Identification of a Signature Predicting Prognosis of Breast Cancer Patients Through Integrative Analysis of Single-cell and Bulk RNA Sequencing Data

Hanghang Chen
Southern Medical University
Tian Tian
Southern Medical University
Haihua Luo
Southern Medical University
Jinming Li
Southern Medical University
Yong Jiang (jiang48231@163.com)
Southern Medical University

Research Article

Keywords: Breast Cancer, Prognosis, Single Cell, Tumour Immune Microenvironment

Posted Date: January 10th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1226555/v1

Abstract

Background: Due to the high heterogeneity, it is still a challenge to precisely predict the prognosis of breast cancer (BRCA) patients. This study aimed to explore the crucial genes in tumourigenesis and prognosis-related molecular characteristics to predict the prognosis of breast cancer through a comprehensive analysis of single cell and bulk RNA-sequencing data.

Methods: Gene expression data and corresponding clinical data of breast cancer from The Cancer Genome Atlas (TCGA) and Gene Expression Ominibus (GEO) datasets were collected. First, differentially expressed genes (DEGs) among trajectories from 412 cells of 11 BRCA patients were identified though PCA, t-SNE and single cell trajectory analyses. Second, we further explore the DEGs between normal and tumour tissues from the TCGA database. Then the intersection DEGs were used to construct a signature and the 4 independent GEO datasets were merged to validate the signature. Finally, gene set enrichment analysis (GSEA) and a website TIDE was employed to compare the immune-related cells, pathways, scores of immune escape with riskscores of our signature.

Results: A prognostic signature comprising 20 genes was developed to divide patients into high-risk and low-risk groups, and its prognostic performance was excellent in two independent patient cohorts (*n*=1053, 620). The high-risk group generally had lower levels of immune cell infiltration, lower activity of immune pathway activity and higher scores of immune escape than did the low-risk group.

Conclusion: A 20-gene signature was successfully constructed to assess the prognosis of breast cancer patients. The signature was also found closely associated with the tumour immune microenvironment and immune escape.

Introduction

Breast cancer is one of the most common cancers worldwide, accounting for 30% of cancer diagnoses in females in 2020[1]. Although considerable progress has been made in therapies, innovative methods are still needed to identify high-risk patients and treatment plans should be individualized due to the tumour heterogeneity.

Distinct molecular subtypes of breast cancer means distinct prognosis[2] and response to various treatment modalities[3]. Traditional classification methods including biological characteristics, such as estrogen receptors (ER), progesterone receptors (PR) and human epidermal growth factor receptor 2 (HER2) status, may have limitations for personalized treatment strategies[4]. The DNA microarray[5] and next-generation sequencing (NGS)[6] technologies may be the most significant advances in cancer treatment, prevention, and screening over the past few decades. Individual gene analysis could provide a alternative choice to predict breast cancer prognosis in addition to clinicopathological features[7].

Due to the fact that there are diverse populations of cells in any single tissue, RNA-sequencing with bulk tissue represents a weighted average expression profiles of the whole population. Two years after the first application of RNA-seq to bulk tissue-level, the first single-cell RNA-sequencing (scRNA-seq) protocol were published in 2009[8]. The advent of single-cell RNA sequencing provides unprecedented opportunities to further explore the cell-specific transcriptome in individual cells and cell–cell interactions in tissues[9]. In recent years, rapid progress in the development of scRNA-seq has provided insights into the heterogeneity of hundreds of thousands of cells in multiple tumours[10-12]. By combining single-cell analysis with the bulk RNA-sequencing[13], scRNA-seq will continue to facilitate the understanding of dynamic gene expression in tumourigenesis and cell differentiation.

In the present study, we aimed to identify the crucial genes in cell differentiation and construct a scoring model based on these genes to predict the prognosis of BRCA patients. We identified the DEGs among trajectories of 412 cells from the GEO database and the DEGs between normal and tumour tissues from the TCGA database. Then we utilized the intersection DEGs from the two datasets to construct a robust signature which was validated via multiple approaches. The signature could effectively predict the prognosis of BRCA patients and indicate immune infiltration. Our findings suggest a potential connection between the crucial genes in the signature, the tumourigenesis and prognosis of BRCA patients, which has seldom been reported earlier to date.

Materials And Methods

Datasets

The single cell RNA-seq data of 412 tumour cells from 11 female BRCA patients were downloaded from the GEO database (, GSE75688). The bulk RNA-seq data of female BRCA patients (included 111 normal tissues and 1053 tumour tissues) and the corresponding clinical data were downloaded from the TCGA data portal (). In addition, the breast cancer RNA expression data with paired clinical and follow-up information of one external validation cohort (including 620 samples) were downloaded from the GEO database (GSE20685+GSE20711+GSE42568+GSE88770).

Data filtration and normalization

We used “Seurat” R package (version 4.0.4) to calculate the “nFeature_RNA”, “nCount_RNA”, “percent.mt” of the single cell RNA-seq data. Then, cells were filtered out with the threshold (nFeature_RNA > 50 & percent.mt < 50). We normalized the expression matrix of each cell with LogNormalize function. The 1500 most variable genes among cells were identified through the FindVariableFeatures function in the Seurat package and were utilized in the downstream analyses.

Principal component analysis (PCA) and t-distributed stochastic neighbor embedding (tSNE)

Subsequently, linear dimensionality reduction analysis was presented by the RunPCA (npcs = 20) in the Seurat package. ScoreJackStraw function was used to calculate the p-value of every pc (dims = 1:15). Non-linear dimensionality reduction analyses were presented by using the FindNeighbors, FindClusters and RunTSNE functions (pcSelect=10). Then we identified the marker genes of each cell cluster through the FindAllMarkers function in the Seurat package (screening criteria:|log fold change (FC)| ≥1 and adjusted p-value ≤0.05) . The package of “SingleR” (version 1.7.1) was used for cell-type annotation.

Pseudotime trajectory analysis

To further investigate the potential relationship among different cellular clusters and the potential crucial genes in cell differentiation, we performed pseudotime trajectory analyses by using the “Monocle” R package (version 2.21.1). Cells were ordered along the trajectory and visualized in a reduced dimensional space. The DEGs among trajectories were identified according to the given criteria (|log 2 FC| ≥1 and FDR<0.05).

Functional enrichment analysis of the DEGs

To further explore the gene functions and pathways of the DEGs, GO and KEGG analyses were performed by applying the “clusterprofiler” (version 4.1.3), "org.Hs.eg.db" (version 3.13.0) R packages.

Identification of the DEGs between tumour and normal tissues

We compared the DEGs among trajectories in 111 normal and 1053 tumour tissues from the TCGA database and further identified DEGs which were used in the downstream analyses.

Construction and validation of a prognostic signature

We further employed univariate Cox regression analysis (R package “survival”) to evaluate the correlations between each gene and survival status in the TCGA cohort. The Cox regression model (R package “glmnet”, version 4.1-2) was then utilized to develop the prognostic model. The risk score was calculated using the following formula: risk score = expression of Gene 1 $*$ β1+ expression of Gene 2 $*$ β2+ ..expression of Gene n $*$ βn, where β represents the regression coefficient of the genes in the signature. The BRCA patients in the TCGA cohort and GEO cohort were divided into low- and high-risk groups according to the median risk score, and oversurvival (OS) was compared between the two groups

via Kaplan–Meier analysis (“survival” R packages, version 3.2-11). The “survminer” (version 0.4.9) and “time-ROC” (version 0.4) R packages were employed to perform 1-year, 3-year, and 5-year ROC curve analyses.

Independent prognostic analysis of the risk score

The clinical data (age, stage, TNM) of the patients in the TCGA and the GEO cohorts were extracted and analysed in combination with the risk score in our independent regression model. Univariate and multivariate Cox regression models were employed for the analysis by the “survival” R package.

Investigating the relationships between expression of the genes in our signature and particular cells

We further investigated and displayed the relationships between expression of the genes in our signature and particular cells through FeaturePlot and DotPlot in the Seurat package.

Survival analyses of the genes in the signature

To explore the independent prognostic value of the genes in the signature, we performed survival analyses of the genes in the signature and plotted Kaplan–Meier survival curves of prognosis-related genes in the signature with the “survival” and “survminer” R packages.

The differences of the mutational status between high- and low-risk groups

We explored the top 20 genes most frequently mutate in the TCGA cohort and displayed the differences between high- and low-risk groups.

Gene set enrichment analysis

Gene set enrichment analysis (GSEA) was performed using clusterProfiler R package (version 4.1.3) [14] to explore the biological pathways regarding to riskscore of our signature.

Correlations between riskscore and immune checkpoint and activities

A few of genes in our signature are highly expressed in immune-related cells, so we explored the correlations between the riskscore and immune checkpoints, activities with the “corrplot” R package.

Calculating the score of tumour immune escape

To explore the potential function for our signature in immunotherapy，we calculate the immune escape score of TCGA patients at a website called TIDE (Tumour Immune Dysfunction and Exclusion,)[15]. Then we compared the immune escape score between high-riskscore and low-riskscore groups.

Statistical analysis

Single-factor analysis of variance was applied to compare the gene expression levels between the normal and tumour tissues. The Pearson chi-square test was used to compare the categorical variables. The Kaplan–Meier method with a two-sided log-rank test was employed to compare the OS of patients between groups. Univariate and multivariate Cox regression models were used to evaluate the independent prognostic value of the risk model. All statistical analyses were carried out with R software (v4.1.0), and a value of $p < 0.05$ was selected as statistically significant.

Results

The quality control and preprocessing of single-cell RNA-seq

The general workflow of this study is displayed in Fig. 1. The “nFeature_RNA”, “nCount_RNA”, “percent.mt” of the single cell RNA-seq data were displayed in Fig. S1A. There is no significant correlation between percent.mt and nCount_RNA while the correlation between nFeature_RNA and nCount_RNA is positive (Fig. S1B). The most variable genes between tumour cells were calculated and displayed in a volcano plot (Fig. S1C). The top 1500 variable genes were displayed with the red dots and the top 10 significant DEGs was labeled (Fig. S1C). After PCA analysis, dimension of all tumour cells were reduced and displayed in Fig. S1D. 11 different colors represent tumours from different patients. The top 30 characteristic genes of the first four PCs were visualized in the Fig. S1E. We calculated p-values of every PC (dims: 1-15, Fig. S1F) and selected 1-10 PCs in the downstream analyses.

Cell clustering analysis

We performed tSNE to further identified the marker genes of each cell cluster which provided nice visualization for distinguishing different cell clusters (Fig. 2A). Subsequently we annotated cell types by exploiting the cell markers and labeled them in Fig. 2B. Through Pseudotime trajectory analysis, we displayed the developmental process between the cell subpopulations of single cells and showed how

clusters were defined by expression of genes at the beginning of cell fate progressions. All cells were ordered along pseudo-time to establish a common pseudotime axis (Fig. 2C, D). We also annotated cell clusters and types by exploiting the cell markers and labeled them in Fig. 2E, F. Then we compare different trajectories and displayed all significant DEGs between in STable 1.

GO and KEGG mainly indicated immune and inflammatory response

GO and KEGG pathway enrichment analyses of DEGs indicated that the genes were mainly associated with the immune- and inflammation-related activities such as antigen processing and presentation, and virus infection (Fig. 3A-F).

Identification of 157 DEGs between normal and tumour tissues

We compared the 339 DEGs among trajectories in 111 normal and 1053 tumour tissues from the TCGA database and further identified 157 DEGs. They were displayed in a heatmap (Fig. S2).

Identification of a 20-gene signature in the TCGA cohort

After univariate Cox regression analysis, 29 genes that met the criteria of $p<0.05$ were retained for further analysis. Among them, 10 genes were associated with increased risk with HRs >1, while the other 19 genes were protective genes with HRs <1 (Fig. 4). By performing Cox regression analysis, a 20-gene signature was constructed. The risk score was calculated using the data in Table 1.

Table 1

The genes involved in the signature and their coefficients.

No.	Gene name	Coef	No.	Gene name	Coef
1	LIMCH1	0.231177284382533	11	PSMB8	-0.666443350442645
2	ABRACL	0.306190466953879	12	CCR7	0.177570702490536
3	MDK	-0.160390820289339	13	BCL2A1	-0.362345417414752
4	SDC1	0.198276495436628	14	GMFG	0.378359128956913
5	CLEC3A	0.0785617360450329	15	HSPB8	0.119812852617582
6	PMAIP1	-0.118183133869402	16	HSPH1	0.26174024222534
7	NUDT19	0.251669376270532	17	MORF4L2	0.339675654431019
8	RGS2	-0.136596848127705	18	TFF1	-0.100258283031602
9	PSMB9	0.621798003179615	19	RAC2	-0.30540190431927
10	IGHG1	-0.0747069021989824	20	CXCL13	-0.121340461203996

Validation of the risk signature

Patients from the TCGA dataset were stratified into low- and high-risk groups based on the median. A notable difference in OS was detected between the low- and high-risk groups (Fig. 5A). Time-dependent receiver operating characteristic (ROC) analysis was applied to evaluate the sensitivity and specificity of the prognostic model, and the area under the ROC curve (AUC) was 0.769 for 1-year survival, 0.795 for 3-year survival, and 0.756 for 5-year survival (Fig. 5C).

External validation of the risk signature

Based on the median risk score from the TCGA cohort, 620 patients from the GEO dataset were divided into low- and high-risk groups. A notable difference in OS time was detected between the low- and high-risk groups (Fig. 5B). The area under the ROC curve (AUC) was 0.756 for 1-year survival, 0.719 for 3-year survival, and 0.660 for 5-year survival (Fig. 5D).

The risk model was an independent prognostic factor

Univariate Cox regression analysis indicated that the risk score was an independent factor capable of predicting poor survival for the TCGA cohort(HR=1.387, 95% CI: 1.301–1.478, Fig. 5E). The multivariate analysis also revealed that, after adjusting for other confounding factors, the risk score was a prognostic factor (HR=1.340, 95% CI: 1.245–1.442, Fig. 5F) for patients with BRCA in TCGA cohort.

The relationships between genes in our signature and particular cell types

We further investigated the relationships between the expression of the genes in our signature and particular clusters (Fig. 6A) and cell subpopulations (Fig. 6B). Most of the genes are highly expressed in cluster 0, 2, 7 and Epithelial_cells, T_cells, DC.

Identification of prognosis-related genes in the signature

Eight genes were related to prognosis independently in the signature and we plotted Kaplan–Meier survival curves of the eight genes (Fig. S3). Seven genes are positively related to the prognosis while only gene(SDC1) is negatively.

The mutational status of high-risk group is generally lower

We identified and displayed the top 20 genes most frequently mutate in the TCGA cohort. The mutational status of high-risk group(Fig. 7A) is generally lower than low-risk group(Fig. 7B).

GSEA of the riskscore

The significant pathways associated with high riskscore are mainly enriched in metabolism activities such as ascorbate_and_aldarate_metabolism, drug_metabolism_cytochrome_p450 (the top 5 significant pathways are shown in Fig. 8A. The significant pathways associated with low riskscore are mainly enriched in immune passways such as allograft_rejection, primary_immunodeficiency (the top 5 significant pathways are shown in Fig. 8B.

Correlations between riskscore and immune checkpoints, immune cells, immune-related activities

The correlations between riskscore and immune checkpoints were displayed in Fig. 8C. The correlations between riskscore and immune activities were displayed in Fig. 8D. The red circle indicate positive correlations and the blue circle indicate negative correlations (*p<0.05).

The TIDE score is significantly different between groups

The immune escape score (TIDE) is higher in high-riskscore group compared with low-riskscore group (***p<0.001, Fig. 8E).

Discussion

Increasing studies suggested that some genes play crucial roles in tumourigenesis and tumour progression[16, 17]. Hence, based on the identification of DEGs through a comprehensive analysis of single cell and bulk RNA-sequencing data, we constructed a signature to predict the prognosis of breast cancer patients. We also validated the signature in two independent cohorts and proved its efficacy. Subsequently, as most of the genes in the signature highly expressed in immune-related cell, we explored the relationships between riskscore and immune checkpoints, immune activities, immune escape. The results mostly indicate negative relationships, which seems well-reasoned and warrants an in-depth understanding.

In the last decade, NGS technology emerged and contributed tremendous achievements in cancer diagnosis[18] and facilitated our understanding at the molecular and cellular levels. With lower costs and increased qualities, NGS has become increasingly feasible in routine clinical practice today[18]. Bulk RNA-seq is used to identify the DEGs between normal and tumour tissues. However, Bulk RNA-seq is measured by an average readout from heterogeneous tissues, the results may be confusing and misleading. Single-cell technologies provide a high-resolution view into cell-type-specific expression[19]. In this study we endeavored to conduct an integrative analysis of bulk tissues and single cell RNA-sequencing and screen out the crucial genes in cell differentiation and tumourigenesis. When faced with huge amounts of highly complex NGS data, bioinformatics has become an indispensable method[20]. To explore the possibility in clinical practice, NGS data and bioinformatics methods were utilized to construct a simplified DEGs-related signature that could not only effectively predict the prognosis of breast cancer patients but also associated with immune checkpoints, immune activities, immune escape.

We supposed that the DEGs through the single cell analyses represent the developmental process between the cell subpopulations and cell fate progressions. So we conducted GO and KEGG analyses to explore the potential functions of the DEGs. One of the most noticeable functions is that antigen processing and presentation. Professional antigen-presenting cells (APCs), including macrophages and dendritic cells (DCs) are central to the activation of T cells and therefore direct the adaptive immune response[21]. Moreover, the GSEA of riskscore shows that primary immunodeficiency and immune-related passways enrich in low-risk group. So our study suggests that macrophages and dendritic cells (DCs) may play crucial roles in fighting against tumourigenesis. The riskscore of our signature also negatively correlates with some immune checkpoint such as PDCD1, CTLA4 and immune activities and could predict the immune escape to some extent. Currently, the mechanisms of immune escape remain poorly understood. There are a number of factors that may contribute to immune escape such as regulatory cells, defective antigen presentation, immune suppressive mediators, tolerance and immune deviation, apoptosis[22]. How to identify patients who may be vulnerable to immune escape in clinical practice still remains a daunting task for doctors. Our study is an encouraging result which could provide some references for clinical immunotherapies.

Eight genes in the signature indicate independent prognostic value and they deserve further attention and analyses. It reported that chemokine C-X-C motif ligand 13 (CXCL13) could regulate lymphocyte infiltration in the tumour microenvironment and play an important role in the growth and metastasis of solid tumours[23]. The expression level of Glia maturation factor γ (GMFG) significantly related to the poor prognosis and sensitivity of some chemotherapy drugs in breast cancer patients[24], which is consistent with our study. Phorbol-12-myristate-13-acetate-induced protein 1 (PMAIP1), induced by p53 or some anti-cancer drugs, down-regulated in pancreatic cancer and may be a candidate tumour suppressor gene[25]. To our surprise, a previous study showed that the overexpression of proteasome subunit beta type-8(PSMB8) and PSMB9 indicate better survival and improved response to immune-checkpoint inhibitors of melanoma patients[26]. Elevated expression of PSMB8 was also related to increased tumour size, and perineural invasion of gastric cancer cells[27] and could be a candidate marker to predict responsiveness to radiation therapy in rectal cancer[28]. However, it is a pity that there has been very few studies about the two genes. Ras-related C3 botulinum toxin substrate 2 (RAC2) is a small signaling GTPase which could activate Wnt signaling pathway[29] and may function as a promising prognostic biomarker of clear cell renal cell carcinoma[30]. Syndecan-1 (SDC1), frequently overexpressed in breast cancer, is a heparin sulfate proteoglycan and could serve as a prognostic indicator[31]. Trefoil factor 1 (TFF1), over expressed in about 50% of human breast cancers[32], acts as a tumour suppressor gene and predicts a better outcome for patients[33]. To our knowledge, almost all these studies are consistent with our findings which offer new candidate therapeutics.

Collectively, these findings highlight the potential value of our signature and remain to be fully elucidated. Despite the drawbacks that clinical and molecular subtypes were not analysed separately and there was a lack of experimental data due to the availability of clinical specimens, we believe our studies might lay the foundation for renewed understanding of cancer initiation and evolution, provide some clues for clinical decision making and expand the tools available for immunotherapy.

Conclusion

In this study, we succeeded in constructing a 20-gene signature that could predict the prognosis of breast cancer patients. The signature was also found to be closely associated with the tumour immune microenvironment and immune escape.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

This study has been approved by all authors for publication.

Competing interests

The authors declare that they have no competing interests.

Funding

This project was supported by National Natural Science Foundation of China (81971895) , Special Support Plan for Outstanding Talents of Guangdong Province (2019JC05Y340) ,

Guangdong Provincial Science and Technology Projects (2016A020216015) .

Authors' contributions

Hanghang Chen: Software, Validation, Visualization, Writing – Original Draft. Tian Tian: Data curation. Haihua Luo: Resources. Jinming Li: Writing-review & editing. Yong Jiang: Conceptualization, Methodology.

ACKNOWLEDGMENTS

We thank the patients and investigators who participated in TCGA and GEO for providing data.

Availability of data and material

The datasets are available from TCGA (https://portal.gdc.cancer.gov/) and Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/).

References

1. Siegel RL, Miller KD, Jemal A: **Cancer statistics, 2020**. *CA Cancer J Clin*2020, **70**(1):7-30.
2. Desmedt C, Haibe-Kains B, Wirapati P, Buyse M, Larsimont D, Bontempi G, Delorenzi M, Piccart M, Sotiriou C: **Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes**. *Clin Cancer Res*2008, **14**(16):5158-5165.
3. Rouzier R, Perou CM, Symmans WF, Ibrahim N, Cristofanilli M, Anderson K, Hess KR, Stec J, Ayers M, Wagner P*et al*: **Breast cancer molecular subtypes respond differently to preoperative chemotherapy**. *Clin Cancer Res*2005, **11**(16):5678-5685.
4. Yersal O, Barutca S: **Biological subtypes of breast cancer: Prognostic and therapeutic implications**. *World J Clin Oncol*2014, **5**(3):412-424.
5. Abd-Elnaby M, Alfonse M, Roushdy M: **Classification of breast cancer using microarray gene expression data: A survey**. *J Biomed Inform*2021, **117**:103764.

6. Hong M, Tao S, Zhang L, Diao LT, Huang X, Huang S, Xie SJ, Xiao ZD, Zhang H: **RNA sequencing: new technologies and applications in cancer research**. *J Hematol Oncol*2020, **13**(1):166.
7. Lal S, McCart Reed AE, de Luca XM, Simpson PT: **Molecular signatures in breast cancer**. *Methods*2017, **131**:135-146.
8. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A*et al*: **mRNA-Seq whole-transcriptome analysis of a single cell**. *Nat Methods*2009, **6**(5):377-382.
9. Chen H, Ye F, Guo G: **Revolutionizing immunology with single-cell RNA sequencing**. *Cell Mol Immunol*2019, **16**(3):242-249.
10. Chung W, Eum HH, Lee HO, Lee KM, Lee HB, Kim KT, Ryu HS, Kim S, Lee JE, Park YH*et al*: **Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer**. *Nat Commun*2017, **8**:15081.
11. Fu K, Hui B, Wang Q, Lu C, Shi W, Zhang Z, Rong D, Zhang B, Tian Z, Tang W*et al*: **Single-cell RNA sequencing of immune cells in gastric cancer patients**. *Aging (Albany NY)*2020, **12**(3):2747-2763.
12. Zheng H, Pomyen Y, Hernandez MO, Li C, Livak F, Tang W, Dang H, Greten TF, Davis JL, Zhao Y*et al*: **Single-cell analysis reveals cancer stem cell heterogeneity in hepatocellular carcinoma**. *Hepatology*2018, **68**(1):127-140.
13. Malikic S, Jahn K, Kuipers J, Sahinalp SC, Beerenwinkel N: **Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data**. *Nat Commun*2019, **10**(1):2750.
14. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L*et al*: **clusterProfiler 4.0: A universal enrichment tool for interpreting omics data**. *Innovation (N Y)*2021, **2**(3):100141.
15. Fu J, Li K, Zhang W, Wan C, Zhang J, Jiang P, Liu XS: **Large-scale public data reuse to model immunotherapy response and resistance**. *Genome Med*2020, **12**(1):21.
16. Rivlin N, Brosh R, Oren M, Rotter V: **Mutations in the p53 Tumor Suppressor Gene: Important Milestones at the Various Steps of Tumorigenesis**. *Genes Cancer*2011, **2**(4):466-474.
17. Berx G, Van Roy F: **The E-cadherin/catenin complex: an important gatekeeper in breast cancer tumorigenesis and malignant progression**. *Breast Cancer Res*2001, **3**(5):289-293.
18. Kamps R, Brandao RD, Bosch BJ, Paulussen AD, Xanthoulea S, Blok MJ, Romano A: **Next-Generation Sequencing in Oncology: Genetic Diagnosis, Risk Prediction and Cancer Classification**. *Int J Mol Sci*2017, **18**(2).
19. Packer J, Trapnell C: **Single-Cell Multi-omics: An Engine for New Quantitative Models of Gene Regulation**. *Trends Genet*2018, **34**(9):653-665.
20. Oliver GR, Hart SN, Klee EW: **Bioinformatics for clinical next generation sequencing**. *Clin Chem*2015, **61**(1):124-135.
21. Vyas JM, Van der Veen AG, Ploegh HL: **The known unknowns of antigen processing and presentation**. *Nat Rev Immunol*2008, **8**(8):607-618.
22. Vinay DS, Ryan EP, Pawelec G, Talib WH, Stagg J, Elkord E, Lichtor T, Decker WK, Whelan RL, Kumara H*et al*: **Immune evasion in cancer: Mechanistic basis and therapeutic strategies**. *Semin Cancer*

*Biol*2015, **35 Suppl**:S185-S198.

23. Kazanietz MG, Durando M, Cooke M: **CXCL13 and Its Receptor CXCR5 in Cancer: Inflammation, Immune Response, and Beyond**. *Front Endocrinol (Lausanne)*2019, **10**:471.
24. Yang Y, He X, Tang QQ, Shao YC, Song WJ, Gong PJ, Zeng YF, Huang SR, Zhou JY, Wan HF*et al*: **GMFG Has Potential to Be a Novel Prognostic Marker and Related to Immune Infiltrates in Breast Cancer**. *Front Oncol*2021, **11**:629633.
25. Ishida M, Sunamura M, Furukawa T, Lefter LP, Morita R, Akada M, Egawa S, Unno M, Horii A: **The PMAIP1 gene on chromosome 18 is a candidate tumor suppressor gene in human pancreatic cancer**. *Dig Dis Sci*2008, **53**(9):2576-2582.
26. Kalaora S, Lee JS, Barnea E, Levy R, Greenberg P, Alon M, Yagel G, Bar Eli G, Oren R, Peri A*et al*: **Immunoproteasome expression is associated with better prognosis and response to checkpoint therapies in melanoma**. *Nat Commun*2020, **11**(1):896.
27. Kwon CH, Park HJ, Choi YR, Kim A, Kim HW, Choi JH, Hwang CS, Lee SJ, Choi CI, Jeon TY*et al*: **PSMB8 and PBK as potential gastric cancer subtype-specific biomarkers associated with prognosis**. *Oncotarget*2016, **7**(16):21454-21468.
28. Ha YJ, Tak KH, Kim CW, Roh SA, Choi EK, Cho DH, Kim JH, Kim SK, Kim SY, Kim YS*et al*: **PSMB8 as a Candidate Marker of Responsiveness to Preoperative Radiation Therapy in Rectal Cancer Patients**. *Int J Radiat Oncol Biol Phys*2017, **98**(5):1164-1173.
29. Xia P, Gao X, Shao L, Chen Q, Li F, Wu C, Zhang W, Sun Y: **Down-regulation of RAC2 by small interfering RNA restrains the progression of osteosarcoma by suppressing the Wnt signaling pathway**. *Int J Biol Macromol*2019, **137**:1221-1231.
30. Liu Y, Cheng G, Song Z, Xu T, Ruan H, Cao Q, Wang K, Bao L, Liu J, Zhou L*et al*: **RAC2 acts as a prognostic biomarker and promotes the progression of clear cell renal cell carcinoma**. *Int J Oncol*2019, **55**(3):645-656.
31. Cui X, Jing X, Yi Q, Long C, Tian J, Zhu J: **Clinicopathological and prognostic significance of SDC1 overexpression in breast cancer**. *Oncotarget*2017, **8**(67):111444-111455.
32. Buache E, Etique N, Alpy F, Stoll I, Muckensturm M, Reina-San-Martin B, Chenard MP, Tomasetto C, Rio MC: **Deficiency in trefoil factor 1 (TFF1) increases tumorigenicity of human breast cancer cells and mammary tumor development in TFF1-knockout mice**. *Oncogene*2011, **30**(29):3261-3273.
33. Yi J, Ren L, Li D, Wu J, Li W, Du G, Wang J: **Trefoil factor 1 (TFF1) is a potential prognostic biomarker with functional significance in breast cancers**. *Biomed Pharmacother*2020, **124**:109827.

Figures

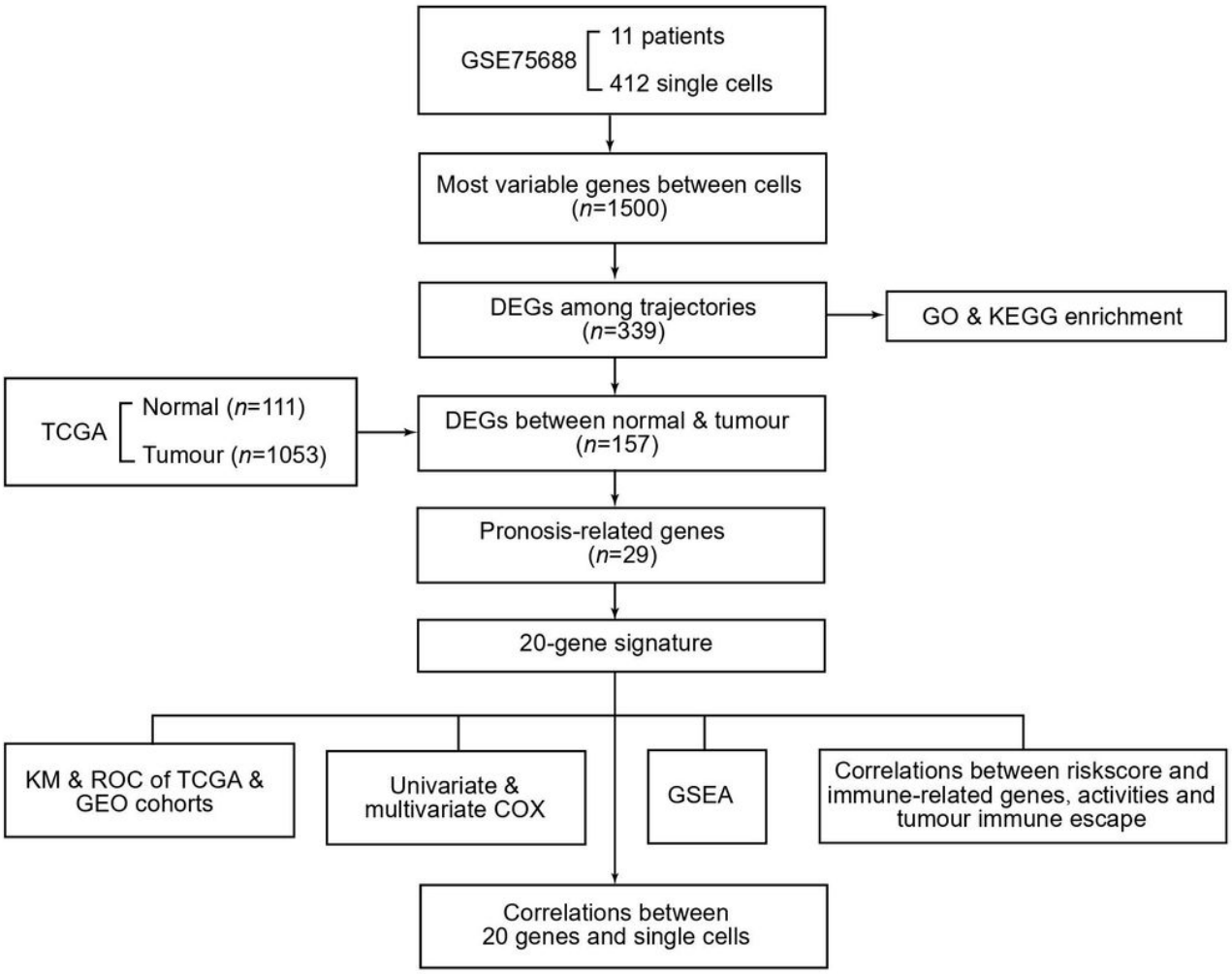

Figure 1

The flowchart of data processing and analyses.

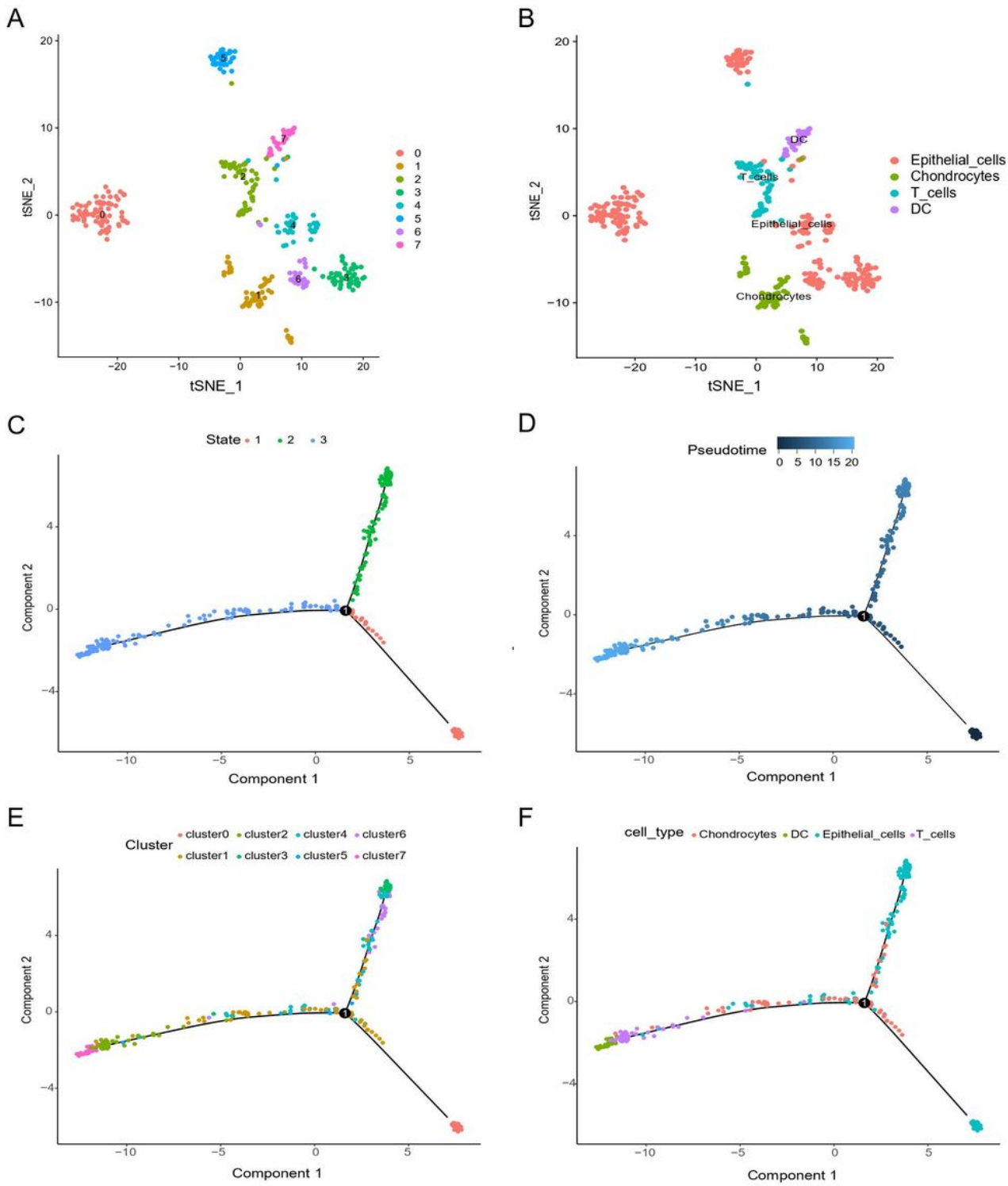

Figure 2

Cell clusters and types were annotated and cell progressions were analyzed. **A** Eight different cell clusters were identified by performing tSNE. **B** Cell types were further annotated and labeled by exploiting the cell markers. **C** All cells were ordered along pseudo-time to establish a common pseudotime axis which displayed the developmental process between the cell subpopulations of single cells. Different colors

represent different states. **D** The deeper the color, the earlier the beginning of cell progressions. Cell clusters(**E**) and types(**F**) were annotated and labeled by exploiting the cell markers.

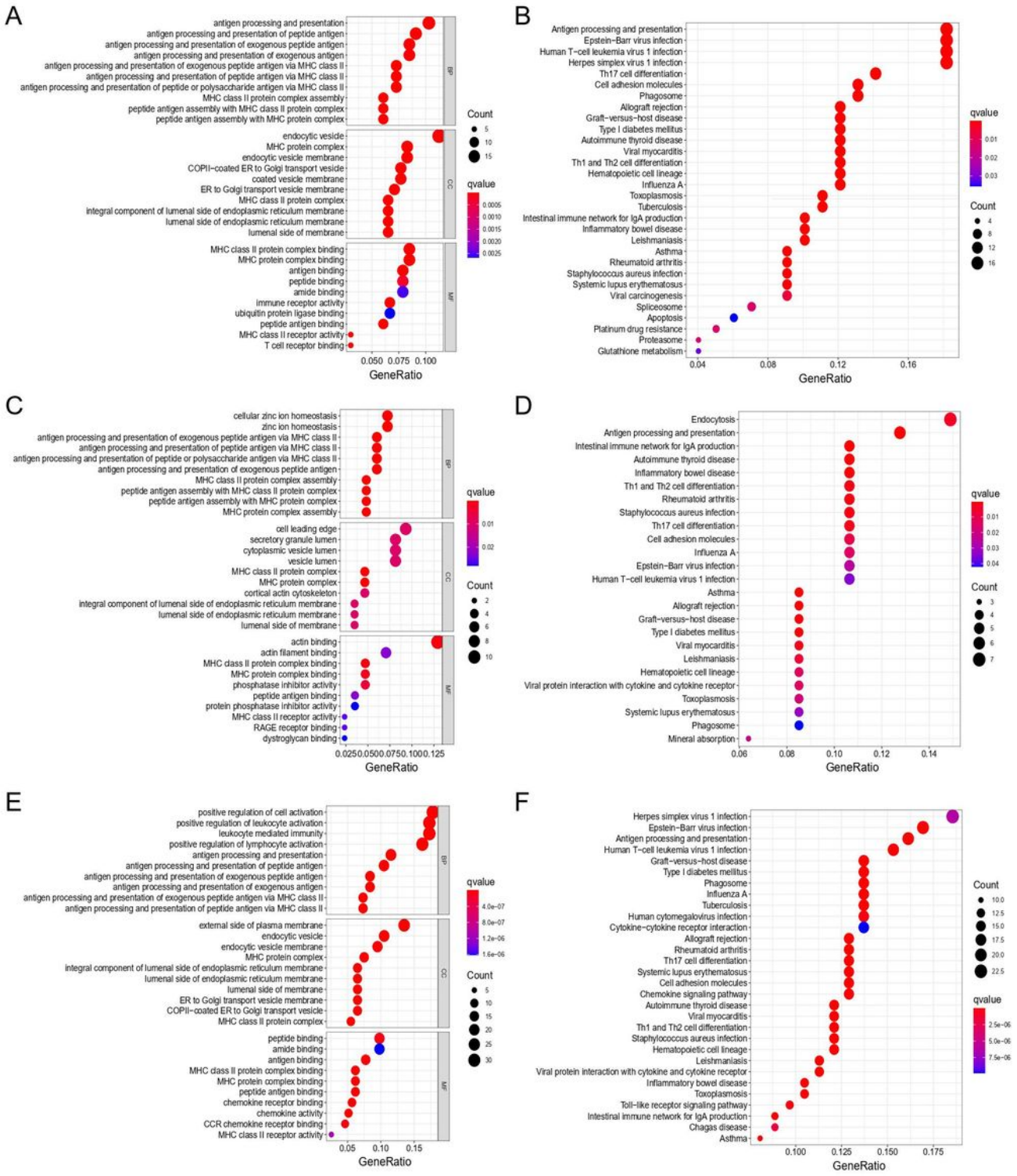

Figure 3

GO and KEGG mainly indicated immune and inflammatory responses. GO(**A, C, E**) and KEGG(**B, D, F**) pathway enrichment analyses of DEGs among three different trajectories indicated that the genes were

mainly associated with the immune- and inflammation-related activities such as antigen processing and presentation, and virus infection.

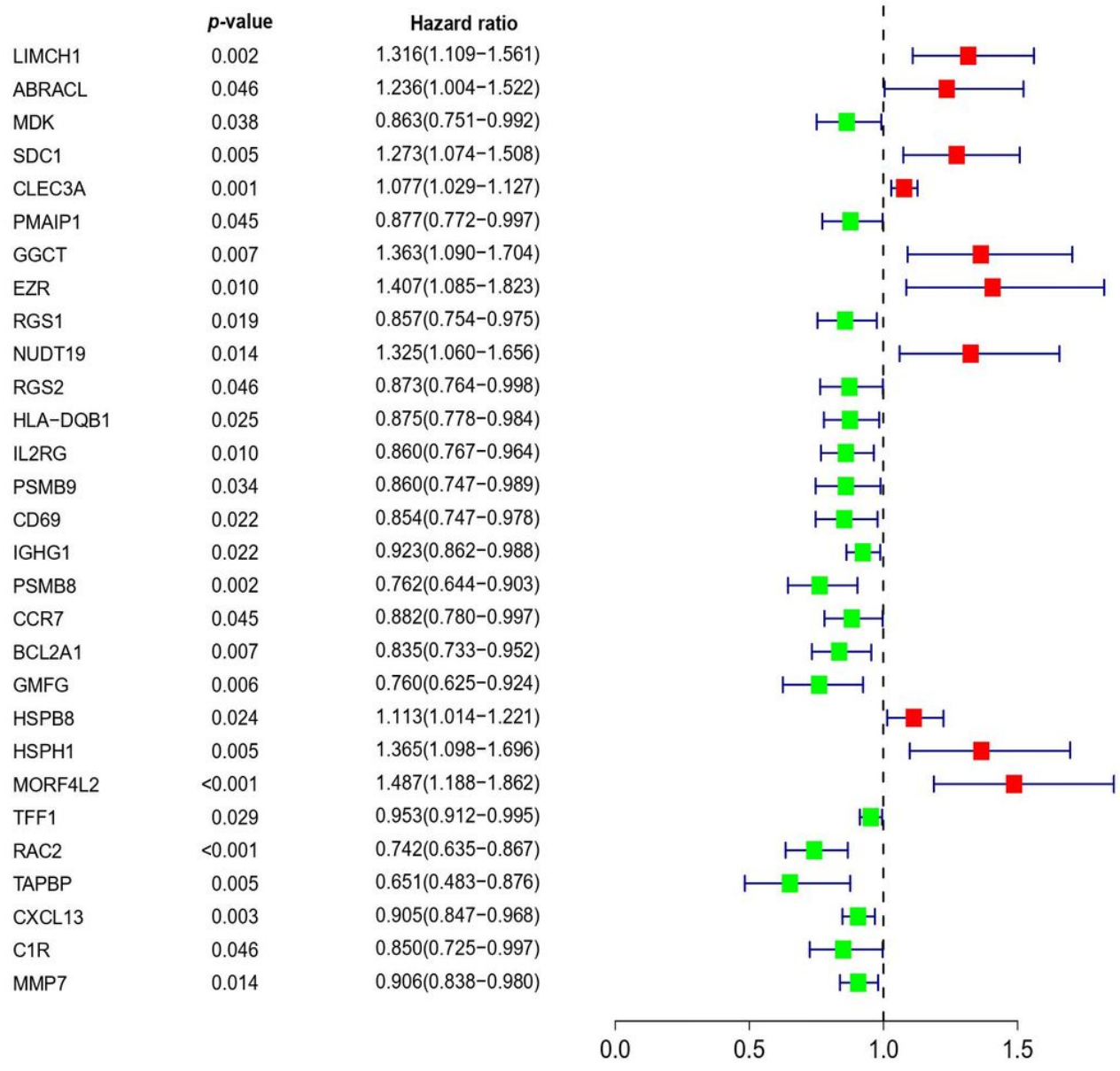

Figure 4

prognosis-related genes were further identified. 29 statistically significant prognosis-related genes were retained for further analysis. Among them, 10 genes were associated with increased risk with HRs >1, while the other 19 genes were protective genes with HRs <1.

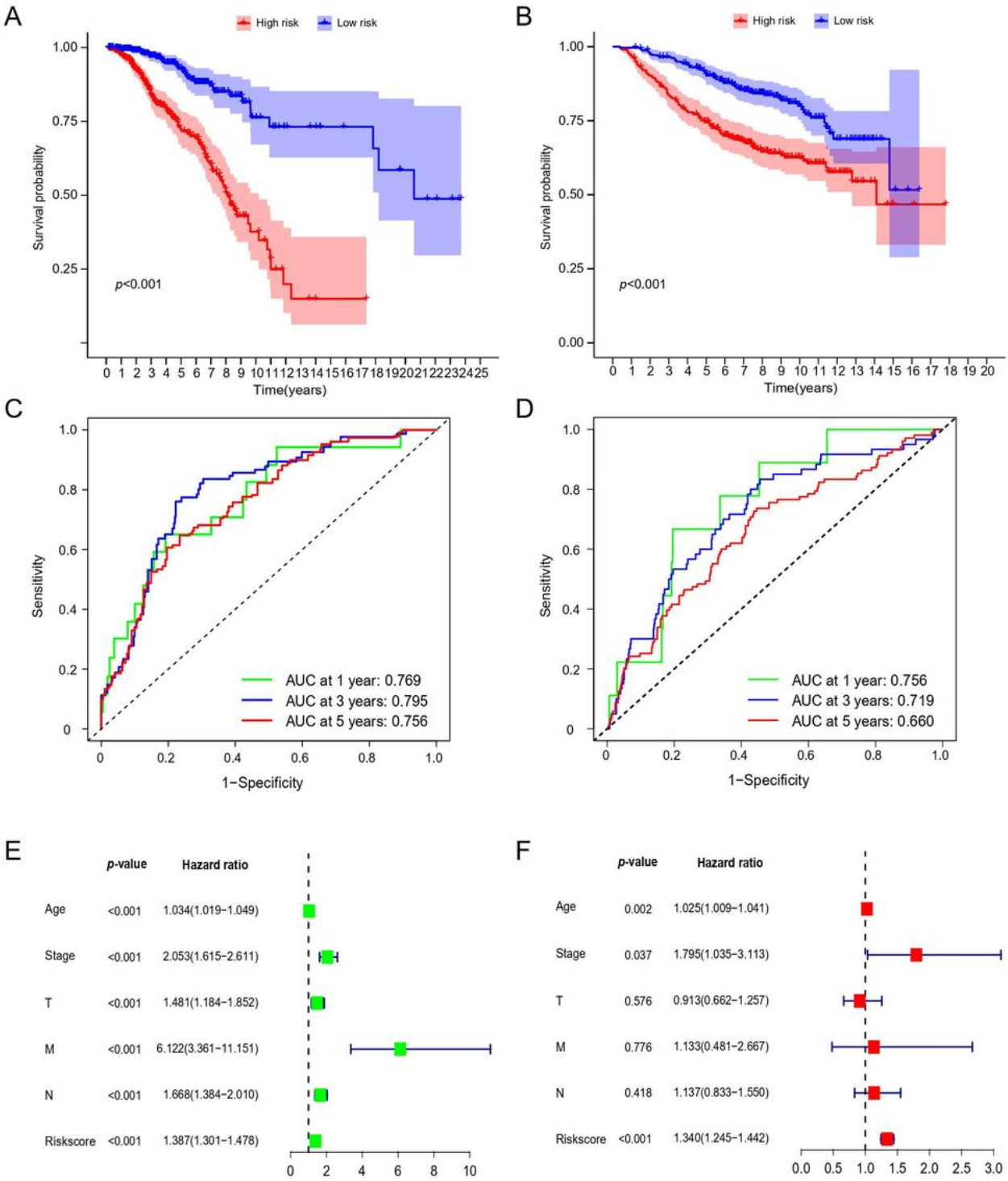

Figure 5

Validation of risk signature. Kaplan–Meier curves for the OS of patients in the high- and low-risk groups of the TCGA cohort(**A**) and the GEO cohort(**B**). ROC curves showed the predictive efficiency of the risk scores in the TCGA cohort(**C**) and the GEO cohort(**D**). **E** The univariate analysis indicated that the risk score was an independent factor capable of predicting survival for the TCGA cohort. **F** The multivariate

analysis also revealed that, after adjusting for other confounding factors, the risk score was a prognostic factor for the TCGA cohort.

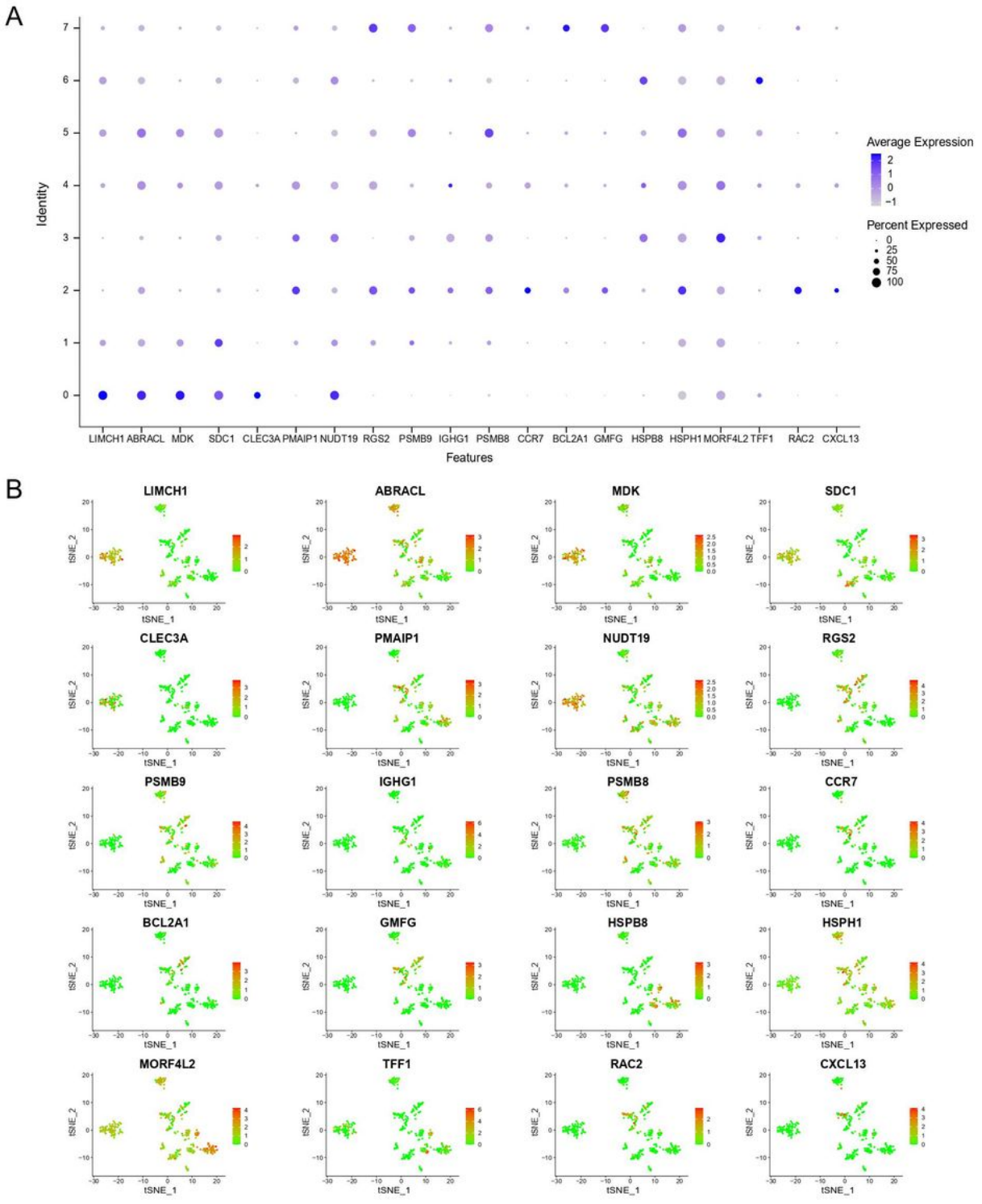

Figure 6

The relationships between genes in our signature and particular cell types. **A** The relationships between the expression of the genes in our signature and particular clusters. The size of the dots represent the

expression percent of the genes in different clusters and the depth of colors represent the expression quantity. **B** The distribution of the genes in our signature in different cell subpopulations. The red dots indicate high expression and the green dots indicate the opposite.

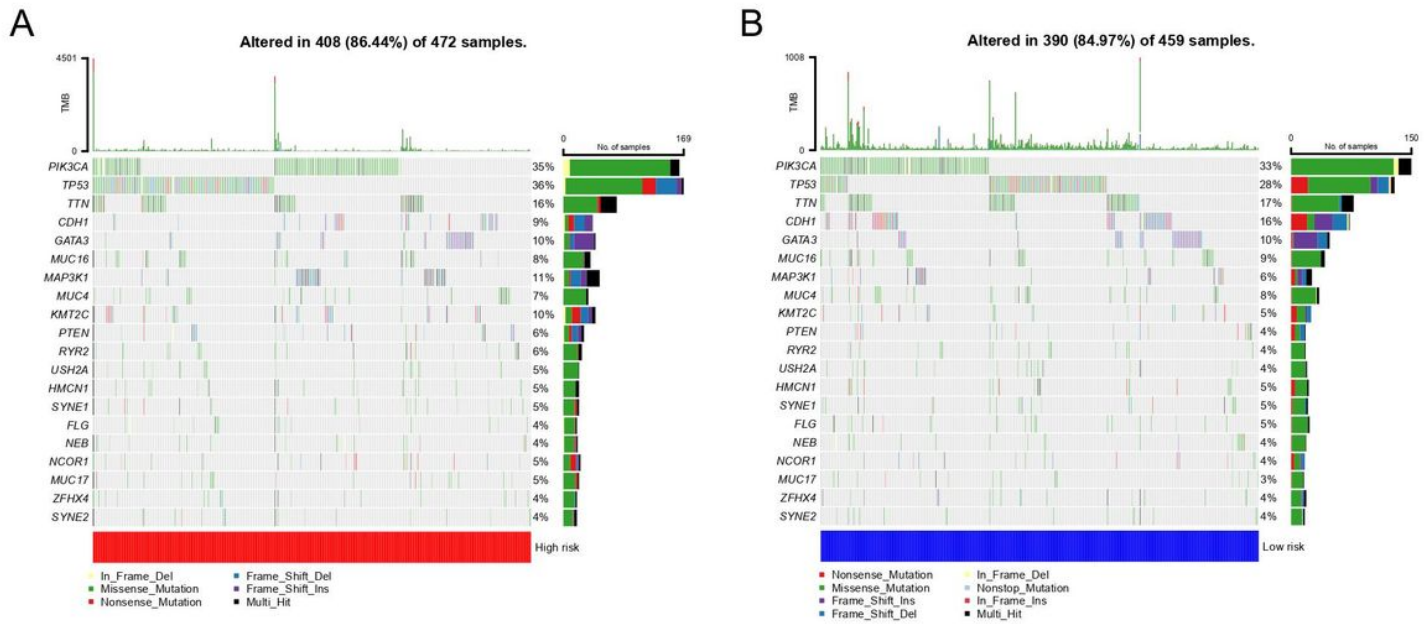

Figure 7

The mutational status of high- and low-risk groups

The top 20 genes most frequently mutate in the TCGA cohort were displayed. The mutational status of high-risk group(**A**) is generally lower than low-risk group(**B**). The nodes and edges of different colors represent different types of mutation.

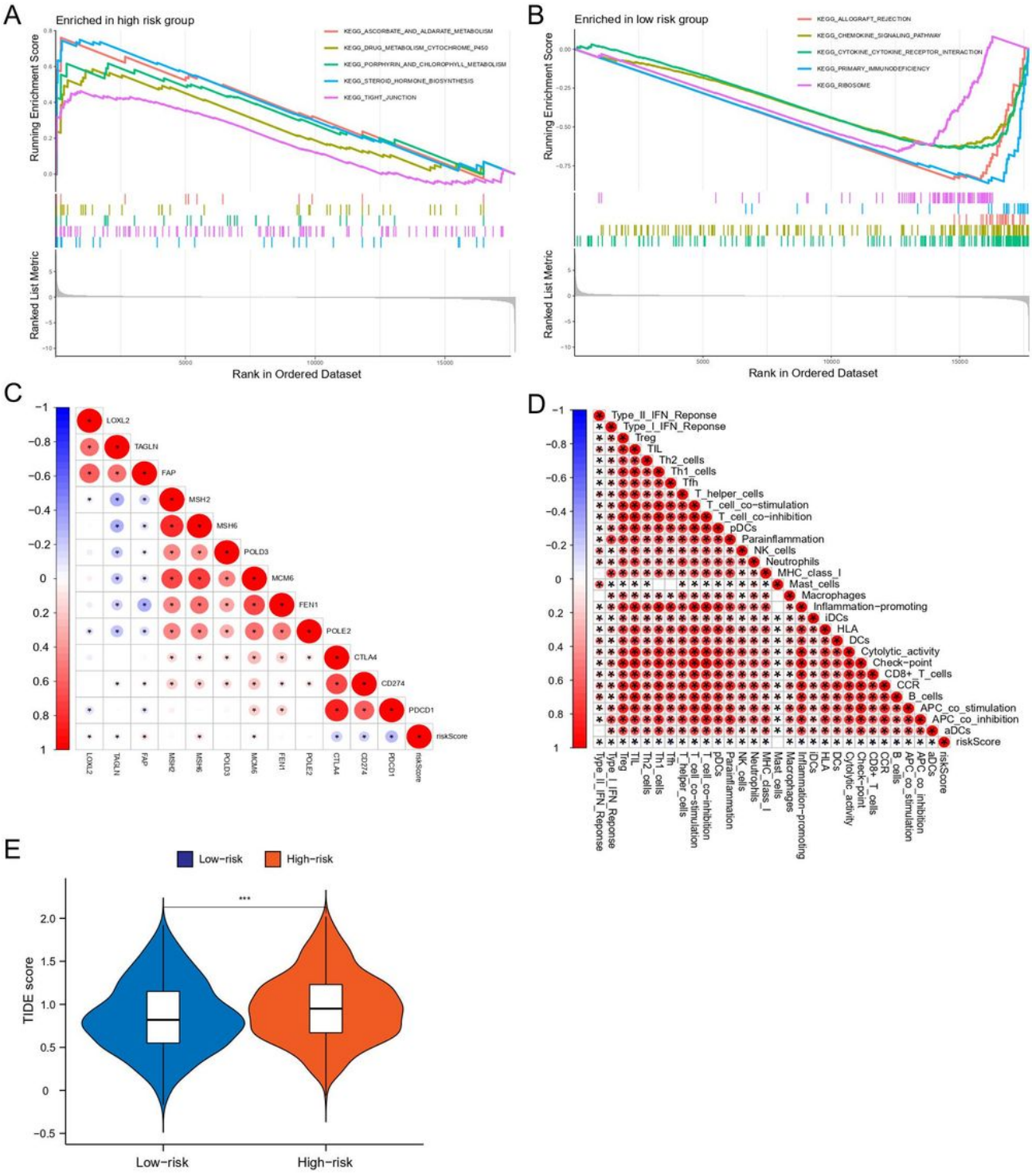

Figure 8

The exploration of the immune correlations of the signature. **A** The top 5 significant pathways associated with high riskscore. **B** The top 5 significant pathways associated with low riskscore. The correlations between riskscore and immune checkpoints(**C**), immune cells and immune-related passways(**D**). The red circle indicate positive correlations and the blue circle indicate negative correlations (*p<0.05). E The

immune escape score (TIDE) is higher in high-riskscore group compared with low-riskscore group (***p<0.001).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS1.docx
- FigureS1.pdf
- FigureS2.pdf
- FigureS3.pdf