetics* 13 (1), 24. doi:10.1186/s13148-021-01019-3

Etcheverry, A., Aubry, M., de Tayrac, M., Vauleon, E., Boniface, R., Guenot, F., et al. (2010). DNA methylation in glioblastoma: Impact on gene expression and clinical outcome. *BMC Genomics* 11, 701. doi:10.1186/1471-2164-11-701

Fan, X., Fu, Y., Zhou, X., Sun, L., Yang, M., Wang, M., et al. (2020). Single-cell transcriptome analysis reveals cell lineage specification in temporal-spatial patterns in human cortical development. *Sci. Adv.* 6 (34), eaaz2978. doi:10.1126/sciadv.aaz2978

Galstyan, A., Markman, J. L., Shatalova, E. S., Chiechi, A., Korman, A. J., Patil, R., et al. (2019). Blood-brain barrier permeable nano immunoconjugates induce local immune responses for glioma therapy. *Nat. Commun.* 10 (1), 3850. doi:10.1038/s41467-019-11719-3

Gangoso, E., Southgate, B., Bradley, L., Rus, S., Galvez-Cancino, F., McGivern, N., et al. (2021). Glioblastomas acquire myeloid-affiliated transcriptional programs via epigenetic immunoediting to elicit immune evasion. *Cell.* 184 (9), 2454–2470.e26. e26. doi:10.1016/j.cell.2021.03.023

Geisenberger, C., Mock, A., Warta, R., Rapp, C., Schwager, C., Korshunov, A., et al. (2015). Molecular profiling of long-term survivors identifies a subgroup of glioblastoma characterized by chromosome 19/20 co-gain. *Acta Neuropathol.* 130 (3), 419–434. doi:10.1007/s00401-015-1427-y

Giacopelli, B., Zhao, Q., Ruppert, A. S., Agyeman, A., Weigel, C., Wu, Y. Z., et al. (2019). Developmental subtypes assessed by DNA methylation-iPLEX forecast the natural history of chronic lymphocytic leukemia. *Blood* 134 (8), 688–698. doi:10.1182/blood.2019000490

Gilbert, A. N., Anderson, J. C., Duarte, C. W., Shevin, R. S., Langford, C. P., Singh, R., et al. (2018). Combinatorial drug testing in 3D microtumors derived from GBM patient-derived xenografts reveals cytotoxic synergy in pharmacokinomics-informed pathway interactions. *Sci. Rep.* 8 (1), 8412. doi:10.1038/s41598-018-26840-4

Gong, P. J., Shao, Y. C., Huang, S. R., Zeng, Y. F., Yuan, X. N., Xu, J. J., et al. (2020). Hypoxia-associated prognostic markers and competing endogenous RNA Co-expression networks in breast cancer. *Front. Oncol.* 10, 579868. doi:10.3389/fonc.2020.579868

Guo, F., Yan, L., Guo, H., Li, L., Hu, B., Zhao, Y., et al. (2015). The transcriptome and DNA methylome landscapes of human primordial germ cells. *Cell.* 161 (6), 1437–1452. doi:10.1016/j.cell.2015.05.015

Han, M., Wang, S., Fritah, S., Wang, X., Zhou, W., Yang, N., et al. (2020). Interfering with long non-coding RNA MIR22HG processing inhibits glioblastoma

progression through suppression of Wnt/β-catenin signalling. *Brain* 143 (2), 512–530. doi:10.1093/brain/awz406

Hardy, T., Zeybel, M., Day, C. P., Dipper, C., Masson, S., McPherson, S., et al. (2017). Plasma DNA methylation: A potential biomarker for stratification of liver fibrosis in non-alcoholic fatty liver disease. *Gut* 66 (7), 1321–1328. doi:10.1136/gutjnl-2016-311526

Hollstein, P. E., Eichner, L. J., Brun, S. N., Kamireddy, A., Svensson, R. U., Vera, L. I., et al. (2019). The AMPK-related kinases SIK1 and SIK3 mediate key tumor-suppressive effects of LKB1 in NSCLC. *Cancer Discov.* 9 (11), 1606–1627. doi:10.1158/2159-8290.CD-18-1261

Hua, W., Li, C., Yang, Z., Li, L., Jiang, Y., Yu, G., et al. (2015). Suppression of glioblastoma by targeting the overactivated protein neddylation pathway. *Neuro. Oncol.* 17 (10), 1333–1343. doi:10.1093/neuonc/nov066

Huang, J., Luo, F., Shi, M., Luo, J., Ma, C., Li, S., et al. (2021). Construction and validation of a metabolic gene-associated prognostic model for cervical carcinoma and the role on tumor microenvironment and immunity. *Aging (Albany NY)* 13 (23), 25072–25088. doi:10.18632/aging.203723

Jacobs, V. L., Landry, R. P., Liu, Y., Romero-Sandoval, E. A., and De Leo, J. A. (2012). Propentofylline decreases tumor growth in a rodent model of glioblastoma multiforme by a direct mechanism on microglia. *Neuro. Oncol.* 14 (2), 119–131. doi:10.1093/neuonc/nor194

Jin, S., Guerrero-Juarez, C. F., Zhang, L., Chang, I., Ramos, R., Kuan, C. H., et al. (2021). Inference and analysis of cell-cell communication using CellChat. *Nat. Commun.* 12 (1), 1088. doi:10.1038/s41467-021-21246-9

Klughammer, J., Kiesel, B., Roetzer, T., Fortelny, N., Nemc, A., Nenning, K. H., et al. (2018). The DNA methylation landscape of glioblastoma disease progression shows extensive heterogeneity in time and space. *Nat. Med.* 24 (10), 1611–1624. doi:10.1038/s41591-018-0156-x

Kosti, A., de Araujo, P. R., Li, W. Q., Guardia, G. D. A., Chiou, J., Yi, C., et al. (2020). The RNA-binding protein SERBP1 functions as a novel oncogenic factor in glioblastoma by bridging cancer metabolism and epigenetic regulation. *Genome Biol.* 21 (1), 195. doi:10.1186/s13059-020-02115-y

Krug, K., Jaehnig, E. J., Satpathy, S., Blumenberg, L., Karpova, A., Anurag, M., et al. (2020). Proteogenomic landscape of breast cancer tumorigenesis and targeted therapy. *Cell.* 183 (5), 1436–1456. doi:10.1016/j.cell.2020.10.036

Lauria, A., Peirone, S., Giudice, M. D., Priante, F., Rajan, P., Caselle, M., et al. (2020). Identification of altered biological processes in heterogeneous RNA-sequencing data by discretization of expression profiles. *Nucleic Acids Res.* 48 (4), 1730–1747. doi:10.1093/nar/gkz1208

Leimkuhler, N. B., Gleitz, H. F. E., Ronghui, L., Snoeren, I. A. M., Fuchs, S. N. R., Nagai, J. S., et al. (2021). Heterogeneous bone-marrow stromal progenitors drive myelofibrosis via a druggable alarmin axis. *Cell. Stem Cell.* 28 (4), 637–652.e8. doi:10.1016/j.stem.2020.11.004

Lennon, N. J., Adalsteinsson, V. A., and Gabriel, S. B. (2016). Technological considerations for genome-guided diagnosis and management of cancer. *Genome Med.* 8 (1), 112. doi:10.1186/s13073-016-0370-4

Li, R., Yin, Y. H., Jin, J., Liu, X., Zhang, M. Y., Yang, Y. E., et al. (2020). Integrative analysis of DNA methylation-driven genes for the prognosis of lung squamous cell carcinoma using MethylMix. *Int. J. Med. Sci.* 17 (6), 773–786. doi:10.7150/ijms.43272

Liu, S., Wang, Z., Zhu, R., Wang, F., Cheng, Y., and Liu, Y. (2021). Three differential expression analysis methods for RNA sequencing: Limma, EdgeR, DESeq2. *J. Vis. Exp.* 2021 (175). doi:10.3791/62528

Liu, Y., Li, L., Li, Y., and Zhao, X. (2020). Research progress on tumor-associated macrophages and inflammation in cervical cancer. *Biomed. Res. Int.* 2020, 6842963. doi:10.1155/2020/6842963

Lopez-Bertoni, H., Johnson, A., Rui, Y., Lal, B., Sall, S., Malloy, M., et al. (2022). Sox2 induces glioblastoma cell stemness and tumor propagation by repressing TET2 and deregulating 5hmC and 5mC DNA modifications. *Signal Transduct. Target. Ther.* 7 (1), 37. doi:10.1038/s41392-021-00857-0

Lu, I. N., Dobersalske, C., Rauschenbach, L., Teuber-Hanselmann, S., Steinbach, A., Ullrich, V., et al. (2021). Tumor-associated hematopoietic stem and progenitor cells positively linked to glioblastoma progression. *Nat. Commun.* 12 (1), 3895. doi:10.1038/s41467-021-23995-z

Ma, S., Song, W., Xu, Y., Si, X., Zhang, D., Lv, S., et al. (2020). Neutralizing tumor-promoting inflammation with polypeptide-dexamethasone conjugate for microenvironment modulation and colorectal cancer therapy. *Biomaterials* 232, 119676. doi:10.1016/j.biomaterials.2019.119676

Manuyakorn, A., Paulus, R., Farrell, J., Dawson, N. A., Tze, S., Cheung-Lau, G., et al. (2010). Cellular histone modification patterns predict prognosis and treatment response in resectable pancreatic adenocarcinoma: Results from RTOG 9704. *J. Clin. Oncol.* 28 (8), 1358–1365. doi:10.1200/JCO.2009.24.5639

Nehoff, H., Parayath, N. N., McConnell, M. J., Taurin, S., and Greish, K. (2015). A combination of tyrosine kinase inhibitors, crizotinib and dasatinib for the treatment of glioblastoma multiforme. *Oncotarget* 6 (35), 37948–37964. doi:10.18632/oncotarget.5698

Northcott, P. A., Buchhalter, I., Morrissy, A. S., Hovestadt, V., Weischenfeldt, J., Ehrenberger, T., et al. (2017). The whole-genome landscape of medulloblastoma subtypes. *Nature* 547 (7663), 311–317. doi:10.1038/nature22973

Ogino, S., Nowak, J. A., Hamada, T., Phipps, A. I., Peters, U., Milner, D. A., Jr., et al. (2018). Integrative analysis of exogenous, endogenous, tumour and immune factors for precision medicine. *Gut* 67 (6), 1168–1180. doi:10.1136/gutjnl-2017-315537

Oh, S., Yeom, J., Cho, H. J., Kim, J. H., Yoon, S. J., Kim, H., et al. (2020). Integrated pharmaco-proteogenomics defines two subgroups in isocitrate dehydrogenase wild-type glioblastoma with prognostic and therapeutic opportunities. *Nat. Commun.* 11 (1), 3288. doi:10.1038/s41467-020-17139-y

Park, A. K., Kim, P., Ballester, L. Y., Esquenazi, Y., and Zhao, Z. (2019). Subtype-specific signaling pathways and genomic aberrations associated with prognosis of glioblastoma. *Neuro. Oncol.* 21 (1), 59–70. doi:10.1093/neuonc/noy120

Pyonteck, S. M., Akkari, L., Schuhmacher, A. J., Bowman, R. L., Sevenich, L., Quail, D. F., et al. (2013). CSF-1R inhibition alters macrophage polarization and blocks glioma progression. *Nat. Med.* 19 (10), 1264–1272. doi:10.1038/nm.3337

Riaz, N., Havel, J. J., Makarov, V., Desrichard, A., Urba, W. J., Sims, J. S., et al. (2017). Tumor and microenvironment evolution during immunotherapy with nivolumab. *Cell.* 171 (4), 934–949. doi:10.1016/j.cell.2017.09.028

Roy, R., Winteringham, L. N., Lassmann, T., and Forrest, A. R. R. (2019). Expression levels of therapeutic targets as indicators of sensitivity to targeted therapeutics. *Mol. Cancer Ther.* 18 (12), 2480–2489. doi:10.1158/1535-7163.MCT-19-0273

Ryan, K. J., White, C. C., Patel, K., Xu, J., Olah, M., Replogle, J. M., et al. (2017). A human microglia-like cellular model for assessing the effects of neurodegenerative disease gene variants. *Sci. Transl. Med.* 9 (421), eaai7635. doi:10.1126/scitranslmed.aai7635

Shaim, H., Shanley, M., Basar, R., Daher, M., Gumin, J., Zamler, D. B., et al. (2021). Targeting the αv integrin/TGF-β axis improves natural killer cell function against glioblastoma stem cells. *J. Clin. Invest.* 131 (14), 142116. doi:10.1172/JCI142116

Stewart, E., Federico, S. M., Chen, X., Shelat, A. A., Bradley, C., Gordon, B., et al. (2017). Orthotopic patient-derived xenografts of paediatric solid tumours. *Nature* 549 (7670), 96–100. doi:10.1038/nature23647

Tao, W., Zhang, A., Zhai, K., Huang, Z., Huang, H., Zhou, W., et al. (2020). SATB2 drives glioblastoma growth by recruiting CBP to promote FOXM1 expression in glioma stem cells. *EMBO Mol. Med.* 12 (12), e12291. doi:10.15252/emmm.202012291

Tawil, N., Bassawon, R., Meehan, B., Nehme, A., Montermini, L., Gayden, T., et al. (2021). Glioblastoma cell populations with distinct oncogenic programs release podoplanin as procoagulant extracellular vesicles. *Blood Adv.* 5 (6), 1682–1694. doi:10.1182/bloodadvances.2020002998

Theodoris, C. V., Li, M., White, M. P., Liu, L., He, D., Pollard, K. S., et al. (2015). Human disease modeling reveals integrated transcriptional and epigenetic mechanisms of NOTCH1 haploinsufficiency. *Cell.* 160 (6), 1072–1086. doi:10.1016/j.cell.2015.02.035

Tian, J., Cai, Y., Li, Y., Lu, Z., Huang, J., Deng, Y., et al. (2021). CancerImmunityQTL: A database to systematically evaluate the impact of genetic variants on immune infiltration in human cancer. *Nucleic Acids Res.* 49 (D1), D1065–D1073. doi:10.1093/nar/gkaa805

Vitucci, M., Irvin, D. M., McNeill, R. S., Schmid, R. S., Simon, J. M., Dhruv, H. D., et al. (2017). Genomic profiles of low-grade murine gliomas evolve during progression to glioblastoma. *Neuro. Oncol.* 19 (9), 1237–1247. doi:10.1093/neuonc/nox050

Wang, Q., Hu, B., Hu, X., Kim, H., Squatrito, M., Scarpace, L., et al. (2017). Tumor evolution of glioma-intrinsic gene expression subtypes associates with immunological changes in the microenvironment. *Cancer Cell.* 32 (1), 42–56. doi:10.1016/j.ccell.2017.06.003

Wang, S., Wang, R., Gao, F., Huang, J., Zhao, X., and Li, D. (2022). Pan-cancer analysis of the DNA methylation patterns of long non-coding RNA. *Genomics* 114 (4), 110377. doi:10.1016/j.ygeno.2022.110377

Wang, Z., Sun, D., Chen, Y. J., Xie, X., Shi, Y., Tabar, V., et al. (2020). Cell lineage-based stratification for glioblastoma. *Cancer Cell.* 38 (3), 366–379. doi:10.1016/j.ccell.2020.06.003

Wilkerson, M. D., and Hayes, D. N. (2010). ConsensusClusterPlus: A class discovery tool with confidence assessments and item tracking. *Bioinformatics* 26 (12), 1572–1573. doi:10.1093/bioinformatics/btq170

Xie, H., Yuan, C., Li, J. J., Li, Z. Y., and Lu, W. C. (2021). Potential molecular mechanism of TNF superfamily-related genes in glioblastoma multiforme based on transcriptome and epigenome. *Front. Neurol.* 12, 576382. doi:10.3389/fneur.2021.576382

Xie, Y., He, L., Lugano, R., Zhang, Y., Cao, H., He, Q., et al. (2021). Key molecular alterations in endothelial cells in human glioblastoma uncovered through single-cell RNA sequencing. *JCI Insight* 6 (15), 150861. doi:10.1172/jci.insight.150861

Xu, N., Wu, Y. P., Ke, Z. B., Liang, Y. C., Cai, H., Su, W. T., et al. (2019). Identification of key DNA methylation-driven genes in prostate adenocarcinoma: An integrative analysis of TCGA methylation data. *J. Transl. Med.* 17 (1), 311. doi:10.1186/s12967-019-2065-2

Yang, C., Huang, X., Li, Y., Chen, J., Lv, Y., and Dai, S. (2021). Prognosis and personalized treatment prediction in TP53-mutant hepatocellular carcinoma: An *in silico* strategy towards precision oncology. *Brief. Bioinform.* 22 (3), bbaa164. doi:10.1093/bib/bbaa164

Yao, M., Ventura, P. B., Jiang, Y., Rodriguez, F. J., Wang, L., Perry, J. S. A., et al. (2020). Astrocytic trans-differentiation completes a multicellular paracrine feedback loop required for medulloblastoma tumor growth. *Cell.* 180 (3), 502–520. doi:10.1016/j.cell.2019.12.024

Yi, G. Z., Huang, G., Guo, M., Zhang, X., Wang, H., Deng, S., et al. (2019). Acquired temozolomide resistance in MGMT-deficient glioblastoma cells is associated with regulation of DNA repair by DHC2. *Brain* 142 (8), 2352–2366. doi:10.1093/brain/awz202

Yoshihara, K., Shahmoradgoli, M., Martinez, E., Vegesna, R., Kim, H., Torres-Garcia, W., et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* 4, 2612. doi:10.1038/ncomms3612

Zhao, X., Ji, J., Wang, S., Wang, R., Yu, Q., and Li, D. (2021). The regulatory pattern of target gene expression by aberrant enhancer methylation in glioblastoma. *BMC Bioinforma.* 22 (1), 420. doi:10.1186/s12859-021-04345-8

Zhao, Z., Zhang, K. N., Wang, Q., Li, G., Zeng, F., Zhang, Y., et al. (2021). Chinese glioma genome Atlas (CGGA): A comprehensive resource with functional genomic data from Chinese glioma patients. *Genomics Proteomics Bioinforma.* 19 (1), 1–12. doi:10.1016/j.gpb.2020.10.005

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 21 September 2022
DOI 10.3389/fgene.2022.984297

OPEN ACCESS

EDITED BY
Geng Chen,
Stemirna Therapeutics Co., Ltd., China

REVIEWED BY
Yutian Zou,
Sun Yat-sen University Cancer Center (SYSUCC), China
Zheng Chen,
Fudan University, China

*CORRESPONDENCE
Ting Yue,
dr_tingyue@163.com

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

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

RECEIVED 01 July 2022
ACCEPTED 01 September 2022
PUBLISHED 21 September 2022

CITATION
Li J, Wu Z, Wang S, Li C, Zhuang X, He Y, Xu J, Su M, Wang Y, Ma W, Fan D and Yue T (2022), A necroptosis-related prognostic model for predicting prognosis, immune landscape, and drug sensitivity in hepatocellular carcinoma based on single-cell sequencing analysis and weighted co-expression network.
Front. Genet. 13:984297.
doi: 10.3389/fgene.2022.984297

A necroptosis-related prognostic model for predicting prognosis, immune landscape, and drug sensitivity in hepatocellular carcinoma based on single-cell sequencing analysis and weighted co-expression network

Jingjing Li[1,2†], Zhi Wu[3†], Shuchen Wang[2], Chan Li[4], Xuhui Zhuang[2], Yuewen He[2], Jianmei Xu[4], Meiyi Su[5], Yong Wang[2], Wuhua Ma[2], Dehui Fan[4,5] and Ting Yue[6*]

[1]Department of Anesthesiology, Jincheng People's Hospital, Jincheng, Shanxi, China, [2]Department of Anesthesiology, The First Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou, China, [3]Department of General Surgery, Jincheng People's Hospital, Jincheng, Shanxi, China, [4]The Fifth Clinical Medical School, Guangzhou University of Chinese Medicine, Guangzhou, China, [5]Department of Rehabilitation, GuangDong Second Traditional Chinese Medicine Hospital, Guangzhou, China, [6]Department of Oncology Rehabilitation, Jincheng People's Hospital, Jincheng, Shanxi, China

Background: Hepatocellular carcinoma (HCC) is a highly lethal cancer and is the second leading cause of cancer-related deaths worldwide. Unlike apoptosis, necroptosis (NCPS) triggers an immune response by releasing damage-related molecular factors. However, the clinical prognostic features of necroptosis-associated genes in HCC are still not fully explored.

Methods: We analyzed the single-cell datasets GSE125449 and GSE151530 from the GEO database and performed weighted co-expression network analysis on the TCGA data to identify the necroptosis genes. A prognostic model was built using COX and Lasso regression. In addition, we performed an analysis of survival, immunity microenvironment, and mutation. Furthermore, the hub genes and pathways associated with HCC were localized within the single-cell atlas.

Results: Patients with HCC in the TCGA and ICGC cohorts were classified using a necroptosis-related model with significant differences in survival times between high- and low-NCPS groups ($p < 0.05$). High-NCPS patients expressed more immune checkpoint-related genes, suggesting immunotherapy and some chemotherapies might prove beneficial to them. In addition, a single-cell sequencing approach was conducted to investigate the expression of hub genes and associated signaling pathways in different cell types.

Conclusion: Through the analysis of single-cell and bulk multi-omics sequencing data, we constructed a prognostic model related to necroptosis and explored the relationship between high- and low-NCPS groups and immune cell infiltration in HCC. This provides a new reference for further understanding the role of necroptosis in HCC.

KEYWORDS

prognostic model, hepatocellular carcinoma, necroptosis, therapy, nomogram

Introduction

Primary liver cancer is the sixth most common cancer in the world and the second leading cause of cancer-related death (Yang et al., 2019). Hepatocellular carcinoma (HCC) is the most common type of primary liver cancer (Chaudhary et al., 2019). Most HCC patients are diagnosed at an advanced stage. The gold standard treatments, including tumor resection, local ablation with radiofrequency, and sometimes liver transplantation, have low success rates with high relapse rates and short survival times (Dhanasekaran et al., 2016). Additionally, patients with HCC who present with similar tumor, lymph node, and metastasis (TNM) stage have different clinical outcomes, and there are few current effective prognostic indicators.

Recent research has demonstrated the importance of the tumor microenvironment (TME) in promoting tumor aggressiveness (Altorki et al., 2019). The survival of patients with various malignancies can be prolonged by immune checkpoint inhibitors. However, many patients with HCC currently often respond poorly to immune checkpoint inhibitors, which may be due to low mutational loads, acquiring new immune checkpoints, and producing immunosuppressive factors (Riley et al., 2019). Therefore, there is a need to identify new biomarkers for HCC as well as to comprehend their significance in TME.

Programmed cell death has a strong impact on the characterization of the TME ecosystem (Chevrier et al., 2017). Resistance to apoptosis, a problem affecting cancer development, is one of the hallmarks of cancer (Reyna et al., 2017). In the process of cancer cell resistance to death, growth signals are overactivated, the metabolism is reprogrammed, and a change in the immune microenvironment occurs (Sahin et al., 2017). Inducing cancer cell death is becoming increasingly popular as a potential cancer treatment method (Bersuker et al., 2019). HCC cells can die by several different mechanisms, including apoptosis and necroptosis (NCPS) (Yuan et al., 2019). Both mechanisms play a significant role in homeostasis, inflammation, anti-infection, and tumorigenesis (Karki et al., 2021; Koren and Fuchs, 2021).

Necroptosis was once believed to be the "accidental death" of cells. However, current research indicates that necroptosis is distinct from conventional apoptosis (González-Juarbe et al., 2017). Necroptosis leads to membrane destabilization, which subsequently precedes swelling and lysis of cells, resulting in the release of intracellular constituents (González-Juarbe et al., 2017). Inhibited caspase 8 and receptor-interacting serine/threonine protein kinase 1 (RIPK1) are both involved in necroptosis pathway activation via recruitment and activation of receptor-interacting serine/threonine protein kinase 3 (RIPK3) (Alvarez-Diaz et al., 2016). Necroptosis occurs when caspase 8 is inactivated or absent, resulting in the activation and autophosphorylation of RIPK1 and RIPK3(Tanzer et al., 2017). During this process, the cell membrane ruptures, and the contents are released, stimulating an immune response (Kalliolias and Ivashkiv, 2016). Necroptosis becomes attractive as an alternative to apoptosis for killing tumor cells if apoptosis fails to kill them (Kalliolias and Ivashkiv, 2016). As well, the immune microenvironment is positively impacted by necroptosis (Gong et al., 2019).

Interestingly, the role of necroptosis in cancer is complex. In general, high levels of necroptosis result in strong adaptive immune responses that inhibit the progression of tumors. The recruitment of strong immune responses may also contribute to tumor progression (Koo et al., 2015; Najafov et al., 2017). Moreover, the inflammatory response may contribute to tumorigenesis and metastasis, as well as generate an immunosuppressive tumor microenvironment. Guo et al. (2022) has shown that loss of key necroptosis gene significantly reduces clinical symptoms of liver injury and fibrosis. Necroptosis has completely opposite effects on different types of cancer, the mechanism of which is still unclear. With the emergence of immune checkpoint therapy, changes in the immune microenvironment resulting from necroptosis are also important to consider. There is therefore a need to investigate the relationship between necroptosis and HCC.

Here, we downloaded the data of HCC patients from the TCGA and ICGC databases, as well as two single-cell datasets, GSE125449 (Ma et al., 2019) and GSE151530 (Ma et al., 2021), and one microarray dataset, GSE76427 (Grinchuk et al., 2018) from the GEO database. The TCGA cohort was used for model building. The ICGC cohort and GSE76427 were used to validate the results of our analysis. Two single-cell sequencing datasets, GSE125449 and GSE151530, were chosen for single-cell analysis because of their relatively large sample size and inclusion of clinical data. Through

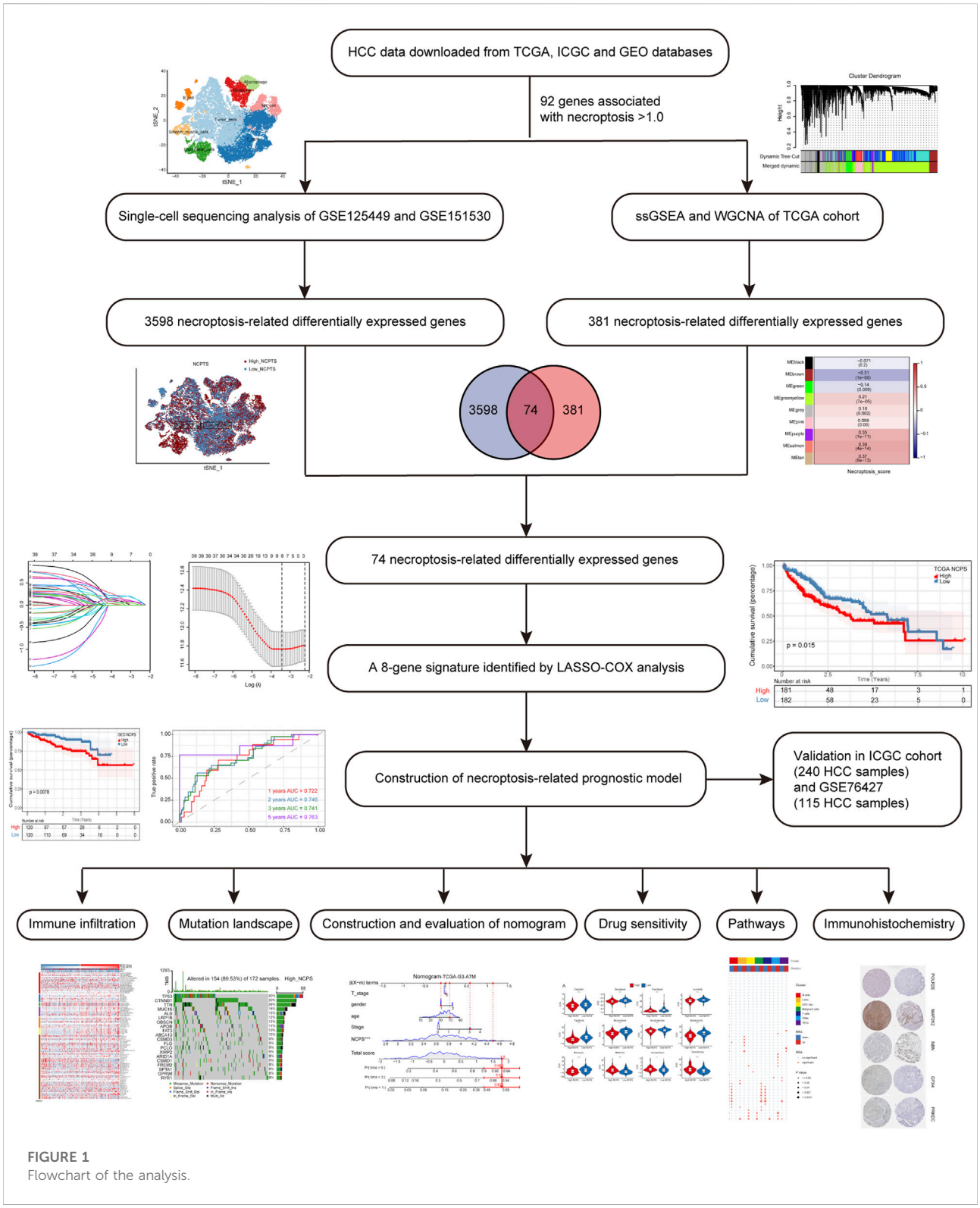


FIGURE 1
Flowchart of the analysis.

comprehensive bioinformatic analysis, we developed a prognostic model based on necroptosis and classified HCC patients into low- and high-risk groups, the results of which were significantly different. Furthermore, we explored the potential value of the signature in guiding the tumor mutational load, immune microenvironment, and drug sensitivity.

Methods

Download and processing of transcriptome data

This flowchart illustrates the key steps in the analysis (Figure 1). The data of HCC were downloaded from TCGA (https://portal.gdc.cancer.gov/) as a training cohort (Grossman et al., 2016). Count data and TPM data of HCC were extracted using R software (4.2.0), and a total of 363 tumor samples with complete clinical data were obtained. The HCC dataset was downloaded through ICGC (https://dcc.icgc.org/) database as a validation cohort, and the count data type and TPM data type of HCC were extracted, and a total of 240 tumor samples were obtained with complete clinical information (Zhang et al., 2019). GSE76427, measured using the Illumina HumanHT-12 V4. 0 expression beadchip, contained 115 HCC samples (Grinchuk et al., 2018). The raw CEL files for GSE76427 were downloaded from the GEO database. More details of the data processing are in Supplementary Material S1, S2.

Download and processing of single-cell data

The single-cell datasets GSE125449 and GSE151530 for HCC were downloaded from the GEO database (https://www.ncbi.nlm.nih.gov/geo/) (Barrett et al., 2013). The GSE125449 dataset contains nine HCC samples and the GSE151530 dataset contains 32 HCC samples. We performed quality control on the data of all samples. We retained cells with genes expressed in at least 10 cells, less than 10% of mitochondrial genes, more than 200 genes, less than 5% hemoglobin genes, less than 50% ribosomal genes, and expression between 200 and 7000. We set a limit of 3000 highly variable genes. Next, we normalized all samples, removed batch effects, and integrated them by SCT. Then, using the tSNE method with the "DIMS" parameter set to 20, the dimensionality of the data was reduced. Cell clustering was then carried out using the "KNN" method with a resolution of 2.0. Subsequently, the cells were annotated with the Human Primary Cell Atlas (HPCA) from the "SingleR" package as a reference dataset (Mabbott et al., 2013). Finally, the proportion of NCPS-related genes in each cell can be calculated using the "PercateFeatureSet" function.

Identification of NCPS-related genes

In the GeneCards database (https://www.genecards.org/), 614 genes associated with necroptosis were identified (Safran et al., 2021). A total of 92 genes were identified that had an association score of greater than 1.0 with necroptosis (Supplementary Material S3). Then, the NCPS-related genes were scored for each sample by the combined analysis of ssGSEA (Single Sample Gene Set Enrichment Analysis) and WGCNA (Weighted Co-Expression Network Analysis). The log2 processed data were used for ssGSEA analysis.

ssGSEA

Gene sets enriched in a sample are often quantified by using the ssGSEA method with "GSVA" package (version: 1.44.2) (Bindea et al., 2013; Hänzelmann et al., 2013). In this study, ssGSEA analysis was utilized to determine the NCPS-related scores of each patient with HCC.

WGCNA

WGCNA analysis is one method used in systems biology for determining patterns of genetic association among diverse samples (Langfelder and Horvath, 2008). In addition to identifying highly covariant genomes, WGCNA analysis can be used to identify potential biomarkers or therapeutic targets based on the correlation between genomes and phenotypes. In this study, gene modules associated with NCPS scores in HCC were found by "WGCNA" package (version: 1.71), and genes associated with necroptosis were obtained. Non-gray modules were identified by setting a soft threshold of eight, a minimum number of module genes of 80, and combining modules that had similarities of less than 0.3.

Construction of NCPS-related prognostic model

First, univariate COX analysis was used to identify NCPS-related genes with prognostic values by using the "survival" package (version: 3.3-1). Next, a prognostic model was developed based on the least absolute shrinkage and selection operator (LASSO) regression for NCPS-related genes by using the "glmnet" package (version: 4.1-4) (Lossos et al., 2004; Friedman et al., 2010). In this way, the NCPS score could be calculated for each HCC sample by the formula. Gene expression levels were weighted by their respective coefficients of LASSO regression to calculate the NCPS score. The formula was as follows:

$$NCPS\,score = \sum_{i=1}^{n} Coef_i \times Exp_i \quad (1)$$

where n, Exp_i, $Coef_i$, represented the number, the expression value, and the coefficient of each selected gene, respectively. According to the median value of the TCGA-HCC cohort, patients could be classified into low- and high-risk groups. Thereafter, we assessed the accuracy of the model by comparing prognostic differences between the two groups.

Validation of NCPS-related prognostic model

The ICGC cohort and GSE76427 were selected as the external validation cohorts. According to the formula of the prognostic model, NCPS scores for each sample were calculated, and patients were categorized based on their median NCPS scores into high-risk and low-risk groups. We then conducted a survival analysis comparing the high- and low-NCPS groups. Receiver operating characteristic (ROC) curves were utilized to evaluate the model's accuracy by using the "timeROC" package (version: 0.4) (Li et al., 2018). To determine whether the model grouped HCC patients more effectively, principal component analysis (PCA) was performed using the "PCAtools" package (version: 2.8.0) and "scatterplot3d" package (version: 0.3-41).

Immune infiltration and mutation landscape

We performed immune infiltration analysis of HCC patients in the TCGA database using immune cell infiltration algorithms from the IOBR package (version: 0.99.9) (Zeng et al., 2021). Next, we examined the differences in the levels of immune cell infiltration between the two NCPS groups and presented the immune cells with different levels of infiltration as a heat map. Also, the expression of immune checkpoint-related genes in the various NCPS subgroups was visualized by a boxplot. We identified the top 20 genes with the highest mutation rates by comparing the mutation rates between groups with high and low NCPS scores.

Nomogram

Using clinical data and NCPS values, a nomogram was developed in this study to assess the probability of mortality in patients with HCC using the "rms" package (version: 6.3-0) and "regplot" package (version: 1.1). This nomogram was evaluated by using prognostic ROC curves and decision curve analysis (DCA) to determine its accuracy in predicting patient outcomes. The DCA analysis was performed using the "ggDCA" package (version: 1.1) (Vickers and Elkin, 2006).

Drug sensitivity, immunohistochemistry, pathways

To improve personalized treatment, we calculated half maximal inhibitory concentrations (IC50) using the "pRRophetic" package (version: 0.5) and compared these data between high-risk and low-risk groups (Geeleher et al., 2014). Low IC_{50} values indicate greater drug effectiveness. The Human Protein Atlas (HPA) database (version: 21.1, http://www.proteinatlas.org/) is the most comprehensive database for assessing protein distribution in human tissues (Uhlén et al., 2015).

HPA database was used to obtain prognostic gene expression data. Immunohistochemical staining images of normal and HCC tissues were used to analyze the protein expression of genes. In addition, we performed an enrichment analysis of pathways associated with different cell types in the single-cell data and then mapped the significantly different pathways to tSNE plots for visualization. Pathway enrichment analysis was performed using the "irGSEA" package (version: 1.1.2). Finally, the pathways associated with HCC in the TCGA dataset were analyzed.

Statistical analysis

Statistical analysis was performed using the R software (version 4.2.0). Continuous data were analyzed using Mann-Whitney tests, and categorical data were analyzed using Fisher's exact tests. Pearson correlation coefficient was used to estimate the correlation between continuous variables. The Kaplan-Meier method was used for survival analysis. The Log-rank test was used to determine the significance of differences. All statistical analyses were considered significant if the *p*-values were less than 0.05.

Results

Annotation of single-cell sequencing data and identification of differentially expressed genes associated with NCPS

We first analyzed the single-cell sequencing datasets GSE125449 and GSE151530 for HCC to integrate different samples. As shown in Figure 2A, there were no significant batch effects in the 23 samples, and further analysis of the results could be conducted. The K-Nearest Neighbor (KNN) algorithm was used to divide all cells into 43 clusters (Figure 2B). After entering 92 genes related to NCPS using the "PercateFeatureSet" function, a percentage of the genes associated with NCPS was calculated for each cell. Cells were classified based on the median percentage of NCPS genes and represented as tSNE plots (Figure 2C). We then identified eight distinct cell types based on the expression of surface markers of different cell types in different clusters. They were T cells, macrophages, endothelial cells, NK cells, monocytes, smooth muscle cells, B cells, and tumor cells (Figure 2D). The surface markers for eight types of cells are shown in Figure 2E. Furthermore, we identified 3598 genes that were differentially expressed between high- and low-NCPS groups

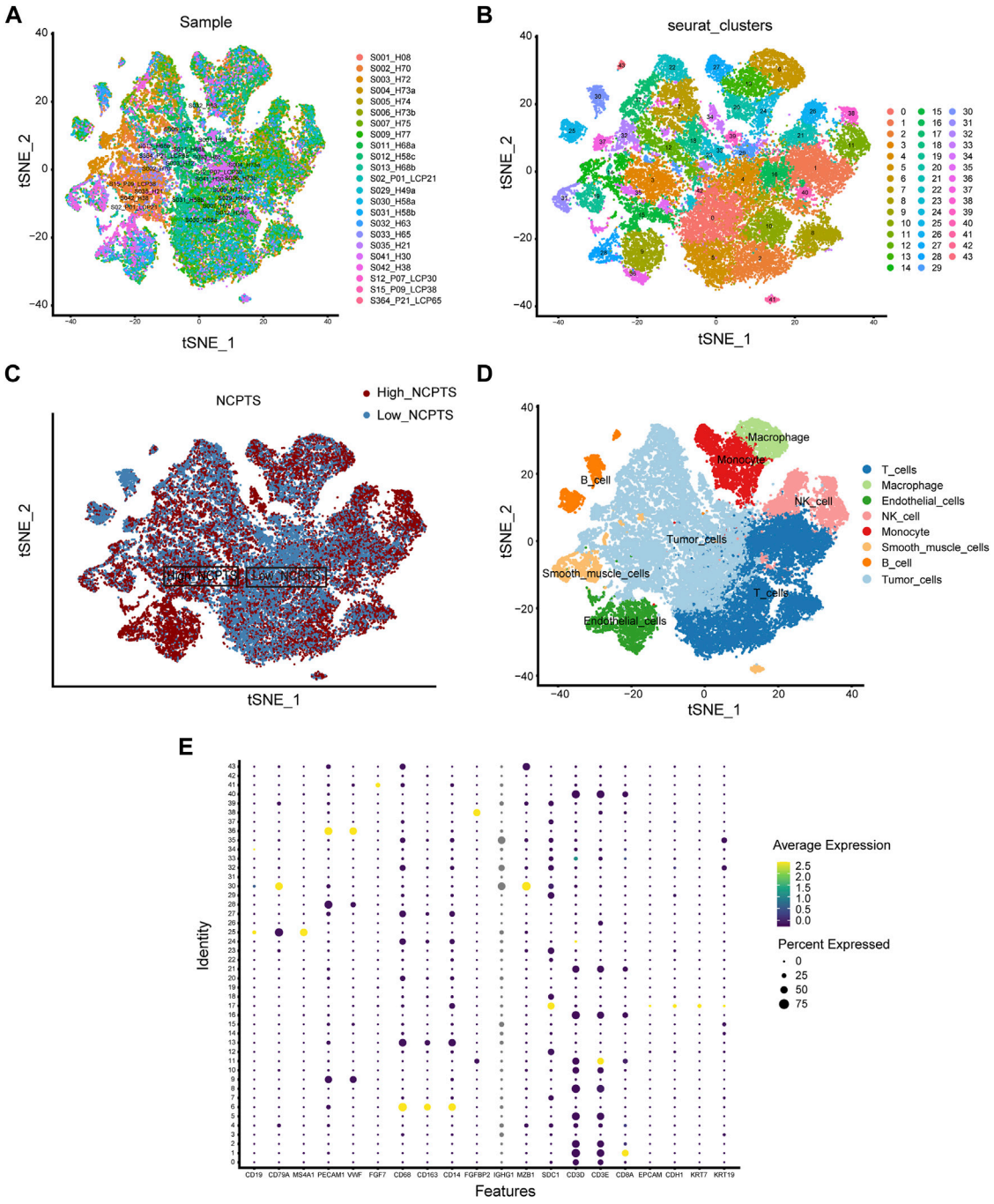


FIGURE 2
Single-cell analysis. **(A)** There were no significant batch effects in the 23 samples. **(B)** Dimensionality reduction and cluster analysis. **(C)** Percentage of genes in each cell that are involved in necroptosis. **(D)** Annotation of cells in accordance with their surface marker genes. **(E)** A list of the surface markers of the eight types of cells.

(Supplementary Material S4). Using the WGCNA analysis of 363 samples from the TCGA cohort, we have obtained gene modules associated with necroptosis. In total, eight non-gray modules were identified by setting a soft threshold of 8 (Figure 3A). As shown in Figure 3B, MEsalmon, MEtan, and MEpurple were strongly associated with the NCPS score. Further analysis was performed on the genes in these three modules.

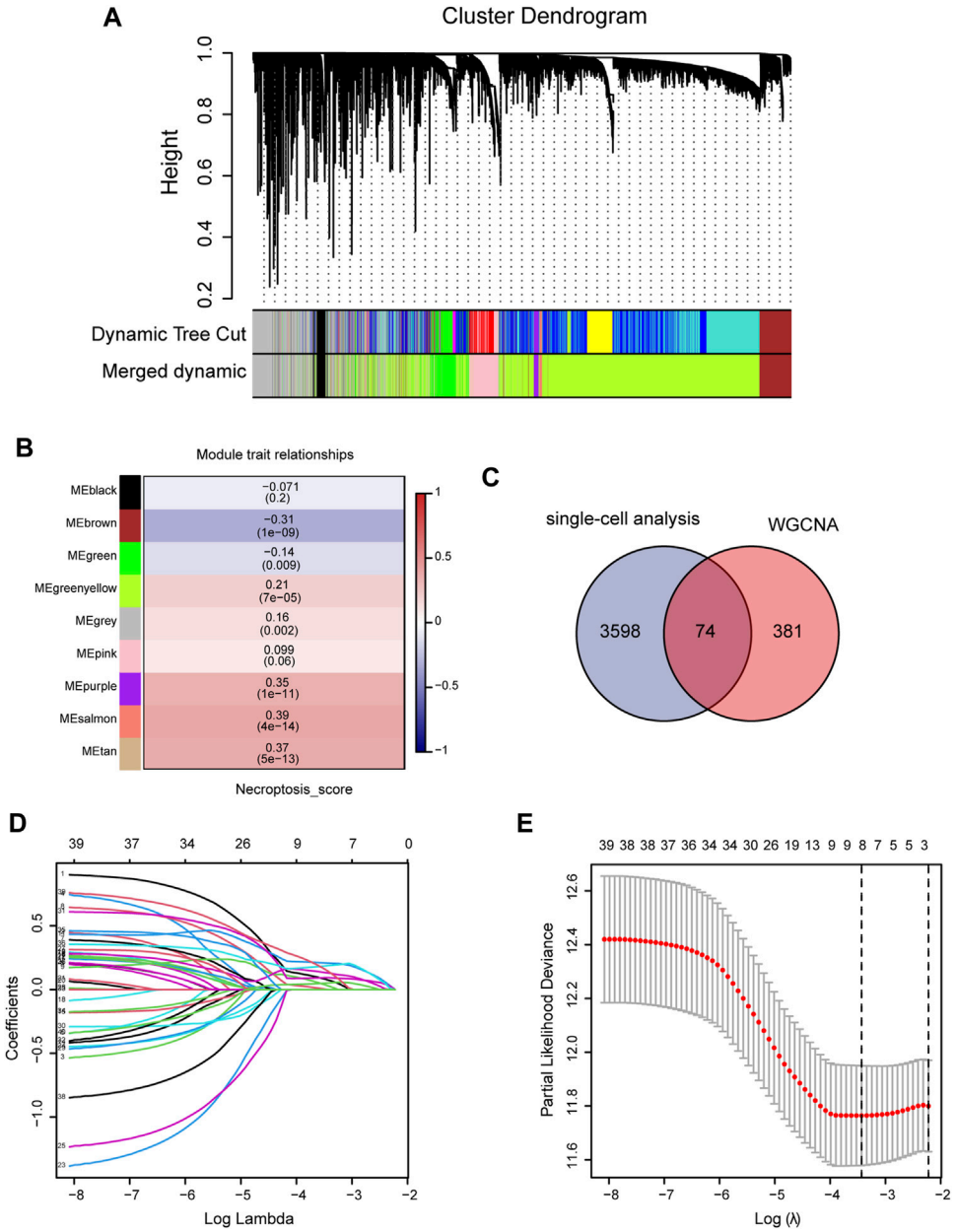


FIGURE 3
Construction of prognostic model. **(A,B)** WGCNA screening for modules relating to necroptosis. **(C)** The intersection between differential genes identified by single-cell analysis and genes identified by WGCNA. **(D,E)** Using Lasso regression, the final genes were selected for the prognostic model.

The NCPS-related prognostic model could be used to classify HCC patients and predict their prognosis

An intersection was drawn between differential genes derived from single-cell analysis and genes identified by WGCNA. In Figure 3C, 74 genes are shown as candidates for the next step in the analysis (Supplementary Material S5). Based on univariate COX analysis within the TCGA cohort, 45 genes have been identified as significantly associated with prognosis. The LASSO regression analysis employed a random seed of 2022, and the results indicated that gene contraction stabilized with minimal partial likelihood deviation when the number of genes included was 8 (Figures 3D,E). Table 1 summarizes the results of the Lasso regression for each of these genes. The prognostic model was constructed from eight genes, including RAD21, NBN, PRKDC,

TABLE 1 Eight genes were identified by lasso regression to construct a prognostic model.

ID	Coef	Hazard_ratio	Low_CI	High_CI	*p*_value
RAD21	0.07523793	1.61760529	1.18401058	2.20998607	0.00252028
NBN	0.02974632	1.69458991	1.18308239	2.42724849	0.00401446
PRKDC	0.2080255	1.68776122	1.27783985	2.22918228	0.00022688
MAP2K2	0.13063826	1.83095476	1.27010213	2.6394691	0.00119002
RIPK2	0.05091183	1.61290563	1.22724908	2.11975272	0.00060646
BOP1	0.0940779	1.48888676	1.1924913	1.85895175	0.00044089
POLR2E	0.18787475	2.33187249	1.40355261	3.87418988	0.00108012
GPX4	0.1444125	1.69395714	1.12709302	2.54592189	0.01122826

MAP2K2, RIPK2, BOP1, POLR2E, and GPX4. As follows was the prognostic model.

$$\begin{aligned}NCPS = {} & RAD21*0.07523793 + NBN*0.02974632 \\ & + PRKDC*0.2080255 + MAP2K2*0.13063826 \\ & + RIPK2*0.05091183 + BOP1*0.0940779 \\ & + POLR2E*0.18787475 + GPX4*\ 0.1444125\end{aligned}$$

Based on median values, patients were divided into high- and low-risk groups. Figure 4A showed that the high-NCPS group in the training cohort had a worse prognosis ($p = 0.015$). Figure 4B demonstrated that patients with high-NCPS had worse outcomes than those with low-NCPS in the validation cohort ($p = 0.0078$). ROC curves were generated for both the training and validation cohorts to test the prognosis assessment ability. As shown in Figure 4C, the area under the curve (AUC) values were 0.722, 0.746, 0.741, and 0.763 at 1, 2, 3, and 5 years in the TCGA cohort, respectively. In the validation cohort, AUC values were 0.730, 0.653, 0.625 and 0.623 at 1, 2, 3 and 5 years, respectively (Figure 4D). In the GSE76427 cohort, the results showed that the high-NCPS group had a worse prognosis ($p = 0.0037$), and AUC values were 0.682, 0.696, and 0.778 at 2, 3 and 5 years, respectively (Supplementary Material S6).

Based on these results, the NCPS-related prognostic model was found to be accurate in predicting the outcomes of patients in all three cohorts. Furthermore, PCA was performed on the eight genes included in all three cohorts, and the results were similar. The results showed that the model performed well in classifying HCC patients (Figures 4E,F).

The nomogram could be more reliable in predicting patient outcomes than other indicators

Combining clinical information and NCPS scores, we constructed a nomogram that allows us to assess patients' prognoses. In Figure 5A, the estimated mortality rates for patients with the high-NCPS score "TCGA-G3-A7M9" were 0.626, 0.92, and 0.984 at 1, 3, and 5 years based on gender, age, T-stage, and total stage (Table 2). Based on the low-NCPS score, the estimated mortality rates for patients with "TCGA-DD-AADS" were 0.0389, 0.102, and 0.153 at 1, 3, and 5 years based on sex, age, T-stage, and total stage (Figure 5B). In Supplementary Material S7, NCPS scores and clinical characteristics of 363 patients from the TCGA-HCC dataset are presented. Accordingly, a clinical decision could be based on assessing a patient's risk and guiding their subsequent treatment. Furthermore, the accuracy of the nomogram was assessed through ROC analysis, which showed AUCs of 0.75, 0.67, and 0.68 for 1, 3, and 5 years, respectively (Figure 5C). In addition, we assessed the utility of the model to support clinical decision-making by using decision curve analysis (DCA) and reported the net clinical benefit of the model. The results showed that the nomogram is better than other clinical indicators, indicating that the nomogram is effective in predicting the patient's prognosis (Figure 5D).

Survival analysis and cellular localization of the eight hub genes

Survival analysis was performed for each of the eight hub genes. Compared with patients with low expression, those with high levels of RAD219 ($p = 0.0078$), RIPK2 ($p = 0.005$), BOP1 ($p = 0.0038$), POLR2E ($p = 0.02$), and MAP2K2 ($p = 0.017$) had significantly poorer outcomes (Figure 6A). To investigate the expression of the eight hub genes in various cell types, we conducted a single-cell sequencing analysis. As shown in Figures 6B–J, RAD21, BOP1, POLR2E, and PRKDC were mainly expressed in tumor cells, RIPK2 was mainly expressed in monocytes, MAP2K2 was mainly expressed in tumor cells and macrophages, NBN was mainly expressed in macrophages and monocytes, and GPX4 was mainly expressed in tumor cells and T cells.

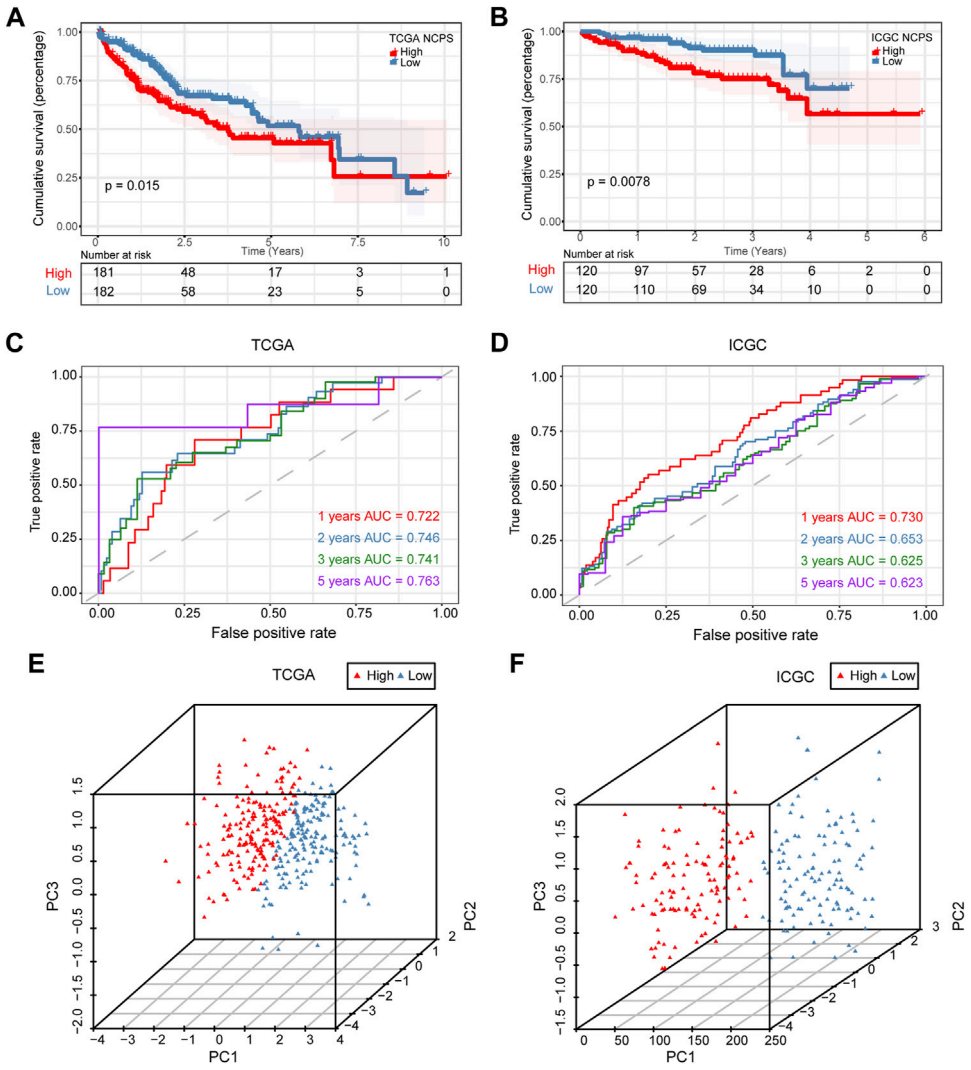


FIGURE 4
Validation of prognostic model. **(A)** Survival analysis of the training set showed significantly poorer outcomes for NCPS high group (p = 0.015). **(B)** Survival analysis results in the validation set were similar to those in the training set (p = 0.0078). **(C)** ROC curve of the training set. **(D)** ROC curve of the validation set. **(E,F)** 3D-PCA analysis in the training set and validation set.

The NCPS scores are positively correlated with the levels of immune cell infiltration and the expression of immune checkpoint genes

As shown in the above analysis, patient outcomes varied significantly within the NCPS subgroups. To explore the reasons for this and inform immunotherapy, comparisons of the levels of immune cell infiltration between the various groups were conducted.

As shown in Figure 7A, six different immune infiltration algorithms have been used to estimate the relationship between necroptosis and immune cells. Specifically, the three algorithms of MCP counter, Quanti-seq, and TIMER clearly demonstrated that there were more immune cell infiltrations in the high-NCPS group, including macrophages, NK cells, T cells, monocytes, B cells, and dendritic cells. We then investigated the expression of genes associated with immune checkpoints. Figure 7B demonstrated that many immune checkpoint genes, such as PDCD1 and CTLA4, were more highly expressed in the high NCPS group. High NCPS patients were likely to have a higher degree of immune infiltration. However, patients with high-NCPS may suffer from low response states due to high levels of immune checkpoint genes, and immune checkpoint inhibitors may be of greater benefit to patients with such conditions. In

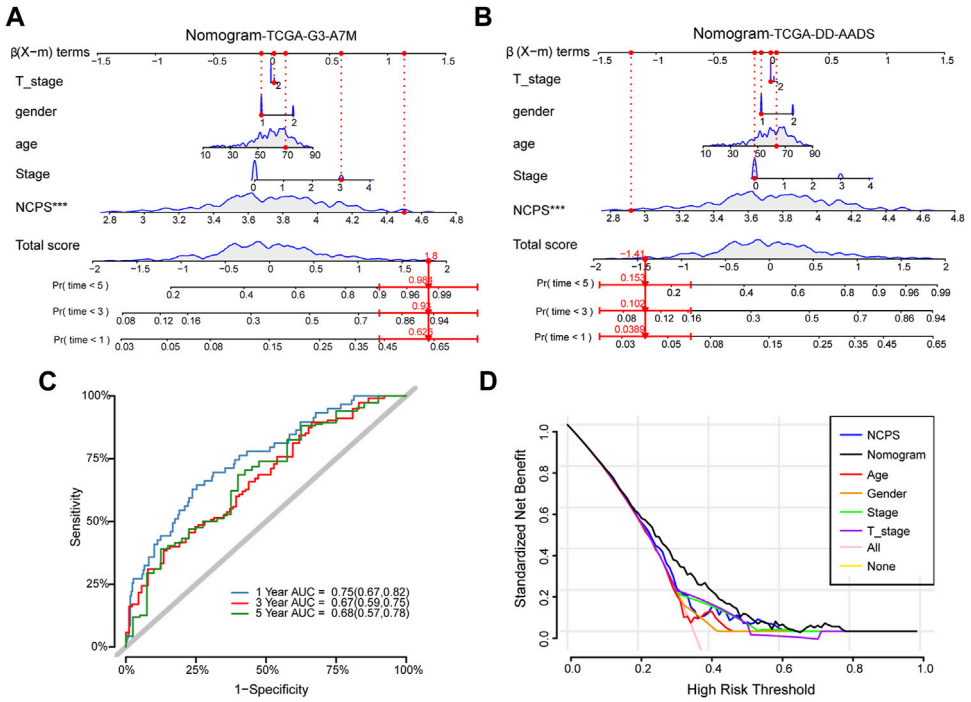


FIGURE 5
Construction of the nomogram. **(A)** High-NCPS patient "TCGA-G3-A7M9": Mortality rates were estimated to be 0.626, 0.92, and 0.984 at 1, 3, and 5 years, respectively. **(B)** Low-NCPS patient "TCGA-DD-AADS": Mortality rates in 1, 3, and 5 years were estimated to be 0.0389, 0.102, and 0.153, respectively. **(C)** The ROC curve for the nomogram. **(D)** DCA analysis showed that the nomogram was more effective than other clinical indicators.

TABLE 2 Comparison of clinical data from patients with high- and low-NCPS in the TCGA-HCC dataset.

Patient ID	TCGA-G3-A7M9	TCGA-DD-AADS
NCPS score	4.501	2.920
Gender	Male	male
Status	Dead	Alive
Age (year)	70.104	63.636
M_stage	MX	M0
N_stage	NX	N0
Stage	Stage IIIB	Stage I
T_stage	T3b	T1
Survival time (year)	0.153	1.299
1-year mortality rates	CI: 0.626 (0.432, 0.82)	CI: 0.0389 (0.0234, 0.0646)
3-year mortality rates	CI: 0.93 (0.782, 0.99)	CI: 0.102 (0.0617, 0.165)
5-year mortality rates	CI: 0.984 (0.906, 0.999)	CI: 0.153 (0.094, 0.243)

addition, we examined the immune infiltration results obtained by different algorithms. The QuantTIseq algorithm showed that patients with high-NCPS levels had more macrophages, B cells, and T cells (Figure 7C).

A high-NCPS score is associated with a greater incidence of gene mutations

According to the NCPS scores in the high- and low-group, 20 of the top mutated genes were identified. As shown in Figures 8A,B, the incidence of mutations in the 20 most frequently mutated genes was 89.53% (High NCPS) and 82.76 % (Low NCPS) for the two groups. In the high-NCPS group, the highest mutation rates were PT53 (40%), CTNNB1 (30%), and TTN (29%). In the low NCPS group, TTN (25%), CTNNB1 (24%), and PT53 (21%) were the mutations with the highest rates. A higher incidence of mutations was observed in the high-NCPS group as compared to the low-NCPS group. Mutations were analyzed for eight hub genes (Supplementary Material S8). The highest Variant Classification shown in Figure 8C was Missense Mutation. Single nucleotide polymorphism (SNP) was the highest Variant Type (Figure 8D). Figure 8E indicated that an average of 100 genes were mutated in each sample. Figure 8F showed that the top three base mutation types of single nucleotide variants (SNVs) were C>T, C>A, and T>C. In addition, we analyzed the correlation between pairs of mutated genes. Figure 8G showed a strong co-relation between FLG and OBSCN ($p < 0.0001$, OR = 8.803), FAT3 and DNAH7 ($p = 0.00064$, OR = 6.925). There was a strong mutually exclusive relationship between CTNNB1 and TP53 ($p =$

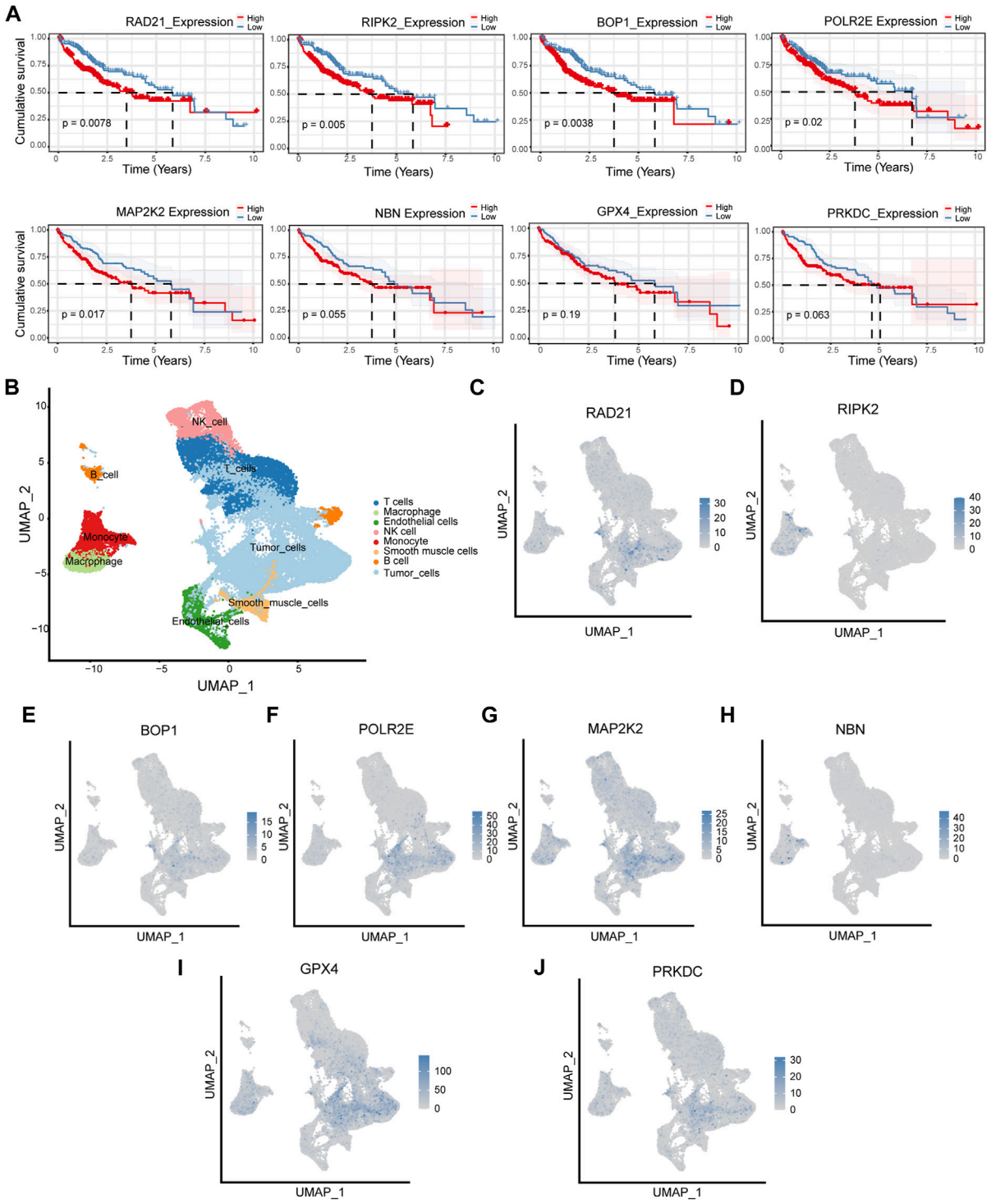


FIGURE 6
Survival analysis and cellular localization of the eight hub genes. **(A)** The survival analysis of eight hub genes in the TCGA cohort. **(B–J)** The expression of eight hub genes in different types of cells.

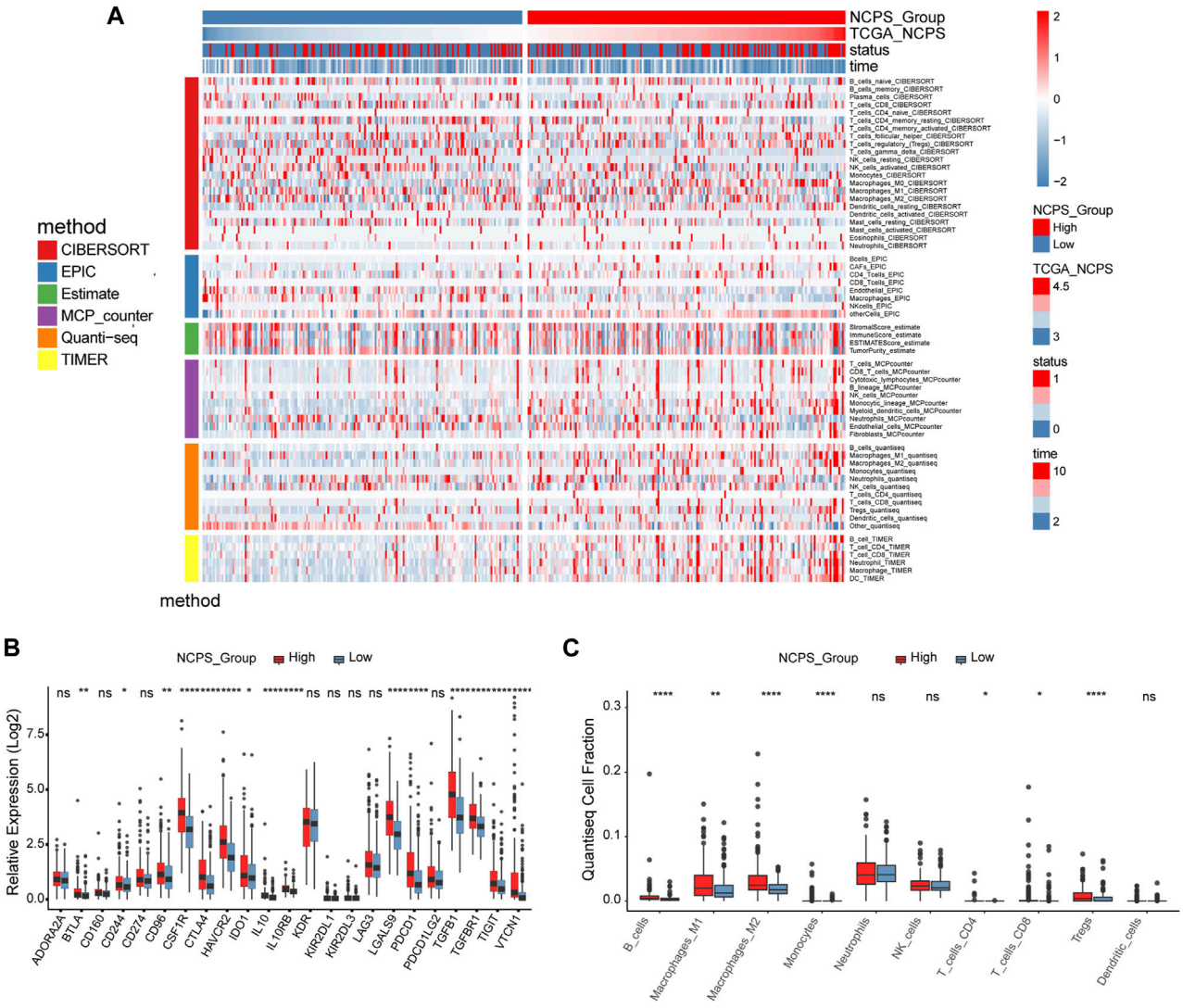


FIGURE 7
Immune infiltration analysis of TCGA cohort. **(A)** Heat map of immune cell infiltration in high and low NCPS groups with six immune infiltration algorithms. **(B)** Expression of immune checkpoint genes in high- and low-NCPS groups. **(C)** Results of the quanTIseq algorithm.

0.00811, OR = 0.459), AXIN1 and CTNNB1 (p = 0.00733, OR = 0.109) (Supplementary Material S9).

Drug sensitivity of HCC and hub gene protein expression are positively correlated with NCPS scores

Based on the "pRRophetic" package, we assessed the sensitivity of different NCPS subgroups to drugs commonly used as a treatment for HCC. The high-risk group showed higher sensitivity to cisplatin, docetaxel, paclitaxel, sunitinib, tipifarnib, bexarotene, bicalutamide, bortezomib, and bleomycin, while the low-risk group showed higher sensitivity to metformin, camptothecin, temsirolimus (Figure 9A). The immunohistochemical analysis of the HPA database showed that protein products with high NCPS-related genes were expressed at higher levels in HCC samples compared to normal tissues (Figure 9B).

Pathway enrichment and localization in single-cell sequencing data

Pathway enrichment analysis of single-cell data revealed that HALLMARK OXIDATIVE PHOSPHORYLATION was

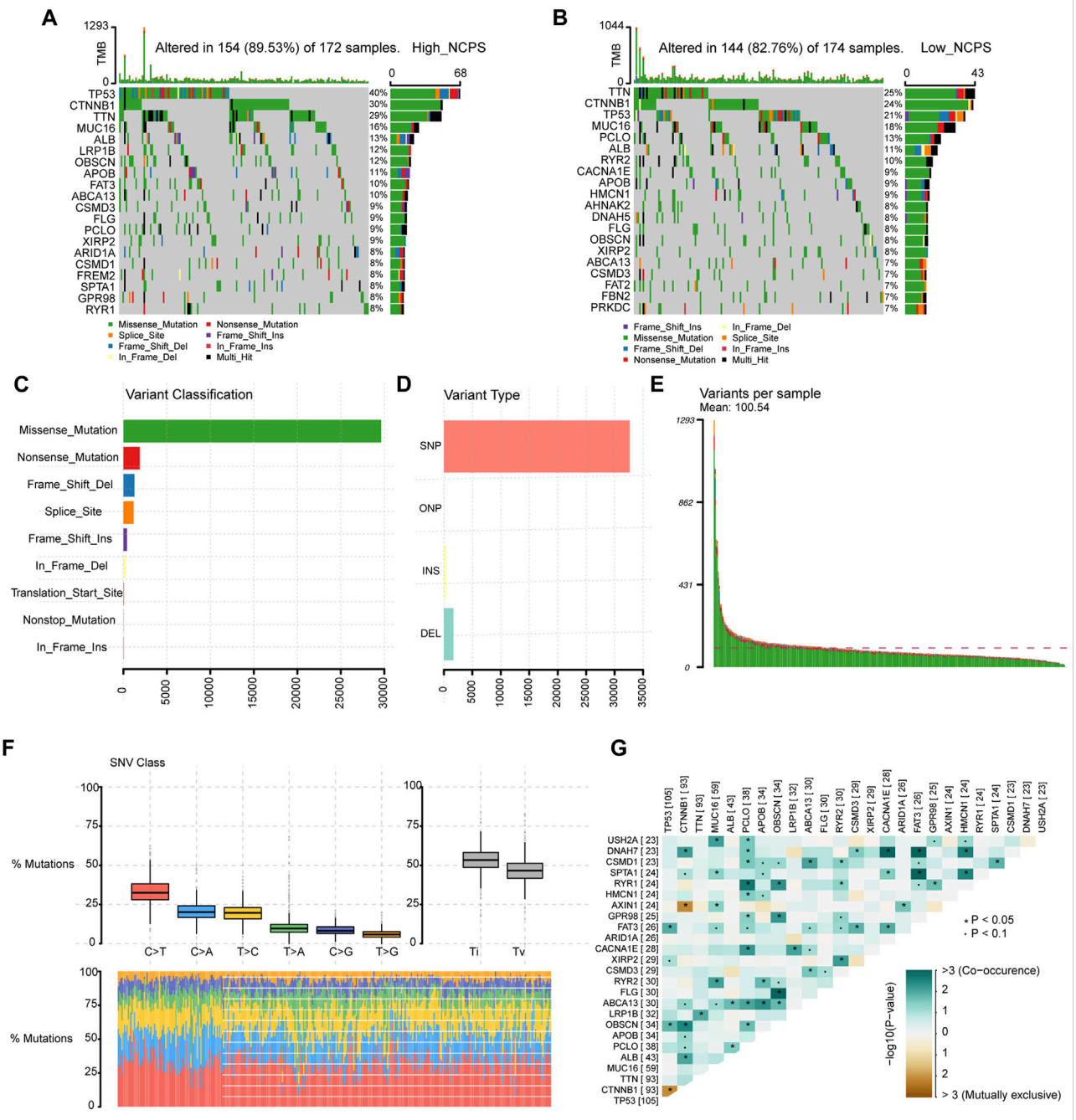


FIGURE 8
Mutation landscape of TCGA cohort. **(A,B)** Mutated genes in high- and low-NCPS groups. **(C,D)** Classification and types of variants. **(E)** Mutations in each sample. **(F)** The base mutation types of SNVs. **(G)** The correlation between pairs of mutated genes.

upregulated in Malignant cells but downregulated in T cells, TECs, and B cells. HALLMARK ALLOGRAFT REJECTION was downregulated in Malignant cells, upregulated in T cells, downregulated in CAFs, and upregulated in TAMS. HALLMARK-TNFA -SIGNALING-VIA-NFKB was downregulated in Malignant cells and upregulated in TAMs. HALLMARK-TGF-BETA-SIGNALING was upregulated in TECs (Figure 10). In addition, we explored the expression of these signaling pathways in different cell types by single-cell sequencing analysis (Figures 11A–D) and profiled the pathways associated with disease (Figure 11E).

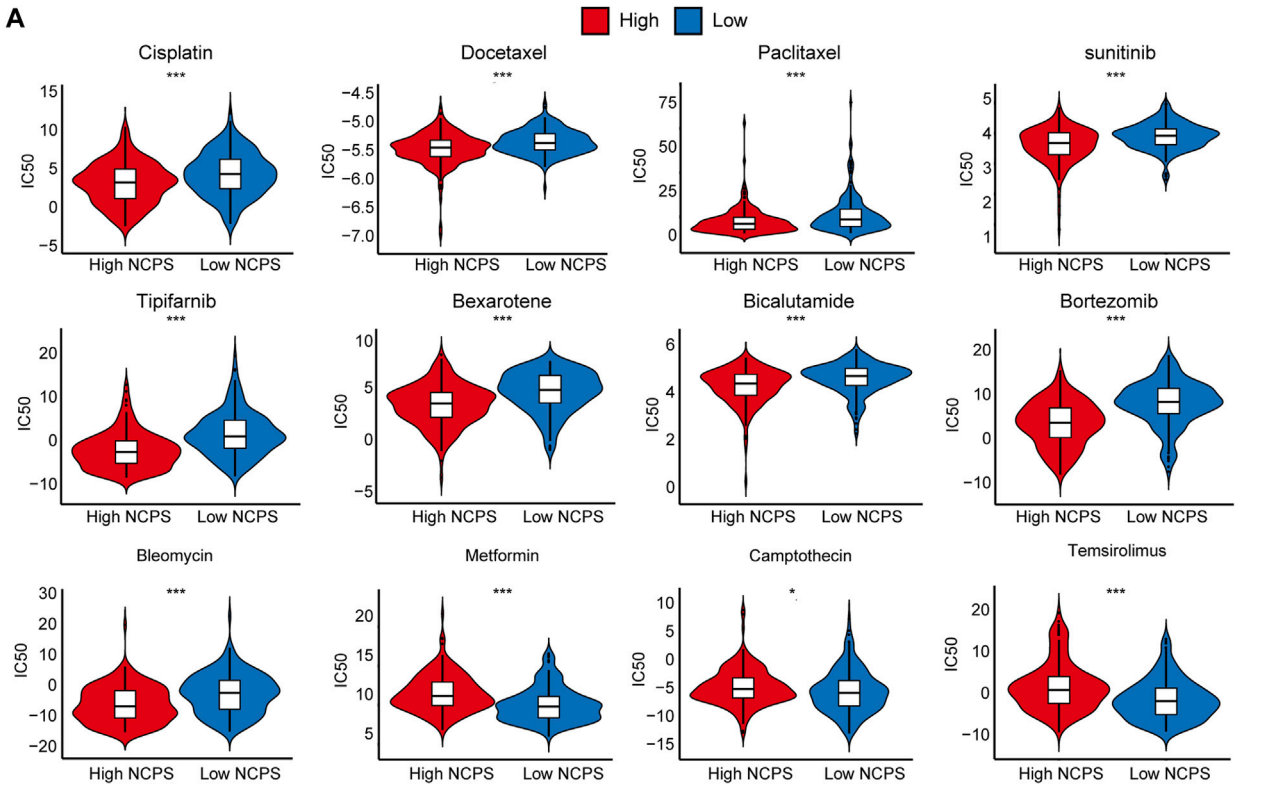


FIGURE 9
Drug sensitivity and Immunohistochemical analysis. **(A)** Drug sensitivity analysis in. high- and low-NCPS groups. **(B)** Immunohistochemical analysis of eight hub genes.

Discussion

With increasing incidence, HCC has become the second leading cause of cancer-related deaths (Bray et al., 2018). Due to lifestyle changes, HCC has become the fastest growing cancer in developed countries, but the response to antitumor therapy is relatively poor. Approximately 50% of HCC patients receive systemic therapy, traditionally with first-line sorafenib or lenvatinib. In the past 5 years, immune checkpoint inhibitors have completely altered the treatment regimen for HCC and improved the prognosis (Llovet et al., 2022). The immune microenvironment plays a significant role in the progression of HCC, and HCC with high- and low-necroptosis respond differently to immune checkpoint inhibitor therapy. However, at present, there are no validated biomarkers to aid in clinical decision-making in this regard.

Immune checkpoint inhibitors are used because immune cells can receive inhibitory signals by activating immune checkpoint molecules. By activating immune checkpoint molecules to receive inhibitory signals, their activity and

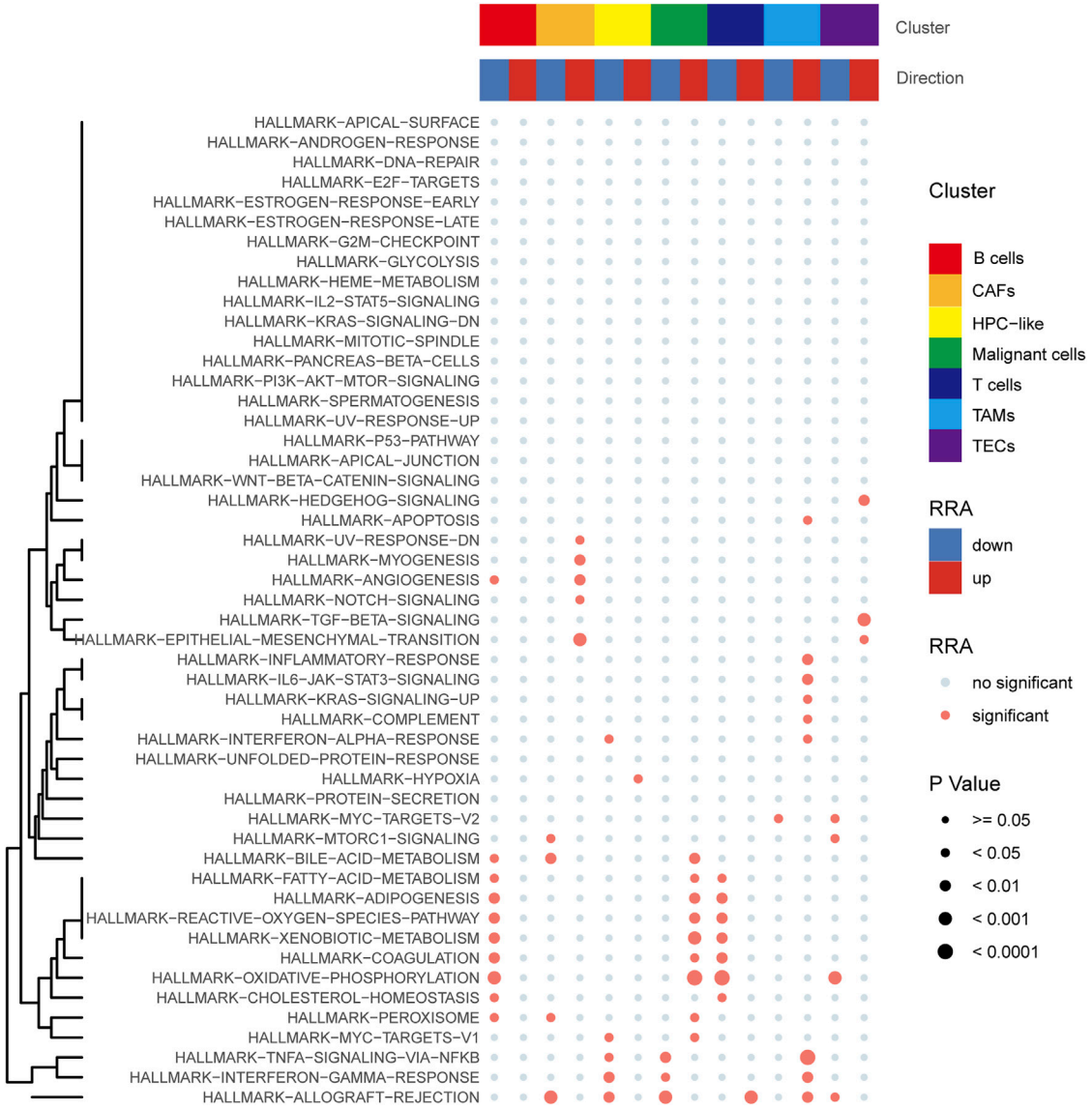


FIGURE 10
Pathway enrichment analysis of single-cell data.

proliferation are blocked (Huang and Chang, 2019). These immune checkpoints can be used by cancer cells, leading to impaired immune surveillance (Liu and Qin, 2019). PD-1, PD-L1, and cytotoxic T cell antigen 4 (CTLA-4) are the main immune checkpoints that have been targeted by monoclonal antibodies.

Utilizing comprehensive data analysis on HCC datasets from TCGA, ICGC, and GEO databases, we built a prognostic profile for NCPS-related genes associated with HCC. We calculated risk scores to identify high- and low-risk groups of patients with HCC. All three cohorts both showed that the high-risk group did significantly worse than the low-risk group in HCC. Xie et al. (2022) found similar results in triple-negative breast cancer, indicating that the higher the NCPS score, the larger the tumor and the worse the prognosis. Furthermore, the ROC curve revealed that this feature might be accurate in predicting the prognosis of patients with HCC at 1, 3, and 5 years. Based on the immune microenvironment analysis, immunotherapy was more likely to be effective in necroptosis with higher expression levels. The low response to immunotherapy of HCC could be attributed in part to the low mutational load and the generation of new immune checkpoints (Ricciuti et al., 2019; Scheiner et al., 2022). Therefore, it becomes fascinating to explore the immune

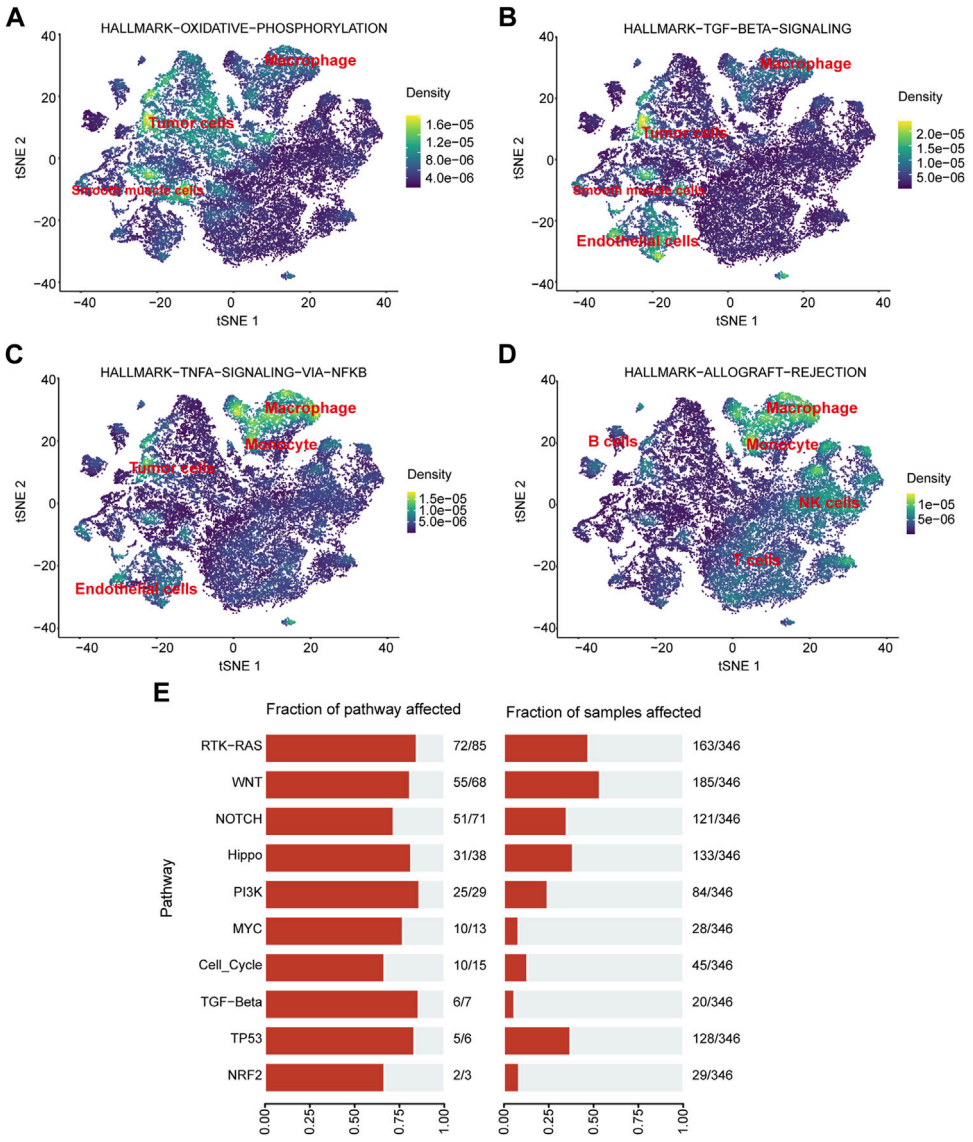


FIGURE 11
Pathway enrichment analysis. **(A–D)** Localization of different pathways in the single-cell dataset. **(E)** The number of pathways enriched in the TCGA cohort.

microenvironment of HCC. Necroptosis may play an important role in TME by the release of inflammatory molecules during the induction of apoptosis. However, it remains unclear whether necroptosis plays a role in HCC.

Necroptosis is a necrotic programmed cell death that is powerfully immunogenic and participates in a complex interplay of autophagy and apoptosis (Gong et al., 2017). There is growing evidence that necroptosis plays an important role in prognosis, disease progression and tumor metastasis, and immune surveillance in cancer patients (Gong et al., 2019). Targeting necroptosis through immune checkpoint is also emerging as a new approach in tumor therapy.

The role of necroptosis in cancer is complex. It is still unclear exactly what role necroptosis plays in cancer. In general, high expression of necroptosis elicits strong adaptive immune responses that can inhibit tumor progression (Yatim et al., 2015). However, these recruited strong immune responses may also promote tumor progression. The inflammatory response may promote tumorigenesis and metastasis, as well as may generate an immunosuppressive tumor microenvironment (Seifert and Miller, 2017). Therefore, it is essential to investigate the molecular mechanisms and physiopathological aspects of necroptosis, as well as its interaction with immunity. In addition, it is imperative to

discover the correlation between specific necroptosis markers and the prognosis of HCC. This is to unravel the confusion of necroptosis correlation in HCC and further develop targeted antitumor therapeutic drugs. In this study, combining single-cell analysis and second-generation sequence analysis, we were able to identify a significant difference between NCPS groups in terms of immune cell infiltration in HCC. Significant differences were observed between the high- and low-NCPS groups. In addition, the study findings indicated that a high level of NCPS group corresponds to a high level of immune checkpoint gene expression. Therefore, patients with HCC who have a high NCPS are more likely to respond to immunotherapy.

The datasets GSE125449 and GSE151530 have been initially explored to reveal changes in the immune microenvironment of HCC. Among the published results, GSE125449 reveals different degrees of heterogeneity of malignant cells within and between tumors and different TME landscapes by single-cell sequencing techniques. GSE151530 provides insights into the collective behavior of HCC cell communities by single-cell sequencing and potential tumor evolution in response to therapy drivers. We first classified HCC cells into two groups based on their NCPS scores by analyzing single cells of GSE125449 and GSE151530. This provided a reference for us to study the heterogeneity of necroptosis in HCC. Based on these two cell populations, we calculated the differentially expressed genes, which then served as a basis for constructing a prognostic model. For the validation of the prognostic model, survival data from the ICGC dataset was analyzed.

Our study has some limitations. First, a comprehensive analysis of HCC tissues is needed to fully validate how the eight NCPS-related genes are involved in the development of HCC. This was not examined in the current study. Second, further validation with larger patient datasets is needed better to estimate the accuracy of the model's predictions. Finally, further experimental evidence is needed to fully understand the role and mechanisms of eight NCPS-related genes in HCC.

Conclusion

Through the analysis of single-cell and bulk multi-omics sequencing data, we constructed a prognostic model related to necroptosis and explored the relationship between high- and low-necroptosis groups and immune cell infiltration in HCC. This provides a new reference for further understanding the role of necroptosis in HCC. This may be useful in developing new therapeutic targets for the treatment of HCC. However, further molecular experiments are required to confirm the present findings.

Data availability statement

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

Author contributions

TY: Contributed to conception and design of the study; JL: Bioinformatics analysis and data mining; ZW and SW: Drafted the manuscript; CL: Statistical analysis; MS and JX: Literature review; YW and DF: Guidance for Bioinformatics analysis and data mining; YH and XZ: Review the manuscript; WM: Edited the manuscript. All the authors read and approved the final manuscript.

Funding

This work was supported by the Jincheng People's Hospital (Nos. JSY-2021D007, JSY-2021D009) and the Traditional Chinese Medicine Bureau of Guangdong Province (20203002).

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.984297/full#supplementary-material

References

Altorki, N. K., Markowitz, G. J., Gao, D., Port, J. L., Saxena, A., Stiles, B., et al. (2019). The lung microenvironment: An important regulator of tumour growth and metastasis. *Nat. Rev. Cancer* 19 (1), 9–31. doi:10.1038/s41568-018-0081-9

Alvarez-Diaz, S., Dillon, C. P., Lalaoui, N., Tanzer, M. C., Rodriguez, D. A., Lin, A., et al. (2016). The pseudokinase MLKL and the kinase RIPK3 have distinct roles in autoimmune disease caused by loss of death-receptor-induced apoptosis. *Immunity* 45 (3), 513–526. doi:10.1016/j.immuni.2016.07.016

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: Archive for functional genomics data sets-update. *Nucleic Acids Res.* 41, D991–D995. doi:10.1093/nar/gks1193

Bersuker, K., Hendricks, J. M., Li, Z., Magtanong, L., Ford, B., Tang, P. H., et al. (2019). The CoQ oxidoreductase FSP1 acts parallel to GPX4 to inhibit ferroptosis. *Nature* 575 (7784), 688–692. doi:10.1038/s41586-019-1705-2

Bindea, G., Mlecnik, B., Tosolini, M., Kirilovsky, A., Waldner, M., Obenauf, A. C., et al. (2013). Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity* 39 (4), 782–795. doi:10.1016/j.immuni.2013.10.003

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 (6), 394–424. doi:10.3322/caac.21492

Chaudhary, K., Poirion, O. B., Lu, L., Huang, S., Ching, T., and Garmire, L. X. (2019). Multimodal meta-analysis of 1, 494 hepatocellular carcinoma samples reveals significant impact of consensus driver genes on phenotypes. *Clin. Cancer Res.* 25 (2), 463–472. doi:10.1158/1078-0432.CCR-18-0088

Chevrier, S., Levine, J. H., Zanotelli, V. R. T., Silina, K., Schulz, D., Bacac, M., et al. (2017). An immune atlas of clear cell renal cell carcinoma. *Cell* 169 (4), 736–749. doi:10.1016/j.cell.2017.04.016

Dhanasekaran, R., Venkatesh, S. K., Torbenson, M. S., and Roberts, L. R. (2016). Clinical implications of basic research in hepatocellular carcinoma. *J. Hepatol.* 64 (3), 736–745. doi:10.1016/j.jhep.2015.09.008

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

Geeleher, P., Cox, N., and Huang, R. S. (2014). pRRophetic: an R package for prediction of clinical chemotherapeutic response from tumor gene expression levels. *PloS One* 9 (9), e107468. doi:10.1371/journal.pone.0107468

Gong, Y.-N., Guy, C., Olauson, H., Becker, J. U., Yang, M., Fitzgerald, P., et al. (2017). ESCRT-III acts downstream of MLKL to regulate necroptotic cell death and its consequences. *Cell* 169 (2), 286–300. doi:10.1016/j.cell.2017.03.020

Gong, Y., Fan, Z., Luo, G., Yang, C., Huang, Q., Fan, K., et al. (2019). The role of necroptosis in cancer biology and therapy. *Mol. Cancer* 18 (1), 100. doi:10.1186/s12943-019-1029-8

González-Juarbe, N., Bradley, K. M., Shenoy, A. T., Gilley, R. P., Reyes, L. F., Hinojosa, C. A., et al. (2017). Pore-forming toxin-mediated ion dysregulation leads to death receptor-independent necroptosis of lung epithelial cells during bacterial pneumonia. *Cell Death Differ.* 24 (5), 917–928. doi:10.1038/cdd.2017.49

Grinchuk, O. V., Yenamandra, S. P., Iyer, R., Singh, M., Lee, H. K., Lim, K. H., et al. (2018). Tumor-adjacent tissue co-expression profile analysis reveals pro-oncogenic ribosomal gene signature for prognosis of resectable hepatocellular carcinoma. *Mol. Oncol.* 12 (1), 89–113. doi:10.1002/1878-0261.12153

Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., et al. (2016). Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* 375 (12), 1109–1112. doi:10.1056/NEJMp1607591

Guo, R., Jia, X., Ding, Z., Wang, G., Jiang, M., Li, B., et al. (2022). Loss of MLKL ameliorates liver fibrosis by inhibiting hepatocyte necroptosis and hepatic stellate cell activation. *Theranostics* 12 (11), 5220–5236. doi:10.7150/thno.71400

Hänzelmann, S., Castelo, R., and Guinney, J. (2013). Gsva: Gene set variation analysis for microarray and RNA-seq data. *BMC Bioinforma.* 14, 7. doi:10.1186/1471-2105-14-7

Huang, P.-W., and Chang, J. W.-C. (2019). Immune checkpoint inhibitors win the 2018 Nobel Prize. *Biomed. J.* 42 (5), 299–306. doi:10.1016/j.bj.2019.09.002

Kalliolias, G. D., and Ivashkiv, L. B. (2016). TNF biology, pathogenic mechanisms and emerging therapeutic strategies. *Nat. Rev. Rheumatol.* 12 (1), 49–62. doi:10.1038/nrrheum.2015.169

Karki, R., Sundaram, B., Sharma, B. R., Lee, S., Malireddi, R. K. S., Nguyen, L. N., et al. (2021). ADAR1 restricts ZBP1-mediated immune response and PANoptosis to promote tumorigenesis. *Cell Rep.* 37 (3), 109858. doi:10.1016/j.celrep.2021.109858

Koo, G.-B., Morgan, M. J., Lee, D.-G., Kim, W.-J., Yoon, J.-H., Koo, J. S., et al. (2015). Methylation-dependent loss of RIP3 expression in cancer represses programmed necrosis in response to chemotherapeutics. *Cell Res.* 25 (6), 707–725. doi:10.1038/cr.2015.56

Koren, E., and Fuchs, Y. (2021). Modes of regulated cell death in cancer. *Cancer Discov.* 11 (2), 245–265. doi:10.1158/2159-8290.CD-20-0789

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

Li, L., Greene, T., and Hu, B. (2018). A simple method to estimate the time-dependent receiver operating characteristic curve and the area under the curve with right censored data. *Stat. Methods Med. Res.* 27 (8), 2264–2278. doi:10.1177/0962280216680239

Liu, X., and Qin, S. (2019). Immune checkpoint inhibitors in hepatocellular carcinoma: Opportunities and challenges. *Oncologist* 24 (1), S3–S10. doi:10.1634/theoncologist.2019-IO-S1-s01

Llovet, J. M., Castet, F., Heikenwalder, M., Maini, M. K., Mazzaferro, V., Pinato, D. J., et al. (2022). Immunotherapies for hepatocellular carcinoma. *Nat. Rev. Clin. Oncol.* 19 (3), 151–172. doi:10.1038/s41571-021-00573-2

Lossos, I. S., Czerwinski, D. K., Alizadeh, A. A., Wechser, M. A., Tibshirani, R., Botstein, D., et al. (2004). Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes. *N. Engl. J. Med.* 350 (18), 1828–1837. doi:10.1056/NEJMoa032520

Mabbott, N. A., Baillie, J. K., Brown, H., Freeman, T. C., and Hume, D. A. (2013). An expression atlas of human primary cells: Inference of gene function from coexpression networks. *BMC Genomics* 14, 632. doi:10.1186/1471-2164-14-632

Najafov, A., Chen, H., and Yuan, J. (2017). Necroptosis and cancer. *Trends Cancer* 3 (4), 294–301. doi:10.1016/j.trecan.2017.03.002

Reyna, D. E., Garner, T. P., Lopez, A., Kopp, F., Choudhary, G. S., Sridharan, A., et al. (2017). Direct activation of BAX by BTSA1 overcomes apoptosis resistance in acute myeloid leukemia. *Cancer Cell* 32 (4), 490–505. doi:10.1016/j.ccell.2017.09.001

Ricciuti, B., Kravets, S., Dahlberg, S. E., Umeton, R., Albayrak, A., Subegdjo, S. J., et al. (2019). Use of targeted next generation sequencing to characterize tumor mutational burden and efficacy of immune checkpoint inhibition in small cell lung cancer. *J. Immunother. Cancer* 7 (1), 87. doi:10.1186/s40425-019-0572-6

Riley, R. S., June, C. H., Langer, R., and Mitchell, M. J. (2019). Delivery technologies for cancer immunotherapy. *Nat. Rev. Drug Discov.* 18 (3), 175–196. doi:10.1038/s41573-018-0006-z

Safran, M., Rosen, N., Twik, M., BarShir, R., Stein, T. I., Dahary, D., et al. (2021). "The GeneCards suite," in *Practical guide to life science databases*. Editors I. Abugessaisa and T. Kasukawa (Singapore: Springer Nature Singapore), 27–56.

Sahin, I. H., Askan, G., Hu, Z. I., and O'Reilly, E. M. (2017). Immunotherapy in pancreatic ductal adenocarcinoma: An emerging entity? *Ann. Oncol.* 28 (12), 2950–2961. doi:10.1093/annonc/mdx503

Scheiner, B., Pomej, K., Kirstein, M. M., Hucke, F., Finkelmeier, F., Waidmann, O., et al. (2022). Prognosis of patients with hepatocellular carcinoma treated with immunotherapy - development and validation of the CRAFITY score. *J. Hepatol.* 76 (2), 353–363. doi:10.1016/j.jhep.2021.09.035

Seifert, L., and Miller, G. (2017). Molecular pathways: The necrosome-A target for cancer therapy. *Clin. Cancer Res.* 23 (5), 1132–1136. doi:10.1158/1078-0432.CCR-16-0968

Tanzer, M. C., Khan, N., Rickard, J. A., Etemadi, N., Lalaoui, N., Spall, S. K., et al. (2017). Combination of IAP antagonist and IFNγ activates novel caspase-10- and RIPK1-dependent cell death pathways. *Cell Death Differ.* 24 (3), 481–491. doi:10.1038/cdd.2016.147

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. *Sci. (New York, N.Y.)* 347 (6220), 1260419. doi:10.1126/science.1260419

Vickers, A. J., and Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Med. Decis. Mak.* 26 (6), 565–574. doi:10.1177/0272989X06295361

Xie, J., Tian, W., Tang, Y., Zou, Y., Zheng, S., Wu, L., et al. (2022). Establishment of a cell necroptosis index to predict prognosis and drug sensitivity for patients with triple-negative breast cancer. *Front. Mol. Biosci.* 9, 834593. doi:10.3389/fmolb.2022.834593

Yang, W., Ma, Y., Liu, Y., Smith-Warner, S. A., Simon, T. G., Chong, D. Q., et al. (2019). Association of intake of whole grains and dietary fiber with risk of hepatocellular carcinoma in US adults. *JAMA Oncol.* 5 (6), 879–886. doi:10.1001/jamaoncol.2018.7159

Yatim, N., Jusforgues-Saklani, H., Orozco, S., Schulz, O., Barreira da Silva, R., Reis e Sousa, C., et al. (2015). RIPK1 and NF-κB signaling in dying cells determines cross-priming of CD8$^+$ T cells. *Sci. (New York, N.Y.)* 350 (6258), 328–334. doi:10.1126/science.aad0395

Yuan, J., Amin, P., and Ofengeim, D. (2019). Necroptosis and RIPK1-mediated neuroinflammation in CNS diseases. *Nat. Rev. Neurosci.* 20 (1), 19–33. doi:10.1038/s41583-018-0093-1

Zeng, D., Ye, Z., Shen, R., Yu, G., Wu, J., Xiong, Y., et al. (2021). Iobr: Multi-Omics immuno-oncology biological research to decode tumor microenvironment and signatures. *Front. Immunol.* 12, 687975. doi:10.3389/fimmu.2021.687975

Zhang, J., Bajari, R., Andric, D., Gerthoffert, F., Lepsa, A., Nahal-Bose, H., et al. (2019). The international cancer genome consortium data portal. *Nat. Biotechnol.* 37 (4), 367–369. doi:10.1038/s41587-019-0055-9

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 10 October 2022
DOI 10.3389/fgene.2022.1021978

OPEN ACCESS

EDITED BY
Xingdong Chen,
Fudan University, China

REVIEWED BY
Aimin Jiang,
The First Affiliated Hospital of Xi'an Jiaotong University, China
Giuseppe Jurman,
Bruno Kessler Foundation (FBK), Italy

*CORRESPONDENCE
Qing Zhou,
bayyzq@sina.com

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

RECEIVED 18 August 2022
ACCEPTED 26 September 2022
PUBLISHED 10 October 2022

CITATION
Hao L, Chen Q, Chen X and Zhou Q (2022), Integrated analysis of bulk and single-cell RNA-seq reveals the role of MYC signaling in lung adenocarcinoma. *Front. Genet.* 13:1021978. doi: 10.3389/fgene.2022.1021978

Integrated analysis of bulk and single-cell RNA-seq reveals the role of MYC signaling in lung adenocarcinoma

Lu Hao[1], Qiuyan Chen[1], Xi Chen[2] and Qing Zhou[2]*

[1]Science and Education Department, Shenzhen Baoan Shiyan People's Hospital, Shenzhen, China, [2]Central Laboratory, The People's Hospital of Baoan Shenzhen, The Second Affiliated Hospital of Shenzhen University, Shenzhen, China

MYC is one of the well-known oncogenes, and its important role in cancer still remains largely unknown. We obtained lung adenocarcinoma (LUAD) multi-omics data including genome, transcriptome, and single-cell sequencing data from multiple cohorts. We calculated the GSVA score of the MYC target v1 using the ssGSEA method, and obtained the genes highly correlated with this score by Spearman correlation analysis. Subsequent hierarchical clustering divided these genes into two gene sets highly associated with MYC signaling (S1 and S2). Unsupervised clustering based on these genes divided the LUAD samples into two distinct subgroups, namely, the MYC signaling inhibition group (C1) and activation group (C2). The MCP counter package in R was used to assess tumor immune cell infiltration abundance and ssGSEA was used to calculate gene set scores. The scRNA-seq was used to verify the association of MYC signaling to cell differentiation. We observed significant differences in prognosis, clinical characteristics, immune microenvironment, and genomic alterations between MYC signaling inhibition and MYC signaling activation groups. MYC-signaling is associated with genomic instability and can mediate the immunosuppressive microenvironment and promote cell proliferation, tumor stemness. Moreover, MYC-signaling activation is also subject to complex post-transcriptional regulation and is highly associated with cell differentiation. In conclusion, MYC signaling is closely related to the genomic instability, genetic alteration and regulation, the immune microenvironment landscape, cell differentiation, and disease survival in LUAD. The findings of this study provide a valuable reference to revealing the mechanism of cancer-promoting action of MYC in LUAD.

KEYWORDS

MYC, lung adenocarcinoma, prognosis, tumor immunity, tumor stemness, cell proliferation, cell differentiation

Introduction

Lung cancer is the most common malignant tumor of the respiratory system, and the basic and clinical research on lung cancer is increasingly attracting attention (Mok et al., 2019; Song et al., 2020; Song et al., 2022). Although the current research on the pathogenesis of lung cancer has made great progress, but the clinical treatment effect of lung cancer is still not satisfactory, and the long-term survival rate of lung cancer still has great room for improvement. With the deepening of research, molecular biology has been widely used in the field of lung cancer research, which not only provides many new methods for lung cancer research, but also makes the diagnosis and treatment of lung cancer into a new stage. According to the different biological characteristics, lung cancer is often divided into small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) in clinical, among which the latter accounts for about 85% of all lung cancer patients (Cheng et al., 2019). Lung adenocarcinoma (LUAD) is the most common pathological subtype of lung cancer (Kim et al., 2019; Song et al., 2021). Lung cancer is a highly heterogeneous tumor, and lung cancer occurrence is a multi-gene, multi-factor joint regulation, multi-stage and multi-step process (Rajagopalan et al., 2018).A large number of molecular abnormalities and the mechanism of action remain to be explored.

The MYC gene family and its products are involved in the regulation of cell growth, differentiation, and programmed death, and play important roles in the formation of various tumors (Vo et al., 2016). Previous studies have shown that MYC can affect the cell cycle progression, and its amplification and overexpression can lead to c-Myc proto-oncogene activation, which subsequently promotes tumorigenesis and progression (King et al., 2016; Lee et al., 2016).It can also regulate the expression of VEGF, to control the angiogenesis (Thompson et al., 2017). MYC, acting as a transcription factor, can regulate the expression of a large number of genes in tumors. It can act as an amplifier that globally upregulates the expression of protein-coding genes within cancer cells. So with a slight MYC expression disorder, it is possible to promote cancer cell evolution (Jing et al., 2016; Wang et al., 2019; Poh et al., 2019). Recent advances in high-throughput sequencing technologies, such as whole-genome sequencing, have allowed us to analyze tumors in unprecedented depth, especially with the single-cell sequencing (scRNA-seq) technologies that have emerged in recent years (Jiang et al., 2022; Becht et al., 2018; Zhang et al., 2021). Among them, scRNA-seq is a new technology for high-throughput sequencing of mRNA at the single-cell level, studying the overall level of gene expression for individual cells. Given the non-negligible and important role of MYC in cancer cell growth, proliferation, and differentiation, this study innovativelyused LUAD multi-omics data from multiple cohorts to systematicallyinvestigate the relevance of transcriptional profile expression, genome instability, genetic alteration and regulation, immune microenvironment landscape, cell differentiation, and disease survival in Halkmark MYC target V1 gene sets by integrating bulk and single-cell RNA sequencing data. Figure 1 showed the workflow of this study. This study indicated significant differences in prognosis, clinical characteristics, immune microenvironment, and genomic alterations between MYC signaling inhibition and MYC signaling activation groups. MYC-signaling is associated with genomic instability and can mediate the immunosuppressive microenvironment and promote cell proliferation, tumor stemness. Moreover, MYC-signaling activation is also subject to complex post-transcriptional regulation and is highly associated with cancer cell differentiation. Take together, the findings of this study provide a valuable reference to revealing the mechanism of cancer-promoting action of MYC in LUAD.

Materials and methods

Data sources and sample collection

Expression profile data (tpm, counts, miRNA isoform) and clinical information for TCGA-LUAD (n = 516) were downloaded from the GDC (https://portal.gdc.cancer.gov/). To avoid batch effects, the counts and tpm data that we used were directly derived from the STAR-counts workflow type, and were subsequently log2-transformed on the TPM data. Mutation data and copy number variation (CNV) data for the TCGA-LUAD dataset were also downloaded from the cBioPortal (www.cbioportal.org/). Three independent LUAD cohorts were collected from the GEO database (https://www.ncbi.nlm.nih.gov/geo/) asexternal validations, respectively, GSE68465 (n = 443) (Shedden et al., 2008), GSE72094 (n = 442) (Schabath et al., 2016), and GSE31210 (n = 226) (Okayama et al., 2012; Yamauchi et al., 2012). For processing the GEO data, we refer to the method of (Song et al. 2022). The LUAD GEO dataset included in this study was mainly considered based on the sample size. The above three LUAD cohorts have a substantial number of cases, which is an important basis for their inclusion in this study. The clinicopathological parameters of LUAD patients in the TCGA and GEO cohorts should be provided in the Supplementary Table S1.

Definition of signature genes

Pathways for MSigDB database were acquired using the"msigdbr" package in R. The enrichment score for the pathway "HALLMARK_MYC_TARGETS_V1" was calculated using the ssGSEA algorithm of the "GSVA" package in R, and

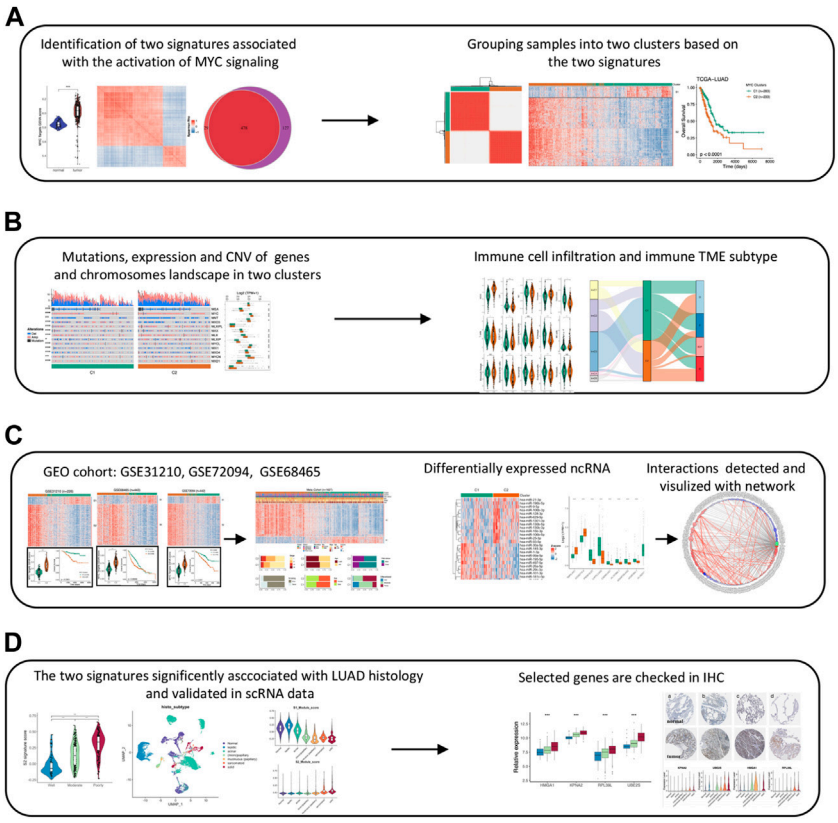


FIGURE 1
The work flow chart of this study. **(A)** Identifying MYC signaling related genes and clustering LUAD samples. **(B)** Analyzing the differences in multiple levels (Genome, transcriptome, and immune infiltration) between two clusters. **(C)** Validating the robustness of the two MYC signatures and constructing the network of differentially expressed lncRNAs and miRNAs. **(D)** Verifying signatures association with LUAD cell differention in scRNAseq data and IHC.

the genes highly correlated with this score were obtained by Spearman correlation analysis, with the threshold set as: Rho> 0.5 and adj.*p* value <1e-3. Genes were subsequently filtered using univariate cox regression analysis and log rank test with *p* value≤0.05 as the threshold, and the two lists of genes obtained were set as intersection, and shared genes were considered as survival-related genes. We divided the resulting gene set into S1 signature set and S2 signature set by hierarchical clustering. All the genes in S1 were negatively correlated with HALLMARK_MYC_TARGETS_V1 pathway score and HR < 1, while S2 was all positively correlated with HALLMARK_MYC_TARGETS_V1 pathway score and HR > 1. Therefore, S1 signature was thought to be associated with MYC signaling inhibition, and S2 signature is associated with MYC signaling activation.

Classifying samples with consensus clustering

Samples were consistently clustered using the "ConsensusClusterPlus" package in R (Qiu et al., 2021), with the parameters set to: distance = "euclidean", clusterAlg = "km," maxK = 5, reps = 100, pItem = 0.8, and the remaining parameters took the default values. And sampels could be most distinctly classified when k = 2. After checking the expression level of the two signatures we previously identified in these two cluters, reasonably, we defined samples with highly expressed S1 signature genes as group C1 (MYC signaling inhibition group). Conversely, samples with highly expressed S2 signature genes were defined as group C2 (MYC signaling activation group).

Analysis of the genomic variability

We used "data_mutations_extended.txt" downloaded from ciBioPortal to analyse the mutation landscape of two clusters. Non-silenced SNV was analyzed using the "maftools" R package. We focused onMYC gene family (MYC, MYCN, MYCL) and pathway core genes and genes listed as cancer driver genes by OncoKB(https://www.oncokb.org/cancerGenes). The Fisher test of genes mutated in at least 30 samples were also performed using the mafCompare algorithm to yield genes with significant differences in mutation frequency in the two groups. Copy number variations of related genes were analysized using "data_CNA.txt" data from ciBioPortal. Among them, the CNV state of genes is divided into −2, −1, 0, 1, 2, and 0 represents no CNV, 1 and 2 represent copy number amplification, and -1 and −2 represent copy number loss. Copy number variation at chromosome level were directly extracted from the data_clinical_sample.txt, and only the top 10 most significantly variated chromosome arms between C1 and C2 clusters were visualized. All the statistics of genomic variability were performed with two-sided Fisher's exact test.

Description of the tumor microenvironment

The scores of 10 typical immune cells, including T cells, CD8 T cells, CTL, B cells, NK, and monocytes, were calculated using MCP-counter. From a previous study (Bagaev et al., 2021), data including purity, intratumor heterogeneity, aneuploidy score, homologous recombination defects, BCR.Shannon, TCR.Shannon, M1/M2 macrophage were obtained. To further evaluate the impact of MYC on the immune microenvironment, we used TIDE (http://tide.dfci.harvard.edu/) to calculate the scores of TIL for MDSC, CAF, and M2, as well as two indicators related to immunotherapy response: T-cell dysfunction and exclusion (Jiang et al., 2018).

Differential expression analysis of genes (including mRNA, lncRNA, miRNA) and the construction of CeRNA network

Using the "DESeq2" R package, the differential expression analysis was performed (Zhao et al., 2021). The threshold was set to adj.*p* value <0.001 and |log2FoldChange|> 0.5. And the resulting log2FC and adj.*p* value were used as the colors and sizes of the nodes in the subsequent network graph drawing, respectively. Circular nodes represents lncRNA, and square nodes represents miRNA. The selected lncRNA-miRNA interaction, MYC/MYCN and-ncRNA interaction, and miRNA-MYC/MYCN interaction were predicted using the online tool RNAInter (http://www.rnainter.org/) and mirWalk (http://mirwalk.umm.uni-heidelberg.de/). For the prediction results, the drawing was performed using the "igraph" R package (Mora and Donaldson, 2011). Gray lines represent all possible interactions between ncRNA and MYC/MYCN, and red lines indicate possible interactions between ncRNA.

Analysis of the scRNA-seq data

The expression matrix of scRNA-seq and the clinical information (such as histological type) of the samples were downloaded from the website (https://doi.org/10.24433/CO.0121060.v1) (Kumar and Song, 2022). The data contained a total of 114,489 cells from 10 LUAD samples and 10 normal lung tissue samples, and used 10x genomics for sequencing. Genes below expression in 100 cells were filtered out using the "Seurat" R package (Kumar and Song, 2022). Low-quality cells were filtered out by the criteria where the number of expressed genes was greater than 100 and less than 6,000 and the proportion of mitochondrial gene expression was less than 20. After defining and isolating epithelial cells from single cell expression profiling data of total cells, again, samples with less than 100 epithelial cells were filtered out for subsequent analysis. The top 15 principal components were used after PCA dimension reduction. Eventually we obtained 3,684 normal epithelial cells from the normal samples, and 15,477 malignant epithelial cells from the tumor samples. The signature module score was calculated using the AddModuleScore function.

The interaction network of the Signature gene and immunohistochemical

Genes with significant differential expression between the two groups (C1 and C2 groups) and the genes belonging to S1/S2 signature set were included in the potential nodes. Associations between nodes were obtained by correlation analysis, and edges with lower associations were filtered out. The visualization was then performed using the "igraph" R package.Circle size represents the -log10 (*p*-value), and the circle color indicates the log2FoldChange for the difference analysis after C1/C2 grouping. Pictures of IHC staining derived from normal samples and LUAD samples were selected on the HPA website (https://www.proteinatlas.org/) to verify the relationship of key genes to cell differentiation. Here, "HPAanalyze" R package was used to download the high definition IHC pictures (Tran et al., 2019).

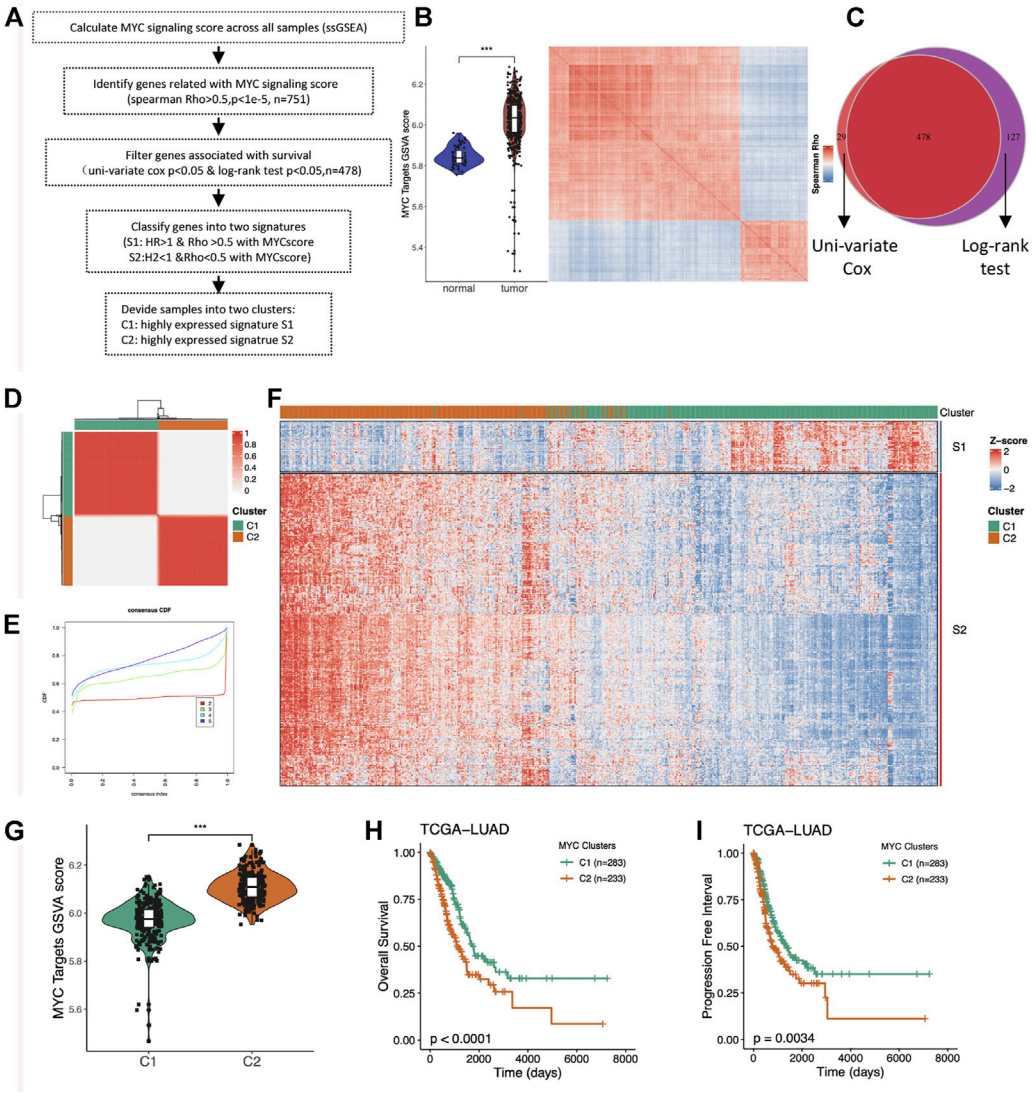


FIGURE 2
Identification of the two MYC signaling-associated signatures and unsupervised consistent clustering. **(A)** Flow chart of the sample classification. **(B)** Comparison of MYC Target GSVA scores between normal and tumor tissues in LUAD (right); Heatmap showing the hierarchical clustering of MYC Target GSVA scores-associated genes (left). **(C)** Venny plot showing survival-associated genes obtained by univariate cox analysis and log-rank test. **(D)** Unsupervised clustering divided the LUAD samples into two distinct subgroups (k = 2). **(E)** CDF and consensus index. **(F)** Heatmap showing the expression distribution of the Signature genes between the two clusters. **(G)** Comparison of MYC Target GSVA scores between the two distinct subgroups (C1 and C2). **(H)** Comparison of overall survival (OS) between C1 and C2. **(I)** Comparison of progress-free interval (PFI) between C1 and C2.

Statistical analysis

All statistical analysis was done using R. Where the KM survival analysis was performed by log rank test using the "survival" and "survminer" R packages, and the univariate and multivariate cox were done using the basis function coxph. We filtered out samples with less than 30 days of follow-up date before performing a survival analysis. Student's *t*-Test was used to compare the differences in gene expression levels between clusters. A *p*-value of less than 0.05 was considered statistically significant. Heatmaps were all plotted using the "ComplexHeatmap" R package.

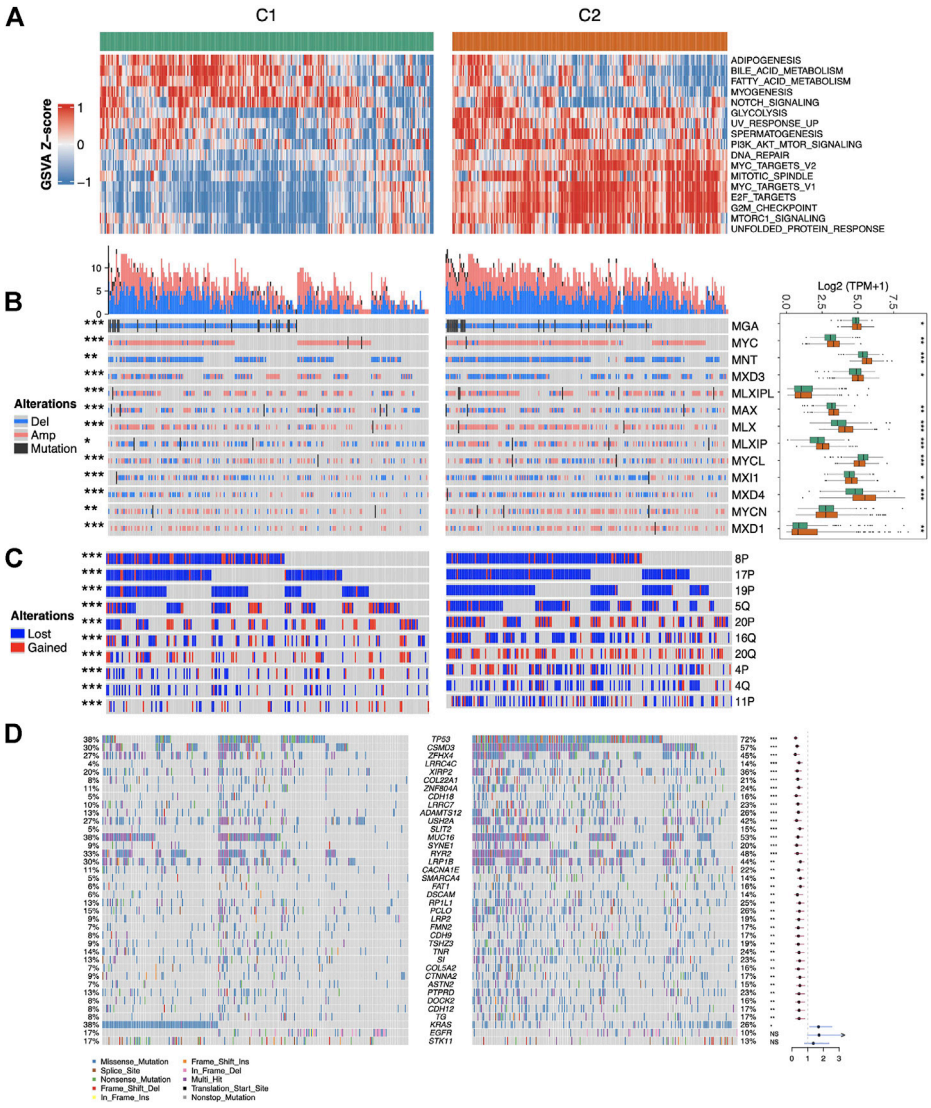


FIGURE 3
Association of MYC signaling with Hallmark pathways and genomic variations. **(A)** Heatmap of pathway scores with significant differences ($p < 0.001$) between C1 and C2. **(B)** Mutations, copy number variations and expression of core members of the MYC pathway between C1 and C2 samples. **(C)** Variation at top10 chromosome arm levels with significant differences between C1 and C2. **(D)** Waterfall plot showing the distribution of mutation characteristics of commonly mutated genes in the C1 and C2, and differently mutated genes between two groups by fisher test.

Results

Identification of the two MYC signaling-associated signatures

The level of MYC signaling activation cannot be simply judged by MYC gene expression and copy number amplification. Considering that the genes regulated by the same pathway are similar in their expression patterns, we assessed the degree of MYC signaling activation by looking at the overall expression levels of the MYC target genes, and obtained all the highly correlated genes by similarity analysis. We performed the subsequent analysis as to Figure 2A. We first evaluated the enrichment score of the MYC_TARGET_V1 pathway by ssGSEA algorithm, then found the genes highly related with the score through correlation analysis, and performed hierarchical clustering (Figure 2B). As we expected, these genes could be divided into two groups that were highly concordant, with one group being highly positively correlated with MYC_TARGET_V1 and the other group being highly negatively correlated. To further screen for key genes, we further filtered out 478 survival-related genes by

intersection using univariate cox analysis and log rank test of the obtained genes (Figure 2C; Supplementary Table S2). Importantly, both cluster genes negatively associated with MYC_TARGET_V1 were associated with better prognosis, and both genes positively associated with MYC_TARGET_V1 were associated with worse prognosis. Therefore, we defined these two cluster gene sets as S1 and S2, respectively.

The MYC signaling-associated signature could divide LUAD patients into two clinical clusters

Considering that the S1 and S2 genes have significantly different characteristics, we subsequently performed unsupervised consistent clustering of LUAD samples based on the expression of the signature genes, and finally obtained two clusters of samples (Figures 2D,E). One group of samples highly expressed the S1 signature gene, while the other group also highly expressed the S2 signature gene (Figure 2F), so we named it as the corresponding two C1 and C2 groups. Group C2 was the MYC signaling activation group, and group C1 was the MYC signaling inhibition group. The MYC scores were significantly different between the two groups (Figure 2G). In addition, we also found significant differences in OS (Figure 2H) and PFI (Figure 2I). This suggests important roles of MYC signaling in LUAD.

Association of MYC signaling with Hallmark pathways and genomic variations

The association of MYC signaling with oncogenic pathways and genomic variants remains unclear, therefore, we investigated the GSVA score differences in Hallmark pathways between MYC signaling activation (C2) and inhibition (C1) groups. As shown in Figure 3A, in addition to the MYC and cell-cycle-related pathways, the pathways such as glycolysis and PI3K were also up-regulated in C2. Copy number variation (CNV) in all MYC pathway core genes were significantly different between C1 and C2 (fisher exact test $p < 0.05$). Specifically, these genes developed CNV more frequently in C2, and MXD3 was both primarily lost in C2 and mostly amplified in C1. MLXIP was the opposite. This suggested that CNV changes were important causes of MYC pathway activation. Meanwhile, besides MLXIP and MYCN, other genes also differed in their expression between C1 and C2. Interestingly, although MXD3 experienced more copy number loss in C2, its expression remained higher in C2 (Figure 3B). We also examined CNV differences in chromosome levels between C1 and C2 (Figure 3C). Not surprisingly, multiple chromosomes-level CNV differences exist between C1 and C2. In addition to occurring more frequently in C2, the types of variants occurring also varied, such as 5q being more amplified in C1. In LUAD, mutations in many key genes play a crucial role in tumor development. They are known as the driver genes. We examined the mutation situation between C1 and C2. The results showed that besides KRAS, EGFR, STK11 (these genes were thought to be mutually exclusive to MYC pathway activation in previous studies (Zhang et al., 2016; Mollaoglu et al., 2017)), most genes were more mutated and higher in C2 (Figure 3D). Overall, the results of this study indicate that MYC signaling is closely related with oncogenic pathways and genomic variants.

MYC-signaling associates with genomic instability, mediates the immunosuppressive microenvironment, and promotes cell proliferation, and tumor stemness

In the above analysis, we found that the MYC signaling activation group was significantly different from the inhibition group in terms of genetic mutations. From this, we further investigated the differences in genomic instability scores between the two groups. We curated a list of genomic instability scores from a previous study (Bagaev et al., 2021), which was composed of the mutation burden score, the aneuploidy score, and the HRD score. The mutation burden score was non-silent mutations per Mb. The aneuploidy score reported the total number of arm-level amplifications and deletions and was computed using ABSOLUTE. Our results indicate that the MYC signaling activation group presents a higher genomic instability score than the MYC signaling inhibition group (Figure 4A). In addition, we also found that the MYC signaling activation group also showed higher intratumoral heterogeneity, IFN-gamma response and M1/M2 macrophages and lower TCR shannon, while the tumor purity and BCR shannon did not be significantly different between the two groups (Figure 4A). To analyze the effect of MYC signaling on the immune cells and the tumor microenvironment, we calculated the infiltration levels of the 10 immune cells using the MCP counter R package and performed a statistical test with a *t*-test (Figure 4B). We found that the cells mediating tumor killing ($CD8^+$ T cells, NK cells) had a higher infiltration abundance in the MYC signaling activation group. As important as the infiltration abundance of the immune cells in mediating the tumor immune response is the functional status of the immune cells, so we also evaluated indicators that reflect the immune function of LUAD with the online web tool TIDE (http://tide.dfci.harvard.edu/). The results showed that the tumor immune dysfunction score was significantly lower in the MYC signaling activation group when compared to the MYC signaling inhibition group (Figure 4C). This further suggests the importance of MYC signaling in mediating the tumor

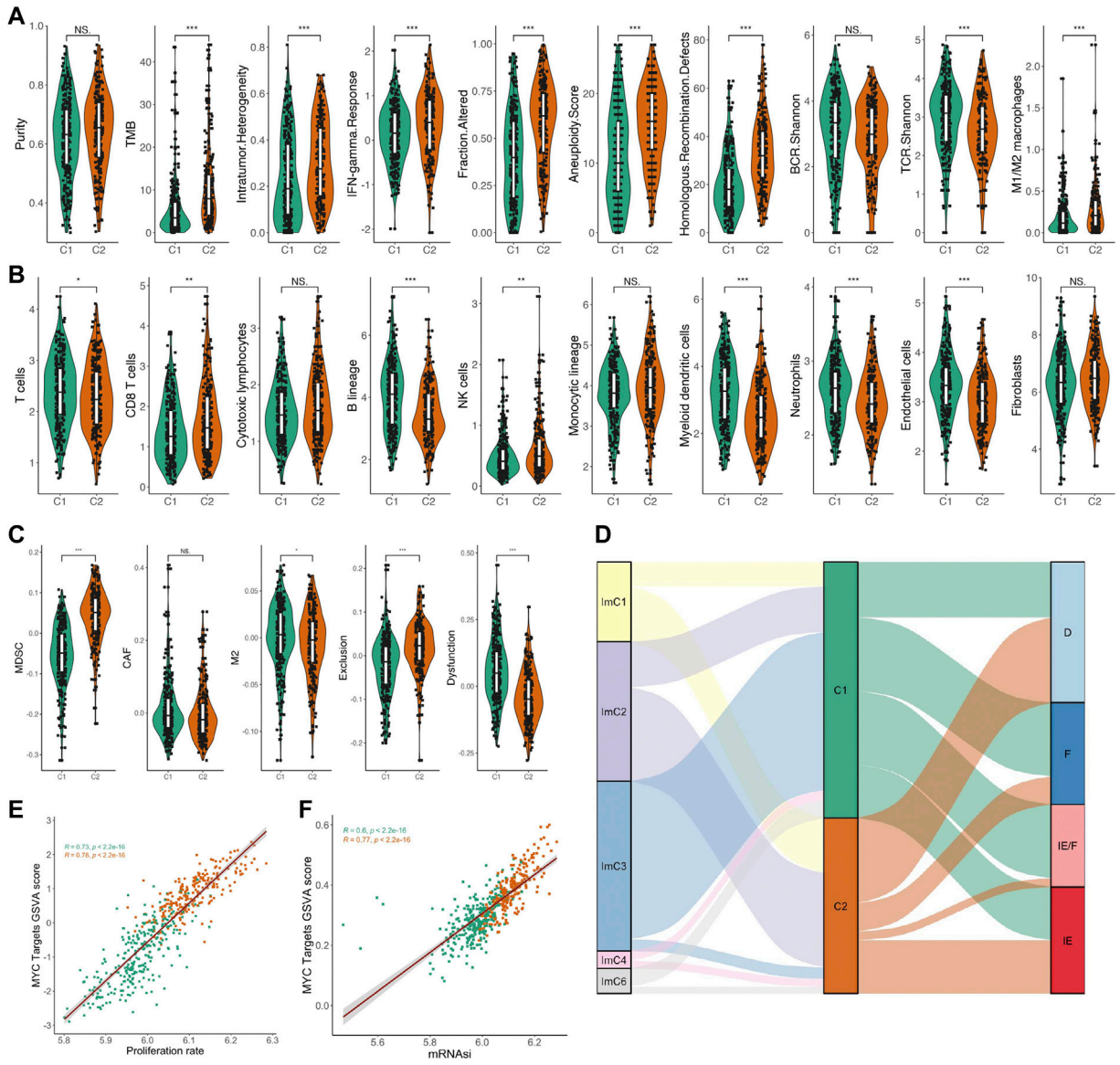


FIGURE 4
Relationship between MYC signaling and genomic instability score, immune microenvironment, cell proliferation, and tumor stemness. **(A)** Comparison of the purity, TMB, intratumor heterogeneity, IFN-gamma response, fraction altered, aneuploidy score, homologous recombination defects, BCR.Shannon, TCR.Shannon, M1/M2 macrophage between C1 and C2. **(B)** Comparison of the abundance of immune cell infiltration between C1 and C2. **(C)** Comparison of the scores of TIL for MDSC, CAF, and M2, as well as two indicators related to immunotherapy response: T-cell dysfunction and exclusion between C1 and C2. **(D)** Association of C1/C2 with two immune microenvironment types. **(E)** Correlation of the MYC Target GSVA score and cell proliferation. **(F)** Correlation of the MYC Target GSVA score and tumor stemness.

immunosuppressive microenvironment. For the tumor immune microenvironment, David Sacks et al. classified cancer samples into immune subtype in C1-C6 (Sacks et al., 2018). Similarly, Alexander Bagaev et al. defined the pan cancer sample of TCGA as four isoforms: IE, IE/F, D, and F (Bagaev et al., 2021). We explored the association between both C1/C2 groups and the tumor microenvironment of these two different differentiation methods. Coincidentally, our data suggest that ImC3 has a largely overlapping relationship with C1 (Figure 4D). This further highlights the association of MYC signaling with the tumor immunosuppressive microenvironment. Incidentally, we also explored the relationship between MYC signaling and cell proliferation and tumor stemness. Surprisingly, the MYC score had a significant correlation with both (Figures 4E,F).

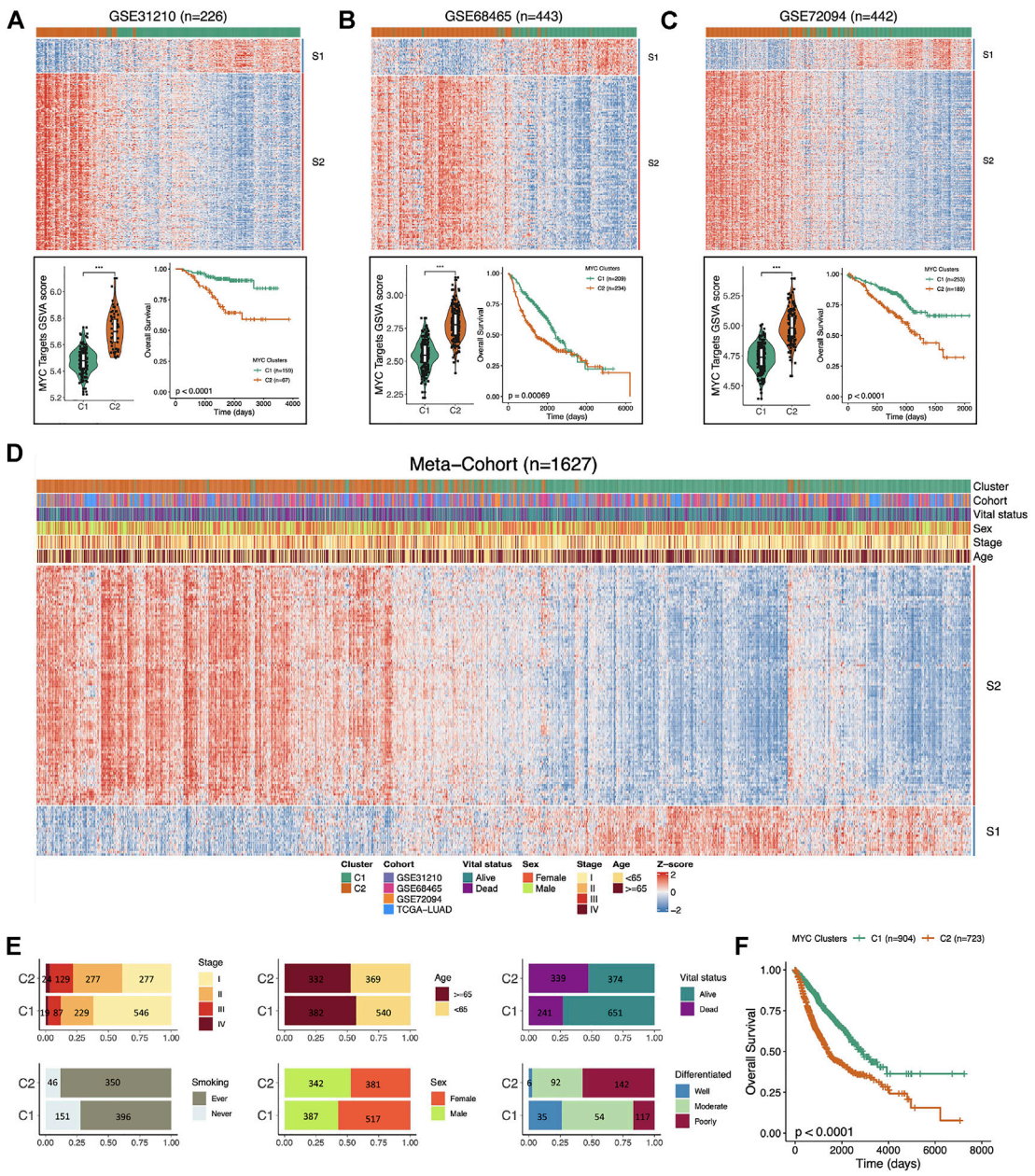


FIGURE 5
Independent validation of MYC-signaling grouping and prognosis. **(A)** The upper part the heatmap showing the expression distribution of the Signature genes between the two clusters in GSE31210. The Lower part: comparison of MYC Target GSVA scores between the two distinct subgroups (C1 and C2) in GSE31210 (right); Comparison of overall survival (OS) between C1 and C2 in GSE31210 (left). **(B)** The upper part: the heatmap showing the expression distribution of the Signature genes between the two clusters in GSE68465. The Lower part: comparison of MYC Target GSVA scores between the two distinct subgroups (C1 and C2) in GSE68465 (right); Comparison of overall survival (OS) between C1 and C2 in GSE68465 (left). **(C)** The upper part: the heatmap showing the expression distribution of the Signature genes between the two clusters in GSE72094. The Lower part: comparison of MYC Target GSVA scores between the two distinct subgroups (C1 and C2) in GSE72094 (right); Comparison of overall survival (OS) between C1 and C2 in GSE72094 (left). **(D)** The expression trend of signature genes and the distribution of the clinical characteristics of LUAD patients in the C1 and C2 in the meta-cohort (n = 1,627). **(E)** The distribution of the clinical characteristics (stage, smoking, age, sex, vital status, and grade) of LUAD patients in the C1 and C2. **(F)** Comparison of overall survival (OS) between C1 and C2 in the meta-cohort.

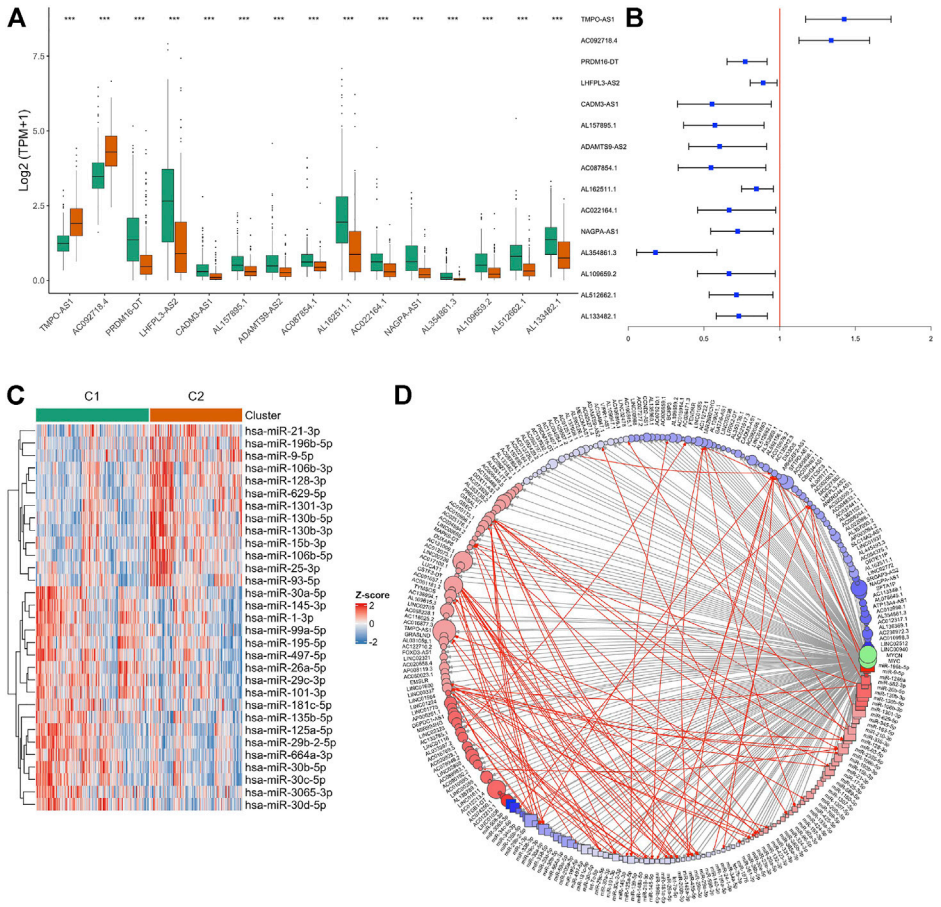


FIGURE 6
Identification of non-coding RNA associated with MYC signaling. **(A)** lncRNAs differentially expressed between the C1 and C2 groups. **(B)** Univariate Cox analysis revealed the relationship between these lncRNAs and prognosis. **(C)** miRNAs differentially expressed between the C1 and C2 groups. **(D)** Construction of CeRNA networks associated to MYC signaling. The resulting log2FC and adj.*p* value were used as the colors and sizes of the nodes in the subsequent network graph drawing, respectively. Circle represents lncRNA, and square represents miRNA.Gray lines represent all possible interactions between ncRNA and MYC/MYCN, and red lines indicate possible interactions between ncRNA.

Independent validation of MYC-signaling grouping and prognosis

To verify that the two signature (S1 and S2) we defined were stable on dividing LUAD samples into C1 and C2 groups according to MYC-signaling activation levels, we used three independent GEO datasets and a meta-cohort including 1,627 caces. The results showed that the MYC-signaling grouping was robust, which could efficiently classify samples into MYC-signaling activation group (C2) and MYC-signaling inhibition (C1), and were always highly correlated with patient prognosis (Figures 5A–D). Subsequently, we also investigated the distribution of clinical characteristics between the two groups, and we found that the MYC signaling activation group had more dead patients, who had later staging and poor cell differentiation, as shown in Figure 5E. Furthermore, Figure 5F also further confirmed that MYC C2 patients had a shorter OS.

MYC-signaling activation was subject to complex post-transcriptional regulation

In both the TCGA and GSE31210 data, some samples were still classified into the MYC-signaling inhibition group (C1) even with MYC experiencing copy number amplification. This suggested that MYC-signaling activation was complex regulated. Coincidentally, we found 15 lncRNAs in these two signature gene sets (S1 and S2), of which 13 belong to S1 and 2 belong to S2 (Figure 6A). And the univariate cox analysis suggested that they were all associated with prognosis (Figure 6B). In addition to lncRNA, miRNA may also play an important role in regulating MYC-signaling activation. Therefore, we also explored the differentially expressed miRNAs between the two groups (Figure 6C), and found 31 miRNAs were differentially expressed between groups. lncRNA and miRNA, mRNA may regulate gene expression

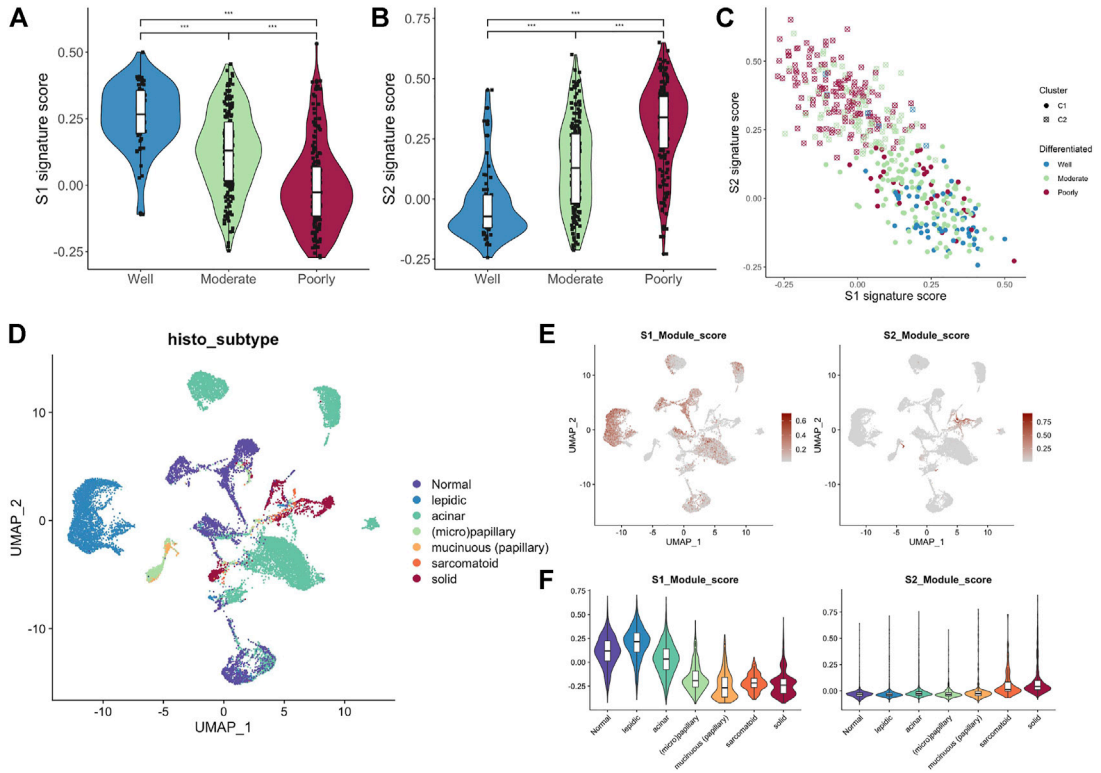


FIGURE 7
MYC-signaling is highly correlated with cell differentiation. **(A)** The relation between the S1 signature score and cell differentiation in GSE68465. **(B)** The relation between the S2 signature score and cell differentiation in GSE68465. **(C)** The connection between S1 signature score, S2 signature score, MYC signaling, and cell differentiation in GSE68465. **(D)** UMAP analysis identifies cell populations of different tissue subtypes **(E,F)** The average expression level of S1/S2 signature genes in each single cell was calculated by the AddMouduleScore algorithm.

through the CeRNA mechanism, and may also independently affect protein expression through other mechanisms such as acetylation. So we constructed a potential MYC/MYCN expression regulatory network (Figure 6D).

MYC-signaling was highly correlated with cell differentiation

Histologically, samples with highly differentiated tumor cells were highly concentrated in C1, while those with poorly differentiated cells were highly concentrated in C2. Moreover, the S1/S2 signature score can independently distinguish the tumor cell differentiation level (Figures 7A–C). To further test the significance of these two signature, epithelial cells from LUAD samples and normal lung tissue at different differentiation levels were isolated and analyzed separately. After dimensionality reduction by PCA and UMAP, we obtained 3,684 normal epithelial cells, and 15,477 malignant epithelial cells (Figure 7D). The average expression level of S1/S2 signature in each single cell was calculated by the AddMouduleScore algorithm (Figures 7E,F). The results were highly consistent with the previous findings. We found eight genes that were highly associated with cell differentiation were significantly differentially expressed in samples with different levels of differentiation in GSE68465 and showed consistent changes with the degree of differentiation (Figure 8A). Among them, CYP4B1, SUSD2, NFIX, and SYNE1 were highly expressed in normal lung epithelial cells and highly differentiated epithelial cells (Figures 8B–E bottom). The IHC staining also indicated that they had a higher expression in the normal (Figures 8B–E top) relative to the LUAD samples (Figures 8B–E middle). KPNA2, UBE2S, HMGA1, and RPL39L were highly expressed in poorly differentiated lung epithelial cells (Figures 9A–D bottom). The IHC staining also indicated that they had a lower expression in the normal (Figures 9A–D top) relative to the LUAD samples (Figures 9A–D middle). These results indicated that MYC-signaling was highly correlated with cancer cell differentiation.

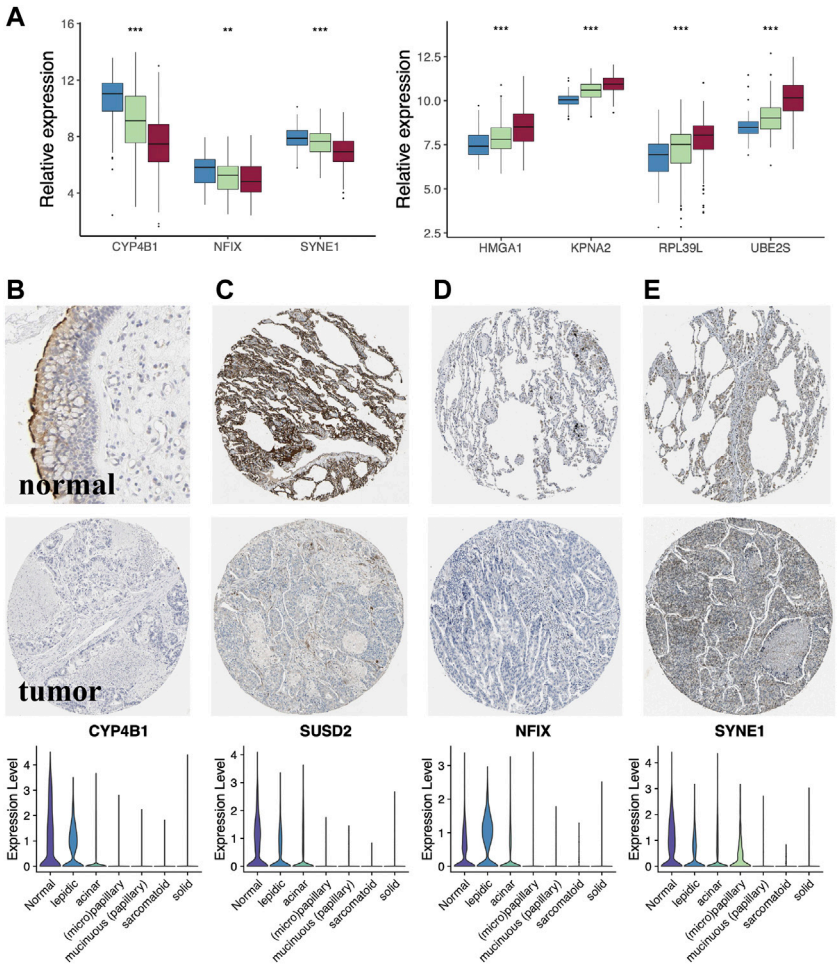


FIGURE 8
The key genes of MYC-signaling signature. **(A)** Genes highly associated with tumor cell differentiation. **(B–E)** Expression level of the key genes (CYP4B1, SUSD2, NFIX and SYNE1) in LUAD tumor tissues, normal tissues as well as in single cells with different degrees of differentiation.

Discussion

In this study, we examined the tumor MYC_TARGET_V1 score in multiple large LUAD cohorts, and its correlation with transcriptional profile expression, genomic instability, genetic alteration and regulation, immune microenvironment landscape, cell differentiation, and disease survival. MYC, acting as a transcription factor, and a slight disturbance of MYC expression may promote cancer cell evolution. To investigate the level of MYC signaling activation, we analyzed the expression levels of the MYC gene family and pathway core genes. We found that these genes were mostly significantly different between the two groups. Further investigating the copy number variation of the core MYC pathway genes between the two groups, we found that they did not show significant differences in the copy number variation. This implies that the activation of MYC signaling is epigenetically regulated, for example, DNA methylation (Panopoulos et al., 2017).It has been shown that the turnover of Myc proteins is determined by a cascade of phosphorylation and ubiquitination events (Liu et al., 2019; Parang et al., 2017).Notably, there is still a lack of evidence on whether MYC is regulated by ncRNA. In contrast, MYC, as a transcription factor, can regulate the activation and expression of ncRNA, for example, the miR-15 and let-7 (Adams and Eischen, 2016). In the study from Hou et al.(Zhang et al., 2022), they found that the MYC/MAX-trans-activated LINC00958 could promote the malignant behavior of LUAD by recruiting HOXA1 and inducing oncogenic reprogramming.To further clarify the pathways in which MYC is involved, we calculated the enrichment scores of the 50 Hallmark pathways in MsigDB by the ssGSEA algorithm, and found differences in multiple Hallmark pathway enrichment scores between the two groups, a finding that was also

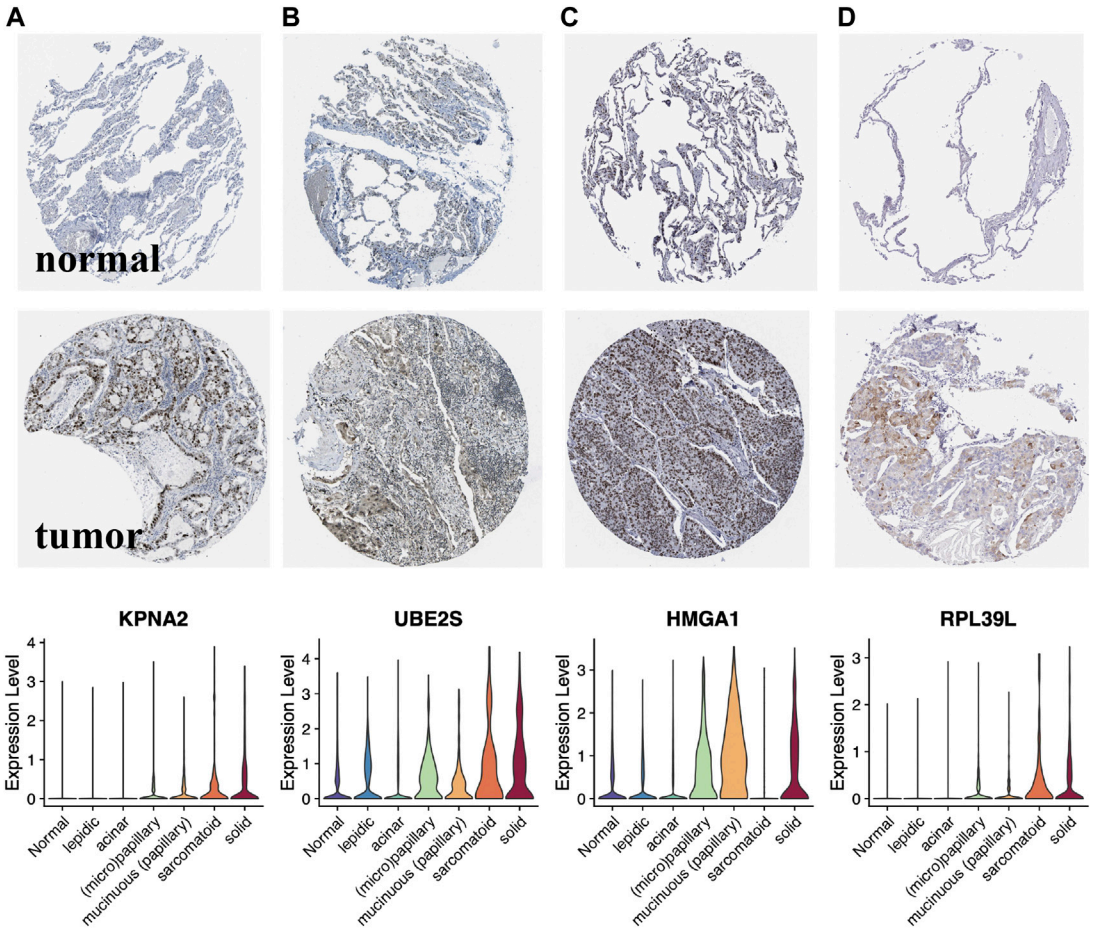


FIGURE 9
Lower expression genes in the normalrelative to the LUAD samples of MYC-signaling signature. **(A–D)** Expression level of the key genes (KPNA2, UBE2S, HMGA1 and RPL39L) in LUAD tumor tissues, normal tissues as well as in single cells with different degrees of differentiation.

consistent with previous studies. c-Myc, an important member of the MYC gene family, acts as a proto-oncogene localized to chromosome 8q24.1 and can be activated by chromosomal amplification, translocation, and rearrangement (Xu-Monette et al., 2016). In the above analysis, we also found that the MYC signaling activation group was significantly different from the inhibition group in terms of gene mutations. We further investigated the relationship between MYC signaling and the genome instability. We found that the MYC signaling activation group presented higher genomic instability scores. This result was not surprising, as reported in previous studies (Hao et al., 2016; H. Song et al., 2020).

More and more researchers have noticed the close link between tumor immune microenvironment and cancer occurrence and progression (Wang et al., 2016; Yuan et al., 2018; Tekpli et al., 2019). The MYC gene was also reported to be involved in the immune regulation of multiple tumors (Han et al., 2019; Swaminathan et al., 2020). In this study, we found that the cells mediating the tumor-killing effect had a higher infiltration abundance in the MYC signaling activation group. Further investigating the functional status of the immune cells, we found that the tumor immune dysfunction score was significantly lower in the MYC signaling activation group as compared to the MYC signaling inhibition group. This further suggests the importance of MYC signaling in mediating the tumor immunosuppressive microenvironment.Moreover, the association of MYC with immune checkpoints is also slowly being revealed. For example, Thongsuksai et al. (Sunpaweravong et al., 2022)found that NSCLC tissues significantly express more c-Myc and PD-L1 compared to the matched normal respiratory epithelium, highlighting the important role of these key drivers in tumorigenesis. Laura Soucek and his colleagues (Masso-Valles et al., 2020) suggested that MYC, MYCL and MYCN might be therapeutic targets for lung cancer and that elevated Myc levels were also associated with treatment resistance, there may be significant opportunities for the combination of Myc inhibitors

with immunotherapies. It is well known that cancer occurrence is closely associated with the uncontrolled clonal proliferation of cells (Chung et al., 2019). As a well-known prooncogenic gene, MYC has been reported in mediating cell proliferation (Feist et al., 2018). However, its relationship between it and cell proliferation and tumor stemness in LUAD also needs to be further clarified. Our study showed a significant positive correlation between cell proliferation rate as well as tumor stemness and MYC score, and further highlights its non-negligible role in regulating LUAD cell proliferation and maintaining tumor stemness. Previous studies (Ireland et al., 2020; Patel et al., 2021) have revealed the key role of MYC in small cell lung cancer (SCLC) from a genomics perspective. Trudy G. Oliver et al. (Ireland et al., 2020) defined different SCLC molecular isoforms, based on the expression of ASCL1, NEUROD1, POU2F3, or YAP1. They used mouse and human models with time-series single-cell transcriptomic analysis to reveal the dynamic evolution of MYC-driven SCLC isoforms, finding that in neuroendocrine cells, MYC activated Notch to dedifferentiate tumor cells, promoting the temporal transition of SCLC from ASCL1 + to NEUROD1 + to YAP1 + state. The study by Hideo Watanabe and his colleagues (Patel et al., 2021)has also revealed the previously undescribed roles of the historically defined general oncogenes c-Myc and L-Myc for regulating lineage plasticity across molecular subtypes and histological subclasses. From the data currently available, MYC in SCLC seems to be studied more fully compared with LUAD. Therefore, it is still important to further investigate the potential role of MYC in LUAD from multi-omics data.

Overall, we used information from up to 1,600 samples of multiple LUAD cohorts to represent the important role of MYC signaling in LUAD from multiple dimensions of transcriptional profile expression, genomic instability, genetic alteration and regulation, immune microenvironment landscape, cell differentiation, and disease survival.This provides a valuable reference for deeply revealing the mechanism of cancer-promoting action of MYC in LUAD. However, like many other studies, the present study has some limitations. First, this study was a retrospective study and it was difficult to completely eliminate selective bias; second, although the important role of MYC in LUAD was described from multiple perspectives using multiple large study cohorts of LUAD and multiple bioinformatics approaches, further validation of the underlying experiments was lacking.

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

LH and QZ conceived and designed the study. LH and QZ analyzed the data. LH and QZ wrote and edited the manuscript. XC and QC participated in the study design and provided critical insights. All authors have approved the manuscript.

Funding

This work was supported by Shenzhen Baoan Shiyan People's Hospital Fund (2020SY07) and Science and Technology Bureau of Baoan (2021JD067).

Acknowledgments

The authors hereby express their gratitude to all participants who supported the study, especially the database providers who provided the data for the analysis.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1021978/full#supplementary-material

References

Adams, C. M., and Eischen, C. M. (2016). Histone deacetylase inhibition reveals a tumor-suppressive function of MYC-regulated miRNA in breast and lung carcinoma. *Cell Death Differ.* 23 (8), 1312–1321. doi:10.1038/cdd.2016.9

Bagaev, A., Kotlov, N., Nomie, K., Svekolkin, V., Gafurov, A., Isaeva, O., et al. (2021). Conserved pan-cancer microenvironment subtypes predict response to immunotherapy. *Cancer Cell* 39 (6), 845–865.e7. doi:10.1016/j.ccell.2021.04.014

Becht, E., McInnes, L., Healy, J., Dutertre, C. A., Kwok, I., Ng, L. G., et al. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* 37, 38–44. doi:10.1038/nbt.4314

Cheng, G., Zhang, Q., Pan, J., Lee, Y., Ouari, O., Hardy, M., et al. (2019). Targeting lonidamine to mitochondria mitigates lung tumorigenesis and brain metastasis. *Nat. Commun.* 10 (1), 2205. doi:10.1038/s41467-019-10042-1

Chung, S., Vail, P., Witkiewicz, A. K., and Knudsen, E. S. (2019). Coordinately targeting Cell-Cycle checkpoint functions in integrated models of pancreatic cancer. *Clin. Cancer Res.* 25 (7), 2290–2304. doi:10.1158/1078-0432.CCR-18-1620

Feist, M., Schwarzfischer, P., Heinrich, P., Sun, X., Kemper, J., von Bonin, F., et al. (2018). Cooperative STAT/NF-κB signaling regulates lymphoma metabolic reprogramming and aberrant GOT2 expression. *Nat. Commun.* 9 (1), 1514. doi:10.1038/s41467-018-03803-x

Han, H., Jain, A. D., Truica, M. I., Izquierdo-Ferrer, J., Anker, J. F., Lysy, B., et al. (2019). Small-molecule MYC inhibitors suppress tumor growth and enhance immunotherapy. *Cancer Cell* 36 (5), 483–497. doi:10.1016/j.ccell.2019.10.001

Hao, T., Gaerig, V. C., and Brooks, T. A. (2016). Nucleic acid clamp-mediated recognition and stabilization of the physiologically relevant MYC promoter G-quadruplex. *Nucleic Acids Res.* 44 (22), 11013–11023. doi:10.1093/nar/gkw1006

Ireland, A. S., Micinski, A. M., Kastner, D. W., Guo, B., Wait, S. J., Spainhower, K. B., et al. (2020). MYC drives temporal evolution of small cell lung cancer subtypes by reprogramming neuroendocrine fate. *Cancer Cell* 38 (1), 60–78. doi:10.1016/j.ccell.2020.05.001

Jiang, A., Wang, J., Liu, N., Zheng, X., Li, Y., Ma, Y., et al. (2022). Integration of Single-Cell RNA sequencing and bulk RNA sequencing data to establish and validate a prognostic model for patients with lung adenocarcinoma. *Front. Genet.* 13, 833797. doi:10.3389/fgene.2022.833797

Jiang, P., Gu, S., Pan, D., Fu, J., Sahu, A., Hu, X., et al. (2018). Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nat. Med.* 24 (10), 1550–1558. doi:10.1038/s41591-018-0136-1

Jing, H., Hu, J., He, B., Negron, A. Y., Stupinski, J., Weiser, K., et al. (2016). A SIRT2-Selective inhibitor promotes c-Myc oncoprotein degradation and exhibits broad anticancer activity. *Cancer Cell* 29 (4), 607. doi:10.1016/j.ccell.2016.03.011

Kim, J. W., Marquez, C. P., Kostyrko, K., Koehne, A. L., Marini, K., Simpson, D. R., et al. (2019). Antitumor activity of an engineered decoy receptor targeting CLCF1-CNTFR signaling in lung adenocarcinoma. *Nat. Med.* 25 (11), 1783–1795. doi:10.1038/s41591-019-0612-2

King, B., Boccalatte, F., Moran-Crusio, K., Wolf, E., Wang, J., Kayembe, C., et al. (2016). The ubiquitin ligase Huwe1 regulates the maintenance and lymphoid commitment of hematopoietic stem cells. *Nat. Immunol.* 17 (11), 1312–1321. doi:10.1038/ni.3559

Kumar, S., and Song, M. (2022). Overcoming biases in causal inference of molecular interactions. *Bioinformatics* 38, 2818–2825. doi:10.1093/bioinformatics/btac206

Lee, J. K., Phillips, J. W., Smith, B. A., Park, J. W., Stoyanova, T., McCaffrey, E. F., et al. (2016). N-Myc drives neuroendocrine prostate cancer initiated from human prostate epithelial cells. *Cancer Cell* 29 (4), 536–547. doi:10.1016/j.ccell.2016.03.001

Liu, P. Y., Tee, A. E., Milazzo, G., Hannan, K. M., Maag, J., Mondal, S., et al. (2019). The long noncoding RNA lncNB1 promotes tumorigenesis by interacting with ribosomal protein RPL35. *Nat. Commun.* 10 (1), 5026. doi:10.1038/s41467-019-12971-3

Masso-Valles, D., Beaulieu, M. E., and Soucek, L. (2020). MYC, MYCL, and MYCN as therapeutic targets in lung cancer. *Expert Opin. Ther. Targets* 24 (2), 101–114. doi:10.1080/14728222.2020.1723548

Mok, T., Wu, Y. L., Kudaba, I., Kowalski, D. M., Cho, B. C., Turna, H. Z., et al. (2019). Pembrolizumab versus chemotherapy for previously untreated, PD-L1-expressing, locally advanced or metastatic non-small-cell lung cancer (KEYNOTE-042): A randomised, open-label, controlled, phase 3 trial. *Lancet* 393 (10183), 1819–1830. doi:10.1016/S0140-6736(18)32409-7

Mollaoglu, G., Guthrie, M. R., Bohm, S., Bragelmann, J., Can, I., Ballieu, P. M., et al. (2017). MYC drives progression of small cell lung cancer to a variant neuroendocrine subtype with vulnerability to aurora kinase inhibition. *Cancer Cell* 31 (2), 270–285. doi:10.1016/j.ccell.2016.12.005

Mora, A., and Donaldson, I. M. (2011). IRefR: An R package to manipulate the iRefIndex consolidated protein interaction database. *BMC Bioinforma.* 12, 455. doi:10.1186/1471-2105-12-455

Okayama, H., Kohno, T., Ishii, Y., Shimada, Y., Shiraishi, K., Iwakawa, R., et al. (2012). Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. *Cancer Res.* 72 (1), 100–111. doi:10.1158/0008-5472.CAN-11-1403

Panopoulos, A. D., Smith, E. N., Arias, A. D., Shepard, P. J., Hishida, Y., Modesto, V., et al. (2017). Aberrant DNA methylation in human iPSCs associates with MYC-Binding motifs in a Clone-Specific manner independent of genetics. *Cell Stem Cell* 20 (4), 505–517. doi:10.1016/j.stem.2017.03.010

Parang, B., Kaz, A. M., Barrett, C. W., Short, S. P., Ning, W., Keating, C. E., et al. (2017). BVES regulates c-Myc stability via PP2A and suppresses colitis-induced tumourigenesis. *Gut* 66 (5), 852–862. doi:10.1136/gutjnl-2015-310255

Patel, A. S., Yoo, S., Kong, R., Sato, T., Sinha, A., Karam, S., et al. (2021). Prototypical oncogene family Myc defines unappreciated distinct lineage states of small cell lung cancer. *Sci. Adv.* 7 (5), eabc2578. doi:10.1126/sciadv.abc2578

Poh, B., Koso, H., Momota, H., Komori, T., Suzuki, Y., Yoshida, N., et al. (2019). Foxr2 promotes formation of CNS-embryonal tumors in a Trp53-deficient background. *Neuro. Oncol.* 21 (8), 993–1004. doi:10.1093/neuonc/noz067

Qiu, C., Shi, W., Wu, H., Zou, S., Li, J., Wang, D., et al. (2021). Identification of molecular subtypes and a prognostic signature based on Inflammation-Related genes in colon adenocarcinoma. *Front. Immunol.* 12, 769685. doi:10.3389/fimmu.2021.769685

Rajagopalan, D., Tirado-Magallanes, R., Bhatia, S. S., Teo, W. S., Sian, S., Hora, S., et al. (2018). TIP60 represses activation of endogenous retroviral elements. *Nucleic Acids Res.* 46 (18), 9456–9470. doi:10.1093/nar/gky659

Sacks, D., Baxter, B., Campbell, B., Carpenter, J. S., Cognard, C., Dippel, D., et al. (2018). Multisociety consensus quality improvement revised consensus statement for endovascular therapy of acute ischemic stroke. *Int. J. Stroke* 13 (6), 612–632. doi:10.1177/1747493018778713

Schabath, M. B., Welsh, E. A., Fulp, W. J., Chen, L., Teer, J. K., Thompson, Z. J., et al. (2016). Differential association of STK11 and TP53 with KRAS mutation-associated gene expression, proliferation and immune surveillance in lung adenocarcinoma. *Oncogene* 35 (24), 3209–3216. doi:10.1038/onc.2015.375

Shedden, K., Taylor, J. M., Enkemann, S. A., Tsao, M. S., Yeatman, T. J., Gerald, W. L., et al. (2008). Gene expression-based survival prediction in lung adenocarcinoma: A multi-site, blinded validation study. *Nat. Med.* 14 (8), 822–827. doi:10.1038/nm.1790

Song, C., Guo, Z., Yu, D., Wang, Y., Wang, Q., Dong, Z., et al. (2020). A prognostic nomogram combining Immune-Related gene signature and clinical factors predicts survival in patients with lung adenocarcinoma. *Front. Oncol.* 10, 1300. doi:10.3389/fonc.2020.01300

Song, C., Lu, Z., Lai, K., Li, D., Hao, B., Xu, C., et al. (2022). Identification of an inflammatory response signature associated with prognostic stratification and drug sensitivity in lung adenocarcinoma. *Sci. Rep.* 12 (1), 10110. doi:10.1038/s41598-022-14323-6

Song, C., Wu, Z., Wang, Q., Wang, Y., Guo, Z., Li, S., et al. (2021). A combined Two-mRNA signature associated with PD-L1 and tumor mutational burden for prognosis of lung adenocarcinoma. *Front. Cell Dev. Biol.* 9, 634697. doi:10.3389/fcell.2021.634697

Song, H., Liu, D., Dong, S., Zeng, L., Wu, Z., Zhao, P., et al. (2020). Epitranscriptomics and epiproteomics in cancer drug resistance: Therapeutic implications. *Signal Transduct. Target. Ther.* 5 (1), 193. doi:10.1038/s41392-020-00300-w

Sunpaweravong, P., Thongwatchara, P., Chotipanvithayakul, R., Sangkhathat, S., and Thongsuksai, P. (2022). Expression and prognostic significance of c-Myc, ALK, ROS1, BRAF, and PD-L1 among patients with Non-Small cell lung cancer. *Clin. Med. Insights. Oncol.* 16, 11795549221092747. doi:10.1177/11795549221092747

Swaminathan, S., Hansen, A. S., Heftdal, L. D., Dhanasekaran, R., Deutzmann, A., Fernandez, W., et al. (2020). MYC functions as a switch for natural killer cell-mediated immune surveillance of lymphoid malignancies. *Nat. Commun.* 11 (1), 2860. doi:10.1038/s41467-020-16447-7

Tekpli, X., Lien, T., Rossevold, A. H., Nebdal, D., Borgen, E., Ohnstad, H. O., et al. (2019). An independent poor-prognosis subtype of breast cancer defined by a distinct tumor immune microenvironment. *Nat. Commun.* 10 (1), 5499. doi:10.1038/s41467-019-13329-5

Thompson, E. M., Keir, S. T., Venkatraman, T., Lascola, C., Yeom, K. W., Nixon, A. B., et al. (2017). The role of angiogenesis in Group 3 medulloblastoma

pathogenesis and survival. *Neuro. Oncol.* 19 (9), 1217–1227. doi:10.1093/neuonc/nox033

Tran, A. N., Dussaq, A. M., Kennell, T. J., Willey, C. D., and Hjelmeland, A. B. (2019). HPAanalyze: An R package that facilitates the retrieval and analysis of the Human Protein Atlas data. *BMC Bioinforma.* 20 (1), 463. doi:10.1186/s12859-019-3059-z

Vo, B. T., Wolf, E., Kawauchi, D., Gebhardt, A., Rehg, J. E., Finkelstein, D., et al. (2016). The interaction of myc with miz1 defines medulloblastoma subgroup identity. *Cancer Cell* 29 (1), 5–16. doi:10.1016/j.ccell.2015.12.003

Wang, S. Z., Poore, B., Alt, J., Price, A., Allen, S. J., Hanaford, A. R., et al. (2019). Unbiased metabolic profiling predicts sensitivity of high MYC-Expressing atypical Teratoid/Rhabdoid tumors to glutamine inhibition with 6-Diazo-5-Oxo-L-Norleucine. *Clin. Cancer Res.* 25 (19), 5925–5936. doi:10.1158/1078-0432.CCR-19-0189

Wang, W., Kryczek, I., Dostal, L., Lin, H., Tan, L., Zhao, L., et al. (2016). Effector t cells abrogate Stroma-Mediated chemoresistance in ovarian cancer. *Cell* 165 (5), 1092–1105. doi:10.1016/j.cell.2016.04.009

Xu-Monette, Z. Y., Deng, Q., Manyam, G. C., Tzankov, A., Li, L., Xia, Y., et al. (2016). Clinical and biologic significance of MYC genetic mutations in de novo diffuse large b-cell lymphoma. *Clin. Cancer Res.* 22 (14), 3593–3605. doi:10.1158/1078-0432.CCR-15-2296

Yamauchi, M., Yamaguchi, R., Nakata, A., Kohno, T., Nagasaki, M., Shimamura, T., et al. (2012). Epidermal growth factor receptor tyrosine kinase defines critical prognostic genes of stage I lung adenocarcinoma. *PLoS One* 7 (9), e43923. doi:10.1371/journal.pone.0043923

Yuan, J., Levitin, H. M., Frattini, V., Bush, E. C., Boyett, D. M., Samanamud, J., et al. (2018). Single-cell transcriptome analysis of lineage diversity in high-grade glioma. *Genome Med.* 10 (1), 57. doi:10.1186/s13073-018-0567-9

Zhang, J., Song, C., Tian, Y., and Yang, X. (2021). Single-Cell RNA sequencing in lung cancer: Revealing phenotype shaping of stromal cells in the microenvironment. *Front. Immunol.* 12, 802080. doi:10.3389/fimmu.2021.802080

Zhang, T., Guo, L., Creighton, C. J., Lu, Q., Gibbons, D. L., Yi, E. S., et al. (2016). A genetic cell context-dependent role for ZEB1 in lung cancer. *Nat. Commun.* 7, 12231. doi:10.1038/ncomms12231

Zhang, T., Su, F., Lu, Y. B., Ling, X. L., Dai, H. Y., Yang, T. N., et al. (2022). MYC/MAX-Activated LINC00958 promotes lung adenocarcinoma by oncogenic transcriptional reprogramming through HOXA1 activation. *Front. Oncol.* 12, 807507. doi:10.3389/fonc.2022.807507

Zhao, Y., Ma, T., and Zou, D. (2021). Identification of unique transcriptomic signatures and hub genes through RNA sequencing and integrated WGCNA and PPI network analysis in nonerosive reflux disease. *J. Inflamm. Res.* 14, 6143–6156. doi:10.2147/JIR.S340452

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 20 October 2022
DOI 10.3389/fgene.2022.960567

OPEN ACCESS

EDITED BY
Geng Chen,
Stemirna Therapeutics Co., Ltd., China

REVIEWED BY
Feng Jiang,
Fudan University, China
Xiangyi Kong,
Chinese Academy of Medical Sciences and Peking Union Medical College, China

*CORRESPONDENCE
Wei Cui,
cui123@cicams.ac.cn

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

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

RECEIVED 03 June 2022
ACCEPTED 24 August 2022
PUBLISHED 20 October 2022

CITATION
Gao S, Wu X, Lou X and Cui W (2022), Identification of a prognostic risk-scoring model and risk signatures based on glycosylation-associated cluster in breast cancer.
Front. Genet. 13:960567.
doi: 10.3389/fgene.2022.960567

Identification of a prognostic risk-scoring model and risk signatures based on glycosylation-associated cluster in breast cancer

Shengnan Gao[1†], Xinjie Wu[2,3,4†], Xiaoying Lou[1] and Wei Cui[1*]

[1]Department of Clinical Laboratory, National Cancer Center/National Clinical Research Center for Cancer/ State Key Laboratory of Molecular Oncology, Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China, [2]Peking University China-Japan Friendship School of Clinical Medicine, Beijing, China, [3]Department of Orthopedic Surgery, China-Japan Friendship Hospital, Beijing, China, [4]Department of Molecular Medicine and Surgery, Center for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden

Breast cancer is a heterogeneous disease whose subtypes represent different histological origins, prognoses, and therapeutic sensitivity. But there remains a strong need for more specific biomarkers and broader alternatives for personalized treatment. Our study classified breast cancer samples from The Cancer Genome Atlas (TCGA) into three groups based on glycosylation-associated genes and then identified differentially expressed genes under different glycosylation patterns to construct a prognostic model. The final prognostic model containing 23 key molecules achieved exciting performance both in the TCGA training set and testing set GSE42568 and GSE58812. The risk score also showed a significant difference in predicting overall clinical survival and immune infiltration analysis. This work helped us to understand the heterogeneity of breast cancer from another perspective and indicated that the identification of risk scores based on glycosylation patterns has potential clinical implications and immune-related value for breast cancer.

KEYWORDS

breast cancer, glycosylation, prognosis, subtype, biomarkers, immune

Introduction

Breast cancer has reached the highest incidence in women's cancer types, and its lethality has reached second place, followed by lung cancer (Sung et al., 2021). As a heterogeneous disease, breast cancer's multiple subtypes represent different histological origins, prognoses, and therapeutic sensitivity (Perou et al., 2000; Cancer Genome Atlas Network, 2012; Curtis et al., 2012; Marusyk et al., 2012). The pathological markers estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor 2 (HER2) stratified patients with various treatment selecting, such as hormonal therapy

(e.g., Tamoxifen) and HER2-targeted therapy (e.g., Trastuzumab) (Goldhirsch et al., 2013). Of note, HER2 is characterized by poor prognosis and has multiple sites of N-glycosylation, whose presence is linked with function (Peiris et al., 2017). Subsequently, intrinsic molecular subtyping based on expression profile highlights the intricate complexity of this cancer type and the importance of genomic/transcriptomic analyses for prognostic prediction. PAM50 utilizes a 50 genes system that classifies breast cancer into luminal A, luminal B, HER2-enriched, and basal-like subtype that involves not only diverse biological processes but also has prognostic significance (Prat et al., 2012; Prat et al., 2015). The highly heterogeneous of breast cancer requires a strong need for more specific biomarkers and broader alternatives for personalized treatment. Meanwhile, efforts to classify established histological subtypes have been carried out, which identified at least four distinct subtypes of ER-negative and six triple-negative subtypes (Teschendorff et al., 2007; Lehmann et al., 2011). According to recent reports, researchers are seeking a multi-angle classification approach to identify diversified functional clustering and signatures, such as glycolysis (Zhang et al., 2020a; Jiang et al., 2021), autophagy (Zhang et al., 2020b; Jiang et al., 2022), ferroptosis (Wang et al., 2021), stemness (Li et al., 2020), and immune microenvironment (Shen et al., 2020). All these attempts allow us to make more defined and precise characterizations based on new parameters to drive the heterogeneity landscape of breast cancer and put forward new ideas in prognostic prediction and treatment in the future.

Glycosylation is defined as a biosynthetic enzymatic process characterized by the covalent attachment of single sugar or glycans to a wide range of target proteins (Pinho and Reis, 2015; Eichler, 2019). As a post-translational modification, they play an essential role in almost all aspects of the life processes of cells, such as cell cycle, proliferation, and aging (Mallard and Tiralongo, 2017; Gudelj et al., 2018; Gao et al., 2021). The glycosylation pattern is profoundly altered during tumorigenesis. Among them, O-glycan truncation, sialylation, fucosylation, and N-glycan branching are common types of glycosylation in cancer (Drake et al., 2015; Kölbl et al., 2015; Kudelka et al., 2015; Taniguchi and Kizuka, 2015), leading to the occurrence of malignant phenotypes such as cell adhesion, metastasis, epithelial–mesenchymal transitioning, and even the shifting of the tumor microenvironment (Günthert et al., 1991; Rabinovich and Toscano, 2009; Pinho et al., 2011; Paredes et al., 2012; Pinho et al., 2013). Researchers have also identified glycosylation-related molecules as biomarkers for cancer diagnosis and prognostics evaluation. For instance, prostate-specific antigen (PSA) in prostate cancer (Gilgunn et al., 2013), carcinoma antigen 125 (CA125/MUC16) in ovarian cancer (Zurawski et al., 1988), CA19-9 and carcinoembryonic antigen (CEA) in colon cancer (Goldstein and Mitchell, 2005), and aberrantly glycosylated MUC1 (also known as CA15-3) in breast cancer (Kumpulainen et al., 2002). More recent studies have mapped the histopathological orientation and tissue distribution of N-linked glycans in clinical breast cancer tissues (Scott et al., 2019a; Scott et al., 2019b), which deepen the understanding of the heterogeneity of breast cancer from the perspective of glycosylation.

Our study classified breast cancer samples from The Cancer Genome Atlas (TCGA) into three groups based on glycosylation-associated genes and then identified differentially expressed genes under different glycosylation patterns to construct a prognostic model. Finally, a model containing 23 risk signatures was built and performed favorable predicting efficacy in training and testing cohorts, and the evaluation of immune infiltration and immunotherapy response were analyzed as well.

Results

Classification of BRCA based on the glycosylation-related gene sets

Figure 1 shows the workflow of our study. The TCGA column of Table 1. We classified TCGA-BRCA samples (n = 1,104) based on 179 glycosylation-related genes (GRGs) performed by consensus clustering analysis. Related clustering parameters are shown in Figures 2A–C, Supplementary Figure S1A, and Supplementary Figure S2A. Considering the complexity of grouping and the feasibility of subsequent analysis, we choose the optimal grouping when k = 3. Thus, we obtain three glycosylation-based clusters. We used t-SNE (Figure 2D) and PCA (Supplementary Figure S2B) dimensional reduction methods to observe that the samples had favorable overall differences under this grouping. Cluster 3 exhibited shorter overall survival (OS), indicating a poorer prognosis compared with clusters 1 and 2. ($p < 0.05$) (Figure 2E). In brief, this grouping method based on intracellular glycosylation status has specific differences in breast cancer samples and has substantial clinical value.

Screening of differentially expressed genes

We classified BRCA tumor samples into three clusters based on glycosylation patterns. Next, we screened the DEGs of these three clusters using the "Deseq2" R package. Supplementary Figures S2C–E show the PCA map and DEGs heatmap between the three clusters. Figure 4 shows the differential analysis volcano plot of group 1 to group 2 (Figure 4A), group 2 to group 3 (Figure 4B), and group 1 to group 3 (Figure 4C). We made a Venn diagram for the three groups

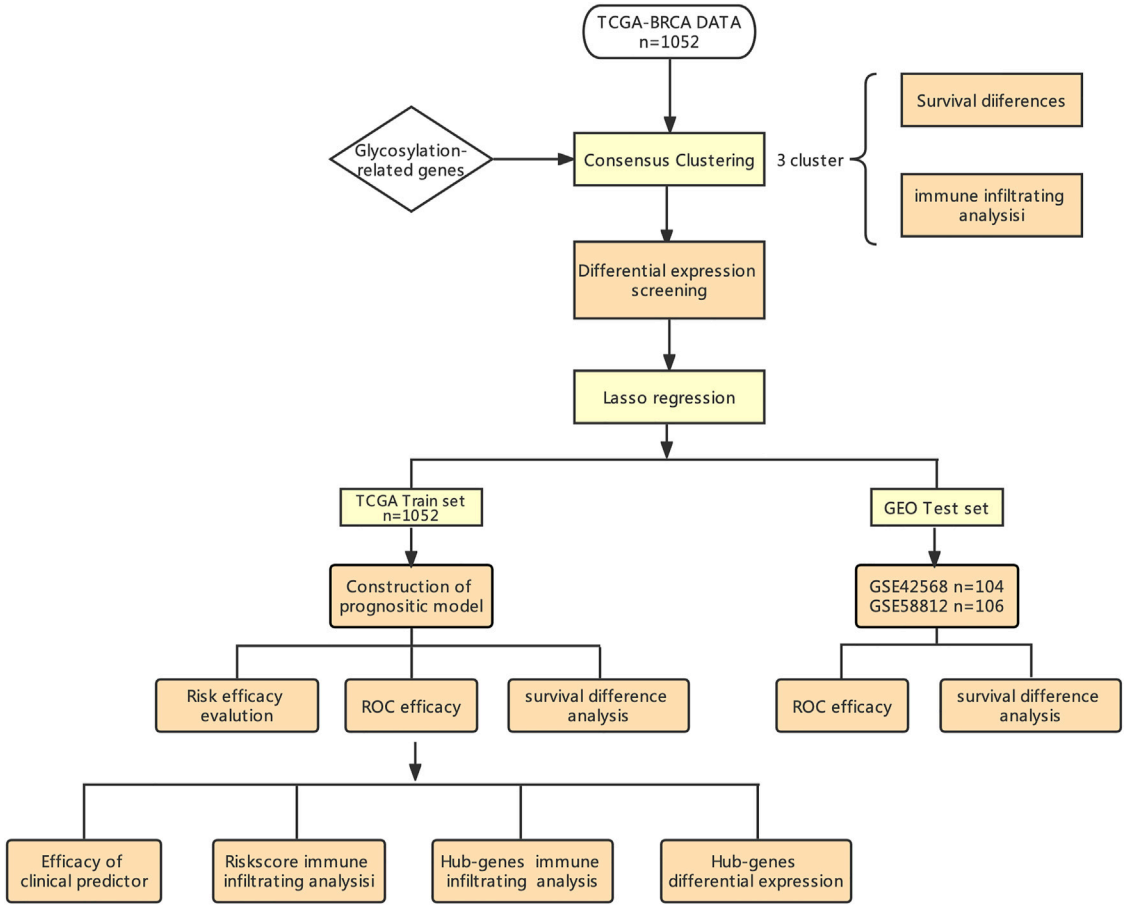


FIGURE 1
Workflow of our study design.

of differential genes to show their overlap (Figure 4D). The genes contained in each unit are shown in Supplementary Table S1, and the genes that show differences under one grouping are included in the next analysis. Finally, 1915 DEGs (Supplementary Table S1) were obtained and used to construct a prognostic risk-scoring model.

Immune characteristics of glycosylation-related groups

To explore the correlation between glycosylation patterns and immune characteristics, we analyzed the immune correlates of the three clusters. Figures 3B and C show significant differences in the immune score, stomal score, and immune cell infiltration. Cluster 3 demonstrated the lowest immune and stomal score and the poorest immune cell infiltration. Cluster 2 had the highest immune score and modest stomal score, and the immune cell infiltration was also the most abundant. Cluster 1 had the mediocre immune score and highest stomal score, and the immune cell infiltration was modest.

Construction and efficacy of risk-scoring model

To further construct a prognostic risk-scoring model without redundant genes, we used lasso regression to narrow down the range of candidate genes. According to mean-square error (Figure 4E) and coefficients (Figure 4F), we opted for the former λ as it results in a better prediction efficiency than the latter λ. Then, we fitted a multivariate Cox proportional hazard model to develop more valuable integrated molecules in the training set. Patients' age, stage, and 23 genes were included in this model, with a concordance index of 0.87 (Log-rank P: 4.48e-43) (Figure 5A). Figure 5B arranged the sample from low to high according to the risk score. The proportion of deaths increased as risk scores rose. The 23 key molecule expression is also shown at

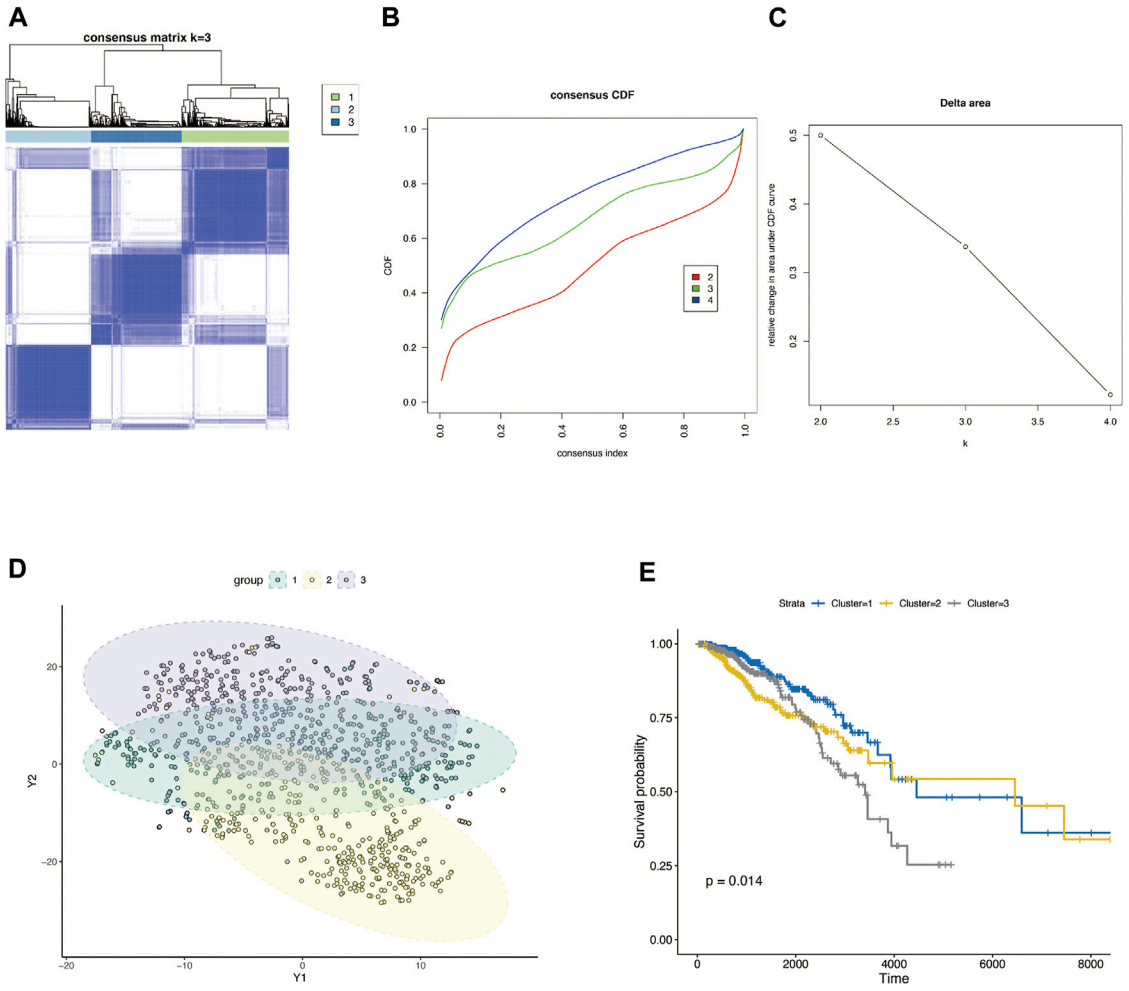


FIGURE 2
Consensus clustering classification of BRCA based on glycosylation-associated genes. **(A)** Optimal cluster distinction by consensus matrix (k = 3). **(B)** Empirical cumulative distribution function (CDF) plot displayed consensus distributions for each k. **(C)** Delta area plot. **(D)** T-SNE clustering of sample distributions based on glycosylation-related genes. **(E)** KM survival analysis of three glycosylation-based groups.

the bottom. Its area under the ROC curve (AUC) in 1, 3, and 5 years prior to death was 0.89, 0.90, and 0.89, respectively (Figure 5C). Kaplan–Meier (KM) analysis showed a significant difference in overall survival ($p < 0.0001$) (Figure 5D).

Validating of risk-scoring model predicting efficacy

We choose two breast cancer cohorts from GEO to validate the efficacy of this protistic model. The GSE42582 column of Table 1. In GSE42568 cohort, AUC in 1, 3, and 5 years prior to death was 0.73, 0.82, and 0.88, respectively (Figure 6A), and KM analysis presents a significant difference ($p < 0.0001$) (Figure 6B). The GSE58812 column of Table 1. In GSE58812 cohort, AUC in 1, 3, and 5 years prior to death was 0.95, 0.77, and 0.79, respectively (Figure 6C), and KM analysis presents a significant difference (p = 6e-04) (Figure 6D).

Risk score related immune infiltration and immunotherapy evaluation

We calculated a risk score for each sample according to the expression levels and regression coefficients and divided the BRCA cohort into low- and high-risk groups by median. To better investigate the interactions between the risk score and the immune microenvironment, we performed the ESTIMATE algorithm and ssGSEA to evaluate the correlation between the prognostic model and immune infiltrating in BRCA patients. Supplementary figure S3A shows PCA clustering of immune

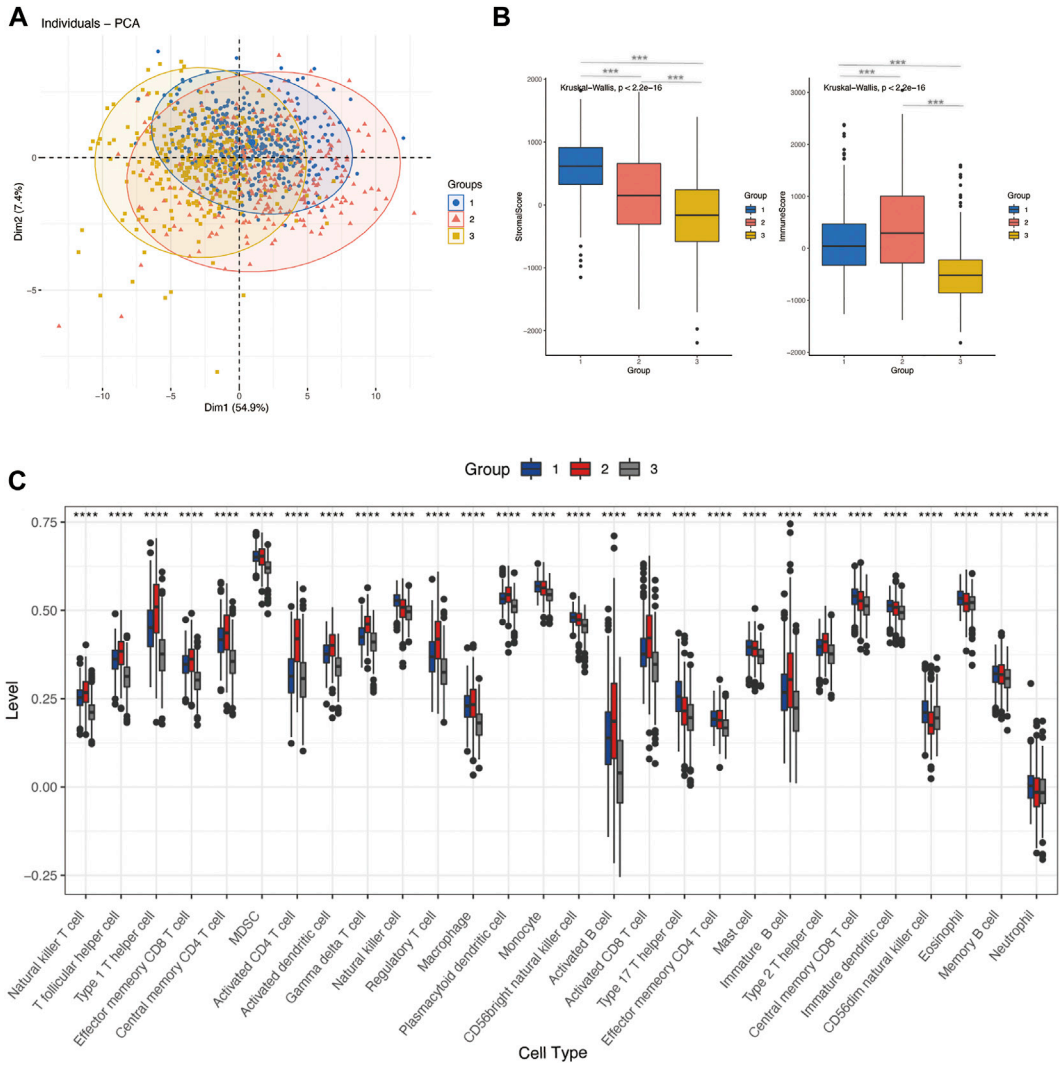


FIGURE 3
Differences in immune characteristics of glycosylation-based groups. **(A)** PCA clustering of sample distributions in immune signatures between three glycosylation-based groups. **(B)** Stomal score and the immune score of glycosylation-based groups (ESTIMATE algorithm). **(C)** Differences in 24 TME infiltration cells between glycosylation-based groups (ssGSEA) (****$p < 0.0001$).

signatures. The low-risk group demonstrates a higher immune score but no difference in the stomal score (Supplementary figure S3B). In terms of immune cell infiltration (Figures 7B, 8C), the risk score was slightly negatively correlated with immune cell level. The low-risk group represents a more significant fraction of activated B cells, eosinophils, mast cells, activated $CD8^+$ T cells, natural killer cells, and effector memory $CD8^+$ T cells but no difference in neutrophils, T follicular helper T cells, type 2 T helper cells, and type 17 T helper cells. Then, we used TIDE, an online tool, to evaluate immune checkpoint blockade (ICB) response for our screened signatures based on the TCGA and PRECOG cohorts. According to Figure 9, the gene set we input obtained almost equivalent area under the curve (AUC) as other predicting scores, especially CD274, CD8, IFNG, and Merck 18.

23 Gene signatures investigation

We further investigated the correlation between 23 gene signatures and immune cell infiltration. Compared with the low-risk group, the high-risk group harbors a low level of SPPL2C, IGKV2D-24, IGLC2, QRFPR, LINC01871, FABP7, AP000851.2, CLIC6, ILOVL2, FYB2, CDHR4, GNG4, TBR1, AC015910.1, and UPK1B and a high level of PXDNL. (Figure 7A). LINC01871, IGLC2, IGKV2D-24, MLIP, LINC01235, and AP000851.2 positively correlated with immune cell infiltration, and GNG4, PXDNL, KCNK3, ELOVL2, FYB2, SPPL2C, CLIC6 negatively correlated with immune cell levels. The main types of immune cells with different infiltrating were activated $CD4^+$ T cells, activated

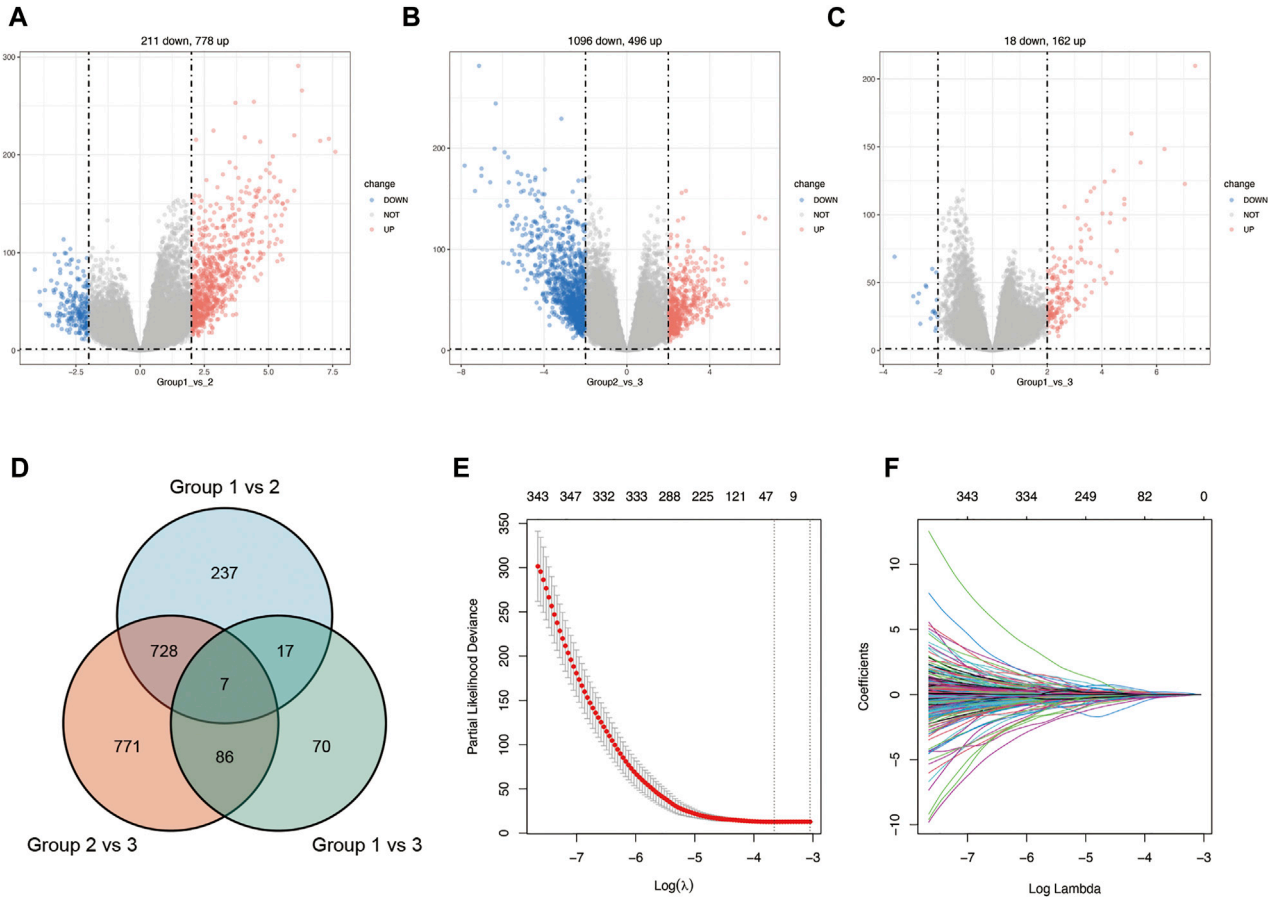


FIGURE 4
Construction of lasso regression model. Volcano plot of differentially expressed genes between cluster 1 vs. cluster 2 **(A)**, cluster 2 vs. cluster 3 **(B)**, and cluster 1 vs. cluster 3 **(C)** in BRCA. **(D)**. Venn diagram of differentially expressed genes between glycosylation-based groups. **(E)**. Cross-validation plot for the penalty term λ based on differentially expressed genes. Vertical bars represent acceptable maximum and minimum λ values with corresponding mean-squared error and the number of covariates. **(F)** Plots for lasso regression coefficients over different values of the penalty parameter λ.

$CD4^+$ T cells, natural killer T cells, activated B cells, activated dendritic cells, and MDSC. (Figure 8A). In addition, LINC01871 and IGLC2 positively correlated with immune checkpoint molecules such as PD-1, PDL1, CTLA4, TIGIT, LAG3, and BTLA and negatively correlated with HAVCR2. FYB2, SPPL2C, ELOVL2, CLIC6, IGKV2D-24, L1CAM, and AP000851.2 (Figure 8B).

Materials and methods

Data collection

The Breast Cancer (BRCA) data from The Cancer Genome Atlas Program (TCGA) was accessed *via* UCSC Xena (http://xena.ucsc.edu/). A total of 179 genes encoding glycosylation enzymes, targets, and regulators were obtained from previous literature (Krushkal et al., 2017) and are listed in Supplementary Table S1.

Consensus clustering analysis based on glycosylation-related genes

BRCA samples from TCGA were grouped into three clusters using the "ConsensusClusterPlus" (version1.60.0) R package (Wilkerson and Hayes, 2010) based on glycosylation-related genes (GRGs) (maxK = 4, innerLinkage = "complete"). "Fpkm" format was used for clustering analysis and "count" for difference analysis. Principal component analysis (PCA) and t-SNE were applied to assess sample clustering using the "FactoMineR" (version2.4) and Rtsne (version0.16) packages. "DESeq2" (version1.36.0) R package was used for screen

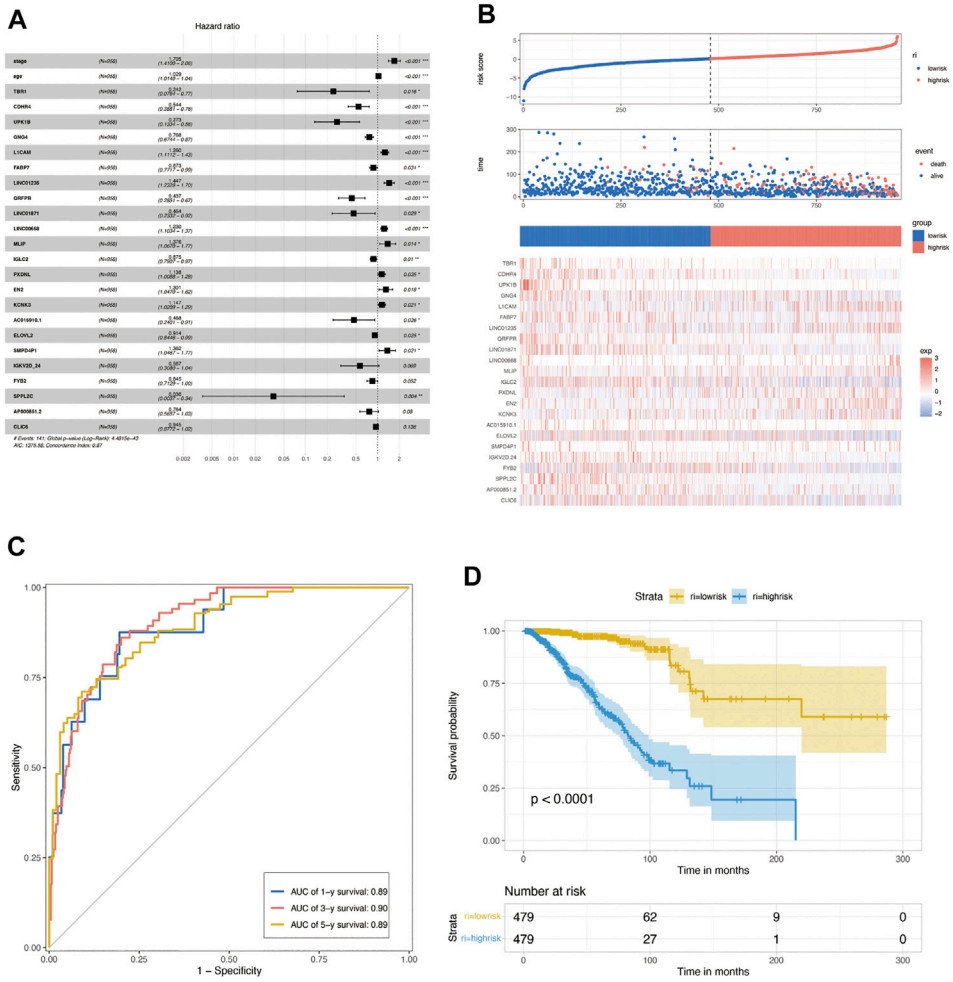


FIGURE 5
Construction of multivariate Cox regression. **(A)** Multivariate Cox proportional hazard model based on lasso de redundant gene set in the TCGA training set. **(B)** Proportion of deaths in the training set in high- and low-risk groups as risk score values increased. Top: red, high-risk; blue, low-risk. Middle: red, death; blue, alive. Bottom: hierarchical clustering of 14 key molecules between low- and high-risk groups. **(C)** COX risk score's time-dependent ROC curves for 1, 3, and 5 years before death in the TCGA training set. In the training set, **(D)** Kaplan–Meier survival analyses for COX low- and high-risk groups. ($p < 0.0001$, log-rank test).

differentially expressed genes (DEGs) between different clusters (|logFC| > 2, FDR <0.05).

The prognostic risk-scoring model constructed through GRGs-based clusters

First, the most minor absolute shrinkage and selection operator (LASSO) removed redundant genes achieved using the "glmnet" (version 4.1-4) R package. Ten-fold cross-validation was used to select the penalty term, λ. The mean-squared error was computed for the test data to measure the fitted models' predictive performance. Then, 38 genes (Supplementary Table S1) were obtained for prognostic Cox regression construction using the "My.stepwise" (version 0.1.0) package to establish the optimal model. Finally, the 23 retained genes were used for calculating risk scores according to the following formula:

$$\text{Risk Score} = \sum_{i=0}^{n} (\text{Coef}i^{*} xi), \quad (1)$$

where Coef$i$ is the coefficient, and xi is the z-score-transformed relative expression value of each selected gene. The time-dependent receiver operating characteristic (ROC) curve evaluated each model's sensitivity and specificity. The "survival" (version 3.3-1) R package was used, and the Kaplan–Meier (KM) overall survival curves between different clusters and risks were performed using the "survival" R package.

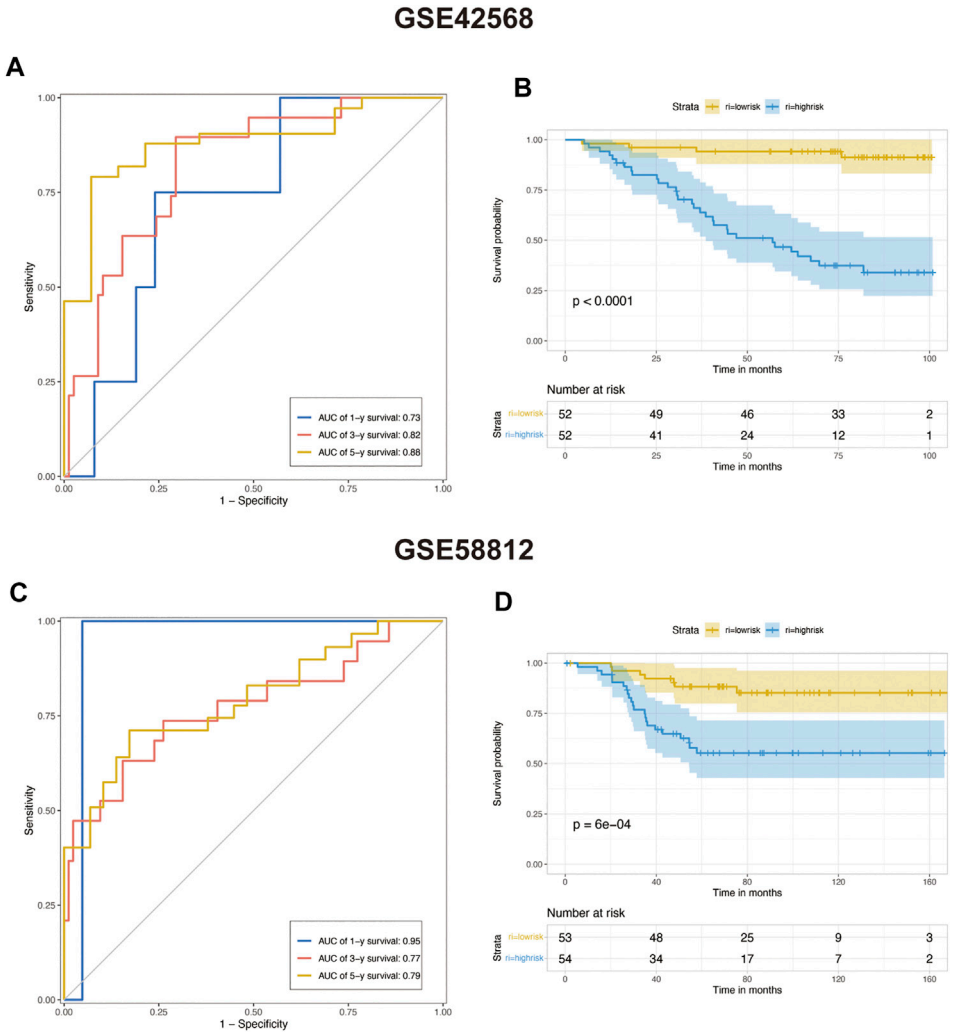


FIGURE 6
Predicting the efficacy of constructed multivariate Cox regression in the testing set. **(A)** COX risk score's time-dependent ROC curves for 1, 3, and 5 years before death in testing cohort GSE42568. **(B)** Kaplan–Meier survival analyses for COX low- and high-risk groups in testing cohort GSE42568 ($p < 0.0001$, log-rank test). **(C)** COX risk score's time-dependent ROC curves for 1, 3, and 5 years before death in testing cohort GSE58812. **(D)** Kaplan–Meier survival analyses for COX low- and high-risk groups in testing cohort GSE58812. ($p < 0.0001$, log-rank test).

Immune infiltrates analysis

The single-sample gene-set enrichment analysis (ssGSEA) was used to establish the relative abundance of 24 cell infiltration, which was analyzed using the "GSVA" (version 1.44.2) package. The ESTIMATE algorithm calculated stomal scores and immune scores of high- versus low-risk groups and different GRGs-based clusters. Immune checkpoint blockade (ICB) predicting evaluation performed by biomarker evaluation module from TIDE (Tumor Immune Dysfunction and Exclusion: harvard.edu)">http://tide.dfci.harvard.edu/) (harvard.edu)), a computational method to model tumor immune evasion and ICB response and resistance regulators.

Hub-genes analysis

Immune Infiltrates differences of prognostic hub-genes were performed using ssGSEA, as mentioned earlier. Checkpoints correlation was analyzed using the 'Hmisc' (version 4.7-0) package. All the statistical significance sets as $p < 0.05$ with two-side. Data processing and visualization were performed using R version 4.1.2.

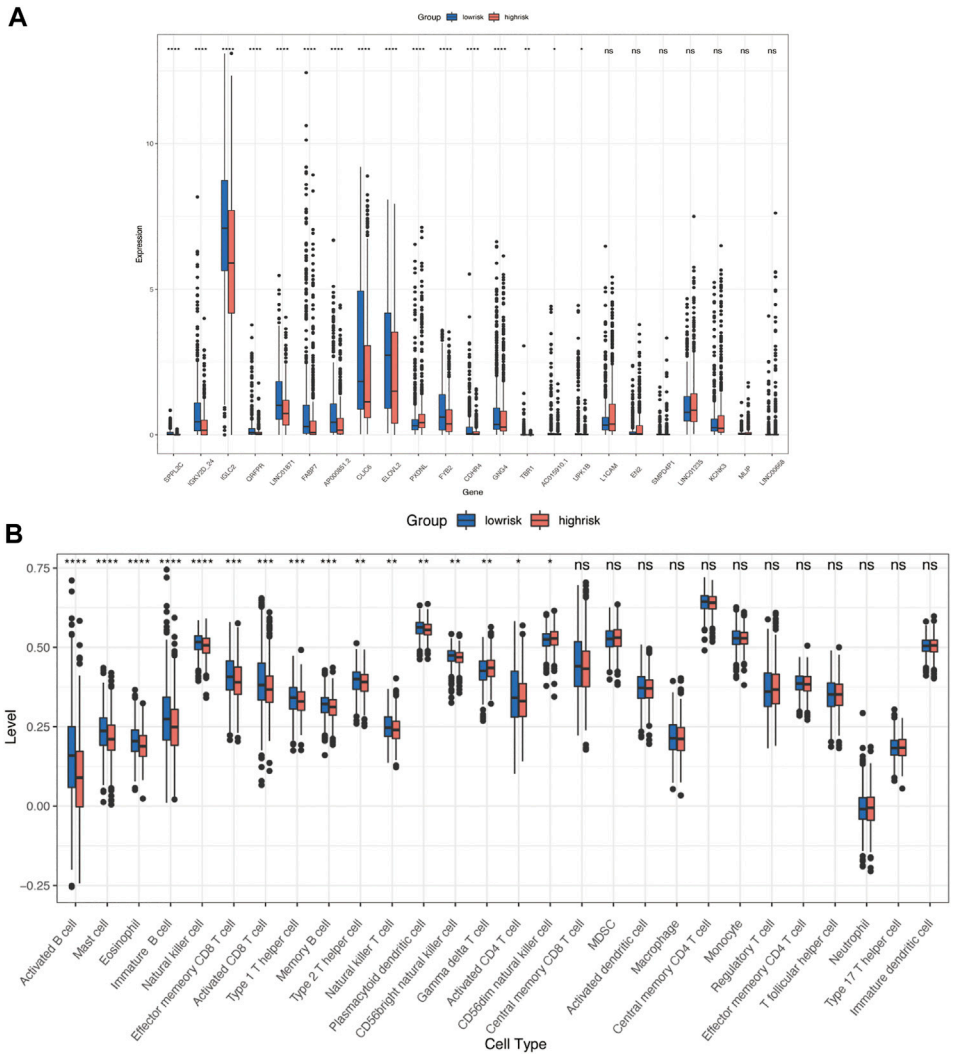


FIGURE 7
Immune characteristics in high- and low-risk groups. **(A)** Risk signatures expression in high- and low-risk groups. **(B)** Differences in 24 TME infiltration cells between high- and low-risk groups (ssGSEA) (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$, ****$p < 0.0001$).

Discussion

The role of glycocalyx–the extracellular carbohydrate coat, has been proposed in breast cancer occurrence and development since the 1950s (Aub et al., 1963). Then, it was noteworthy that plant lectin and carbohydrate motif binding proteins showed a higher affinity for malignant cells than normal cells in the 1960s (Remmele et al., 1986). By the 1980s, biochemists found that the enzyme-linked lectin binding assay could be used to predict tumor differentiation and therapeutic reactivity (Parodi et al., 1982). Shortly afterward, it was widely accepted that glycosylation status alteration could be used as biomarkers for breast cancer prognosis and tumor burden (Springer, 1997; Lin et al., 2002; Duffy et al., 2010). Given the heterogeneity of breast cancer, more recent studies have mapped the histopathological orientation and tissue distribution of glycosylated modifications in clinical breast cancer samples. So far, the influentially changed landscape of glycosylation processes in breast cancer is vividly portrayed.

We obtained a set of glycosylation-related genes containing 181 molecules from previous pieces of literature, including glycosylation pathways, genes encoding glycosylation targets or regulators, and members of cancer pathways affected by glycosylation (Supplementary Table S1) (Krushkal et al., 2017). In our study, TCGA-BRCA tumor samples were divided into three groups. We can consider three different glycosylated states based on these glycosylation-related genes by using consensus clustering analysis. There were significant differences in the expression patterns of glycosylated genes between them, and the survival

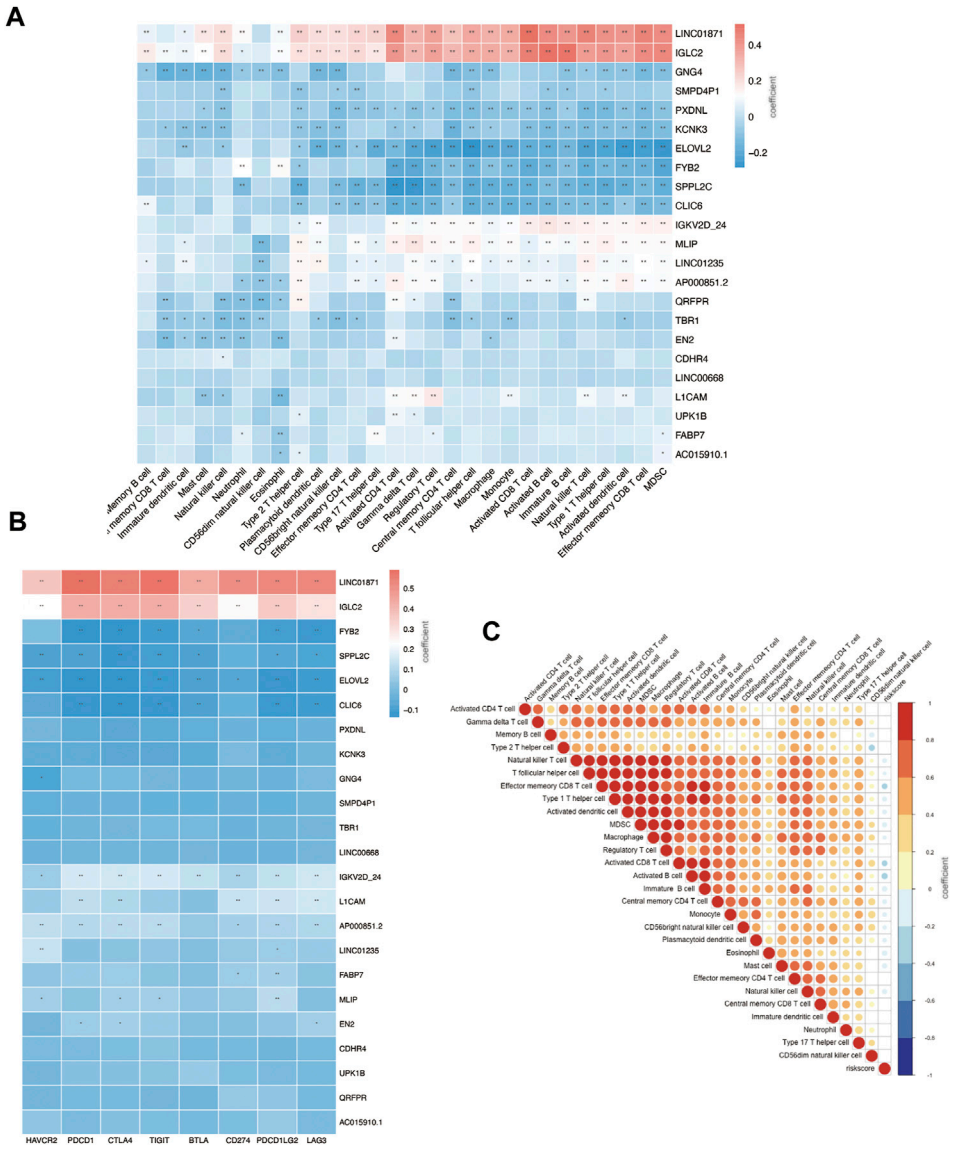


FIGURE 8
Immune infiltration status of prognostic signature. **(A)** Risk genes level in high- and low-risk groups. **(B)** Correlation between risk genes and checkpoint molecule expression. **(C)** Correlation between riskscore and immune infiltration.

analysis also reflected the difference in survival time under different glycosylated states (Figures 2D and E). It is well-documented that an altered "glycan coat" is a distinct hallmark of cancer.

Given that immune cells express a large variety of lectin (glycan-binding receptors), they recognize glycans on the tumor cell. Those immune cells can sense and respond to changes in the glycan signature of their environment. This often leads to tumor immune escape and immunomodulation. Therefore, the glycosylation-related signatures could affect tumor-immune cells' connections within the tumor microenvironment (Rodríguez et al., 2018; Lopes et al., 2021). In addition, a variety of recruited stromal components–transformed parenchyma and the associated stroma–are involved in tumor progression and response to treatment (Arneth, 2019; Hanahan, 2022). We further analyzed the immune characteristics of glycosylation-based groups. According to our results, group 3 demonstrated the lowest immune and stomal score and the poorest immune cell infiltration; group 2 had the highest immune score and modest stomal score, and the immune cell infiltration was also the most abundant. This indicates that group 2 tends to the glycosylation pattern of immune cells, group 1 of stromal cells, and group 3 of malignant cells (Figures 3A–C). In combination with the survival analysis of Figure 2E, we were surprised to find that in terms of glycosylation pattern, the glycosylation mode of tumor cells and

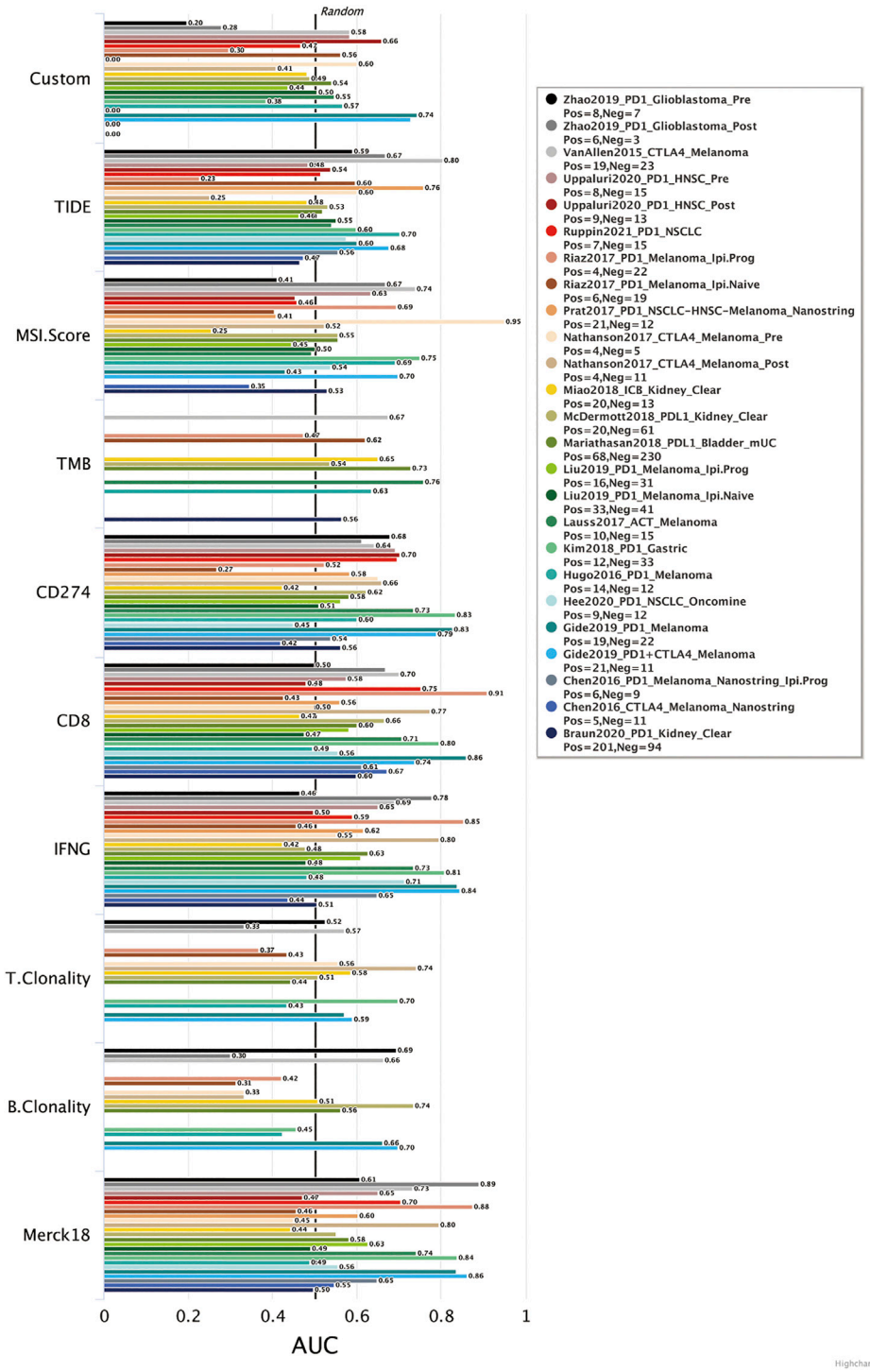


FIGURE 9
Biomarker evaluation from TIDE (Tumor Immune Dysfunction and Exclusion).

immune cells did not show any difference in patient survival, while the glycosylation of stromal cells may have a significant impact on patients' survival. In future explorations of tumor microenvironment glycosylation, focusing on stromal cells may be a more effective research direction. These results prove that the classification based on glycosylation is meaningful and effective,

TABLE 1 Clinical information of TCGA, GSE42586, GSE58812.

	TCGA	GSE42568 PMID: 23740839	GSE58812 PMID: 25887482
Sample			
Tumor	1,109	104	107
Normal	113	17	0
Survival			
Dead	144	35	29
Alive	933	69	78
Age			
<60	575	59	64
≥60	502	45	43
Grade			
I	—	11	—
II	—	40	—
III	—	53	—
Stage			
I	179	11	—
II	609	40	—
III	246	53	—
IV	19	0	—
Unknown	24	0	—
Subtype			
Luminal A	497	—	—
Luminal B	197	—	—
Basal	171	—	—
Her2	77	—	—
Unknown	135	—	—
ER expression			
Positive	—	67	—
Negative	—	34	—

Her2, human epidermal growth factor receptor 2; ER, estrogen receptor.

which helps us to understand the heterogeneity of breast cancer from another perspective. However, at present, the classification samples are limited. Increasing the sample size will help formulate a more stable grouping method and hopefully be applied to clinical prognosis and prediction.

The change of glycosylation pattern in tumor cells and immune microenvironment will affect the expression of other critical genes and make their corresponding bioprocesses abnormal, thus, inducing the transformation of malignant phenotypes, such as proliferation, epithelial–mesenchymal transition, and apoptosis resistance. To identify the prognostic genes influenced by glycosylation processes, we screened the DEGs of these three groups and constructed a predictive risk model through lasso and Cox regression calculation. The final prognostic model containing 23 key molecules achieved exciting performance both in the TCGA training set and testing set GSE42568 and GSE58812 (Figures 5C and D, Figure 6). Using the model algorithm, we calculated a risk score and divided the sample into high- and low-risk groups by the median. This risk score also showed a significant difference in predicting overall clinical survival and immune infiltration (Figures 7B, 8C). Great achievement has been obtained in ICB-based immunotherapies (Chen et al., 2020). In order to obtain better clinical remission and fewer immune-related adverse events, researchers are committed to developing biomarkers to screen an effective population accurately. The reported measures that can be used to predict the efficacy of ICI therapy include immune cell infiltration (Cogdill et al., 2017), protein expressions such as PD-L1 (Teng et al., 2015), mutations and neoantigens (Mcgranahan et al., 2016), and genetic and epigenetic characteristics (Ascierto et al., 2012). On the TIDE prediction website, our gene set shows a favorable performance compared with the existing evaluation methods (Figure 9), which proves that our model has practical proficiency and value for further exploration and improvement in immunotherapy prediction.

Then, we move on to several single prognostic genes. LINC01871 significantly lower expression in the high-risk group and positively correlated with most of the immune cell infiltration (Figure 8). This suggests that LINC01871 may play a protective role in breast cancer. According to a recent review of the literature, LINC01871 has been identified by several studies in breast cancer through bioinformatic measurement involving the cellular phenotype of autophagy (Li et al., 2021; Wu et al., 2021; Jiang et al., 2022; Luo et al., 2022), stemness (Li et al., 2020), immune response (Ma et al., 2020; Mathias et al., 2021), ferroptosis (Xu et al., 2021), and lipid metabolism (Shi et al., 2022). IGLC2 has a similar expression and functional pattern to LINC01871 in our study (Figure 8). Chang et al. (2021) found in a study of triple-negative breast cancer (TNBC) cohort that a high expression of IGLC2 was related to a favorable prognosis for TNBC patients, which is consistent with our results. In addition, IGLC2 is linked with the proliferation, migration, and invasion of MDA-MB-231 cells. Pathway enrichment analysis showed that IGLC2 is related to the extracellular matrix–receptor interaction (Chang et al., 2021). All these features make IGLC2 have the potential to be a biomarker to predict prognosis, even for identifying breast cancer patients who can benefit the most from immune checkpoint blockade treatment. ELOVL2 is another prognostic signature in our results. Studies have shown that long noncoding RNA on its antisense chain (ELOVL2-AS1) correlates with breast cancer prognosis. The predictive efficacy of ELOVL2 needs to be verified in a larger sample size, and its mediated cell function also needs to be further explored.

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

SG conceived the presented idea. XW developed the theory and performed the computations. XL undertook the visualization of results. WC encouraged SG and XW to investigate the tumor microenvironment and supervised the findings of this work. All authors discussed the results and contributed to the final manuscript.

Funding

The work was supported by CAMS Innovation Fund for Medical Sciences (CIFMS) 2021-I2M-1-014.

Acknowledgments

The authors thank Clarke C and Jézéquel P for submitting the dataset to Gene Expression Omnibus. The authors thank Dr. Jianming Zeng (University of Macau) and all the members of his bioinformatics team, Biotrainee, for generously sharing their experience and codes.

Conflict of interest

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

The reviewer XK declared a shared parent affiliation with authors SG, XL, and WC to the handling editor at the time of review.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.960567/full#supplementary-material

References

Arneth, B. (2019). Tumor microenvironment. *Med. Kaunas.* 56 (1), E15. doi:10.3390/medicina56010015

Ascierto, M. L., Kmieciak, M., Idowu, M. O., Manjili, R., Zhao, Y., Grimes, M., et al. (2012). A signature of immune function genes associated with recurrence-free survival in breast cancer patients. *Breast Cancer Res. Treat.* 131 (3), 871–880. doi:10.1007/s10549-011-1470-x

Aub, J. C., Tieslau, C., and Lankester, A. (1963). Reactions of normal and tumor cell surfaces to enzymes. I. Wheat-germ lipase and associated mucopolysaccharides [J]. *Proc. Natl. Acad. Sci. U. S. A.* 50 (4), 613–619. doi:10.1073/pnas.50.4.613

Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours[J]. *Nature* 490 (7418), 61–70. doi:10.1038/nature11412

Chang, Y. T., Tsai, W. C., Lin, W. Z., Wu, C. C., Yu, J. C., Tseng, V. S., et al. (2021). A novel IGLC2 gene linked with prognosis of triple-negative breast cancer. *Front. Oncol.* 11, 759952. doi:10.3389/fonc.2021.759952

Chen, X., Pan, X., Zhang, W., Guo, H., Cheng, S., He, Q., et al. (2020). Epigenetic strategies synergize with PD-L1/PD-1 targeted cancer immunotherapies to enhance antitumor responses. *Acta Pharm. Sin. B* 10 (5), 723–733. doi:10.1016/j.apsb.2019.09.006

Cogdill, A. P., Andrews, M. C., and Wargo, J. A. (2017). Hallmarks of response to immune checkpoint blockade. *Br. J. Cancer* 117 (1), 1–7. doi:10.1038/bjc.2017.136

Curtis, C., Shah, S. P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J., et al. (2012). The genomic and transcriptomic architecture of 2, 000 breast tumours reveals novel subgroups. *Nature* 486 (7403), 346–352. doi:10.1038/nature10983

Drake, R. R., Jones, E. E., Powers, T. W., and Nyalwidhe, J. O. (2015). Altered glycosylation in prostate cancer. *Adv. Cancer Res.* 126, 345–382. doi:10.1016/bs.acr.2014.12.001

Duffy, M. J., Evoy, D., and Mcdermott, E. W. (2010). CA 15-3: Uses and limitation as a biomarker for breast cancer. *Clin. Chim. Acta.* 411 (23-24), 1869–1874. doi:10.1016/j.cca.2010.08.039

Eichler, J. (2019). Protein glycosylation. *Curr. Biol.* 29 (7), R229–r231. doi:10.1016/j.cub.2019.01.003

Gao, Y., Luan, X., Melamed, J., and Brockhausen, I. (2021). Role of glycans on key cell surface receptors that regulate cell proliferation and cell death. *Cells* 10 (5), 1252. doi:10.3390/cells10051252

Gilgunn, S., Conroy, P. J., Saldova, R., Rudd, P. M., and O'Kennedy, R. J. (2013). Aberrant PSA glycosylationa sweet predictor of prostate cancer. *Nat. Rev. Urol.* 10 (2), 99–107. doi:10.1038/nrurol.2012.258

Goldhirsch, A., Winer, E. P., Coates, A. S., Gelber, R. D., Piccart-Gebhart, M., Thurlimann, B., et al. (2013). Personalizing the treatment of women with early breast cancer: Highlights of the st gallen international expert consensus on the primary therapy of early breast cancer 2013. *Ann. Oncol.* 24 (9), 2206–2223. doi:10.1093/annonc/mdt303

Goldstein, M. J., and Mitchell, E. P. (2005). Carcinoembryonic antigen in the staging and follow-up of patients with colorectal cancer. *Cancer Invest.* 23 (4), 338–351. doi:10.1081/cnv-58878

Gudelj, I., Lauc, G., and Pezer, M. (2018). Immunoglobulin G glycosylation in aging and diseases. *Cell. Immunol.* 333, 65–79. doi:10.1016/j.cellimm.2018.07.009

Günthert, U., Hofmann, M., Rudy, W., Reber, S., Zoller, M., Haussmann, I., et al. (1991). A new variant of glycoprotein CD44 confers metastatic potential to rat carcinoma cells. *Cell* 65 (1), 13–24. doi:10.1016/0092-8674(91)90403-l

Hanahan, D. (2022). Hallmarks of cancer: New dimensions. *Cancer Discov.* 12 (1), 31–46. doi:10.1158/2159-8290.CD-21-1059

Jiang, F., Wu, C., Wang, M., Wei, K., and Wang, J. (2022). An autophagy-related long non-coding RNA signature for breast cancer. *Comb. Chem. High. Throughput Screen.* 25 (8), 1327–1335. doi:10.2174/1386207324666210603122718

Jiang, F., Wu, C., Wang, M., Wei, K., and Wang, J. (2021). Identification of novel cell glycolysis related gene signature predicting survival in patients with breast cancer. *Sci. Rep.* 11 (1), 3986. doi:10.1038/s41598-021-83628-9

Kölbl, A. C., Andergassen, U., and Jeschke, U. (2015). The role of glycosylation in breast cancer metastasis and cancer control. *Front. Oncol.* 5, 219. doi:10.3389/fonc.2015.00219

Krushkal, J., Zhao, Y., Hose, C., Monks, A., Doroshow, J. H., and Simon, R. (2017). Longitudinal transcriptional response of glycosylation-related genes,

regulators, and targets in cancer cell lines treated with 11 antitumor agents. *Cancer Inf.* 16, 1176935117747259. doi:10.1177/1176935117747259

Kudelka, M. R., Ju, T., Heimburg-Molinaro, J., and Cummings, R. D. (2015). Simple sugars to complex diseasemucin-type O-glycans in cancer. *Adv. Cancer Res.* 126, 53–135. doi:10.1016/bs.acr.2014.11.002

Kumpulainen, E. J., Keskikuru, R. J., and Johansson, R. T. (2002). Serum tumor marker CA 15.3 and stage are the two most powerful predictors of survival in primary breast cancer. *Breast Cancer Res. Treat.* 76 (2), 95–102. doi:10.1023/a:1020514925143

Lehmann, B. D., Bauer, J. A., Chen, X., Sanders, M. E., Chakravarthy, A. B., Shyr, Y., et al. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Invest.* 121 (7), 2750–2767. doi:10.1172/JCI45014

Li, X., Jin, F., and Li, Y. (2021). A novel autophagy related lncRNA prognostic risk model for breast cancer. *J. Cell. Mol. Med.* 25 (1), 4–14. doi:10.1111/jcmm.15980

Li, X., Li, Y., Yu, X., and Jin, F. (2020). Identification and validation of stemness-related lncRNA prognostic signature for breast cancer. *J. Transl. Med.* 18 (1), 331. doi:10.1186/s12967-020-02497-4

Lin, S., Kemmner, W., Grigull, S., and Schlag, P. M. (2002). Cell surface alpha 2, 6 sialylation affects adhesion of breast carcinoma cells. *Exp. Cell Res.* 276 (1), 101–110. doi:10.1006/excr.2002.5521

Lopes, N., Correia, V. G., Palma, A. S., and Brito, C. (2021). Cracking the breast cancer glyco-code through glycan-lectin interactions: Targeting immunosuppressive macrophages. *Int. J. Mol. Sci.* 22 (4), 1972. doi:10.3390/ijms22041972

Luo, Z., Nong, B., Ma, Y., and Fang, D. (2022). Autophagy related long non-coding RNA and breast cancer prognosis analysis and prognostic risk model establishment. *Ann. Transl. Med.* 10 (2), 58. doi:10.21037/atm-21-6251

Ma, W., Zhao, F., Yu, X., Guan, S., Suo, H., Tao, Z., et al. (2020). Immune-related lncRNAs as predictors of survival in breast cancer: A prognostic signature. *J. Transl. Med.* 18 (1), 442. doi:10.1186/s12967-020-02522-6

Mallard, B. W., and Tiralongo, J. (2017). Cancer stem cell marker glycosylation: Nature, function and significance. *Glycoconj. J.* 34 (4), 441–452. doi:10.1007/s10719-017-9780-9

Marusyk, A., Almendro, V., and Polyak, K. (2012). Intra-tumour heterogeneity: A looking glass for cancer?[J]. *Nat. Rev. Cancer* 12 (5), 323–334. doi:10.1038/nrc3261

Mathias, C., Muzzi, J. C. D., Antunes, B. B., Gradia, D. F., Castro, M. A. A., and Carvalho de Oliveira, J. (2021). Unraveling immune-related lncRNAs in breast cancer molecular subtypes. *Front. Oncol.* 11, 692170. doi:10.3389/fonc.2021.692170

Mcgranahan, N., Furness, A. J., Rosenthal, R., Ramskov, S., Lyngaa, R., Saini, S. K., et al. (2016). Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* 351 (6280), 1463–1469. doi:10.1126/science.aaf1490

Paredes, J., Figueiredo, J., Albergaria, A., Oliveira, P., Carvalho, J., Ribeiro, A. S., et al. (2012). Epithelial E- and P-cadherins: Role and clinical significance in cancer. *Biochim. Biophys. Acta* 1826 (2), 297–311. doi:10.1016/j.bbcan.2012.05.002

Parodi, A. J., Blank, E. W., Peterson, J. A., and Ceriani, R. L. (1982). Dolichol-bound oligosaccharides and the transfer of distal monosaccharides in the synthesis of glycoproteins by normal and tumor mammary epithelial cells. *Breast Cancer Res. Treat.* 2 (3), 227–237. doi:10.1007/BF01806935

Peiris, D., Spector, A. F., Lomax-Browne, H., Azimi, T., Ramesh, B., Loizidou, M., et al. (2017). Cellular glycosylation affects Herceptin binding and sensitivity of breast cancer cells to doxorubicin and growth factors. *Sci. Rep.* 7, 43006. doi:10.1038/srep43006

Perou, C. M., Sørlie, T., Eisen, M. B., van de RijnM.Jeffrey, S. S., Rees, C. A., et al. (2000). Molecular portraits of human breast tumours. *Nature* 406 (6797), 747–752. doi:10.1038/35021093

Pinho, S. S., Figueiredo, J., Cabral, J., Carvalho, S., Dourado, J., Magalhaes, A., et al. (2013). E-Cadherin and adherens-junctions stability in gastric carcinoma: Functional implications of glycosyltransferases involving N-glycan branching biosynthesis, N-acetylglucosaminyltransferases III and V. *Biochim. Biophys. Acta* 1830 (3), 2690–2700. doi:10.1016/j.bbagen.2012.10.021

Pinho, S. S., and Reis, C. A. (2015). Glycosylation in cancer: Mechanisms and clinical implications. *Nat. Rev. Cancer* 15 (9), 540–555. doi:10.1038/nrc3982

Pinho, S. S., Seruca, R., Gärtner, F., Yamaguchi, Y., Gu, J., Taniguchi, N., et al. (2011). Modulation of E-cadherin function and dysfunction by N-glycosylation. *Cell. Mol. Life Sci.* 68 (6), 1011–1020. doi:10.1007/s00018-010-0595-0

Prat, A., Parker, J. S., Fan, C., and Perou, C. M. (2012). PAM50 assay and the three-gene model for identifying the major and clinically relevant molecular subtypes of breast cancer. *Breast Cancer Res. Treat.* 135 (1), 301–306. doi:10.1007/s10549-012-2143-0

Prat, A., Pineda, E., Adamo, B., Galvan, P., Fernandez, A., Gaba, L., et al. (2015). Clinical implications of the intrinsic molecular subtypes of breast cancer. *Breast* 24 (2), S26–S35. doi:10.1016/j.breast.2015.07.008

Rabinovich, G. A., and Toscano, M. A. (2009). Turning 'sweet' on immunity: Galectin-glycan interactions in immune tolerance and inflammation. *Nat. Rev. Immunol.* 9 (5), 338–352. doi:10.1038/nri2536

Remmele, W., Hildebrand, U., Hienz, H. A., VierbuchenM.Behnken, L. J., et al. (1986). Comparative histological, histochemical, immunohistochemical and biochemical studies on oestrogen receptors, lectin receptors, and Barr bodies in human breast cancer. *Virchows Arch. A Pathol. Anat. Histopathol.* 409 (2), 127–147. doi:10.1007/BF00708323

Rodríguez, E., Schetters, S. T. T., and Van Kooyk, Y. (2018). The tumour glyco-code as a novel immune checkpoint for immunotherapy. *Nat. Rev. Immunol.* 18 (3), 204–211. doi:10.1038/nri.2018.3

Scott, D. A., Casadonte, R., Cardinali, B., Spruill, L., Mehta, A. S., Carli, F., et al. (2019a). Increases in tumor N-glycan polylactosamines associated with advanced HER2-positive and triple-negative breast cancer tissues. *Proteomics. Clin. Appl.* 13 (1), e1800014. doi:10.1002/prca.201800014

Scott, D. A., Norris-Caneda, K., Spruill, L., Bruner, E., Kono, Y., Angel, P. M., et al. (2019b). Specific N-linked glycosylation patterns in areas of necrosis in tumor tissues. *Int. J. Mass Spectrom.* 437, 69–76. doi:10.1016/j.ijms.2018.01.002

Shen, Y., Peng, X., and Shen, C. (2020). Identification and validation of immune-related lncRNA prognostic signature for breast cancer. *Genomics* 112 (3), 2640–2646. doi:10.1016/j.ygeno.2020.02.015

Shi, G. J., Zhou, Q., Zhu, Q., Wang, L., and Jiang, G. Q. (2022). A novel prognostic model associated with the overall survival in patients with breast cancer based on lipid metabolism-related long noncoding RNAs. *J. Clin. Lab. Anal.* 36, e24384. doi:10.1002/jcla.24384

Springer, G. F. (1997). Immunoreactive T and Tn epitopes in cancer diagnosis, prognosis, and immunotherapy. *J. Mol. Med.* 75 (8), 594–602. doi:10.1007/s001090050144

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Ca. Cancer J. Clin.* 71 (3), 209–249. doi:10.3322/caac.21660

Taniguchi, N., and Kizuka, Y. (2015). Glycans and cancer: Role of N-glycans in cancer biomarker, progression and metastasis, and therapeutics. *Adv. Cancer Res.* 126, 11–51. doi:10.1016/bs.acr.2014.11.001

Teng, M. W., Ngiow, S. F., Ribas, A., and Smyth, M. J. (2015). Classifying cancers based on T-cell infiltration and PD-L1. *Cancer Res.* 75 (11), 2139–2145. doi:10.1158/0008-5472.CAN-15-0255

Teschendorff, A. E., Miremadi, A., Pinder, S. E., Ellis, I. O., and Caldas, C. (2007). An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. *Genome Biol.* 8 (8), R157. doi:10.1186/gb-2007-8-8-r157

Wang, D., Wei, G., Ma, J., Cheng, S., Jia, L., Song, X., et al. (2021). Identification of the prognostic value of ferroptosis-related gene signature in breast cancer patients. *BMC Cancer* 21 (1), 645. doi:10.1186/s12885-021-08341-2

Wilkerson, M. D., and Hayes, D. N. (2010). ConsensusClusterPlus: A class discovery tool with confidence assessments and item tracking. *Bioinformatics* 26 (12), 1572–1573. doi:10.1093/bioinformatics/btq170

Wu, Q., Li, Q., Zhu, W., Zhang, X., and Li, H. (2021). Identification of autophagy-related long non-coding RNA prognostic signature for breast cancer. *J. Cell. Mol. Med.* 25 (8), 4088–4098. doi:10.1111/jcmm.16378

Xu, Z., Jiang, S., Ma, J., Tang, D., Yan, C., and Fang, K. (2021). Comprehensive analysis of ferroptosis-related LncRNAs in breast cancer patients reveals prognostic value and relationship with tumor immune microenvironment. *Front. Surg.* 8, 742360. doi:10.3389/fsurg.2021.742360

Zhang, D., Zheng, Y., Yang, S., Li, Y., Wang, M., Yao, J., et al. (2020). Identification of a novel glycolysis-related gene signature for predicting breast cancer survival. *Front. Oncol.* 10, 596087. doi:10.3389/fonc.2020.596087

Zhang, R., Zhu, Q., Yin, D., Yang, Z., Guo, J., Zhang, J., et al. (2020). Identification and validation of an autophagy-related lncRNA signature for Patients with breast cancer. *Front. Oncol.* 10, 597569. doi:10.3389/fonc.2020.597569

Zurawski, V. R., Jr., Orjaseter, H., Andersen, A., et al. (1988). Elevated serum CA 125 levels prior to diagnosis of ovarian neoplasia: Relevance for early detection of ovarian cancer. *Int. J. Cancer* 42 (5), 677–680. doi:10.1002/ijc.2910420507

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 01 November 2022
DOI 10.3389/fgene.2022.1009316

OPEN ACCESS

EDITED BY
Rongshan Yu,
Xiamen University, China

REVIEWED BY
Yuanyue Li,
College of Biological Sciences (UC), Davis, United States
Samson Pandam Salifu,
Kwame Nkrumah University of Science and Technology, Ghana

*CORRESPONDENCE
Joanna Polanska,
Joanna.Polanska@polsl.pl

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

RECEIVED 01 August 2022
ACCEPTED 13 October 2022
PUBLISHED 01 November 2022

CITATION
Kujawa T, Marczyk M and Polanska J (2022), Influence of single-cell RNA sequencing data integration on the performance of differential gene expression analysis.
Front. Genet. 13:1009316.
doi: 10.3389/fgene.2022.1009316

Influence of single-cell RNA sequencing data integration on the performance of differential gene expression analysis

Tomasz Kujawa[1], Michał Marczyk[1,2] and Joanna Polanska[1]*

[1]Department of Data Science and Engineering, Silesian University of Technology, Gliwice, Poland, [2]Yale Cancer Center, Yale School of Medicine, New Haven, CT, United States

Large-scale comprehensive single-cell experiments are often resource-intensive and require the involvement of many laboratories and/or taking measurements at various times. This inevitably leads to batch effects, and systematic variations in the data that might occur due to different technology platforms, reagent lots, or handling personnel. Such technical differences confound biological variations of interest and need to be corrected during the data integration process. Data integration is a challenging task due to the overlapping of biological and technical factors, which makes it difficult to distinguish their individual contribution to the overall observed effect. Moreover, the choice of integration method may impact the downstream analyses, including searching for differentially expressed genes. From the existing data integration methods, we selected only those that return the full expression matrix. We evaluated six methods in terms of their influence on the performance of differential gene expression analysis in two single-cell datasets with the same biological study design that differ only in the way the measurement was done: one dataset manifests strong batch effects due to the measurements of each sample at a different time. Integrated data were visualized using the UMAP method. The evaluation was done both on individual gene level using parametric and non-parametric approaches for finding differentially expressed genes and on gene set level using gene set enrichment analysis. As an evaluation metric, we used two correlation coefficients, Pearson and Spearman, of the obtained test statistics between reference, test, and corrected studies. Visual comparison of UMAP plots highlighted ComBat-seq, limma, and MNN, which reduced batch effects and preserved differences between biological conditions. Most of the tested methods changed the data distribution after integration, which negatively impacts the use of parametric methods for the analysis. Two algorithms, MNN and Scanorama, gave very poor results in terms of differential analysis on gene and gene set levels. Finally, we highlight ComBat-seq as it led to the highest correlation of test statistics between reference and corrected dataset among others. Moreover, it does not distort the original distribution of gene expression data, so it can be used in all types of downstream analyses.

KEYWORDS

single-cell RNA sequencing, data integration, batch correction, differential gene expression, joint analysis

1 Introduction

Single-cell RNA sequencing (scRNAseq) is a technique that allows the high-throughput examination of transcriptomes with a single-cell resolution (Lee et al., 2014; Qian et al., 2022). The transcriptome is a dynamic structure that responds rapidly in the form of gene expression to the variety of factors that a cell is subjected to. Moreover, the expression profile can be different in cells of the same type which proves significant cellular heterogeneity (Adil et al., 2021). This heterogeneity is masked in bulk analyses where populations of cells are mixed and sequenced together resulting in signal averages from millions of cells. Single-cell RNA-seq overcomes this barrier and allows the processing of millions of individual cells at a time.

In large projects that involve the processing of many cells data are frequently generated at different times and in different laboratories often equipped with various sequencing platforms (Ming et al., 2022). Combining data generated separately for a consolidated downstream analysis improves statistical power but requires reliable data integration methods. Data integration is also crucial in studies of different omics levels (genomics, proteomics, metabolomics, etc.) to fully understand the molecular complexity of different cell types (Bao et al., 2022). The goal of single-cell data integration is to cluster together cells of similar types; these cells should be intermingled and indistinguishable even if they come from different experiments. In other words, technical differences between datasets should be removed while key biological variations should be preserved. Data integration is a challenging task, especially in large datasets containing highly heterogeneous cell populations. Batch effect removal is a step in which we want to reduce the technical variability in our data that might occur due to differences in sample preparation, sequencing, or processing. Thus, we want to integrate the data that could be assigned to a known batch. Here, we are using the terms data integration and batch correction interchangeably.

There is a variety of distinct algorithms for scRNAseq data integration that are based on different principles and assumptions (Haghverdi et al., 2018; Hie et al., 2019; Lin et al., 2019; Liu et al., 2020; Zhang et al., 2020). An important criterion of the division in terms of our study is based on the output format which can be: (i) full expression matrix; (ii) low-dimensional matrix of embeddings; or (iii) integrated graph. The output type limits the potential downstream applications of integrated data. The full expression matrix is the most versatile format as it could be used in all downstream analyses. On the other hand, a joint embedding is not appropriate for some applications like differential expression analysis or biomarker detection. Hence, the decision about the choice of integration method is crucial and consequential. Another key factor influencing the choice is the main statistical approach that a particular method is based on. We can distinguish two groups here: supervised and unsupervised. The former requires cell-type annotations, and the latter does not rely on data labeling.

The most recent and comprehensive evaluation of scRNAseq integration methods was performed in (Luecken et al., 2022). They evaluated the most popular tools on their ability to remove batch effects while conserving biological information. Their evaluation involved setups with and without cell identity labels and different preprocessing combinations [with/without scaling and highly variable genes (HVGs) selection], as well as the diversity in output formats for each method and task. The conclusion from this work is that there is no single, best integration method and the performance is dependent on the complexity of the integration task (the strength of batch effect, the degree of confounding between batch and biological signals, presence of nuanced biological variation, etc.) (Luecken et al., 2022). Some methods, like BBKNN or Harmony, showed a stronger action towards removing batch effect over conservation of biological variation. For others, like ComBat, MNN, and DESC the trend was in favor of bio-conservation. Deep learning methods that use cell identity information, like scGen or scANVI, preserved biological variation stronger than label-free ones but require larger input data. Generally, HVG selection improved the overall integration performance over the full feature set, except for trajectory and cell-cycle conservation analysis. Scaling the input data typically improved batch removal at a cost of bio-conservation. In another evaluation of scRNAseq data integration methods (Tran et al., 2020), they examined in different simulation scenarios (balanced/unbalanced batches, different dropout rates) the impact of data integration on differential gene expression analysis (DGE analysis), particularly whether it improves the recovery of differentially expressed genes (DEGs). They found that MNN Correct, ZINB-WaVE, ComBat, and scMerge were the top-performing methods. ComBat turned out to be the best method for this task (being one of the worst overall). scMerge had a good balance between DEGs recovery and overall performance.

The above and other benchmarks typically cover a wide range of evaluation aspects such as removal of batch effects and conservation of biological variation, scalability for large datasets, or computational requirements. Regardless of existing comparisons of data integration methods, there is a lack of studies that comprehensively investigate the impact of data integration on differential gene expression analysis using real data. In terms of DGE analysis most studies on benchmarking methods for data integration focus only on

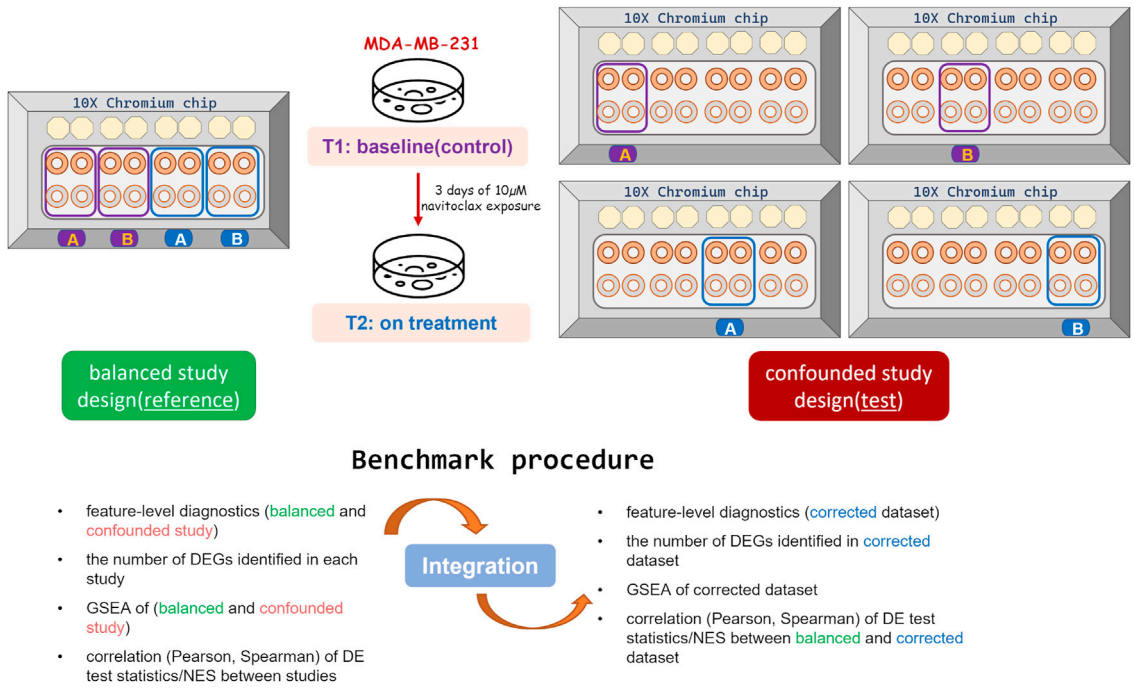


FIGURE 1
Experimental design and benchmark procedure.

overlaps of differentially expressed genes (Chazarra-Gil et al., 2021) not providing deeper insight into the problem. This study aims to fulfill this gap. Using two datasets, one of which requires data integration to correct the confounded design of the study, six integration methods that provide corrected gene expression matrices were compared. The evaluation was done on individual gene level using two different approaches (parametric and non-parametric) and on gene set level (gene set enrichment analysis).

2 Materials and methods

2.1 Data

The datasets used in this study come from two related scRNAseq experiments aimed at investigating of the effects of navitoclax treatment on the transcriptome of triple-negative breast cancer cell line to better understand the process of developing drug resistance (Marczyk et al., 2020; Patwardhan et al., 2021). In both experiments, the same cancer cell line (MDA-MB-231) was used as a model organism and two biological replicates were provided (A and B). Cells were exposed to 10 μM navitoclax and harvested at 3 time points: before the treatment (baseline; T1), after treatment (T2), and after recovery from the treatment (T3).

In both cases, immediately after plate harvesting, cells were trypsinized and a single-cell suspension at a concentration of 1,000 cells/μl with viability above 90% was prepared. Chromium Single Cell 3′ Library and Gel Bead Kit V2 (PN-120237), Chromium Single Cell A Chip Kit (PN-120236), and Chromium i7 Multiplex Kit (PN-120262) were used to prepare single-cell libraries following the manufacturer's instructions. The same sequencer was used—HiSeq 4,000 (Illumina). In the first study (Patwardhan et al., 2021) 6,000 cells per sample were used (two samples were multiplexed on one lane) and 25,000 reads per cell were generated. In other study (Marczyk et al., 2020) 1,500 cells/sample were sequenced in one lane generating 200,000 reads/cell.

To simplify the evaluation procedure only two time points (T1 and T2) from both datasets (experiments) were considered (Figure 1). Each experiment corresponds to a different design. The first experiment corresponds to a balanced study design where cells collected at different time points were split and processed on the same chip, on the same day (Marczyk et al., 2020). Two biological replicates termed replicate "A" and "B" were involved. This dataset serves as a reference. The second experiment corresponds to a confounded study design where cells collected at different time points were processed on different chips/batches (Patwardhan et al., 2021). This dataset termed a test set was corrected using different data integration methods for the removal of the batch effect.

TABLE 1 Selected scRNAseq data integration methods.

Tool	Input	Strategy	Reference
ComBat-seq	raw counts	linear model	Zhang et al. (2020)
limma	logcounts	linear model	Leek et al. (2012)
MNN	logcounts	mutual nearest neighbors (gene expression space)	Haghverdi et al. (2018)
scMerge	logcounts	stably expressed genes + RUV model	Lin et al. (2019)
Seurat	logcounts	canonical correlation analysis + mutual nearest neighbors	Stuart et al. (2019)
Scanorama	raw counts	mutual nearest neighbors (reduced space) + panoraming stiching	Hie et al. (2019)

2.2 Data preprocessing

The quality of raw RNA sequencing reads was assessed with FastQC (Andrews, 2010) and the reads were processed with 10x Genomics Cell Ranger 6.1.1 (Zheng et al., 2017) to generate a gene-cell count matrix. Quality control was performed separately for each dataset at cell- and gene-level. Adaptive, sample-specific thresholds were chosen for the number of UMI counts per cell, the number of genes, and the fraction of mitochondrial counts using median absolute deviation (MAD) from the median. Cells were considered of poor quality if a given metric was more than 3 MADs from the median in the wrong direction. Genes that were expressed in less than 1% of cells for each dataset were removed. Finally, we obtained expression matrices with the following dimensions (cells x genes): 12,402 × 4,180 for reference set (Marczyk et al., 2020) and 12,402 × 21,548 for test set (Patwardhan et al., 2021). Such filtered expression matrices were normalized separately using two approaches: deconvolution (Lun et al., 2016) for non-parametric DGE and transcript per million (TPM) metrics for parametric DGE, both followed by (log2+1)-transformation.

Selection of highly variable genes (HVGs) for each dataset was performed using the SCTransform function with variable features. $n = 5{,}000$ (Hafemeister and Satija, 2019). A common part of 3,620 HVGs was taken as input for data integration. We did not want to be too restrictive with subsampling, as high dimensionality is required for some methods (e.g., to satisfy the orthogonality assumption in MNN detection).

2.3 Data integration methods

Since the goal of this study was to evaluate the applicability of scRNAseq data integration methods in terms of further differential analysis, we selected only the methods that: (i) output full corrected expression matrix; (ii) work in an unsupervised manner as we don't have cell-type labels. Thus, we benchmarked six algorithms (Table 1) and for some of these tested two cases: (i) using all genes; (ii) using only top HVGs.

2.3.1 ComBat-seq

ComBat-seq (Zhang et al., 2020) takes two parameters as input: a raw, untransformed count matrix and a vector describing the annotation of samples into batches. It is also possible to specify biological covariates, whose signals will be preserved in the corrected data. In our case, the technical variable associated with the repetition was used as a batch separation vector and the biological variable was associated with a time point. ComBat-seq uses a negative binomial regression model to estimate batch effects. The computed batch-effect estimators are then used to calculate "batch-free" distributions, i.e., the expected distributions if there were no batch effects in the data based on the model (Zhang et al., 2020). Correction is performed by quantile normalization to make the two distributions (empirical and batch-free) with identical statistical properties. ComBat-seq is the only method that preserves the integer nature of counts making corrected data compatible with various differential expression software (e.g., edgeR, DESeq2).

2.3.2 Limma

Limma (Leek et al., 2012) is another linear method to remove batch effect components from the data. The correction is performed by subtraction of the estimated component from the original data. Limma batch-effect removal function (removeBatchEffect) takes normalized and log-transformed counts as an input. Similarly to ComBat-seq, it allows addition of batch annotations and biological covariates into the model.

2.3.3 Mutual nearest neighbor

MNN searches for mutual nearest neighbors (MNNs) between two datasets or batches in the gene expression space. A pair of MNNs consists of cells present in each batch set of nearest neighbors based on Euclidean distance. These cells are considered to be of the same type/state across batches (Haghverdi et al., 2018). Differences in expression between identified MNNs are used to compute the batch correction vector which is applied to all cells. mnnCorrect function was run with two setups: with all genes and with HVGs. In both cases normalized and log-transformed expression values were used. `merge. order,` argument was specified such as both repetitions from a given

time point were merged first and then combined. Thus, the merging order was as follows: first T1A + T1B and T2A + T2B. Then the summation results were added together. `cos.norm.out`, was set to FALSE to disable cosine normalization before computing corrected expression values to obtain corrected values on the log scale, similar to the input data. The rest parameters were set to default values.

2.3.4 scMerge

ScMerge (Lin et al., 2019) was run in the unsupervised mode as we do not have cell-type information. In this mode, the estimation of batch effects is performed on two levels: (i) identification of stably expressed genes (SEGs) across batches which serve as "negative control genes"; (ii) k-means clustering based on the HVGs followed by the identification of mutual nearest clusters (MNCs) from the batches based on Pearson correlation as the dissimilarity metric. Cells belonging to a pair of MNCs are considered to be of the same type in different batches and serve as pseudo replicates. SEGs and pseudo replicate information are the inputs for scMerge which uses the RUV model to adjust the data. We ran scMerge with three setups of kmeansK parameter: (5, 5, 5, 5), (4, 4, 4, 4) and (4, 4, 3, 3) on (log2+1)-transformed counts.

2.3.5 Seurat v4

Seurat v4 (Stuart et al., 2019) is another method based on the MNN concept (referred there as "anchors"). This method includes two approaches to match anchors across datasets/batches: Canonical Correlation Analysis (CCA) and reciprocal Principal Component Analysis (rPCA). In both cases, the searching of anchors is performed in a shared, reduced subspace obtained by CCA (linear combinations of genes with the maximum correlation between batches) or rPCA (maximum variation between batches). The correction vector is computed similarly to MNN (difference in expression profiles between two cells in each anchor). The batch integration order is derived from hierarchical clustering based on the distance between the datasets. Seurat v4 (version 4.0) was run according to the data integration tutorial on the web (https://satijalab.org/seurat/articles/integration_introduction.html).

2.3.6 Scanorama

In Scanorama (Hie et al., 2019) the nearest neighbor searching is performed in the low-dimensional subspace obtained by randomized singular value decomposition (SVD). The searching is performed across all batches and the priority of dataset merging is determined based on the percentage of matching cells in the batch. This reduces the risk of overcorrection. Scanorama was run using the reticulate R package following the tutorial (https://github.com/brianhie/scanorama). Two setups were evaluated: with all genes as input and using the top 2,000 HVGs based on data dispersion (internally selected by the algorithm).

2.4 Evaluation of data integration methods

2.4.1 Visual inspection of data

UMAP (McInnes and Healy, 2018) was employed for all data visualizations before and after data integration as it performs well at preserving global data structure. UMAP was run with default parameters using runUMAP function from scatter R package (McCarthy et al., 2017).

2.4.2 Differential gene expression analysis: Parametric and non-parametric approaches

Both datasets were processed through the same protocol to find differentially expressed genes using two approaches: the parametric method called MAST (Finak et al., 2015) and the non-parametric method called EMDomics (Nabavi et al., 2016). MAST uses a hurdle model to address bimodal expression distributions in scRNAseq data. The bimodality is manifested in such a way that observed expression is either strongly positive (continuous part) or non-detectable (discrete part). The Hurdle model parameterizes both parts and combines the information from them in the form of gene statistics to infer changes in expression levels. DE testing is performed across the two conditions through the LRT statistic. MAST was applied on the log2 (TPM + 1) expression matrix without including the cellular detection rate (the fraction of genes that are detected with non-zero counts) as a covariate in the model. The following thresholds were used for DEGs identification: an absolute value of log-fold change (LFC) higher than 2, and false discovery rate (FDR) lower than 0.001 (Benjamini–Hochberg method for multiple testing correction was used).

As an alternative when the corrected data do not fit the MAST model, the EMDomics method was used which does not make any assumptions about the data distribution. EMDomics uses the Earth Mover's distance (EMD) to measure the overall difference between the two normalized distributions (gene expression in two conditions/groups). This method is not restricted to finding only differences in mean expression between two conditions but also captures the overall difference in shape (bimodal vs. unimodal expression) between two distributions. EMDomics was applied to log-normalized counts with default parameters. DEGs were identified based on the following thresholds: emd score higher than 2 and FDR smaller than 0.001. In both cases, cells from two replicates (A and B) were compared between two time points (T1 vs. T2).

A receiver operating characteristic (ROC) curve was created by setting different thresholds on *p*-values from statistical tests while estimating DEGs. To calculate performance metrics, a reference dataset (with a balanced study design) was used as a "ground truth", and the sensitivity and specificity of each batch correction method were calculated.

2.4.3 GSEA

Differential expression was also performed at the level of gene sets using gene set enrichment analysis (GSEA). This step was done using the fGSEA R package (Korotkevich et al., 2019).

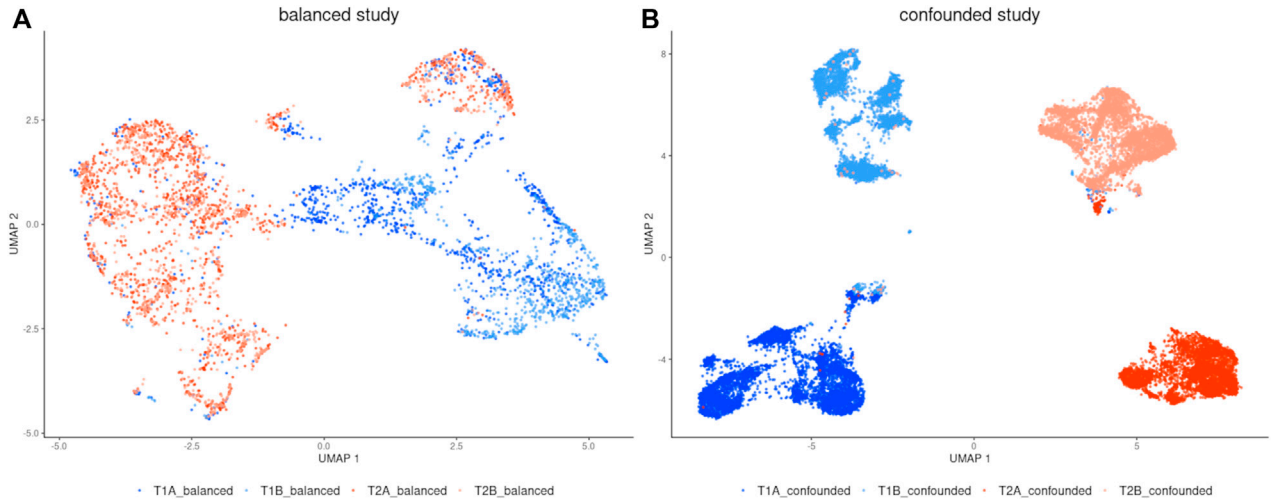


FIGURE 2
UMAP visualization of **(A)** balanced study, **(B)** confounded study (before correction). There are strong batch effects manifested in the confounded study.

DE test statistics obtained by MAST (continuous Z-score: C-component) were used as the ranking metrics. The following gene sets from Molecular Signatures Database (MSigDB) (Liberzon et al., 2015) were tested: Hallmark, Kegg, GO, and REACTOME. The total number of considered gene sets was 12,253; 50 gene sets for Hallmark, 186 for KEGG, 10,402 gene sets for GO, and 1,615 for REACTOME. Gene set was identified as differentially enriched based on a *p*-value lower than 0.05.

2.4.4 Correlation analysis

The correlation analysis was performed both at the level of individual genes (DGE) and gene sets (GSEA). For each data integration method, the previously mentioned DE test statistics were taken: (i) MAST: continuous Z-score (C-component) (ii) EMDomics: emd score (iii) GSEA: normalized enrichment score (NES). The correlation between the balanced (reference) study and the confounded dataset (test set, before correction) was assessed and used as the benchmark for assessing the quality of the data integration (Figure 1). Both, Pearson and Spearman correlation coefficients were calculated.

3 Results

3.1 Comparison of datasets before data integration

To visually examine the batch effect problem, the UMAP algorithm was run separately for each dataset (Figure 2). In a balanced study design (Figure 2A) there is strong segregation of cells along time points while cells from both repetitions are intermingled, which is desired. The opposite situation is observed in the confounded study design (Figure 2B) where together with separation along time points, the cells group by replicates which proves a strong batch effect. The main cause was that the samples in the confounded study were measured on different days.

Next, we calculated the following properties of individual genes at the single-cell level: mean expression, the variance of expression, and detection rate, which is a proportion of expressed cells (Supplementary Figure S1). We observe a typical situation that could be found in scRNAseq data: up to a mean normalized count of around 1, variance and mean are roughly equal as expected under a Poisson model either for balanced or confounded (before correction) study design. Genes with a higher average expression show overdispersion compared to Poisson distribution (Supplementary Figure S1B,C). As expected in scRNAseq data, in both experiments, many genes are expressed in very few cells. All feature-level statistics were comparable between balanced and confounded studies.

3.2 Differential analysis before data integration

The number of DEGs identified with the parametric approach was 965 for balanced and 191 for confounded study. The overlap between the two datasets was 63 genes, from which 43 genes were upregulated, and 20 genes were downregulated in the balanced study, and for the confounded study, the ratio of upregulated to downregulated genes was equal to 20/43. The correlation coefficients were equal to 0.16 (Pearson) and -0.21

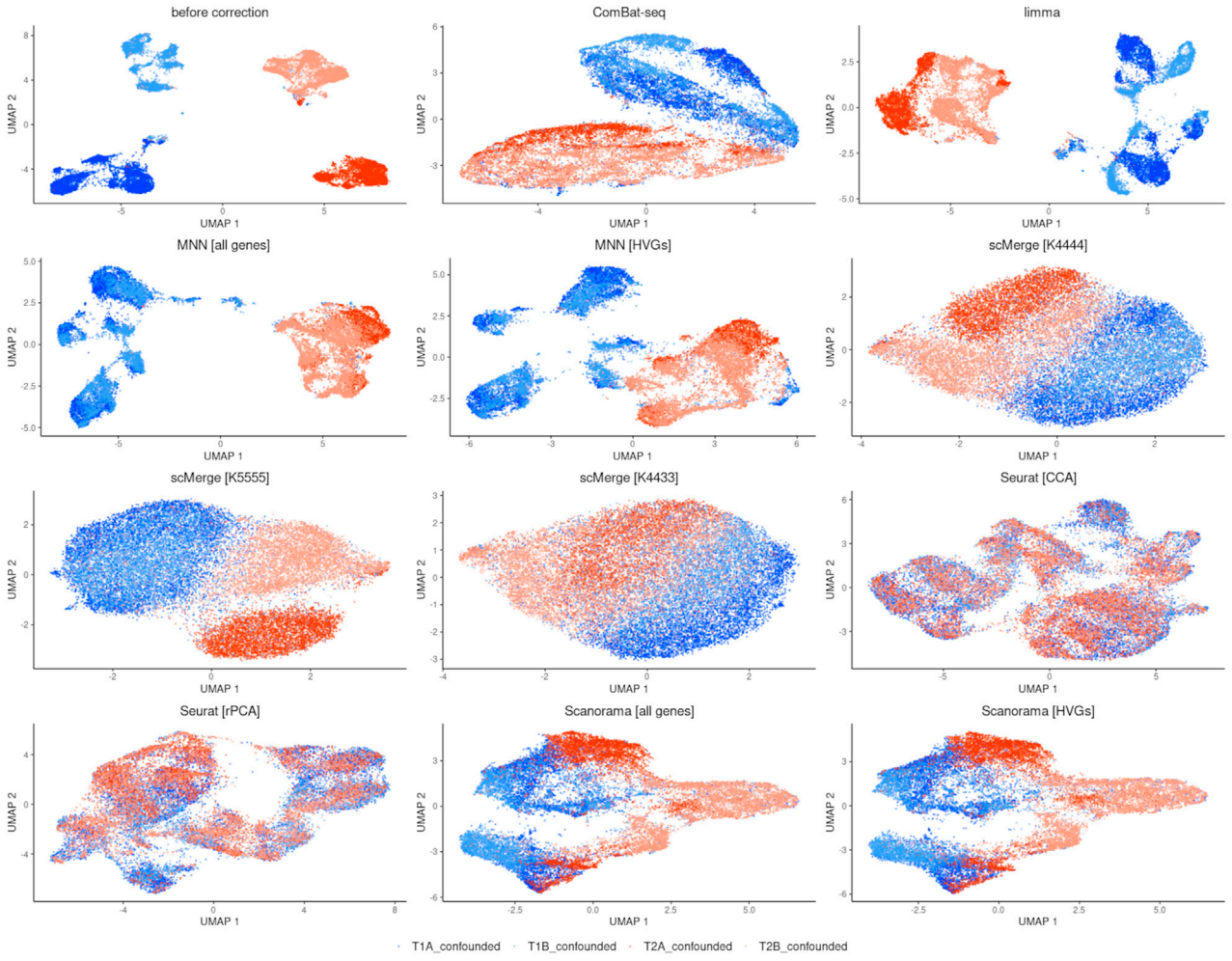


FIGURE 3
UMAP visualization after integration.

(Spearman) and both were significant. After using a non-parametric approach, the number of DEGs was smaller: 80 for balanced and 114 for confounded study. There were no common DEGs between datasets. The correlation between test statistics from both studies was much higher (Spearman: 0.72, Pearson: 0.75) than when the parametric method was used.

3.3 Data integration for batch effect correction

The UMAP plots (Figure 3) show that ComBat-seq might perform best in removing batch effects and preserving biological variation. It produced two strong clusters separated by time point, while the cells from technical repetitions are mixed well. In the case of the limma method, we observe separation by time point, but the repetitions are not mixed well—they seem to have a small tendency to group separately. MNN algorithm improved the separation by time point in both cases when all genes and only the top 3,620 HVGs were taken. However, within the time point T1 cells form characteristic subgroups are observed. scMerge performed visually best with *kmeansK* = (4,4,3,3). In other setups, there is an improvement in separation by time point over no correction, and technical replicates from T1 are well intermingled but not from T2 (replicate A clusters separately from replicate B). Seurat achieved the worst result by mixing all cells together, thus it was not evaluated in further comparisons. Scanorama achieved little improvement no matter if all genes were used or HVGs only.

As before, we counted the feature-level metrics after data integration (Figure 4). Except for ComBat-seq, genes with a higher average expression were not following the raw data distribution after correction (Figure 4A). Moreover, for

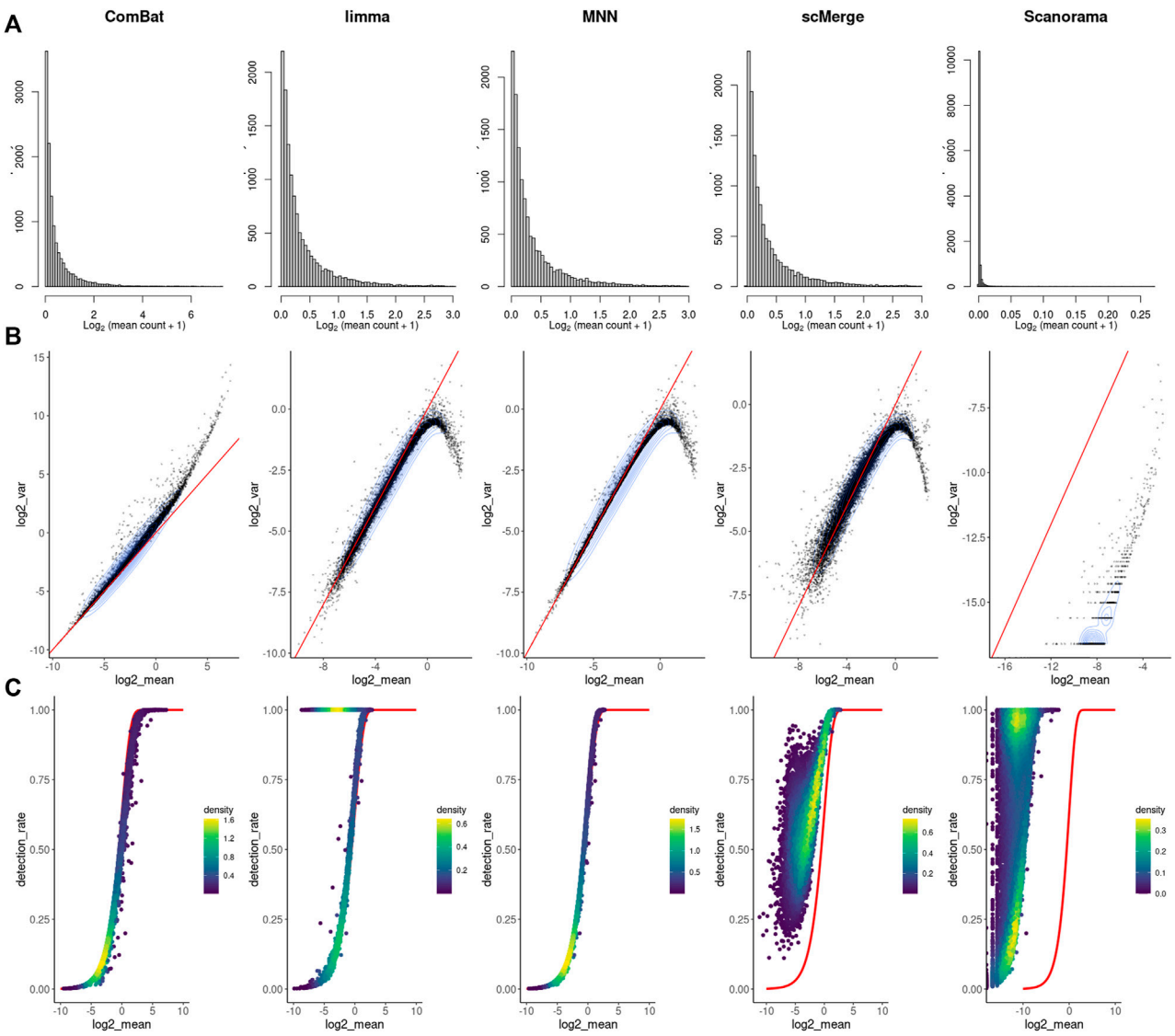


FIGURE 4
Feature-level metrics for corrected study. **(A)** histograms of mean value, **(B)** mean-variance relationship - red line with intercept = 0 and slope = 1, **(C)** mean-detection rate relationship - red line indicates the expected distribution under a Poisson model. Individual points are colored by the number of neighboring points.

MNN, Scanorama negative values started to occur in the corrected matrix. In most cases, the batch effect correction also distorts the characteristic of the scRNAseq data mean-variance relationship (Figure 4B). There is a sharp collapse of the log variance in the upper range of the mean expression (Figure 4B). The association between average expression and detection rate is conserved only for ComBat-seq and MNN (Figure 4C). Limma introduces small expression values to all cells for many low expression genes (dropout rate equal 1), while scMerge and Scanorama consequently increase dropout rate with increased expression of the gene.

3.4 Differential gene expression analysis after data integration

For each method, only the best DEGs finding results were shown from all the setups tested (Figures 5, 6): MNN and Scanorama were run with all genes as input and scMerge with K4444 setting. The number of DEGs identified with MAST (parametric approach) and EMDomics (non-parametric approach) is presented in Table 2. The intersection between different data integration methods and approaches for DEGs finding was small. For the confounded study, the number of identified DEGs was almost identical between the two

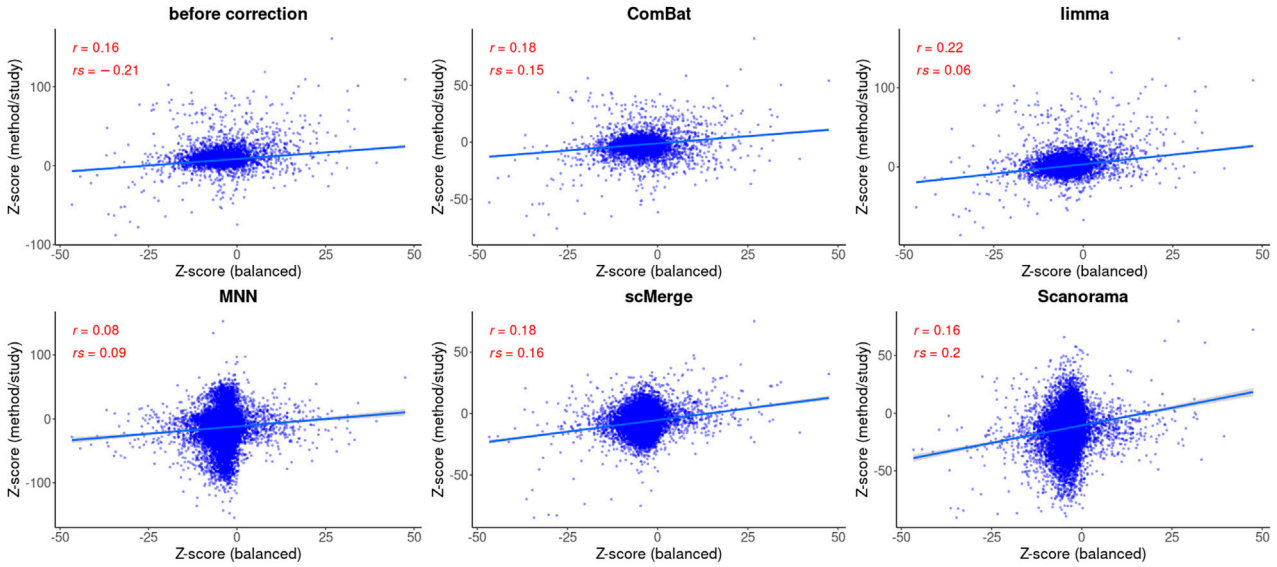


FIGURE 5
Correlation analysis after data integration using MAST statistics. Two correlation coefficients are shown: Pearson (R) and Spearman (ρ) and the corresponding p values. The regression model is fitted (blue line) with confidence intervals (the grey area around the line).

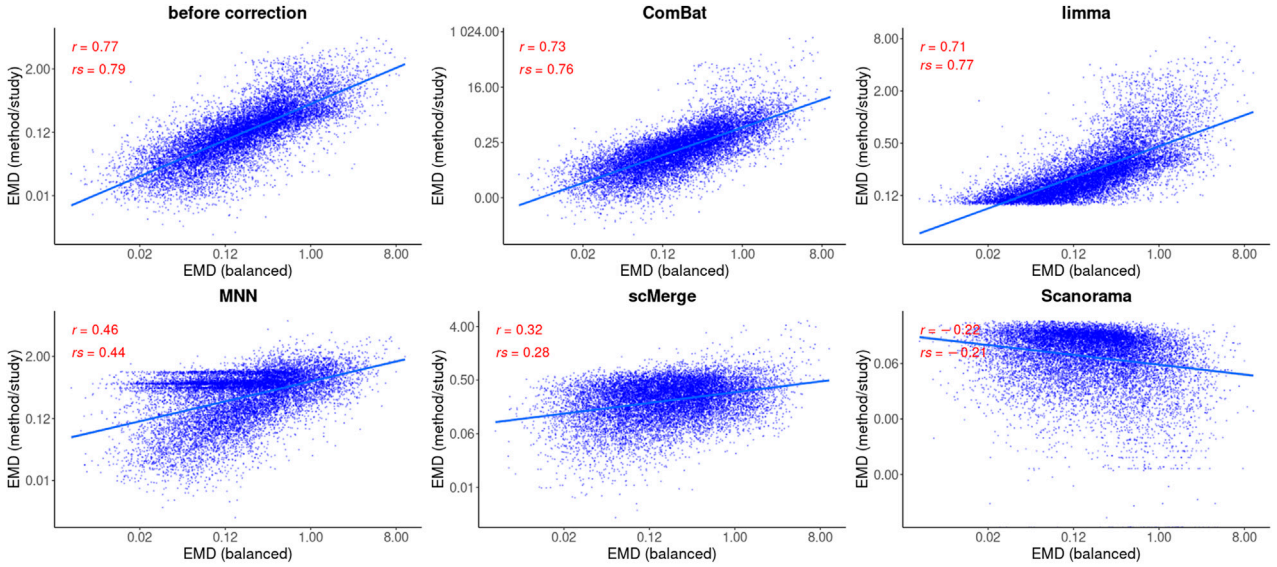


FIGURE 6
Correlation analysis after data integration using EMDomics statistics. Two correlation coefficients are shown: Pearson (R) and Spearman (ρ) and the corresponding p values. Regression model is fitted (blue line) with confidence intervals (grey area around the line).

approaches, but the common part consists of only 62 genes (Table 2). After data integration, only ComBat-seq gave a higher number of DEGs than other methods, mostly when the parametric approach was used. The non-parametric approach identified a significantly larger number of DEGs after correction for other data integration methods.

Based on the Pearson correlation coefficient (R), there is an improvement in the correlation of MAST DE statistics

TABLE 2 Number of DEGs after correction.

Study/tool	[MAST]	[EMDomics]	Intersection
balanced	965	328	246
confounded	191	197	62
ComBat-seq	287	115	114
limma	9	206	9
MNN	6	137	6
scMerge	0	20	0
Seurat	0	0	0
Scanorama	0	0	0

between the reference and the corrected study in the case of ComBat-seq, limma, and scMerge (Figure 5). For MNN and Scanorama, the test statistics themselves were much higher, thus the correlation with the reference is smaller (Figure 5). When the Spearman correlation coefficient is considered (ρ), the correlation is higher for every integration method, and Scanorama, scMerge, and ComBat-seq are the best. For a non-parametric test approach, after data integration, both correlation coefficients were smaller in all cases (Figure 6). However, ComBat-seq and limma showed the smallest decrease, while Scanorama gave negative correlation values. In some cases, rank-based EMDomics gave the same value of test statistic (dots arranged in horizontal lines in Figure 6), which follows from assigning the same expression values for individual genes after batch correction using selected methods (e.g., limma, MNN).

ROC curves calculated for each method and statistical tool (Figure 7) support the findings of correlation analysis. Only for ComBat-seq and limma, the area under the ROC curve was higher than 0.5 (Combat-seq: 0.72 and 0.86; limma: 0.74 and 0.65). The worst method was Scanorama (0.39 and 0.44).

3.5 Gene set analysis after data integration

The number of significantly enriched pathways for selected gene sets is presented in Table 3. Overall, a smaller number of enriched pathways was found after correction. Data integration using ComBat-seq did not improve the correlation coefficients for any of the considered gene sets (Figure 8; Supplementary Table S1), but the dissimilarity was small. The opposite is observed in the case of limma, where the correlation improvement was found for all gene sets and both coefficients. scMerge improved both coefficients for Hallmark and GO and worsened for KEGG and Reactome. MNN and Scanorama worsened the correlation for every gene set.

4 Discussion

We tested six scRNAseq data integration methods against two experimentally derived datasets which, in some sense, are mirror images of each other. Both experiments had the same biological properties such as cell line, drug, time of harvesting, etc. The only difference was in the technical study design; one experiment was designed to minimize the technical variation and was our reference,

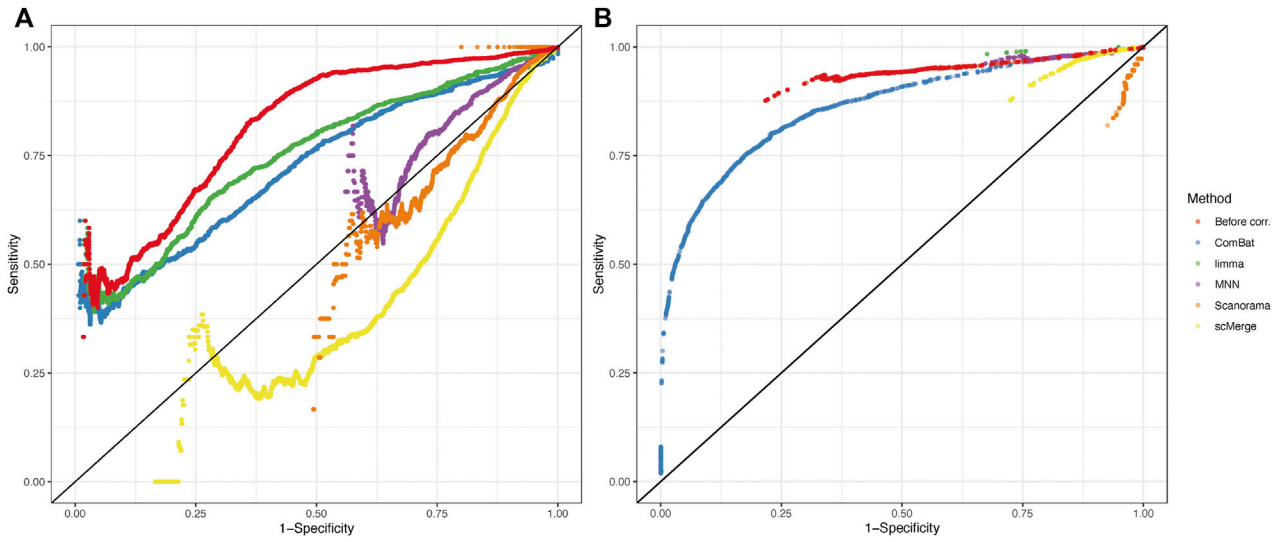


FIGURE 7
Receiver operating characteristic (ROC) curves from results of DEGs analysis using MAST **(A)** and EMDomics **(B)** tests. Color coding represents different data integration methods or no correction (red).

TABLE 3 Number of enriched pathways for selected gene sets.

Study/tool	Hallmark	KEGG	GO	Reactome
balanced	31	19	568	219
confounded	21	14	387	51
ComBat-seq	0	0	25	13
limma	12	7	153	29
MNN	3	8	77	51
scMerge	4	3	46	16
Seurat [CCA]	15	9	326	10
Seurat [rPCA]	16	3	235	71
Scanorama	0	5	59	5

while the other manifested strong batch effects due to the difference in capturing time of each batch. This dataset was corrected for batch removal. Our study was not intended to evaluate many aspects of the batch correction (accuracy, speed, scalability) as other published benchmarks, but is focused on one unexplored so far aspect of scRNAseq data integration which is its impact on DGE analysis in real data scenario. Available benchmarks also address this problem, however, based only on the simulated data scenarios. While these evaluations can easily compute the number of true/false positive DEGs identified in corrected datasets, they do not stress the real challenge behind DGE analysis on batch-corrected datasets by excluding multiple technical and biological factors occurring in real data. For example, R package splatter (Zappia et al., 2017) simulates the batch effect by randomly generating multiplication factors from a log-normal distribution for each gene and group of cells (i.e., batch). However, since all cells within a batch are modified in the same way, parametric statistical tests can easily handle these artificial batch effects by adding covariates to the model. Thus, our study is unique and extends previous comparisons.

In this work, we tried to emphasize the challenge involved in feature-level analyses on corrected gene expression matrices. Indeed, cell-level analyses which are based on computing the distance (clustering or trajectory analysis) are safe to apply to corrected data because all cells are placed in the same coordinate space, which is the idea of data integration. However, integration algorithms give no guarantee to preserve relative differences in gene expression space. Therefore, correction methods may introduce artificial differential expression between cell types or conditions. Moreover, a majority of integration tools change the original nature of scRNAseq data: counts are no longer counts. One exception is ComBat-seq which preserves the integer nature

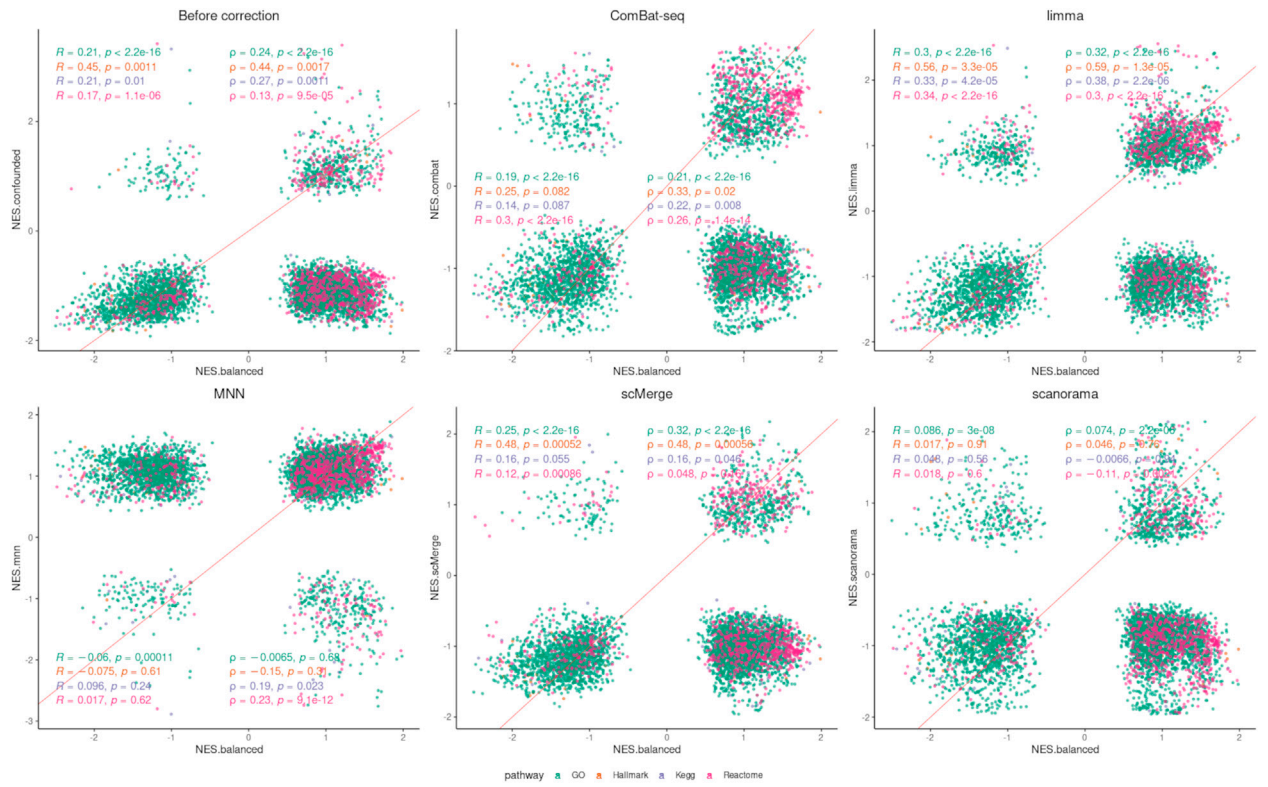


FIGURE 8
Correlation analysis (GSEA) after data integration using NES. Two correlation coefficients are shown: Pearson (R) and Spearman (ρ) and the corresponding *p* values separately for each pathway/gene set.

TABLE 4 Summary of comparison between data integration methods.

Method	Total time [sec]	Single core time [sec]	Total RAM [MiB]	Peak RAM [MiB]	Ease of use	Original data distribution	UMAP separation
ComBat-seq	4,007.2	4,007.2	2,039.5	30,652.1	easy	not changed	good
limma	24.2	24.2	2,039.2	15,302.1	easy	changed	medium
MNN (8 cores)	52,300.6	418,405	2,040.1	28,121.4	easy	changed	medium
scMerge (8 cores)	14,404.9	115,240	2,042.8	28,117.2	easy	changed	medium
Seurat (5 workers)	1,116.3	5,581.5	4,486.6	17,066.4	easy	changed	weak
Scanorama	5,925.9	2,542.7	2,055.4	2,055.6	medium	changed	weak

of counts. Counts preservation is important for the compatibility of a corrected matrix with the available tools for differential expression analysis which may require counts or values equivalent to counts. A natural consequence of subtracting expression during integration (for example in MNN or Scanorama) is negative values in the corrected matrix which are hard to biological interpretation. Moreover, the scale of corrected values can be much different from the original counts which were especially apparent for Scanorama. Therefore, corrected values can no longer be considered as expression measures (of course still higher values reflect higher expression). Model-based methods specifically designed for scRNAseq DGE analysis (parametric approaches) may not work well with corrected data given the fact that many properties of original data are lost, and higher expressed genes are dragged down after correction. Of course, one can attempt to apply some transformations (e.g., Box-Cox transformation) on corrected data, but they are computationally intensive and do not guarantee the intended effect.

In general, gene set enrichment analysis should be more robust against batch correction than gene level analysis but in our case, this was not manifested. ComBat-seq which was best on DGE analysis (in both, number of DEGs and correlation with balanced study) did not improve correlations on the level of gene sets, but it also did not decrease it much.

In terms of computational time, limma was the fastest algorithm, while Scanorama used the least amount of memory (Table 4). MNN ran on 8 processor cores, was much slower than others (even algorithms ran on a single core) and in peaks, it needed almost 30 GB of memory. We summarized all our findings when comparing data integration methods in Table 4. Our evaluations were done on a machine with Intel® Xeon(R) CPU E5-2,650 v3 at 2.30GHz × 40 and 256 GB RAM.

Our study has some limitations. First, the analysis was done on a set of two experiments concerning the same cancer cell line. The results might slightly differ for other organisms. However, since there is no other pair of experimentally derived balanced/confounded studies, it was not possible to test it. Second, different methods have multiple parameters to set. We have chosen default values where possible and tested a few settings for another method, however, we are aware that the optimal settings might not be reached in this study.

Finally, we are rather careful with formulating overall recommendations for the particular method as well as we do not state that DGE analysis should not be performed at all. We rather wanted to highlight the fact that single-cell data integration is one of the current grand challenges (Lahnemann et al., 2020) in omics analyses and better methods might still appear. Nevertheless, we wanted to highlight the ComBat-seq method as it led to the highest correlation of test statistics between reference and corrected dataset among others and it does not distort the original distribution of gene expression, so it can be used in all types of downstream analyses.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: [Reference study]: https://doi.org/10.3390/cancers12092551 [Test study]: https://doi.org/10.1038/s41523-021-00270-4.

Ethics statement

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

Author contributions

MM and JP conceived the concept of the study and supervised the methodology. TK was responsible for the data acquisition, data analysis, and visualization. JP was responsible for financing. All authors wrote and approved the final version of the article.

Funding

This work was financed by the Silesian University of Technology grant no. 02/070/BK22/0033 for maintaining and developing research potential (MM, JP) and co-financed by the European Union through the European Social Fund grant POWR.03.02.00–00-I029 (TK).

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1009316/full#supplementary-material

References

Adil, A., Kumar, V., Jan, A. T., and Asger, M. (2021). Single-cell transcriptomics: Current methods and challenges in data acquisition and analysis. *Front. Neurosci.* 15, 591122. doi:10.3389/fnins.2021.591122

Andrews, S. (2010). *FastQC: A quality control tool for high throughput sequence data [Online]*. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Bao, X., Li, Q., Chen, J., Chen, D., Ye, C., Dai, X., et al. (2022). Molecular subgroups of intrahepatic cholangiocarcinoma discovered by single-cell RNA sequencing–assisted multiomics analysis. *Cancer Immunol. Res.* 10 (7), 811–828. doi:10.1158/2326-6066.cir-21-1101

Chazarra-Gil, R., van Dongen, S., Kiselev, V. Y., and Hemberg, M. (2021). Flexible comparison of batch correction methods for single-cell RNA-seq using BatchBench. *Nucleic Acids Res.* 49 (7), e42. doi:10.1093/nar/gkab004

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, 278. doi:10.1186/s13059-015-0844-5

Hafemeister, C., and Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* 20 (1), 296. doi:10.1186/s13059-019-1874-1

Haghverdi, L., Lun, A. T. L., Morgan, M. D., and Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* 36 (5), 421–427. doi:10.1038/nbt.4091

Hie, B., Bryson, B., and Berger, B. (2019). Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* 37 (6), 685–691. doi:10.1038/s41587-019-0113-3

Korotkevich, G., Sukhov, V., and Sergushichev, A. (2019). Fast gene set enrichment analysis. bioRxiv, 060012. doi:10.1101/060012

Lahnemann, D., Koster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., et al. (2020). Eleven grand challenges in single-cell data science. *Genome Biol.* 21 (1), 31. doi:10.1186/s13059-020-1926-6

Lee, M. C., Lopez-Diaz, 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 (44), E4726–E4735. doi:10.1073/pnas.1404656111

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28 (6), 882–883. doi:10.1093/bioinformatics/bts034

Liberzon, A., Birger, C., Thorvaldsdottir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* 1 (6), 417–425. doi:10.1016/j.cels.2015.12.004

Lin, Y., Ghazanfar, S., Wang, K. Y. X., Gagnon-Bartsch, J. A., Lo, K. K., Su, X., et al. (2019). scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets. *Proc. Natl. Acad. Sci. U. S. A.* 116 (20), 9775–9784. doi:10.1073/pnas.1820006116

Liu, J., Gao, C., Sodicoff, J., Kozareva, V., Macosko, E. Z., and Welch, J. D. (2020). Jointly defining cell types from multiple single-cell datasets using LIGER. *Nat. Protoc.* 15 (11), 3632–3662. doi:10.1038/s41596-020-0391-8

Luecken, M. D., Buttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M. F., et al. (2022). Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* 19 (1), 41–50. doi:10.1038/s41592-021-01336-8

Lun, A. T., Bach, K., and Marioni, J. C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* 17, 75. doi:10.1186/s13059-016-0947-7

Marczyk, M., Patwardhan, G. A., Zhao, J., Qu, R., Li, X., Wali, V. B., et al. (2020). Multi-omics investigation of innate navitoclax resistance in triple-negative breast cancer cells. *Cancers* 12 (9), 2551. doi:10.3390/cancers12092551

McCarthy, D. J., Campbell, K. R., Lun, A. T., and Wills, Q. F. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* 33 (8), 1179–1186. doi:10.1093/bioinformatics/btw777

McInnes, L., and Healy, J. (2018). UMAP: uniform manifold approximation and projection for dimension reduction. *ArXiv* (2018) abs/1802.03426.

Ming, J., Lin, Z., Zhao, J., Wan, X., Tabula Microcebus, C., Yang, C., et al. (2022). FIRM: Flexible integration of single-cell RNA-sequencing data for large-scale multi-tissue cell atlas datasets. *Brief. Bioinform.* 23, bbac167. doi:10.1093/bib/bbac167

Nabavi, S., Schmolze, D., Maitituoheti, M., Malladi, S., and Beck, A. H. (2016). EMDomics: a robust and powerful method for the identification of genes differentially expressed between heterogeneous classes. *Bioinformatics* 32 (4), 533–541. doi:10.1093/bioinformatics/btv634

Patwardhan, G. A., Marczyk, M., Wali, V. B., Stern, D. F., Pusztai, L., and Hatzis, C. (2021). Treatment scheduling effects on the evolution of drug resistance in heterogeneous cancer cell populations. *NPJ Breast Cancer* 7 (1), 60. doi:10.1038/s41523-021-00270-4

Qian, Y., Zhai, E., Chen, S., Liu, Y., Ma, Y., Chen, J., et al. (2022). Single-cell RNA-seq dissecting heterogeneity of tumor cells and comprehensive dynamics in tumor microenvironment during lymph nodes metastasis in gastric cancer. *Int. J. Cancer* 151, 1367–1381. doi:10.1002/ijc.34172

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., 3rd, et al. (2019). Comprehensive integration of single-cell data. *Cell* 177 (7), 1888–1902. doi:10.1016/j.cell.2019.05.031

Tran, H. T. N., Ang, K. S., Chevrier, M., Zhang, X., Lee, N. Y. S., Goh, M., et al. (2020). A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* 21 (1), 12. doi:10.1186/s13059-019-1850-9

Zappia, L., Phipson, B., and Oshlack, A. (2017). Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.* 18, 174. doi:10.1186/s13059-017-1305-0

Zhang, Y., Parmigiani, G., and Johnson, W. E. (2020). ComBat-seq: batch effect adjustment for RNA-seq count data. *Nar. Genom. Bioinform.* 2 (3), lqaa078. doi:10.1093/nargab/lqaa078

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* 8, 14049. doi:10.1038/ncomms14049

frontiers | Frontiers in Genetics

TYPE Case Report
PUBLISHED 10 February 2023
DOI 10.3389/fgene.2022.987175

OPEN ACCESS

EDITED BY
Zhichao Liu,
Boehringer Ingelheim, United States

REVIEWED BY
Kaustav Bera,
Maimonides Medical Center, United States
Kar-Tong Tan,
Harvard University, United States

*CORRESPONDENCE
Donald J. Johann,
djjohann@uams.edu

†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 05 July 2022
ACCEPTED 01 December 2022
PUBLISHED 10 February 2023

CITATION
Ma L, Peterson EA, Shin IJ, Muesse J, Marino K, Steliga MA, Atiq O, Arnaoutakis K, Wardell C, Wooldridge J, Prior F and Johann DJ (2023), An advanced molecular medicine case report of a rare human tumor using genomics, pathomics, and radiomics. *Front. Genet.* 13:987175. doi: 10.3389/fgene.2022.987175

An advanced molecular medicine case report of a rare human tumor using genomics, pathomics, and radiomics

Li Ma[1,2†], Erich A. Peterson[1,3†], Ik Jae Shin[1], Jason Muesse[1], Katy Marino[1], Mathew A. Steliga[1], Omar Atiq[1], Konstantinos Arnaoutakis[1], Christopher Wardell[3], Jacob Wooldridge[3], Fred Prior[3,4] and Donald J. Johann[1,3*]

[1]Winthrop P. Rockefeller Cancer Institute, University of Arkansas for Medical Sciences, Little Rock, AR, United States, [2]Department of Information Science, University of Arkansas at Little Rock, Little Rock, AR, United States, [3]Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR, United States, [4]Department of Radiology, University of Arkansas for Medical Sciences, Little Rock, AR, United States

Background: Pulmonary Sclerosing Pneumocytoma (PSP) is a rare tumor of the lung with a low malignant potential that primarily affects females. Initial studies of PSP focused primarily on analyzing features uncovered using conventional X-ray or CT imaging. In recent years, because of the widespread use of next-generation sequencing (NGS), the study of PSP at the molecular-level has emerged.

Methods: Analytical approaches involving genomics, radiomics, and pathomics were performed. Genomics studies involved both DNA and RNA analyses. DNA analyses included the patient's tumor and germline tissues and involved targeted panel sequencing and copy number analyses. RNA analyses included tumor and adjacent normal tissues and involved studies covering expressed mutations, differential gene expression, gene fusions and molecular pathways. Radiomics approaches were utilized on clinical imaging studies and pathomics techniques were applied to tumor whole slide images.

Results: A comprehensive molecular profiling endeavor involving over 50 genomic analyses corresponding to 16 sequencing datasets of this rare neoplasm of the lung were generated along with detailed radiomic and pathomic analyses to reveal insights into the etiology and molecular behavior of the patient's tumor. Driving mutations (AKT1) and compromised tumor suppression pathways (TP53) were revealed. To ensure the accuracy and reproducibility of this study, a software infrastructure and methodology known as NPARS, which encapsulates NGS and associated data, open-source software libraries and tools including versions, and reporting features for large and complex genomic studies was used.

Conclusion: Moving beyond descriptive analyses towards more functional understandings of tumor etiology, behavior, and improved therapeutic predictability requires a spectrum of quantitative molecular medicine

approaches and integrations. To-date this is the most comprehensive study of a patient with PSP, which is a rare tumor of the lung. Detailed radiomic, pathomic and genomic molecular profiling approaches were performed to reveal insights regarding the etiology and molecular behavior. In the event of recurrence, a rational therapy plan is proposed based on the uncovered molecular findings.

KEYWORDS

pulmonary sclerosing pneumocytoma, molecular profiling, TP53 signaling pathway, genomics, radiomics, pathomics, case report

1 Introduction

Pulmonary Sclerosing Pneumocytoma (PSP) is a relatively uncommon benign tumor of the lung with potential for malignant transformation that is manifested most commonly by metastasis to regional lymph nodes (Zheng et al., 2022). PSP was first reported by Liebow in 1956 (Liebow and Hubbell, 1956), and shows a striking female predominance (female to male ratio 5:1) (Kalhor et al., 2010). Histologically, PSP is primarily composed of 2 cell types (cuboidal epithelial and polygonal stromal cells) and four histological types (hemorrhagic, sclerotic, solid and papillary) (Gao et al., 2020).

Due to the lack of noteworthy clinical or imaging findings, PSP is hard to recognize, and most cases are diagnosed by histopathological analysis (Song et al., 2021). The neoplasm may be confused with other benign nodules like hamartoma, tuberculoma, bronchial cysts, or certain lung cancers (Cheung et al., 2003). Often, patients are asymptomatic and PSP is detected incidentally. Non-specific associated symptoms may include: cough, chest pain, chest tightness and hemoptysis (Cardemil et al., 2004).

Initial studies of PSP focused primarily on analyzing features discovered using conventional X-ray or CT imaging. PSP has been described as a distinct, juxta-pleural nodule with strong and homogeneous enhancement on CT (Im et al., 1994; Xie et al., 2003). Nevertheless, using the above-mentioned techniques, there are no specific or classic imaging findings associated with PSP (Wang et al., 2011).

In recent years, because of the widespread use of next-generation sequencing (NGS), the study of PSP at the molecular-level has emerged. PSP lacks the classic driver gene mutational signatures of lung adenocarcinoma, e.g., EGFR, KRAS; ALK, or ROS1 fusions (Sartori et al., 2007; Pal and Chetty, 2020). A study utilizing whole-exome sequencing to explore genomic modifications in PSP has been performed (Jung et al., 2016). That study confirmed a high frequency of AKT1 point mutations (overall 31 of 68 patients, 46%) including p.E17K. It has been postulated that AKT1 mutations are the genetic hallmark of PSP (Yeh et al., 2020). Another study revealed that irregular activation of the mTOR pathway is a consistent genetic event in PSP (Boland et al., 2021). The PI3K/AKT/mTOR pathway is one of the most frequently activated oncogenic pathways (Porta et al., 2014), and activated AKT phosphorylates mTOR, which activates mTORC1.

This is the first study to use an advanced quantitative molecular medicine approach to provide a more thorough description of PSP. Using a combination of genomics, radiomics (Lambin et al., 2017) and pathomics (Gupta et al., 2019) a comprehensive description of the patient's presentation as well as the molecular determinants of this rare tumor are provided along with a precision medicine therapy plan in case of recurrence.

2 Case presentation

The patient is a pre-menopausal female who was admitted to the hospital because of progressive and severe left sided flank pain over a 1-week duration. The patient was a former smoker (cigarettes, one pack/day) for 7 years, who quit 2 years ago. She currently uses vaping products on a regular basis. The initial clinical suspicion included a possible kidney stone; however, imaging studies were negative for stones, but did reveal a 3 cm mass in the left lower lung. Following a referral to medical oncology a lobectomy of the left lower lung for curative intent was performed by thoracic surgery. Histopathologic features were consistent with pulmonary pneumocytoma cell types, the tumor measured 3.2 cm in greatest dimension, surgical margins were clean, and two hilar/peribronchial lymphnodes were negative for malignancy (stage Ib, p.T2a.N0.M0, NCCN v.3.2022). Also identified were abundant hemosiderin-laden macrophages, compatible with vaping related lung injuries.

3 Methods

3.1 Ethical compliance

This study is part of a clinical trial (NCT02597738) approved by the Institutional Review Board of the University of Arkansas for Medical Sciences (UAMS). As part of this trial, written informed consent was obtained from the patient for research use of clinical specimens and associated data.

3.2 Genomics sample preparation

The QIAGEN QIAseq Human Lung Cancer Panel (DHS-005Z) library prep kit (QIAGEN, 2022) was used for targeted DNA-based assays involving tumor and normal (T/N). Supplementary File 1 in BED format contains the exact regions of interest for the amplicon-based assay. An Illumina HiSeq 3000 was utilized for all NGS studies. The lung cancer panel, which utilizes uniform molecular identifiers (UMIs) was run with a coverage of 3,000x for the tumor and 600x for the germline. Whole genome sequencing (WGS) libraries were constructed using the New England BioLabs (NEB) NEBNext Ultra II DNA library prep kit (NEB, 2022), and sequenced in an ultra-low-pass fashion for copy number analysis (CNA) at ~0.3x coverage for T/N. For RNA-based experiments, the Illumina TruSeq Stranded Total RNA library prep kit (Illumina, 2022) was used. Six biological replicates were utilized for the tumor and six for the normal adjacent lung tissue. Sequencing was targeted at 200M reads for these 12 samples. In summary, six biological replicates of the tumor and adjacent normal lung (12 RNA NGS libraries) were built and sequenced, and four DNA libraries were built and sequenced.

3.3 Genomics molecular profiling

Genomics datasets were processed as previously reported by the NGS Post-pipeline Accuracy and Reproducibility System (NPARS), a reproducible software infrastructure developed by our group (Ma et al., 2021). Three separate pathway analysis tools were utilized and all run using default parameters. For canonical signaling pathway analysis, two traditional pathway analysis tools were used, pathfindR v1.6.3 (Ulgen et al., 2019) and Gene Set Enrichment Analysis (GSEA) v4.2.3 (Aravind et al., 2005). Additionally, an unsupervised pathway analysis tool named Weighted Correlation Network Analysis (WGCNA) v1.71 (Langfelder and Horvath, 2008) was used and then limma (v3.52.1) based methods were employed to further elucidate outputs generated by WGCNA. A normalized RNA-seq gene counts matrix, which was generated by NPARS via DESeq2 v1.36.0 (Love et al., 2014), was used as input for signaling pathway analyses.

3.4 Radiomics

DICOM imaging studies from the initial medical workup were obtained from the UAMS PACS and converted to NIfTI format. Segmentations and visualizations were produced using 3D Slicer v4.13 (Fedorov et al., 2012). Tumor segmentations (performed via thresholding techniques) were produced from CT studies. The border region was segmented by adding a margin of 10 mm to the tumor. Radiomic features were extracted from original images using Pyradiomics (van Griethuysen et al., 2017), both in aggregate for segmentations and as feature maps. A bin width of 25 voxels was used, and feature maps used a kernel radius of 1 voxel and calculated in 2D space. The entropy radiomics feature used is defined by the Image Biomarker Standardization Initiative as intensity histogram entropy (Zwanenburg et al., 2020).

3.5 Pathomics

Whole slide images were acquired using an Aperio CS2 whole slide imaging scanner (Leica Biosystems) at ×40 magnification. Image analysis was performed using the open-source program QuPath (v0.3.2) that included a suite of tools (Bankhead et al., 2017). Representative areas of the slide were annotated by a pathologist, indicating areas of tumor, hemosiderin-laden macrophages, and background lung parenchyma. From these areas, cell nuclei were segmented using StarDist with the *he_heavy_augment* model as described in the QuPath documentation (Schmidt et al., 2018). Cell expansion was enabled to approximate overall cell size. Cell classification was accomplished using the built-in object classifier to train a random trees classifier using the default feature extractor. Features included measurements of area, shape and, color of nuclei, cytoplasm, and overall cell.

4 Results

Figure 1 shows the salient medical imaging for the patient and results from radiomics analyses. Sub-image **(A)**, shows a pre-operative chest CT image with contrast, zoomed to show a more optimal view of the tumor in the left lower lung. Segmentations of the tumor and a 1.0 cm circumferential border were performed. At presentation, the tumor had a maximum diameter of 3.2 cm, minimum diameter of 2.8 cm and a volume of 19.6 cm^3. The median radiodensity of the tumor was 41 HU, approximately midway between the median densities of the kidneys (24 HU) and the liver (58 HU). As reference, the median density of normal lung (alveolar space) is ~ -650.

As part of cancer staging a PET/CT study (**B**) was performed. Raw PET values were converted to standardized uptake values (SUV). The mean SUV in the tumor was 1.3 with a maximum SUV of 2.2. A reference volume of approximately 3 cm was measured in the liver (standard comparison), which had a mean SUV of 1.2 and maximum SUV of 1.5, implying that the tumor had relatively low metabolic activity.

From the CT study, radiomic features were extracted (**C**) and compared between the tumor and surrounding border region representing the tumor microenvironment. Radiomic features are most informative when comparing many similar tumors, but salient information can be inferred from a single case. We extracted the entropy of the segmentations (**C**), which is a measure of the amount of information required to encode the voxels of the image. Entropy measures the randomness of the voxel values, where low values represent more homogenous

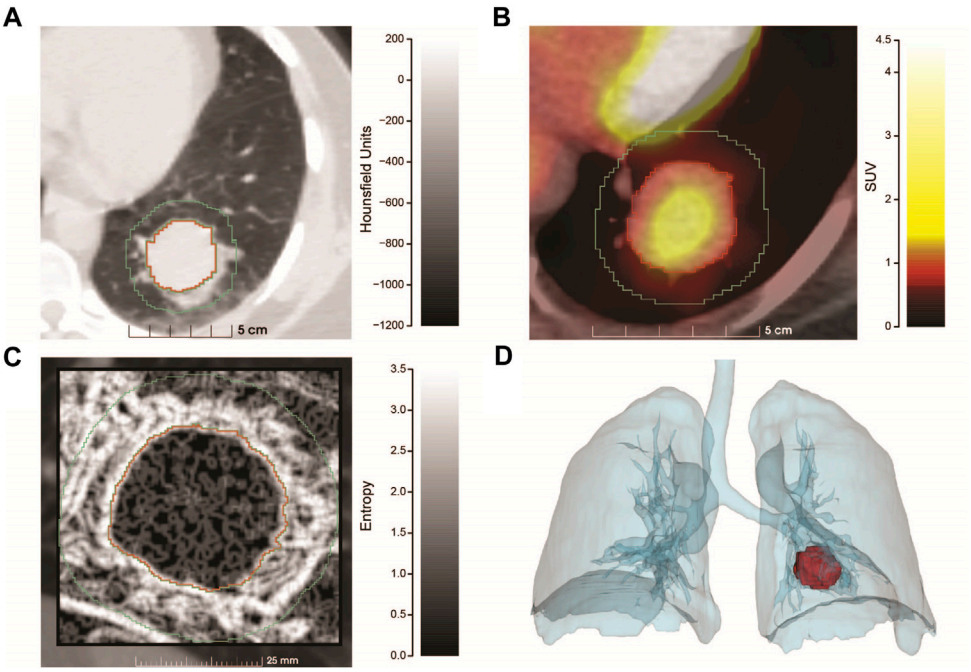


FIGURE 1
Radiomics analysis of the PSP tumor. **(A)** Pre-operative chest CT scan with contrast utilizing lung window settings. The image is an axial projection that has been zoomed to show an optimal view of the tumor that resides in the left lower lung along with a small region of the mediastinum. Tumor segmentation is outlined in red, with the 1 cm border surrounding the tumor proper, outlined in green. The x-axis contains a size scale (cm) and y-axis Hounsfield Units (HU) scale (−1200–200) with shading. **(B)** Combined PET/CT of the tumor (zoomed) at diagnosis. The tumor had a SUV max of 2.2 and SUV mean of 1.3, the x-axis contains a size scale (cm) and y-axis contains the SUV scale (0–4.5) with color coding. **(C)** Feature map showing the entropy of the tumor and 1 cm surrounding region, generated from a sagittal slice of the CT at presentation. The tumor is significantly more homogenous than the surrounding region. The x-axis contains a size scale (mm) and y-axis contains an entropy scale (0–3.5) with shading. **(D)** Volume rendering showing the size and position of the tumor at diagnosis. Produced using segmentations of the lungs and tumor from the PET/CT series.

regions and higher values represent more heterogenous regions. The median entropy of the tumor and border regions were 0.92 and 1.89 respectively, illustrating that the microenvironment (border region) was more complex (heterogeneous). This result was highly statistically significant using a two-sided Wilcoxon test ($p < 2.2 \times 10^{-16}$). Finally, volume rendering showing the size and position of the tumor (**D**) was produced using segmentations of the lungs and tumor from the PET/CT study.

Figure 2 displays the results of pathomics analyses. As background, nuclear segmentation using StarDist performed well overall, with the primary deficiencies being occasional segmentation of large cytoplasmic blebs without a visible nucleus, over-estimation of nuclear size in foamy macrophages, and difficulty distinguishing nuclei from hemosiderin in some hemosiderin-laden macrophages. In Supplementary Figure 1, examples of measurement maps corresponding to cell circularity are shown overlaid onto intermediate magnification photomicrographs of background lung and the tumor. In Figure 2, the pathologist's annotations (**A**) are shown in a low-power (4x) photomicrograph for areas containing tumor (red) and hemosiderin laden macrophages (blue). Density maps for cells classified as tumor (**B**), and as hemosiderin-laden macrophages (**C**) for a region of tissue which was not used for classifier training are displayed separately and then jointly (**D**).

Figure 3 displays a graphic produced by RCircos v1.2.2 (Zhang et al., 2013), which summarizes and integrates the findings of seven genomics methods into a single graphical image. The layout of the RCircos diagram is as follows, from the outmost circle inward this plot contains: i. human chromosomal ideogram, ii. lung cancer targeted 72 gene panel for T/N, iii. RNA expressed mutations from the full transcriptome (represented as a "dot" due to spacing), iv. WGS DNA T/N CNA with the red color representing amplification, black for normal, and deletion as blue, v. Tumor RNA gene expression and, vi. Tumor RNA gene fusions. In our study, 52 total genomic analyses were generated and analyzed, specifically: DNA targeted panel T/N, DNA ultra-low-pass WGS T/N for CNA, RNA studies involving six biological replicates from the tumor and the normal adjacent lung (12 samples) subjected to: 1) RNA expressed mutation analysis, ii) statistical inferencing with DESEq2 (Love et al., 2014), and iii) Fusion analysis via STAR-Fusion (Haas et al., 2019). Supplementary Figure 2 illustrates the tissue specimens and genomic analyses (total of 52) generated.

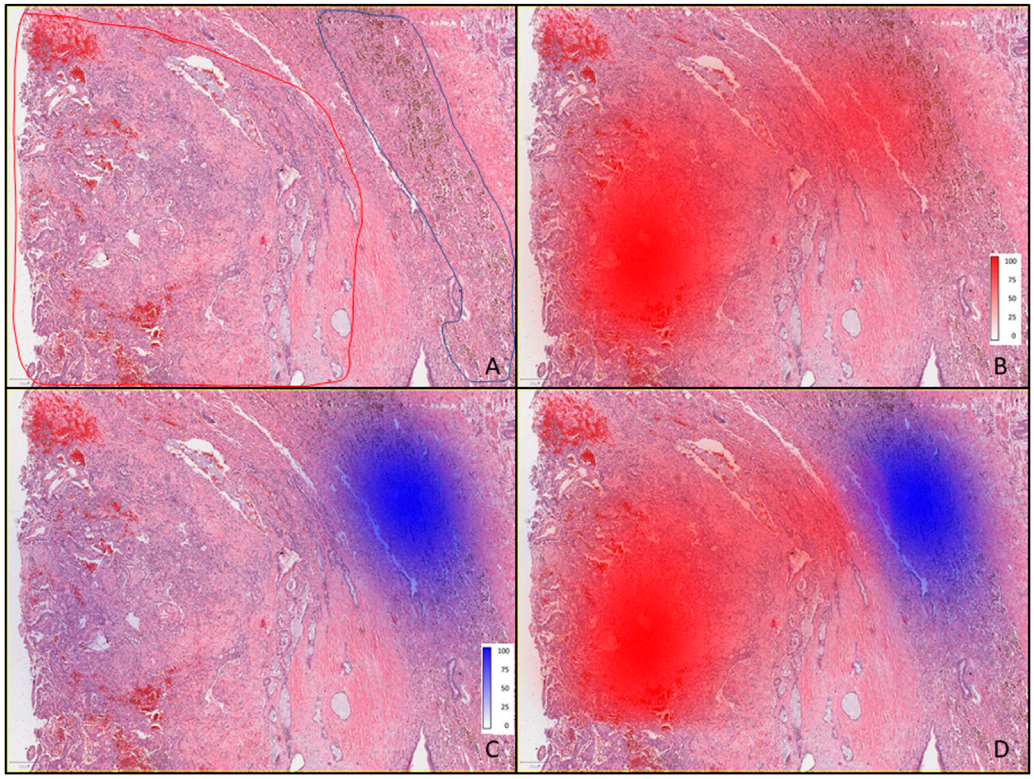


FIGURE 2
Pathomics analysis of the PSP tumor. **(A)** Low-power (4×) photomicrograph showing areas containing tumor (outlined red) and hemosiderin-laden macrophages (outlined blue) as annotated by the pathologist. **(B)** Tumor with red color density maps showing the number of cells per mm^2 as identified by the classifier, and shown as percentages (0–100), where the 100% scale value corresponds to 1660 cells per mm^2, along with intense red coloring. **(C)** Tumor tissue with blue color density maps showing hemosiderin-laden macrophages where the most intense blue color and scale value of 100% corresponds to 349 cells per mm^2 (as identified by the classifier). **(D)** Overlaid density maps for both cell types (same classifier results and color intensity scales as in **(B,C)**.

Examining Figure 3, three somatic non-synonymous DNA mutations were found by the targeted DNA panel: AKT1 p.E17K, NF1 p.H1826Y, APC p.V1822D, with sequencing depths of 6,243 (allelic frequency: 36.75%), 5,809 (6.72%), 9,735 (61.6%) respectively (see Supplementary Table 1 for targeted DNA panel details). The AKT1 mutation is a driver for PSP tumors (Yeh et al., 2020), the findings for NF1 and APC are not drivers. The germline TP53 mutation p.P72R was detected with a depth of 1573 and an allelic frequency of 50%, but this is not indicated to be of significance per ClinVar (TP53). Finally, a TP53 p.K382fs frameshift mutation was found at the low allelic frequency of 0.6% and a depth of 5158; however, the mutation did not pass filter by smCounter2 (Xu et al., 2019) (homopolymer).

Due to RNA-seq experiments covering the entire transcriptome and the use of six biological replicates, a total of 1,119,654 RNA expressed mutations were found to pass filter by HaplotypeCaller (DePristo et al., 2011; Van der Auwera et al., 2013). Using the recommended depth filter of 10 from Guo et al. and limiting mutations to those having a predicted impact of moderate or high, the RNA expressed mutation analysis was further filtered (Guo et al., 2017). After filtering, 8,139 mutations remained for further analysis. Among these mutations, 2,938 of them are found in all six tumor samples (see Supplementary Table 2), and 1,854 mutations are private to specific samples (see Supplementary Table 3). Based on the RNA-seq VCF files of the six tumor samples and the six normal samples, a phylogentic analysis was performed using PHYLIP v3.697 (PHYLIP) (see Supplementary Figure 3). The PHYLIP dendrogram shows a clear separation of tumor vs. normal and with the tumor arising from the normal. The driving mutation found in the DNA study, AKT1 p.E17K was expressed in five of six RNA biological replicates with a depth range of 101–471, and VAF range of 28%–49% (see Supplementary Table 4).

Ultra-low pass WGS experiments revealed copy number variations concentrated in chromosomes 5, 10, 14, 17, 19 and 21 (all amplifications). All the three DNA mutated genes, AKT1, NF1, and APC, were amplified (see Supplementary Table 5; Supplementary Figure 4). A differential gene expression (DGE) analysis was performed by DESeq2 (Love et al., 2014) on the RNA-seq data via NPARS. DGE analysis revealed

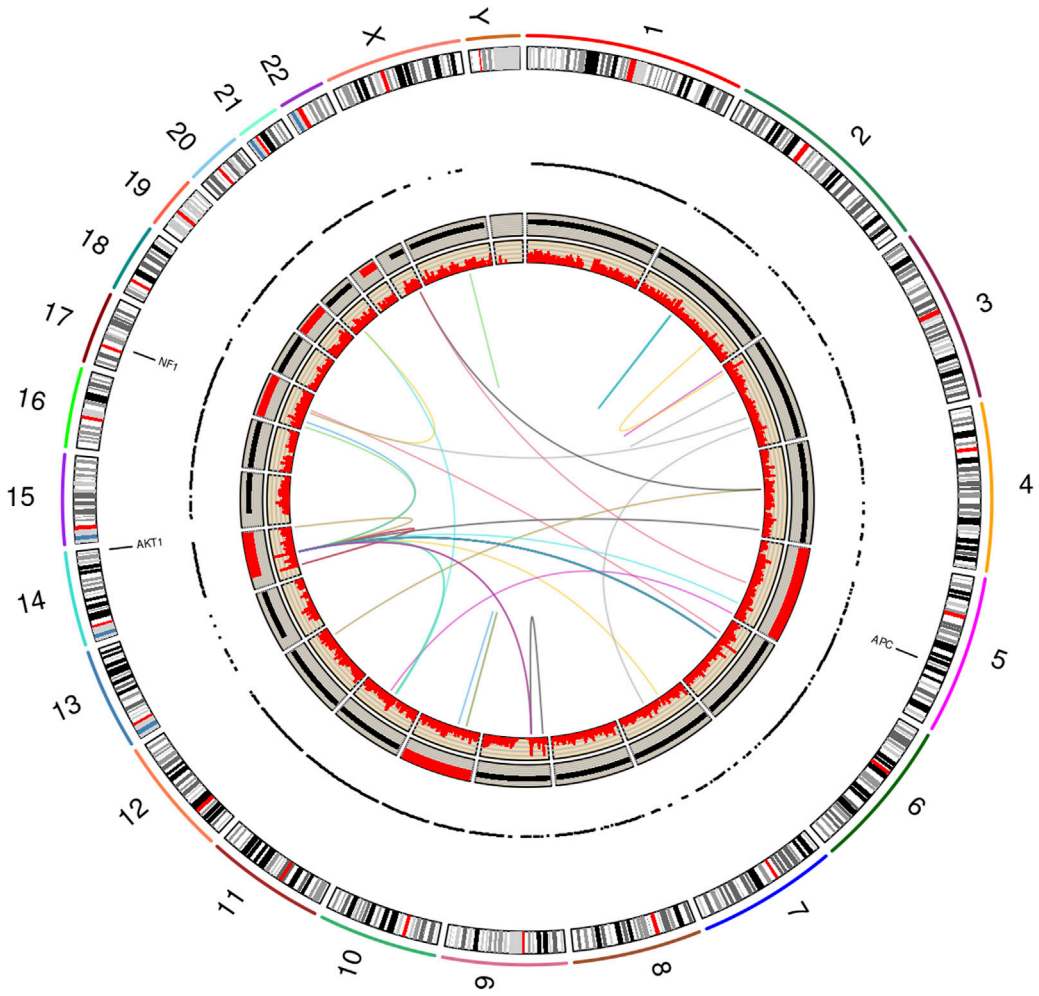


FIGURE 3
RCircos plot produced by the NGS Post-pipeline Accuracy and Reproducibility System (NPARS). This figure summarizes and integrates seven genomics methods into one graphical plot. From the outermost ring inward: (i) human chromosomal ideogram, (ii) DNA panel mutations (tumor vs. germline), (iii) RNA expressed mutations from the full transcriptome, each dot represents a RNA expressed mutation (depth greater than or equal 10), (iv) whole genome DNA copy number variations (tumor vs. germline) with red representing a copy number greater than 2, copy number equal to 2 by black coloring, and a copy number smaller than 2 by blue, (v) RNA gene expression (TPM) and, (vi) RNA gene fusions.

11,646 genes to be significantly differentially expressed (adjusted p-value < 0.1) between the tumor and matched normal adjacent lung replicates (see Supplementary Table 6). A significant finding was the overexpression of MDM2 in the tumor (log2 fold change: 1.33; q-value: 2.93E-11), a key regulator in the TP53 pathway.

RNA-seq gene fusion analysis showed a number of fusion events across the genome (see Supplementary Table 7), with TIMM23-PARGP1 found in all six tumor replicates. However, the TIMM23-to-PARGP1 fusion does not drive PSP, in the literature to-date. The total distinct fusions found across all six tumor replicates and passed by STAR-Fusion was 36.

Using RNA-seq data (tumor and normal adjacent lung biological replicates), both conventional signaling pathway analysis tools, pathfindR and GSEA, found a large number of abnormal candidate pathways. The pathways found to be statistically significant by pathfindR are listed in Supplementary Table 8. The GSEA's most significant pathways are listed in Supplementary Table 9. WGCNA initially clusters genes into significant modules (in this study, there are total of 100 modules). Then using the R package limma v3.52.1, the most significantly differentiated modules were extracted (Ritchie et al., 2015). Next, the most differentiated module (module number 1, containing 5,108 genes), was sent to pathfindR for further analysis. The most significant pathways for genes within module number 1 were identified (see Supplementary Table 10). Comparing the output from these pathway analysis tools, we found that the TP53 signaling pathway to be statistically significant by all three pathway

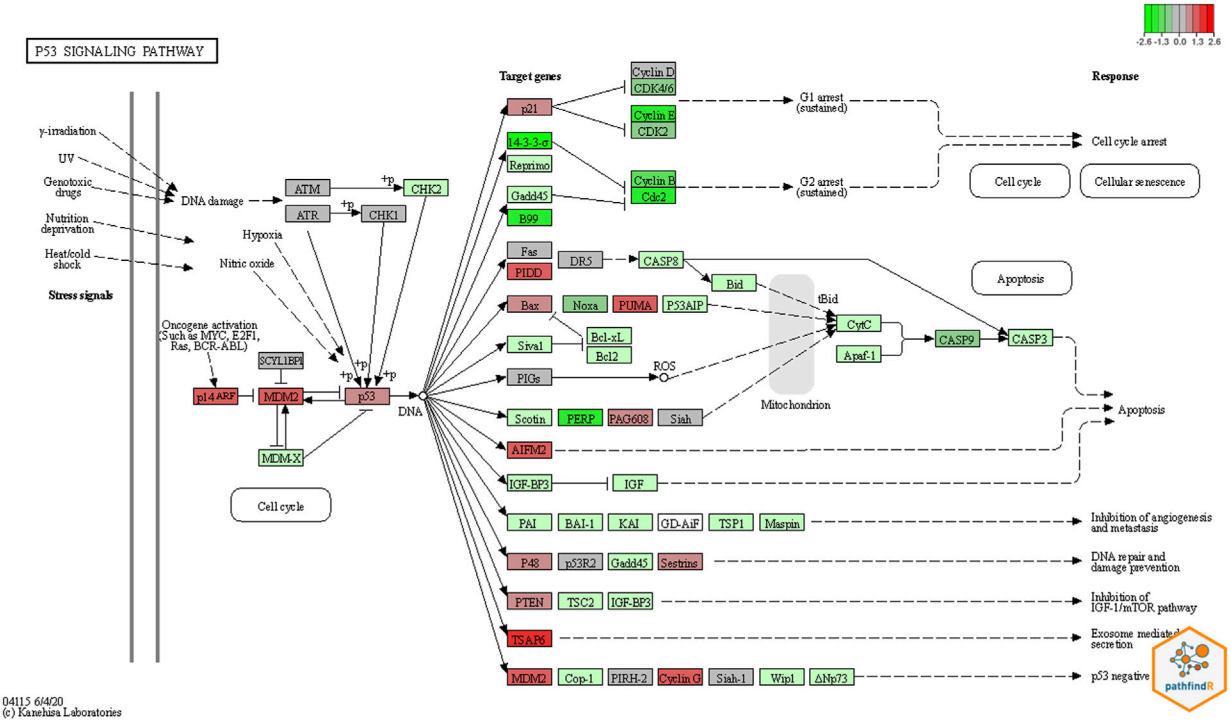


FIGURE 4
Pathway Result by *pathfindR*. The colored plot was generated based on KEGG pathway diagrams through *pathfindR*. Red colors represent upregulated genes, and green colors the down regulated genes. In this plot for the TP53 pathway, MDM2 which is the principal negative regulator of the pathway, was significantly upregulated.

analysis tools, and MDM2 overexpressed. Using pathfindR's KEGG (Kanehisa and Goto, 2000) integration, the TP53 pathway shown in Figure 4.

5 Discussion and conclusion

Why does a relatively young woman develop an unusual tumor in her lung? How is her presentation involving left flank pain related to her pathologic processes? Using genomics, radiomics and pathomics we sought to bring additional clarity to these questions.

The patient presented with severe left flank pain. It is established that disease processes or injuries involving the lower lung may present as flank pain (LeBlond, 2015). The 3D position of the tumor and the proximity to the left lung base is nicely displayed by the radiomics study in Figure 1D. Utilizing segmentation and entropy calculations (Figure 1C) radiomics showed the tumor region to be much more homogeneous vs. a surrounding 1 cm rim representing an inflamed microenvironment, which is now known to be filled with abundant hemosiderin-laden macrophages. Hemosiderin-laden macrophages are an important finding regarding an acute lung injury and indicates alveolar hemorrhage (Beasley, 2010). This finding was also observed and quantified by the pathomics study (Figures 2C,D). The patient's lung injury is related to her vaping practices and may be manifested in left lower lung due to tumor growth and corresponding increased metabolism (Figure 1B).

The first principal genomic finding of this study, was the detection of the AKT1 p.E17K mutation within both the DNA and RNA of the patient's tumor with convincing VAF and depth of coverage. This finding is consistent with previous studies that have shown many PSP cases to harbor AKT1 mutations (Jung et al., 2016; Yeh et al., 2020). There is a growing body of evidence that AKT1 mutations are a hallmark of PSP (Yeh et al., 2020), and this oncogene can be assumed to be the driving mutation for this patient's tumor.

AKT1 is a member of the AKT kinase family. As meaningful down-stream regulators of the PI3K signaling pathway, members of the AKT kinase family play an import role. In all cancers, the PI3K/AKT pathway is considered one of the most frequently deranged (Mundi et al., 2016). Although our signaling study did not find the pathway to be statistically significant, the pathway contains a mutated AKT1, driving tumor proliferation (Yeh et al., 2020), and is a viable drug target.

The second principal genomic finding, was that the TP53 signaling pathway was found to be statistically significant in all three pathway analysis methods. Chief among the alteration of genes in the TP53 pathway is that the p53 inhibitor MDM2 is significantly over-expressed in the patient's tumor. The overexpression of MDM2 in tumors

inhibits p53 and favors an uncontrolled environment for cell proliferation (Chène, 2003; Hou et al., 2019). This helps to explain an additional reason for tumor development. Namely, a dampened response regarding tumor suppressor function by an essential pathway focused on tumor surveillance and eradication.

In the TP53 signaling pathway, p53 and MDM2 proteins form a central hub which is one of the key molecular complexes most frequently connected to other signaling pathways in the cell. The MDM2-p53 hub receives stress inputs, and by forming and changing a large number of other pathways and functions in the cell, p53 responds to the inputs (Levine, 2020). The MDM2-p53 hub is also a negative feedback loop. In this loop, MDM2 is transcriptionally induced by p53, but reciprocally blocks p53 activity (Zhou et al., 2017). According to the colored KEGG pathway plot generated by pathfindR (Figure 4), it is evident that the MDM2 gene is significantly upregulated.

Per standard-of-care guidelines, the patient had a lung surgery for curative intent, but a precision oncology therapy plan was formulated as a precaution in case of tumor recurrence. Active clinical trials enrolling patients that target MDM2 abnormalities and AKT1 p.E17K mutations exist. Regarding MDM2 inhibitors: (i) RO5045337 (Roche), prevents the MDM2 protein from binding to the transcriptional activation domain of p53 (NCI, 2022; Roche, 2022); (ii) siremadlin (HDM201, Novartis), increases the activity of the tumor suppressor p53 by selectively inhibiting the MDM2-p53 interaction (Novartis, 2022; Stein et al., 2022); and, (iii) alrizomadlin (APG-115, Ascentage), restores p53 expression by binding to MDM2 protein (Tolcher et al., 2019; Ascentage, 2022). Regarding the AKT1 finding, there are two small molecule drugs targeting the ATK1 p.E17K mutation being investigated: (i) capivasertib (AZD5363, AstraZeneca), inhibits all three isoforms of AKT by inhibiting downstream signaling of the AKT1 p.E17K mutation, (Chen et al., 2020; Kalinsky et al., 2021; AstraZeneca, 2022); and, (ii) BAY1125976 (Bayer), deactivates full-length AKT1 by binding into an allosteric binding pocket (Politz et al., 2017; Bayer, 2022) (see Supplementary Table 11).

To date, this study provides the most comprehensive analysis of a single human PSP neoplasm by utilizing radiomics, pathomics, and multiple genomic analyses. Using these studies insights are gleaned and discussed that span the patient's initial presentation, tumor development with molecular determinants, and a precision medicine therapy plan is proposed in case of recurrence.

Data availability statement

The original contributions presented in the study are publicly available. The public study report page and summary-level phenotype data may be browsed at dbGaP: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs003154.v1.p1. The Individual-level data and sequence data are now available for download: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs003154.v1.p1. Data dictionaries and variable summaries are available on the dbGaP FTP site: https://ftp.ncbi.nlm.nih.gov/dbgap/studies/phs003154/phs003154.v1.p1/.

Ethics statement

The studies involving human participants were reviewed and approved by the University of Arkansas for Medical Sciences Institutional Review Board. The patients/participants provided their written informed consent to participate in this study.

Author contributions

DJ conceived the project. LM, EP, and DJ devised the experiments. LM and EP performed the software implementation. IS performed genomics laboratory experiments. MS, JM, KM, KA, and OA performed clinical duties and patient care. LM, EP, and DJ performed genomics data analyses. CW and FP performed radiomics analyses. JW and FP performed pathomics analyses. DJ, EP, and LM wrote the manuscript. All authors read and approved the manuscript.

Acknowledgments

The authors would like to acknowledge the financial support of the United States Department of Health and Human Services, Food and Drug Administration, contract HHSF223201610111C through the Arkansas Research Alliance. Funding through the National Cancer Institute 1U24CA215109 is acknowledged by FP and CW.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.987175/full#supplementary-material

References

Aravind, S., Pablo, T., Sayan, M., Amanda, P., 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 (43), 15545–15550. doi:10.1073/pnas.0506580102

Ascentage (2022). *APG-115 in patients with advanced solid tumors or lymphomas*. Bethesda, Maryland: U. S. National Library of Medicine.

AstraZeneca (2022). *Safety, tolerability & potential anti-cancer activity of increasing doses of AZD5363 in different treatment schedules - full text view - ClinicalTrials.gov*. Bethesda, Maryland: U. S. National Library of Medicine.

Bankhead, P., Loughrey, M. B., Fernandez, J. A., Dombrowski, Y., McArt, D. G., Dunne, P. D., et al. (2017). QuPath: Open source software for digital pathology image analysis. *Sci. Rep.* 7 (1), 16878. doi:10.1038/s41598-017-17204-5

Bayer (2022). *Phase I dose escalation study with an allosteric AKT 1/2 inhibitor in patients - full text view - ClinicalTrials.gov*. Bethesda, Maryland: U. S. National Library of Medicine.

Beasley, M. B. (2010). The pathologist's approach to acute lung injury. *Arch. Pathol. Lab. Med.* 134 (5), 719–727. doi:10.1043/1543-2165-134.5.719

Boland, J. M., Lee, H. E., Barr Fritcher, E. G., Voss, J. S., Jessen, E., Davila, J. I., et al. (2021). Molecular genetic landscape of sclerosing pneumocytomas. *Am. J. Clin. Pathol.* 155 (3), 397–404. doi:10.1093/ajcp/aqaa136

Cardemil, G., Fernndez, E., Riffo, P., Reyes, D., Ledezma, R., Mira, M., et al. (2004). Sclerosing hemangioma presenting as a solitary lung nodule. Report of one case. *Rev. Med. Chil.* 132 (7), 853–856. doi:10.4067/s0034-98872004000700010

Chen, Y., Huang, L., Dong, Y., Tao, C., Zhang, R., Shao, H., et al. (2020). Effect of AKT1 (p. E17K) hotspot mutation on malignant tumorigenesis and prognosis. *Front. Cell Dev. Biol.* 8. doi:10.3389/fcell.2020.573599

Chène, P. (2003). Inhibiting the p53–MDM2 interaction: An important target for cancer therapy. *Nat. Rev. Cancer* 3 (2), 102–109. doi:10.1038/nrc991

Cheung, Y. C., Ng, S. H., Chang, J. W. C., Tan, C. F., Huang, S. F., and Yu, C. T. (2003). Histopathological and CT features of pulmonary sclerosing haemangiomas. *Clin. Radiol.* 58 (8), 630–635. doi:10.1016/s0009-9260(03)00177-6

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43 (5), 491–498. doi:10.1038/ng.806

Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J. C., Pujol, S., et al. (2012). 3D slicer as an image computing platform for the quantitative imaging network. *Magn. Reson. Imaging* 30 (9), 1323–1341. doi:10.1016/j.mri.2012.05.001

Gao, Q., Zhou, J., Zheng, Y., Cui, J., and Teng, X. (2020). Clinical and histopathological features of pulmonary sclerosing pneumocytoma with dense spindle stromal cells and lymph node metastasis. *Histopathology* 77 (5), 718–727. doi:10.1111/his.14159

Guo, Y., Zhao, S. L., Sheng, Q. H., Samuels, D. C., and Shyr, Y. (2017). The discrepancy among single nucleotide variants detected by DNA and RNA high throughput sequencing data. *Bmc Genomics* 18, 690. doi:10.1186/s12864-017-4022-x

Gupta, R., Kurc, T., Sharma, A., Almeida, J. S., and Saltz, J. (2019). The emergence of pathomics. *Curr. Pathobiol. Rep.* 7 (3), 73–84. doi:10.1007/s40139-019-00200-x

Haas, B. J., Dobin, A., Li, B., Stransky, N., Pochet, N., and Regev, A. (2019). Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol.* 20 (1), 213. doi:10.1186/s13059-019-1842-9

Hou, H., Sun, D., and Zhang, X. (2019). The role of MDM2 amplification and overexpression in therapeutic resistance of malignant tumors. *Cancer Cell Int.* 19 (1), 216–218. doi:10.1186/s12935-019-0937-4

Illumina: TruSeq stranded total RNA. In.; 2022.

Im, J. G., Kim, W. H., Han, M. C., Han, Y. M., Chung, J. W., Ahn, J. M., et al. (1994). Sclerosing hemangiomas of the lung and interlobar fissures: CT findings. *J. Comput. Assist. Tomogr.* 18 (1), 34–38. doi:10.1097/00004728-199401000-00007

Jung, S. H., Kim, M. S., Lee, S. H., Park, H. C., Choi, H. J., Maeng, L., et al. (2016). Whole-exome sequencing identifies recurrent AKT1 mutations in sclerosing hemangioma of lung. *Proc. Natl. Acad. Sci. U. S. A.* 113 (38), 10672–10677. doi:10.1073/pnas.1606946113

Kalhor, N., Staerkel, G. A., and Moran, C. A. (2010). So-called sclerosing hemangioma of lung: Current concept. *Ann. Diagn. Pathol.* 14 (1), 60–67. doi:10.1016/j.anndiagpath.2009.07.002

Kalinsky, K., Hong, F., McCourt, C. K., Sachdev, J. C., Mitchell, E. P., Zwiebel, J. A., et al. (2021). Effect of capivasertib in patients with an AKT1 E17K-mutated tumor: NCI-MATCH subprotocol EAY131-Y nonrandomized trial. *JAMA Oncol.* 7 (2), 271–278. doi:10.1001/jamaoncol.2020.6741

Kanehisa, M., and Goto, S. (2000). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28 (1), 27–30. doi:10.1093/nar/28.1.27

Lambin, P., Leijenaar, R. T. H., Deist, T. M., Peerlings, J., de Jong, E. E. C., van Timmeren, J., et al. (2017). Radiomics: The bridge between medical imaging and personalized medicine. *Nat. Rev. Clin. Oncol.* 14 (12), 749–762. doi:10.1038/nrclinonc.2017.141

Langfelder, P., and Horvath, S. (2008). Wgcna: an R package for weighted correlation network analysis. *BMC Bioinforma.* 9 (1), 559. doi:10.1186/1471-2105-9-559

LeBlond, R. F. (2015). *DeGowin's diagnostic examination*. New York: McGraw-Hill Education.

Levine, A. J. (2020). P53: 800 million years of evolution and 40 Years of discovery. *Nat. Rev. Cancer* 20 (8), 471–480. doi:10.1038/s41568-020-0262-1

Liebow, A. A., and Hubbell, D. S. (1956). Sclerosing hemangioma (histiocytoma, xanthoma) of the lung. *Cancer* 9 (1), 53–75. doi:10.1002/1097-0142(195601/02)9:1<53:aid-cncr2820090104>3.0.co;2-u

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15 (12), 550. doi:10.1186/s13059-014-0550-8

Ma, L., Peterson, E. A., Shin, I. J., Muesse, J., Marino, K., Steliga, M. A., et al. (2021). NPARS-A novel approach to address accuracy and reproducibility in genomic data science. *Front. Big Data* 4, 725095. doi:10.3389/fdata.2021.725095

Mundi, P. S., Sachdev, J., McCourt, C., and Kalinsky, K. (2016). AKT in cancer: New molecular insights and advances in drug development. *Br. J. Clin. Pharmacol.* 110, 943–956. doi:10.1111/bcp.13021

NCI (2022). *Definition of MDM2 antagonist RO5045337 - NCI drug dictionary - NCI*. Bethesda, Maryland: National Cancer Institute.

NEB: NEBNext \textregistered Ultra™ II DNA library prep kit for Illumina \textregistered. In.; 2022.

Novartis (2022). *Study to determine and evaluate a safe and tolerated dose of HDM201 in patients with selected advanced tumors that are TP53wt*. Bethesda, Maryland: U. S. National Library of Medicine.

Pal, P., and Chetty, R. (2020). Multiple sclerosing pneumocytomas: A review. *J. Clin. Pathol.* 73 (9), 531–534. doi:10.1136/jclinpath-2020-206501

PHYLIP *PHYLIP (the PHYLogeny inference package)*. Available at: https://evolution.genetics.washington.edu/phylip.html.

Politz, O., Siegel, F., Barfacker, L., Bomer, U., Hagebarth, A., Scott, W. J., et al. (2017). BAY 1125976, a selective allosteric AKT1/2 inhibitor, exhibits high efficacy on AKT signaling-dependent tumor growth in mouse models. *Int. J. Cancer* 140 (2), 449–459. doi:10.1002/ijc.30457

Porta, C., Paglino, C., and Mosca, A. (2014). Targeting PI3K/Akt/mTOR signaling in cancer. *Front. Oncol.* 4, 64. doi:10.3389/fonc.2014.00064

QIAGEN: QIAgen panels. In.; 2022.

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

Roche (2022). *A study of RO5045337 in patients with solid tumors*. Bethesda, Maryland: U. S. National Library of Medicine.

Sartori, G., Bettelli, S., Schirosi, L., Bigiani, N., Maiorana, A., Cavazza, A., et al. (2007). Microsatellite and EGFR, HER2 and K-RAS analyses in sclerosing hemangioma of the lung. *Am. J. Surg. Pathol.* 31 (10), 1512–1520. doi:10.1097/PAS.0b013e318032c8cc

Schmidt, U., Weigert, M., Broaddus, C., and Myers, G. (2018). "Cell detection with star-convex polygons," in *Medical image computing and computer assisted intervention – miccai 2018* (Cham: Springer International Publishing), 265–273.

Song, L., Yan, P., and Mo, G. (2021). Sclerosing pneumocytoma: A carcinoma mimicker. *Integr. Cancer Sci. Ther.* 8 (1), 1–3. doi:10.15761/icst.1000352

Stein, E. M., DeAngelo, D. J., Chromik, J., Chatterjee, M., Bauer, S., Lin, C. C., et al. (2022). Results from a first-in-human phase I study of siremadlin (HDM201) in patients with advanced wild-type TP53 solid tumors and acute leukemia. *Clin. Cancer Res.* 28 (5), 870–881. doi:10.1158/1078-0432.CCR-21-1295

Tolcher, A. W., Fang, D. D., Li, Y., Tang, Y., Ji, J., Wang, H., et al. (2019). A phase Ib/II study of APG-115 in combination with pembrolizumab in patients with

unresectable or metastatic melanomas or advanced solid tumors. *Ann. Oncol.* 30, i2. doi:10.1093/annonc/mdz027

TP53 TP53 germline mutation p.P72R. Available at: https://www.ncbi.nlm.nih.gov/clinvar/variation/12351/?new_evidence=false.

Ulgen, E., Ozisik, O., and Sezerman, O. U. (2019). pathfindR: An R package for comprehensive identification of enriched pathways in omics data through active subnetworks. *Front. Genet.* 10, 858. doi:10.3389/fgene.2019.00858

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., et al. (2013). From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* 43 (1), 11. doi:10.1002/0471250953.bi1110s43

van Griethuysen, J. J. M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., et al. (2017). Computational radiomics System to decode the radiographic phenotype. *Cancer Res.* 77 (21), e104–e107. doi:10.1158/0008-5472.CAN-17-0339

Wang, Q. B., Chen, Y. Q., Shen, J. J., Zhang, C., Song, B., Zhu, X. J., et al. (2011). Sixteen cases of pulmonary sclerosing haemangioma: CT findings are not definitive for preoperative diagnosis. *Clin. Radiol.* 66 (8), 708–714. doi:10.1016/j.crad.2011.03.002

Xie, R-M., Zhou, X-H., Lu, P. X., and He, W. (2003). Diagnosis of pulmonary sclerosing hemangioma with incremental dynamic CT: Analysis of 20 cases. *Zhonghua jie he he hu xi za zhi = Zhonghua jiehe he huxi zazhi = Chin. J. Tuberc. Respir. Dis.* 26 (1), 7–9.

Xu, C., Gu, X., Padmanabhan, R., Wu, Z., Peng, Q., DiCarlo, J., et al. (2019). smCounter2: an accurate low-frequency variant caller for targeted sequencing data with unique molecular identifiers. *Bioinformatics* 35 (8), 1299–1309. doi:10.1093/bioinformatics/bty790

Yeh, Y. C., Ho, H. L., Wu, Y. C., Pan, C. C., Wang, Y. C., and Chou, T. Y. (2020). AKT1 internal tandem duplications and point mutations are the genetic hallmarks of sclerosing pneumocytoma. *Mod. Pathol.* 33 (3), 391–403. doi:10.1038/s41379-019-0357-y

Zhang, H., Meltzer, P., and Davis, S. (2013). RCircos: an R package for Circos 2D track plots. *BMC Bioinforma.* 14, 244. doi:10.1186/1471-2105-14-244

Zheng, Q., Zhou, J., Li, G., Man, S., Lin, Z., Wang, T., et al. (2022). Pulmonary sclerosing pneumocytoma : Clinical features and prognosis. *World J. Surg. Oncol.* 20, 140–149. doi:10.1186/s12957-022-02603-4

Zhou, X., Cao, B., and Lu, H. (2017). Negative auto-regulators trap p53 in their web. *J. Mol. Cell Biol.* 9 (1), 62–68. doi:10.1093/jmcb/mjx001

Zwanenburg, A., Vallieres, M., Abdalah, M. A., Aerts, H., Andrearczyk, V., Apte, A., et al. (2020). The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* 295 (2), 328–338. doi:10.1148/radiol.2020191145

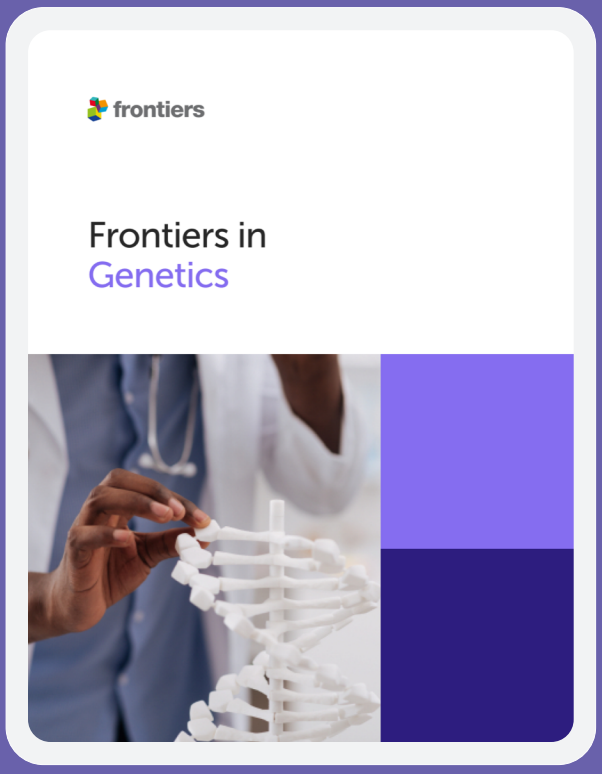
Frontiers in Genetics

Highlights genetic and genomic inquiry relating to all domains of life

The most cited genetics and heredity journal, which advances our understanding of genes from humans to plants and other model organisms. It highlights developments in the function and variability of the genome, and the use of genomic tools.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact