scientific reports

OPEN

Integrated analysis of single-cell and bulk RNA-sequencing to predict prognosis and therapeutic response for colorectal cancer

Liyang Cai[1,4], Xin Guo[1,4], Yucheng Zhang[1,4], Huajie Xie[3], Yongfeng Liu[1], Jianlong Zhou[1], Huolun Feng[1,2]✉, Jiabin Zheng[1]✉ & Yong Li[1]✉

Colorectal cancer (CRC) is a prevalent malignant tumor characterized by high global incidence and mortality rates. Furthermore, it is imperative to comprehend the molecular mechanisms underlying its development and to identify effective prognostic markers. These efforts are crucial for pinpointing potential therapeutic targets and enhancing patient survival rates. Therefore, we develop a novel prognostic model aimed at providing new theoretical support for clinical prognosis evaluation and treatment. We downloaded data from the Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) databases. Subsequently, we performed single-cell analysis and developed a prognostic model associated with colorectal cancer. We divided the scRNA-seq dataset (GSE221575) into 19 cell clusters and classified these clusters into 11 distinct cell types using marker genes. Using univariate Cox regression and LASSO (Least Absolute Shrinkage and Selection Operator) analyses, we developed a prognostic model consisting of 9 genes. Based on our 9-gene model, we divided patients into high-risk and low-risk groups using the median risk score. The high-risk group demonstrated significant positive correlations with M0 macrophages, CD8+T cells, and M2 macrophages. The enrichment analyses indicate significant enrichment of immune-related pathways in the high-risk group, including HEDGEHOG_SIGNALING, Wnt signaling pathway, and cell adhesion molecules. Drug sensitivity analysis revealed that the low-risk group was sensitive to 5 chemotherapeutic drugs, while the high-risk group was sensitive to only 1. Additionally, we developed a highly reliable nomogram for clinical application. This suggests that the risk score derived from our modeling analysis is highly effective for stratifying colorectal cancer samples. This study comprehensively applied bioinformatics methods to construct a risk score model. The model showed good predictive performance, offering potential guidance for individualized treatment of colorectal cancer patients. Furthermore, it may provide valuable insights into the disease's pathogenesis and identify potential therapeutic targets for further research.

Keywords Colorectal cancer, Prognostic model, scRNA-seq, Epithelial cell marker genes

Background

Colorectal cancer is a prevalent malignant tumor, comprising approximately 10% of global cancer diagnoses and cancer-related deaths each year[1], with nearly 9 million deaths annually. Over the past few decades, the incidence of colorectal cancer in high-income countries has stabilized or declined, largely due to increased acceptance of colorectal cancer screening and colonoscopic polypectomy among the elderly[2]. In contrast, there has been a global rise in the incidence of colorectal cancer diagnosed in young people, known as early-onset colorectal cancer[3]. Additionally, factors such as family history, obesity, poor diet (high-fat, low-fiber diets), and long-term inflammatory bowel disease are also considered related to the occurrence of colorectal cancer[4]. Most colorectal

[1]Department of Gastrointestinal Surgery, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou 510080, Guangdong, China. [2]School of Medicine, South China University of Technology, Guangzhou 510006, Guangdong, China. [3]The First Clinical Medical College, Guangdong Medical University, Guangzhou, China. [4]Liyang Cai, Xin Guo, and Yucheng Zhang contributed equally to this work. ✉email: fenghuolun2022@qq.com; zhengjiabin@gdph.org.cn; liyong@gdph.org.cn

cancers originate from polyps, a process that starts with abnormal crypts, evolves into pre-tumor lesions (polyps), and ultimately develops into colorectal cancer within an estimated period of 10–15 years[4]. It is currently believed that the cells of origin for most colorectal cancers are stem cells or stem cell-like cells[5,6]. The emergence of these cancer stem cells stems from the progressive accumulation of genetic and epigenetic alterations that deactivate tumor suppressor genes and activate oncogenes.

Presently, the primary treatment modalities for colorectal cancer encompass surgical resection, chemotherapy, radiotherapy, and targeted therapy. Surgical resection represents the cornerstone of curative treatment for colorectal cancer, with the quality of resection playing a pivotal role in prognosis. Assessment of resection quality can be achieved through objective parameters[7]. As adjuvant therapy, fluoropyrimidine-based chemotherapy can improve the survival rates of resected stage III and high-risk stage II colon cancer (such as high-risk T4, poorly differentiated)[8,9]. Preoperative radiotherapy is beneficial in reducing the risk of local recurrence, with the absolute risk reduction depending on clinical staging and surgical quality[10]. Currently available biomarkers for predicting prognosis and treatment response in CRC patients, such as carcinoembryonic antigen (CEA) and carbohydrate antigen 19 – 9 (CA19-9), have suboptimal sensitivity and specificity[11–13]. Therefore, there is a need for more precise biomarkers. With the advancement of high-throughput genomic screening technologies, such as Next-Generation Sequencing (NGS) and microarray analysis, a multitude of molecular biomarkers and features have been identified. These have potential clinical prognostic and predictive value, identified through comprehensive association and bioinformatics analyses. Notably, several genomics-based biomarkers, such as mismatch repair (MMR) or microsatellite instability (MSI) status, have entered clinical practice and have been validated as predictive markers for adjuvant chemotherapy in stage II CRC patients[14–17].

With the rapid advancements in NGS technologies, an increasing number of studies are employing RNA sequencing (RNA-seq) to analyze gene expression patterns in colorectal cancer. However, RNA-seq is typically conducted in bulk, where the data reflect the average gene expression patterns across a large population of cells[18]. Notably, single-cell RNA sequencing (scRNA-seq) is a cutting-edge sequencing technology that offers detailed insights into the characteristics of individual immune cells or tumor cells[19]. The scRNA-seq indeed highlights intratumoral heterogeneity by revealing different subpopulations of cells within tumors. It also has the capability to quantify and analyze immune cell infiltration patterns within tumor tissues. The level of detail provided by scRNA-seq is not achievable with bulk RNA sequencing, where the averaging of gene expression across all cells in a sample may obscure significant cellular differences. The scRNA-seq is capable of capturing the gene expression signatures of cells in specific functional states, which aids in identifying particular cellular conditions associated with disease progression and therapeutic response. This approach facilitates the discovery of novel biomarkers that may be masked in bulk RNA sequencing data, as they might be expressed only in a minority of cell types.

In this study, we aimed to explore the distribution of different cell types, gene expression characteristics, and their correlations with clinical prognosis in colorectal cancer tissues using both scRNA-seq and gene chip analysis. Additionally, we developed a novel prognostic model leveraging feature genes identified through single-cell analysis(Fig. 1).

Methods

Data source and preprocessing

The GEO database, managed by NCBI, stores gene expression data from various studies, aiding researchers in accessing and analyzing biological and biomedical research datasets. We retrieved the single-cell data files for GSE221575 from the GEO database, specifically selecting datasets from 4 samples that feature comprehensive single-cell expression profiles, essential for our detailed single-cell analysis. We additionally obtained the Series Matrix File data for GSE17536 from the NCBI GEO public database, annotated with GPL570, encompassing expression profiles from 177 patients. Furthermore, we acquired the Series Matrix File data for GSE38832 from the NCBI GEO public database, annotated with GPL570, featuring expression profile data from 122 patients. The TCGA database (https://portal.gdc.cancer.gov/), recognized as the largest repository of cancer-related genetic information, comprehensively archives diverse data types such as gene expression data, copy number variations, SNPs, and beyond. In our study, we accessed raw mRNA expression data specifically for colorectal cancer, encompassing a total of 701 samples, comprising 51 normal samples and 650 tumor samples.

Single-cell analysis

Initially, the expression profiles were imported using the Seurat package. Subsequently, aberrant samples were excluded by evaluating UMI counts, the quantity of genes detected per cell, and the mitochondrial gene fraction. The data were subsequently standardized, normalized, and subjected to PCA (Principal Component Analysis) to achieve linear dimensionality reduction. The optimal number of principal components was identified using an elbow plot. Subsequently, UMAP (Uniform Manifold Approximation and Projection) was employed for nonlinear dimensionality reduction to elucidate the spatial relationships between clusters. Cell types and their associated marker genes within the tissue were identified and annotated through utilization of the CellMarker and PanglaoDB databases, supplemented by consulting pertinent literature sources.

Contribution of different cell subpopulations to colorectal cancer

We characterized the contribution of various cell subpopulations to the disease by evaluating both the cell counts and alterations in gene expression patterns. In summary, our approach involved conducting differential gene expression analysis to pinpoint the top 100 highly expressed genes in control versus tumor samples, treating these genes as feature markers for each group. Subsequently, we computed the differential expression levels and expression proportions of these genes within each cell subtype. To comprehensively evaluate the extent

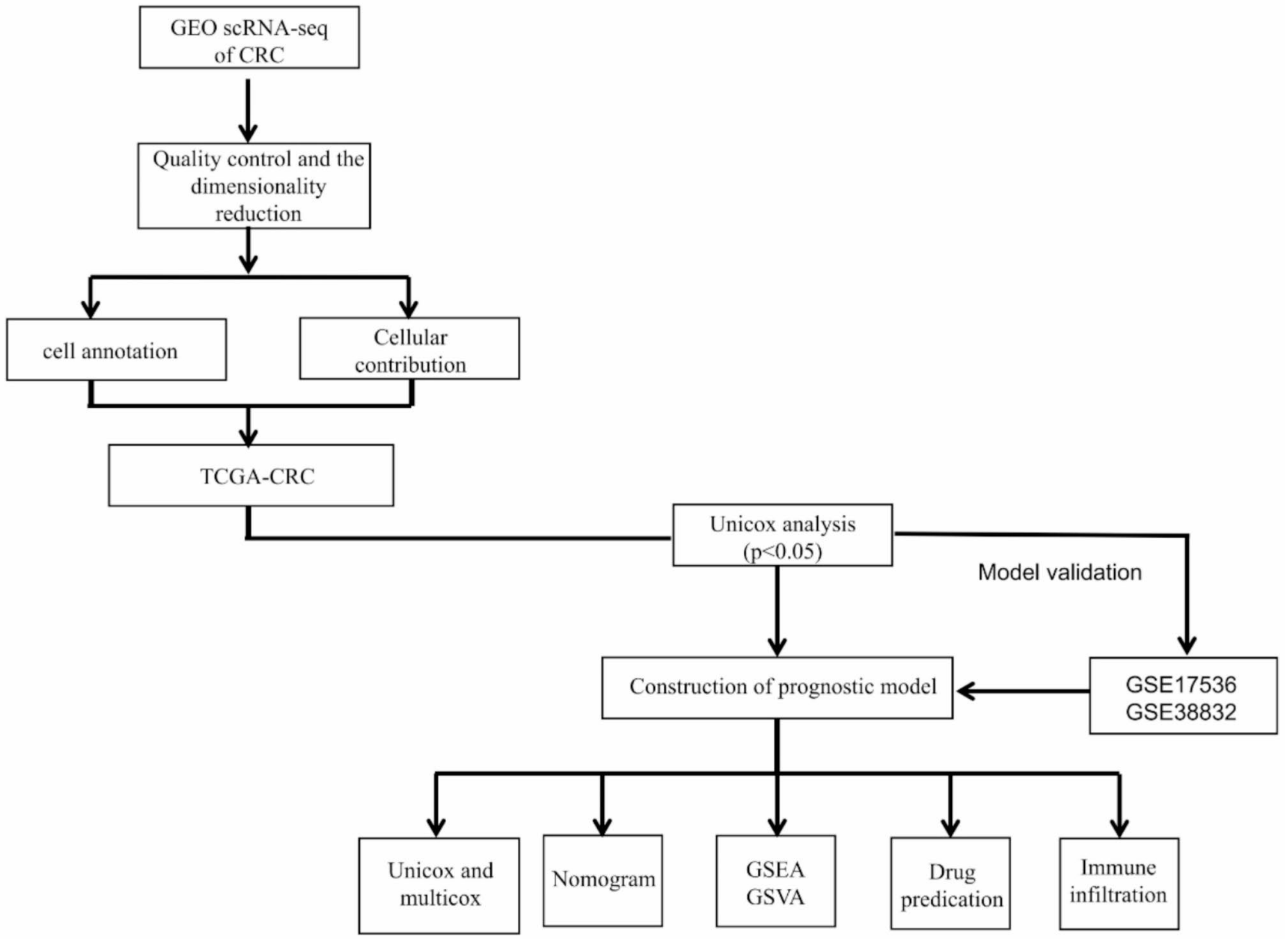


Fig. 1. The flowchart of the study. This flowchart illustrates the research design, data processing, and analysis steps of the study. Initial steps include quality control and dimensionality reduction, followed by cell annotation and cellular contribution analysis. Data is integrated with TCGA-CRC for Unicox analysis ($p < 0.05$) to construct a prognostic model, validated using GSE17536 and GSE38832 datasets. Further analyses include Unicox and multicox, nomogram, GSEA, GSVA, drug prediction, and immune infiltration assessments.

of expression changes of characteristic genes in biological processes, we have defined an index—FCscore—based on the reference[20], which is used to measure the changes in both the number and expression levels of characteristic genes. Specifically, FCscore is derived by calculating the expression differences (fold change, FC) between the disease group and the control group, as well as the changes in gene quantity. The formula for calculating FCscore is FCscore(i, j) = √(FCexp(i, j) * FCprop(j)), where FCexp(i, j) represents the expression fold change of the ith characteristic gene in the jth cluster, and FCscore(i, j) denotes the proportional fold change of the ith characteristic gene in the jth cluster. The FCscore of the ith characteristic gene in the jth cluster is defined as FCexp(i, j) divided by the square root of FCexp(i, j). A higher value of FCscore indicates a more significant change in the cellular subgroup in the disease, suggesting a potentially more important role in disease progression. By calculating the FCscore for each subgroup, we are able to quantify the contribution of each subgroup and reveal which subgroups play a more prominent role in the disease.

Model construction and prognosis

A candidate gene set was identified, and LASSO regression was employed to develop a prognostic model. This involved incorporating the expression values of each selected gene to formulate a risk score formula for every patient. The coefficients derived from the LASSO regression analysis were utilized to weight the contribution of each gene in the risk score calculation. This approach helps in predicting patient outcomes based on the expression levels of the selected genes. The regularization parameter (λ) value was determined using the 'lambda. min' algorithm, with λ set at 0.01681083. Based on the risk score formula derived from the LASSO regression, patients were categorized into low-risk and high-risk groups using the median risk score as the threshold. To assess survival disparities between these groups, Kaplan-Meier analysis(Chi-square test) was conducted, and the results were compared using the log-rank test. LASSO regression analysis and stratified analysis were employed to rigorously examine the impact of the risk score on predicting patient prognosis. The precision of the model's predictions was meticulously evaluated by analyzing Receiver Operating Characteristic (ROC) curves, ensuring a thorough assessment of its predictive power.

Immune cell infiltration analysis

The CIBERSORT method is indeed widely recognized for its application in evaluating immune cell types within biological microenvironments. This approach utilizes support vector regression to perform deconvolution analysis on the expression matrix of immune cell subtypes. With a robust framework built on 547 biomarkers, CIBERSORT effectively discriminates among 22 distinct human immune cell phenotypes. These phenotypes encompass a broad spectrum, encompassing T cells, B cells, plasma cells, and various myeloid subgroups. In this study, we used the CIBERSORT algorithm to analyze patient data, estimating the proportions of 22 immune infiltrating cell types. We then performed correlation analysis between gene expression and immune cell content.

Drug sensitivity analysis

Based on data from the Genomics of Drug Sensitivity in Cancer (GDSC) database, we utilized the R package “pRRophetic” to forecast the sensitivity of each tumor sample to chemotherapy. We obtained IC50 estimates for each specific chemotherapy drug using regression methods and validated the accuracy of these predictions through 10-fold cross-validation on the GDSC training set. In our analysis, default parameters were used throughout, including “combat” for batch effect removal and averaging for handling duplicate gene expression data.

GSVA analysis (gene set difference analysis)

Gene Set Variation Analysis (GSVA) is a non-parametric, unsupervised method for evaluating transcriptome gene set enrichment. GSVA converts gene-level changes to pathway-level changes by comprehensively scoring gene sets of interest, thereby assessing the biological functions of samples. In this study, gene sets were downloaded from the Molecular Signatures Database (v7.0) and scored using the GSVA algorithm to assess potential biological function changes in different samples.

GSEA analysis

Patients were meticulously stratified into high-risk and low-risk cohorts according to the nuanced risk scores generated by the model, followed by an intricate exploration of signal pathway variances between these delineated groups using the powerful analytical tool known as Gene Set Enrichment Analysis (GSEA). The background gene sets were meticulously curated from the comprehensive MsigDB database, specifically tailored for subtype pathway annotation and rigorous differential expression analysis across subtypes. Enriched gene sets achieving statistical significance (adjusted p-value < 0.05) were meticulously prioritized based on their consistency scores. Utilizing GSEA analysis, a widely adopted approach, allowed for an in-depth exploration into the intricate associations linking tumor subtypes with their profound biological implications.

Nomogram model construction

A sophisticated nomogram, meticulously crafted through rigorous regression analysis, elegantly integrates both the nuanced risk scores and intricate clinical symptoms. Scaled line segments were meticulously plotted on a unified plane, each segment proportionally representing the interdependencies among variables within the predictive model. Through the construction of a multivariate regression model, distinct scores were assigned to each tier of influential factors, delineated by their respective contributions (regression coefficients) to the outcome variable. The cumulative total score was then computed by aggregating these individual scores, thereby yielding the predictive value.

Statistical analysis

Survival curves were generated using the Kaplan-Meier method and were compared using the log-rank test. Multivariate analysis was conducted using the Cox proportional hazards model. All statistical analyses were conducted using R (version 4.3.0), with $p < 0.05$ considered statistically significant.

Results

Definition of clusters and dimensionality reduction for visual representation of the cells

First, we read the expression profiles using the Seurat package and filtered out low-expression genes (nFeature_RNA > 200 & nFeature_RNA < 6000 & percent.mt < 25 & nCount_RNA < 40000), resulting in 5,241 cells (Fig. 2A, B). We displayed the top 5 genes with the highest standard deviation among these cells (Fig. 2C). The data were then sequentially processed for standardization, normalization, PCA, and Harmony analysis (Fig. 2D-F).

Using UMAP analysis, we determined the positional relationships between each cluster, identifying 19 cell clusters (Fig. 3A). Further annotation of each subtype in this study revealed that all cell clusters were annotated into the following cell categories: Plasma cell, T cell, Profiling NKT cell, Macrophage cell, B cell, Epithelial cell, Fibroblasts, Goblet cell, Mast cell, EC cell, and Cholangiocytes (Fig. 3B). We also presented a bubble plot of the classic markers for these 11 cell types (Fig. 3C) and a bar plot showing the proportion of cells corresponding to each group (Fig. 3D).

In comparing control versus tumor samples, we conducted screening to identify highly expressed genes. Subsequently, we quantified the differential expression levels and determined the expression proportions of these genes within each cell subtype. The disease contribution was determined by the square root of the FC * PctProp value, with Epithelial cells showing the highest contribution (Fig. 3E). Therefore, we selected highly

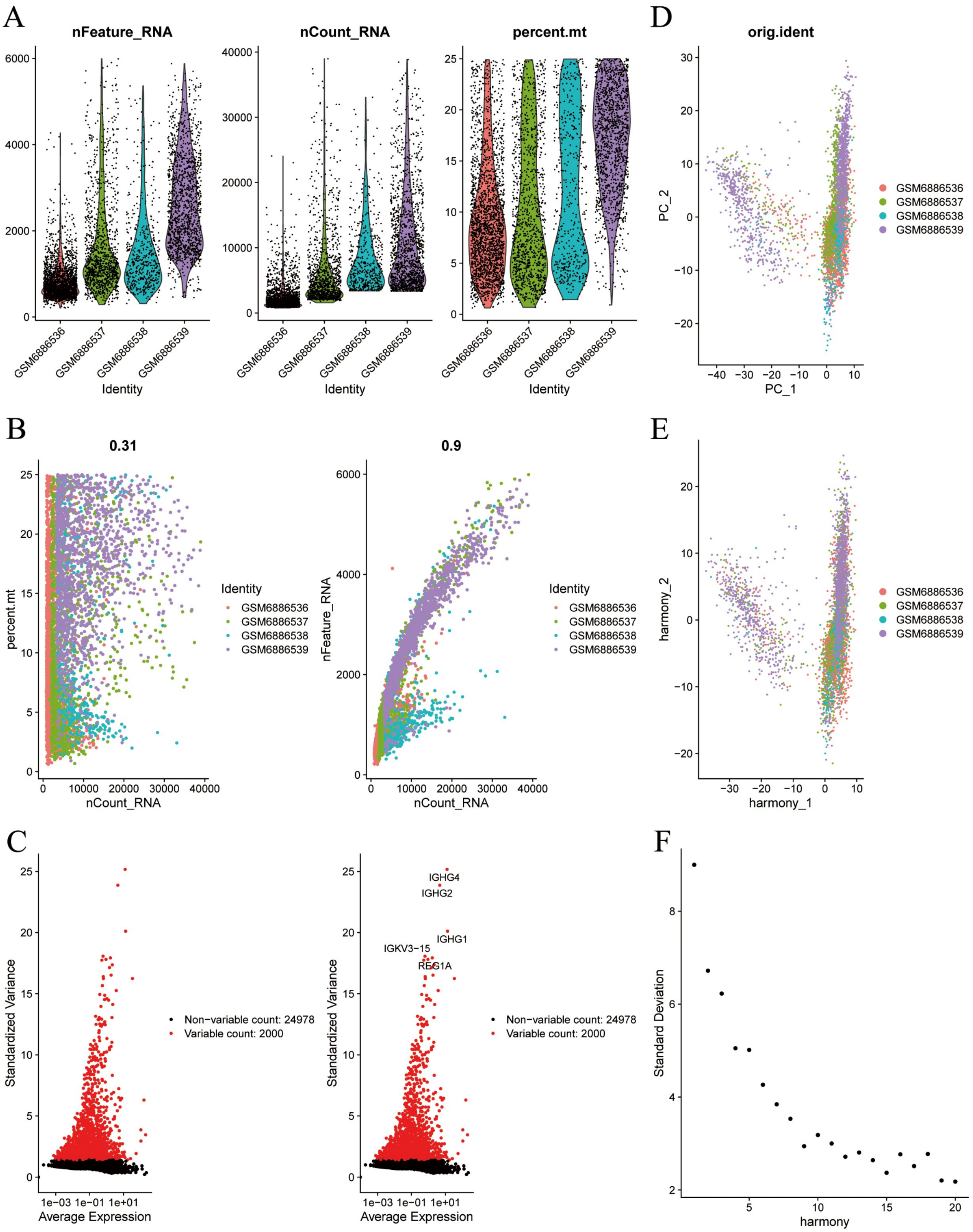


Fig. 2. Single-cell analysis workflow. (**A**,**B**) Quality control of scRNA-seq data. (**C**) Variance plot showing 24,978 genes across all cells, with red dots representing the top 2000 highly variable genes, highlighting the top 5 genes based on standard deviation. (**D**–**F**) Sequential data processing steps including normalization, scaling, PCA, and harmony analysis.

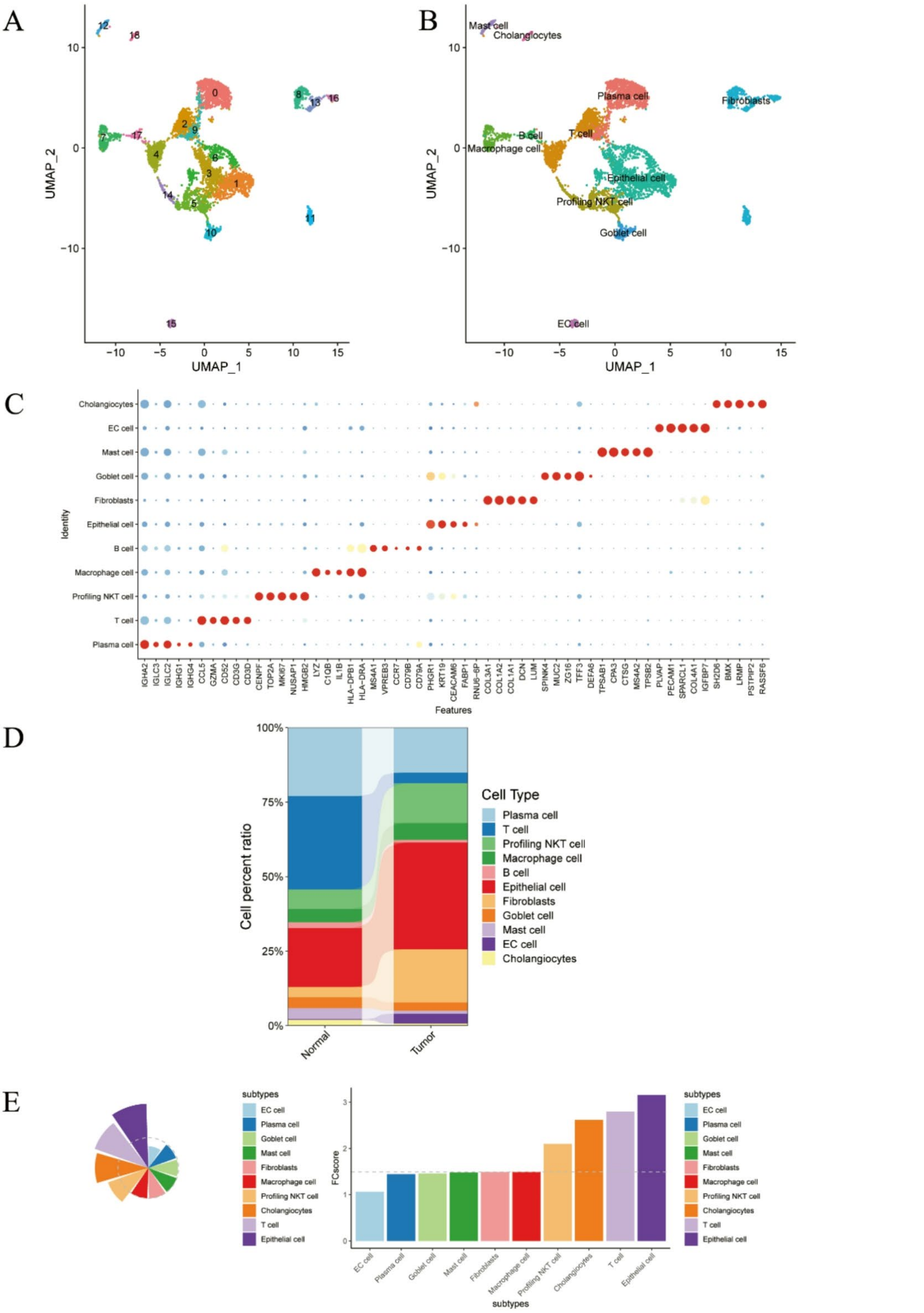


Fig. 3. Single-cell overview of tumor samples. (**A**,**B**) UMAP analysis identifies 19 cell clusters annotated with cell types. (**C**,**D**) Bubble plots display expression profiles of classical markers for 11 cell types; bar charts show proportions of each cell type. (**E**) Identification of significantly upregulated genes in tumor vs. control samples, highlighting Epithelial cells with the highest disease contribution.

expressed genes in control vs. tumor samples with avg_log2FC > 1 and p_val_adj < 0.05 as the candidate gene set for subsequent analysis.

Construction and validation of the predictive model based on epithelial cell marker genes

Using the candidate genes obtained in the previous step, we applied the LASSO regression feature selection algorithm to identify characteristic genes in colorectal cancer. The processed colorectal cancer dataset from the TCGA database, containing patient survival information, was randomly divided into a training set and a test set at a 4:1 ratio. After LASSO regression analysis (Fig. 4A-C), we obtained the optimal risk score value for each sample for subsequent analysis.RiskScore = S100P * (-0.127895799000468) + PIGR * (-0.110505479982379) + RAB11FIP1 * (-0.110168920940582) + USP53 * (-0.0657253153577585) + CDH1 * (-0.0447173205642763) + LGALS4 * (-0.026899354451419) + ATP10B * (-0.0253973538959315) + SLC12A2 * (-0.0148820970575857) + LAMB3 * (0.0964297473594213).

Patients were stratified into high-risk and low-risk groups according to their calculated risk scores, and subsequent survival analysis was performed utilizing Kaplan-Meier curves. In both the training set and the test set, the survival rate was markedly lower in the high-risk group compared to the low-risk group (Fig. 4D-E). Furthermore, ROC curve analysis from both the training and test sets indicated strong validation performance of the model (Fig. 4F-G).

We downloaded processed colorectal cancer patient data with survival information from the GEO database (GSE17536 and GSE38832). Using our model, we predicted the clinical classification of colorectal cancer patients sourced from the GEO database. Subsequently, we assessed survival differences between the predicted groups using Kaplan-Meier analysis to evaluate the stability and predictive accuracy of the model. The findings revealed a significant disparity in survival rates between the high-risk and low-risk groups within the external validation set obtained from the GEO database (Fig. 4H). To validate the accuracy of our model, we conducted ROC curve analysis using the external dataset. The results illustrated robust predictive performance of the model in assessing patient prognosis. (Fig. 4I).

Analysis of immune cell infiltration to explore the impact of risk scores on the immune microenvironment in colorectal cancer

The tumor microenvironment (TME) primarily consists of tumor-associated fibroblasts, immune cells, extracellular matrix, various growth factors, inflammatory factors, specific physicochemical characteristics, and the cancer cells themselves. The TME significantly influences tumor diagnosis, survival outcomes, and clinical treatment sensitivity. By analyzing the relationship between risk scores and tumor immune infiltration, we further investigated the potential molecular mechanisms through which risk scores impact colorectal cancer progression.

The proportions of immune cells in the high-risk and low-risk groups are illustrated in Fig. 5A. Additionally, we compared the differences in immune cell content between these two groups. The results showed that the high-risk group had significantly lower levels of activated dendritic cells, plasma cells, and resting CD4+ memory T cells, while the levels of M0 macrophages and CD8+ T cells were significantly higher (Fig. 5B). Subsequently, we investigated the relationship between the risk score and immune cells. The study results indicated that the risk score was significantly positively correlated with M0 macrophages, CD8+ T cells, and M2 macrophages, and significantly negatively correlated with resting CD4+ memory T cells, activated dendritic cells, eosinophils, and plasma cells (Fig. 5C).

Further analysis to explore the potential molecular mechanisms of risk scores impacting tumor progression

The treatment of early-stage colorectal cancer with surgery combined with chemotherapy has demonstrated clear efficacy. Our study utilized drug sensitivity data from the GDSC database and employed the R package "pRRophetic" to predict the chemotherapy sensitivity of each tumor sample. This approach allowed us to further explore the relationship between risk scores and sensitivity to common chemotherapy drugs. The study results indicated that the risk score was significantly associated with sensitivity to drugs such as AKT inhibitor VIII, Axitinib, BAY 61-3606, BIBW2992, BMS 708,163, and Bicalutamide (Fig. 6A).

Next, we examined the specific signaling pathways involved in the high-risk and low-risk models to investigate the potential molecular mechanisms by which risk scores influence tumor progression. GSVA results showed that the differential pathways between the two groups were primarily enriched in the HEDGEHOG_SIGNALING, WNT_BETA_CATENIN_SIGNALING, and IL6_JAK_STAT3_SIGNALING pathways(Fig. 6B). GSEA results indicated that the pathways involved included the Wnt signaling pathway, cell adhesion molecules, and the MAPK signaling pathway (Fig. 6C). The molecular interaction network among these pathways is illustrated in Fig. 6D.

Construction of nomogram model

In this study, both univariate and multivariate analyses demonstrated that the risk score is an independent prognostic factor for colorectal cancer patients(Fig. 7A-C). Subsequently, samples were stratified into high-risk and low-risk groups based on the median value of the risk score. The results of the regression analysis were visualized using column plots, which demonstrated that the risk score significantly contributes to the scoring process of the nomogram prediction model across all samples (Fig. 7C). Additionally, predictions were made for both the three-year and five-year survival periods in colorectal cancer (Fig. 7D).

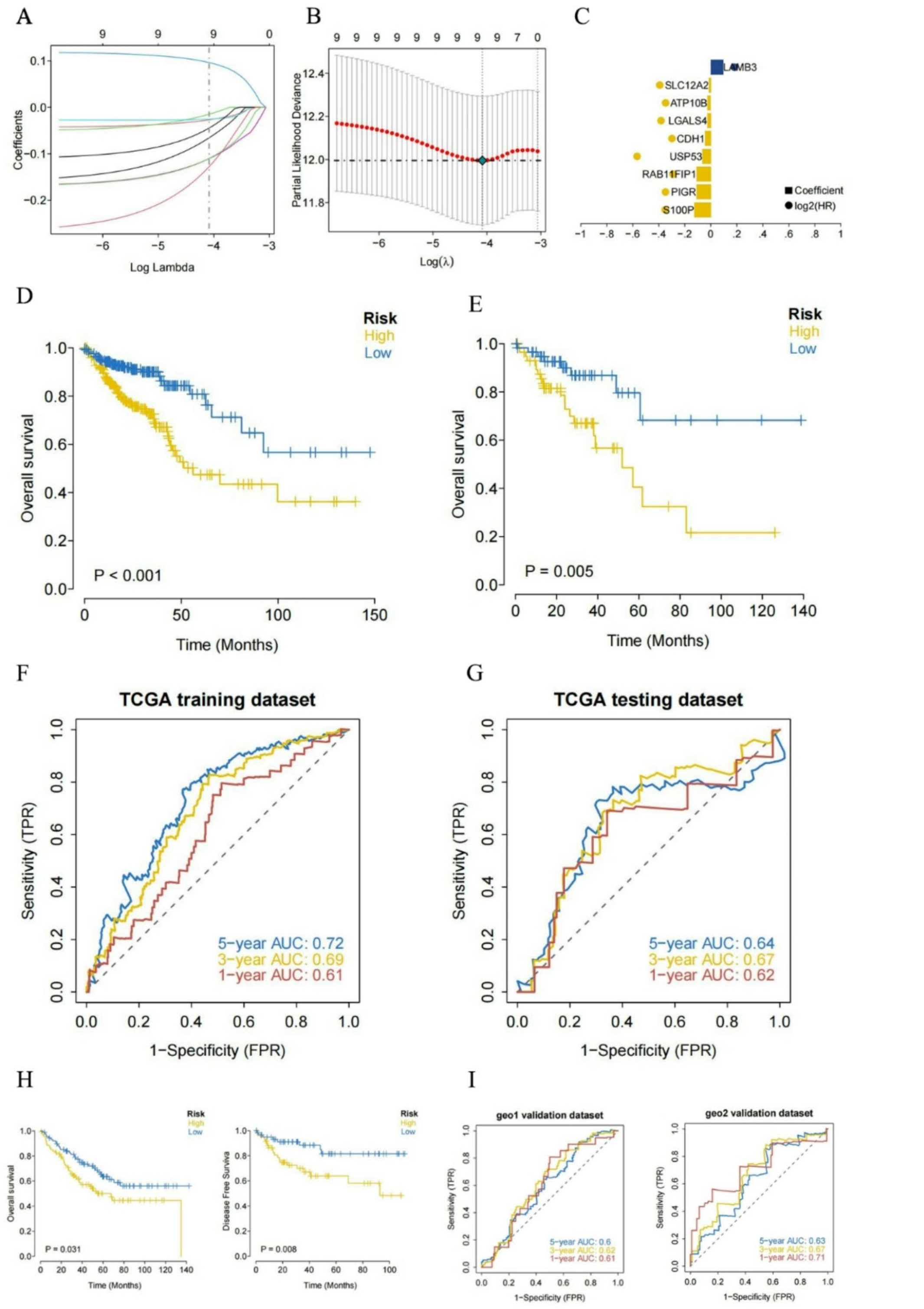


Fig. 4. Construction and prediction of the prognostic model. (**A**–**C**) LASSO regression analysis. (**D**,**E**) Kaplan-Meier curves in the training and testing sets. (**F**,**G**) ROC curves in the training and testing sets. (**H**,**I**) External validation demonstrating strong prognostic performance of the model for patients.

Clinical indicator analysis further demonstrates the applicability of risk score to colorectal cancer samples

Next, we stratified samples based on the values of clinical indicators and displayed the corresponding risk score values using box plots(Fig. 8A-G). Through rank-sum tests, we identified significant differences in risk score distributions among groups defined by clinical indicators such as Fustat, M, N, T, and Stage (p-value < 0.05).

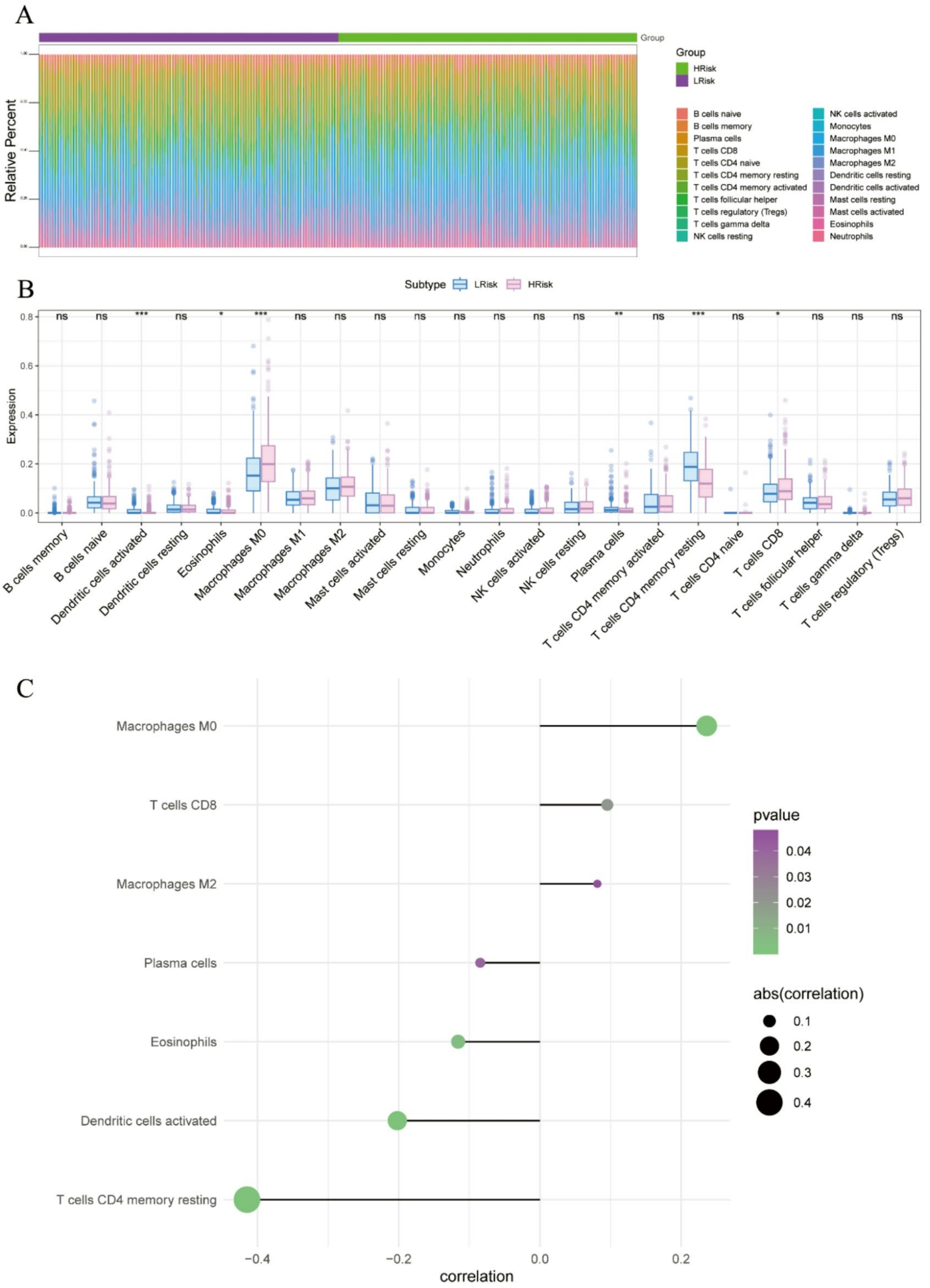


Fig. 5. Immune cell infiltration analysis. (**A**) Proportions of immune cell content between high and low-risk groups. (**B**) Differences in immune cell content between high and low-risk groups. (**C**) Relationship between risk scores and immune cells.

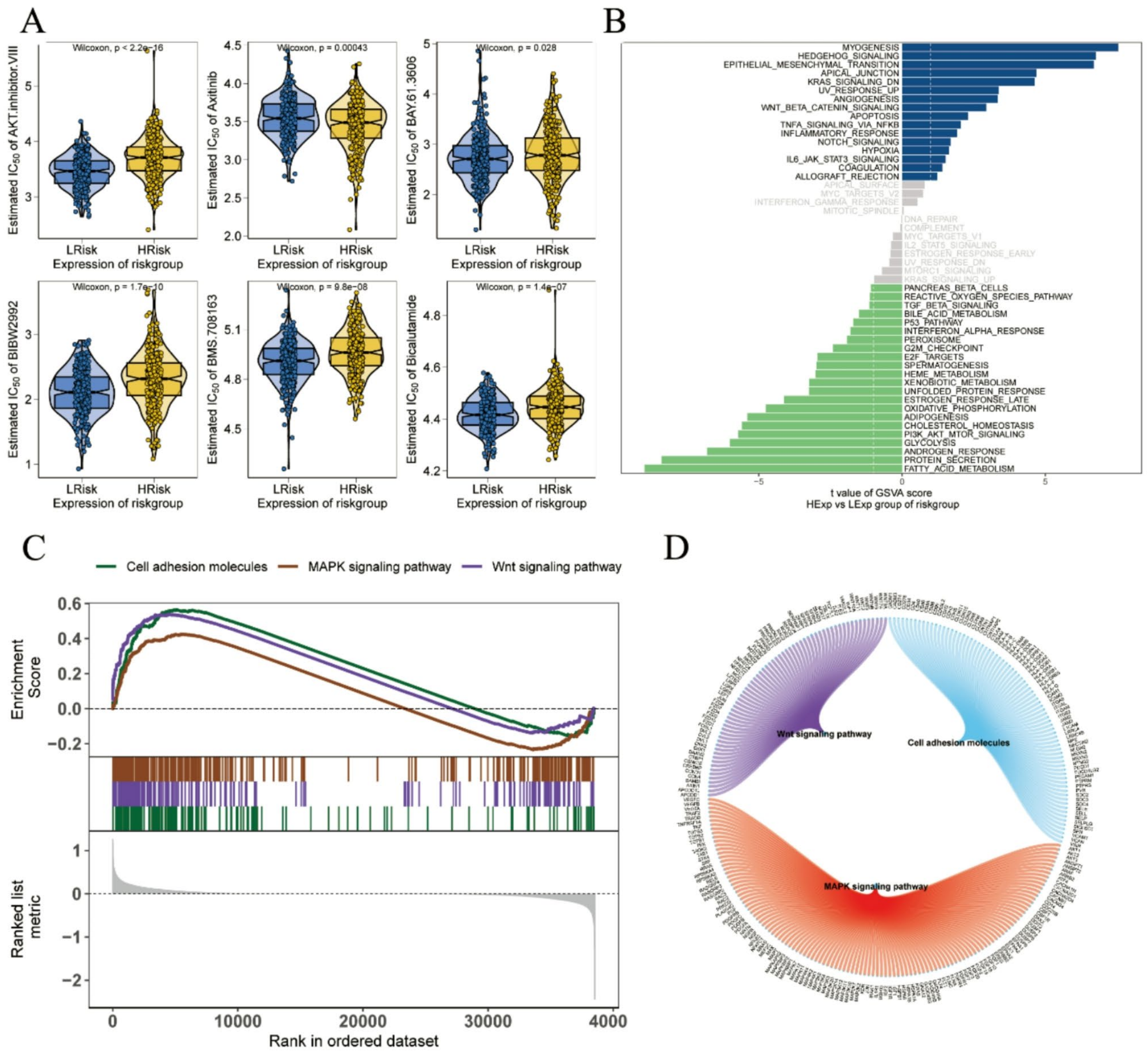


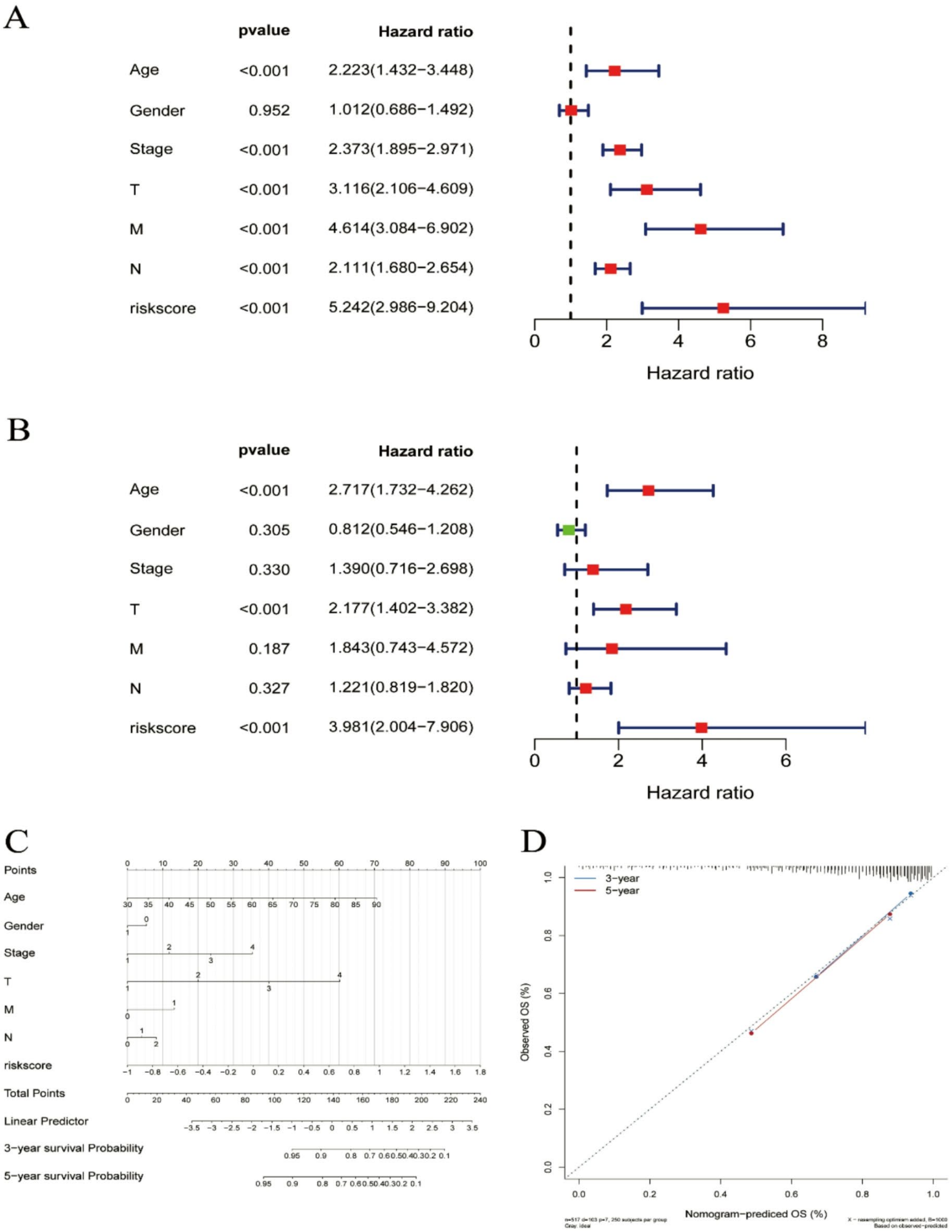
Fig. 6. Potential molecular mechanisms by which the risk score affects tumor progression. (**A**) Relationship between the risk score and sensitivity to common chemotherapy drugs. (**B**,**C**) GSVA and GSEA analyses exploring signaling pathway differences between high and low-risk groups. (**D**) Molecular interaction network among the pathways.

These findings suggest that the risk score derived from the modeling analysis is well-suited for subtyping colorectal cancer samples.

Disscussion

CRC is one of the most common cancers worldwide, with high incidence and mortality rates. It is reported that nearly 1.9 million new cases of CRC and 904,000 CRC-related deaths occur globally each year[1]. Brenner et al. found that patients with early-diagnosed colorectal cancer have a 5-year survival rate exceeding 90%[21]. However, due to inadequate diagnostic methods, colorectal cancer is often diagnosed at advanced stages[22]. Despite significant improvements in diagnosis and treatment, the 5-year survival rate for patients diagnosed with metastatic colorectal cancer remains low, at approximately 12%[23]. Therefore, there is an urgent need to elucidate the molecular mechanisms of colorectal cancer development and to identify novel biomarkers for early detection and prognosis assessment to improve survival outcomes.

The scRNA-seq has emerged as a valuable tool for transcriptomic profiling of various cancer cell types, crucial for identifying potential therapeutic targets. In this study, we utilized colorectal cancer scRNA-seq data from the GEO database to define cellular subpopulations within tumors and characterize their contributions to the disease based on cell numbers and gene expression changes. We then selected marker genes with the



A

	pvalue	Hazard ratio
Age	<0.001	2.223(1.432−3.448)
Gender	0.952	1.012(0.686−1.492)
Stage	<0.001	2.373(1.895−2.971)
T	<0.001	3.116(2.106−4.609)
M	<0.001	4.614(3.084−6.902)
N	<0.001	2.111(1.680−2.654)
riskscore	<0.001	5.242(2.986−9.204)

B

	pvalue	Hazard ratio
Age	<0.001	2.717(1.732−4.262)
Gender	0.305	0.812(0.546−1.208)
Stage	0.330	1.390(0.716−2.698)
T	<0.001	2.177(1.402−3.382)
M	0.187	1.843(0.743−4.572)
N	0.327	1.221(0.819−1.820)
riskscore	<0.001	3.981(2.004−7.906)

Fig. 7. Construction of the nomogram prediction model. (**A**,**B**) Univariate and multivariate analyses identify the risk score as an independent prognostic factor for colorectal cancer patients. (**C**) Regression analysis shows that the risk score significantly contributes to the nomogram prediction model. (**D**) Predictions for three-year and five-year survival rates of colorectal cancer patients.

highest disease relevance from these subpopulations as a candidate gene set for further analysis. This led to the construction of a prognostic risk model with favorable prognostic efficiency, which serves as a biomarker for predicting immunotherapy response. Lei Zheng et al.[24] performed differential and prognostic analyses of genes associated with Cancer-associated fibroblasts (CAFs) and constructed CAF-related signatures to predict

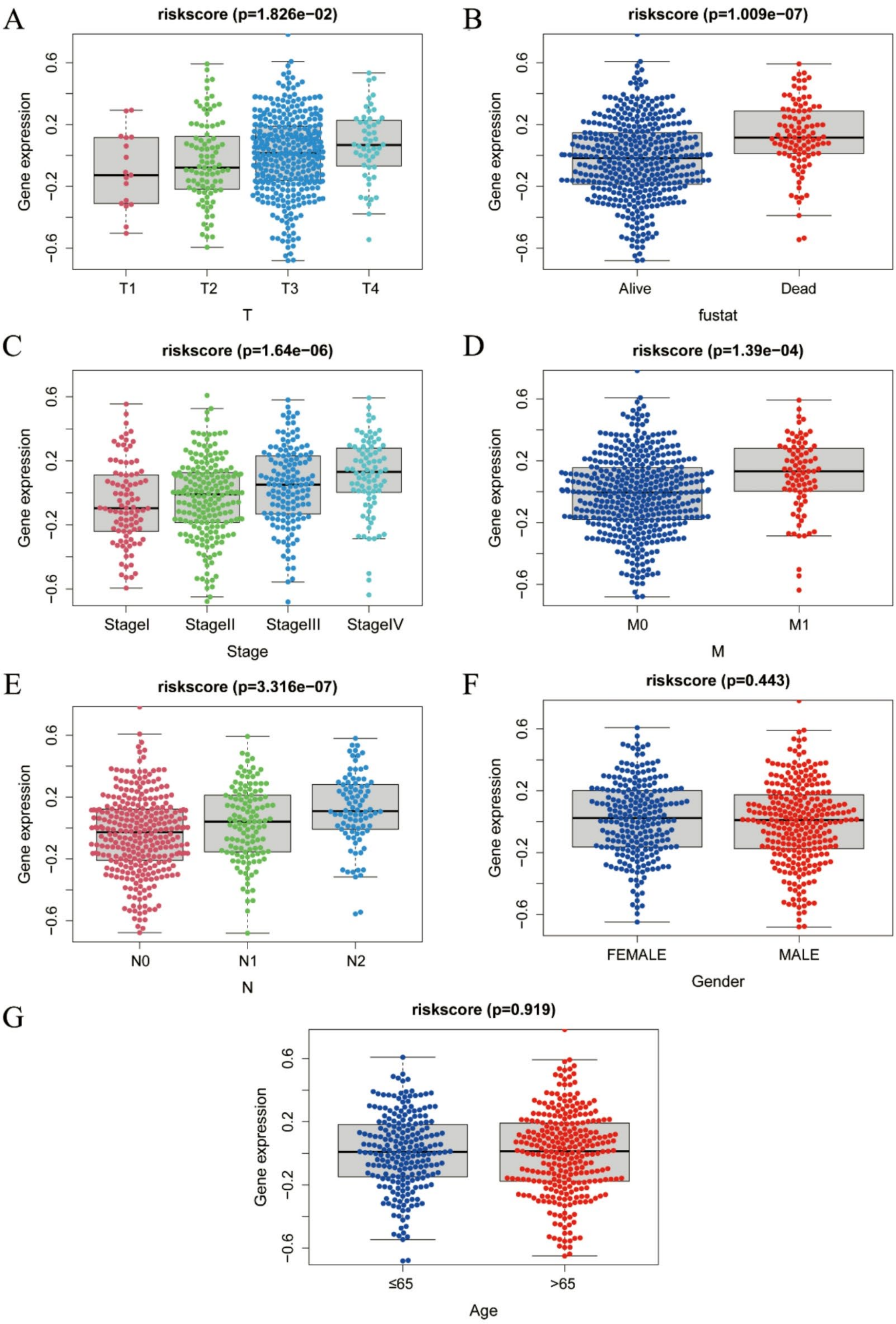


Fig. 8. Clinical indicator analysis further demonstrates the applicability of the risk score to colorectal cancer samples. (**A**–**G**) Box plots display the distribution of risk scores across different groups based on clinical indicator values.

clinical outcomes in individuals with colon adenocarcinoma (COAD) based on machine learning algorithms. Gui et al.[25] also developed a prognostic signature for cervical cancer using transcriptome profiling and clinical data from TCGA, GEO and TISCH database, focusing on cancer-associated fibroblasts (CAFs). Similar viewpoints were also proposed by Juan et al.[26]. They applied scRNA-seq to analyze the heterogeneity of tumor immune cells, developing a 3-gene biomarker (including CLTA, TALDO1, and CSTB) based on tumor immune microenvironment (TIME) heterogeneity to predict survival outcomes and immunotherapy responses. Zheng et al.[27] selected 6 prognosis-related HUB genes from GEO esophageal squamous cell carcinoma (ESCC) and TCGA esophageal cancer datasets, showing significantly increased expression of HUB genes in normal tissues and cells based on scRNA-seq. Further Kaplan-Meier survival analysis and immune infiltration analysis indicated that HUB genes are promising biomarkers for ESCC diagnosis and prognosis. Additionally, studies utilizing scRNA-seq technology have elucidated intercellular interactions in gliomas, identifying autocrine ligand-receptor signaling that significantly impacts prognosis in glioma patients[28]. Collectively, these findings demonstrate that scRNA-seq technology can effectively dissect and identify potential prognostic biomarkers, which are crucial for pinpointing therapeutic targets and improving patient survival outcomes.

In our study, the prognostic signature composed of nine marker genes (S100P, PIGR, RAB11FIP1, USP53, CDH1, LGALS4, ATP10B, SLC12A2, and LAMB3) may provide valuable insights into the molecular mechanisms of CRC. For instance, S100P, a 95-amino acid protein and member of the S100 family, plays a crucial role in regulating cell differentiation, proliferation, migration, apoptosis, and other biological functions by interacting with various signaling proteins such as P53, β-catenin, and nuclear factor-κB (NF-κB). Through these interactions, S100P is involved in tumorigenesis and tumor progression. Research has shown that SIX3 can downregulate S100P via the Wnt/β-catenin signaling pathway, thereby inhibiting cell migration and proliferation[29]. Other studies identified that S100P mRNA levels correlate with the activation status of the PI3K/AKT pathway, a classical pathway involved in promoting cancer migration, invasion, proliferation, and drug resistance[30];[31].

The performance of the prognostic model based on the nine marker genes was validated in both the test and GEO cohorts, yielding consistent results across the two cohorts, indicating good effectiveness and reproducibility of the model. Various validation methods, including univariate, multivariate, and clinical indicator analyses, demonstrated that the nomogram model has high predictive accuracy. Therefore, the nomogram can guide the establishment of personalized examination procedures for CRC patients, promoting the effective utilization of medical resources.

Given that the TME plays a crucial role in anti-tumor responses and significantly influences tumor diagnosis, survival outcomes, and clinical treatment sensitivity[32], we investigated the relationship between risk score and tumor immune infiltration. Firstly, we observed significant decreases in activated dendritic cells, plasma cells, and resting CD4 memory T cells in the high-risk group, suggesting that these patients may be in a relatively immunosuppressive state. Secondly, the study results showed significant positive correlations between the risk score and M0 macrophages, CD8+ T cells, and M2 macrophages, and significant negative correlations with resting CD4+ memory T cells, activated dendritic cells, eosinophils, and plasma cells. This indicates that the TME of the high-risk group may function to reduce inflammation, promote tumor growth, and suppress immunity. Specifically, due to the substantial infiltration of CD8+ T cells and macrophages in the TME of the high-risk group, which suggests a potential inflammatory immune state, adoptive cell transfer therapies may yield promising therapeutic outcomes. In subsequent studies, it would be feasible to employ Chimeric Antigen Receptor T-Cell (CAR-T) immunotherapy or Chimeric Antigen Receptor Macrophages (CAR-M) immunotherapy in the high-risk group. These therapies involve the transduction of artificially designed CAR molecules into T cells or macrophages, endowing them with targeting capabilities to eliminate tumors. However, considering the immunosuppressive environment caused by the extensive infiltration of M2 macrophages, it may be necessary to concurrently activate CD8+ T cells to transform them into M1 macrophages and exert immune functions.

To better guide CRC treatment, we conducted drug sensitivity analyses on different risk groups, studying six common chemotherapy drugs for colorectal cancer. The results indicated that the low-risk group is sensitive to five anticancer drugs, while the high-risk group is sensitive to one anticancer drug. These findings provide a reference for the clinical selection of chemotherapy drugs. Specifically, the high-risk group demonstrates sensitivity to tyrosine kinase inhibitors, represented by Axitinib, suggesting a potential overexpression of EGFR. In clinical practice, it may be feasible to utilize the latest EGFR-targeted drugs, such as Cetuximab and Panitumumab, in patients of the high-risk group to explore their therapeutic efficacy. In subsequent research, we will also employ molecular biological approaches to investigate the expression of EGFR in tumors of the high-risk group and its resistance mutations, with the aim of further elucidating the mechanism of action of EGFR in the occurrence and development of colorectal cancer.

Inevitably, our study has some inherent limitations. Firstly, all cohort studies are retrospective and require further validation in prospective cohort studies. Secondly, further mechanistic studies are needed to reveal the exact role of each gene, and drug sensitivity needs further confirmation through cellular experiments. Thirdly, the number and volume of scRNA-seq samples available in public databases are limited, resulting in an incomplete analysis of clinical and pathological parameters, which may lead to potential biases. Therefore, it is necessary to conduct multicenter, large-sample, prospective double-blind trials for further verification in the future. In addition, the sample size of the GEO datasets may limit the generalizability of the model. Insufficient sample size can lead to unstable performance of the model when applied to new data, especially when the sample distribution is uneven. The distribution variability of the GEO datasets can also affect the model's performance, potentially resulting in poorer performance in specific subpopulations compared to the development dataset. Despite these limitations, our prognostic model demonstrated good consistency and stability across multiple GEO datasets, indicating strong generalizability and applicability to data from different sources. To further enhance the model's generalizability, we will conduct external validation using larger-scale and more representative datasets in future validation studies. Additionally, we will explore the model's performance in different subgroups (e.g., by gender, race/ethnicity) to ensure its fairness and applicability. Lastly, although CRC is often analyzed as a single entity, it is important to recognize that colon and rectal cancers exhibit significant molecular differences and possess distinct biological characteristics. Therefore, it will be necessary in the future to incorporate additional cohorts of colon adenocarcinoma and rectum adenocarcinoma to further refine and validate our model.

Conclusion

This study comprehensively applied various bioinformatics methods to reveal the distribution of different cell types, gene expression characteristics, and their association with clinical prognosis in colorectal cancer tissues. The constructed risk score model demonstrates good predictive performance, offering valuable insights for the personalized treatment of colorectal cancer patients and guiding further exploration of the disease's pathogenesis and therapeutic targets.

Data availability

scRNA-seq and RNA-seq data can be obtained from the GEO and TCGA databases. (GSE221575 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE221575).

Received: 22 July 2024; Accepted: 24 February 2025
Published online: 07 March 2025

References

1. Bray, F. et al. Global Cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Ca-Cancer J. Clin.* **74**, 229–263 (2024).
2. Murphy, C. C., Wallace, K., Sandler, R. S. & Baron, J. A. Racial disparities in incidence of Young-Onset colorectal Cancer and patient survival. *Gastroenterology* **156**, 958–965 (2019).
3. Rho, Y. S. et al. Comparing clinical characteristics and outcomes of Young-Onset and Late-Onset colorectal cancer: an international collaborative study. *Clin. Colorectal Canc* **16**, 334–342 (2017).
4. Dekker, E., Tanis, P. J., Vleugels, J., Kasi, P. M. & Wallace, M. B. *Lancet* **394**, 1467–1480 (2019).
5. Medema, J. P. Cancer stem cells: the challenges ahead. *Nat. Cell. Biol.* **15**, 338–344 (2013).
6. Nassar, D. & Blanpain, C. Cancer stem cells: basic concepts and therapeutic implications. *Annu. Rev. Pathol. -Mech Dis.* **11**, 47–76 (2016).
7. Bondeven, P., Hagemann-Madsen, R. H., Laurberg, S. & Pedersen, B. G. Extent and completeness of mesorectal excision evaluated by postoperative magnetic resonance imaging. *Br. J. Surg.* **100**, 1357–1367 (2013).
8. André, T. et al. Improved overall survival with oxaliplatin, fluorouracil, and leucovorin as adjuvant treatment in stage II or III Colon cancer in the mosaic trial. *J. Clin. Oncol.* **27**, 3109–3116 (2009).
9. Haller, D. G. et al. Capecitabine plus oxaliplatin compared with fluorouracil and folinic acid as adjuvant therapy for stage III Colon cancer. *J. Clin. Oncol.* **29**, 1465–1471 (2011).
10. Ma, B. et al. What has preoperative Radio(Chemo)Therapy brought to localized rectal Cancer patients in terms of perioperative and Long-Term outcomes over the past decades?? A systematic review and Meta-Analysis based on 41,121 patients. *Int. J. Cancer* **141**, 1052–1065 (2017).
11. Alderson, P. & Tan, T. The Use of Cochrane Reviews in Nice Clinical Guidelines. *Cochrane Database Syst Rev.* ED000032 (2011). (2011).
12. Primrose, J. N. et al. Effect of 3 to 5 years of scheduled Cea and Ct Follow-Up to detect recurrence of colorectal cancer: the facs randomized clinical trial. *Jama-J Am. Med. Assoc.* **311**, 263–270 (2014).
13. Shayimu, P. et al. Serum nutritional predictive biomarkers and risk assessment for anastomotic leakage after laparoscopic surgery in rectal Cancer patients. *World J. Gastrointest. Surg.* **16**, 3142–3154 (2024).
14. Pagès, F. et al. International validation of the consensus immunoscore for the classification of Colon cancer: A prognostic and accuracy study. *Lancet* **391**, 2128–2139 (2018).
15. Sargent, D. J. et al. Defective mismatch repair as a predictive marker for lack of efficacy of Fluorouracil-Based adjuvant therapy in Colon cancer. *J. Clin. Oncol.* **28**, 3219–3226 (2010).
16. Le, D. T. et al. Mismatch repair deficiency predicts response of solid tumors to Pd-1 Blockade. *Science* **357**, 409–413 (2017).
17. Shapaer, T. et al. Elevated Bean1 expression correlates with poor prognosis, immune evasion, and chemotherapy resistance in rectal adenocarcinoma. *Discov Oncol.* **15**, 446 (2024).
18. Olsen, T. K. & Baryawno, N. Introduction to Single-Cell Rna sequencing. *Curr. Protoc. Mol. Biol.* **122**, e57 (2018).
19. Torroja, C., Sanchez-Cabo, F. Corrigendum: Digitaldlsorter: Deep-Learning on Scrna-Seq to deconvolute gene expression data. *Front. Genet.* **10**, 1373 (2019).
20. Jin, K. et al. Single-Cell Rna sequencing reveals the Temporal diversity and dynamics of cardiac immunity after myocardial infarction. *Small Methods* **6**, e2100752 (2022).
21. Brenner, H., Kloor, M. & Pox, C. P. *Lancet* **383**, 1490–1502 (2014).
22. Pinsky, P. F. & Doroudi, M. Colorectal Cancer screening. *Jama-J Am. Med. Assoc.* **316**, 1715 (2016).
23. Siegel, R. L. et al. *Ca-Cancer J. Clin.* 73, 17–48 (2023). (2023).
24. Zheng, L. et al. Construction of a novel Cancer-Associated Fibroblast-Related signature to predict clinical outcome and immune response in Colon adenocarcinoma. *Aging (Albany Ny)* **15**, 9521–9543 (2023).
25. Gui, Z. et al. Construction of a novel Cancer-Associated Fibroblast-Related signature to predict clinical outcome and immune response in cervical Cancer. *Transl Oncol.* **46**, 102001 (2024).
26. Lu, J. et al. A novel prognostic model based on Single-Cell Rna sequencing data for hepatocellular carcinoma. *Cancer Cell. Int.* **22**, 38 (2022).
27. Zheng, L., Li, L., Xie, J., Jin, H. & Zhu, N. Six novel biomarkers for diagnosis and prognosis of esophageal squamous cell carcinoma: validated by Scrna-Seq and Qpcr. *J. Cancer* **12**, 899–911 (2021).
28. Yuan, D., Tao, Y., Chen, G. & Shi, T. Systematic expression analysis of Ligand-Receptor pairs reveals important Cell-to-Cell interactions inside glioma. *Cell. Commun. Signal.* **17**, 48 (2019).
29. Liu, S. et al. Trim27 acts as an oncogene and regulates cell proliferation and metastasis in Non-Small cell lung Cancer through Six3-B-Catenin signaling. *Aging (Albany Ny)* **12**, 25564–25580 (2020).
30. De Marco, C. et al. Specific gene expression signatures induced by the multiple oncogenic alterations that occur within the Pten/Pi3K/Akt pathway in lung Cancer. *Plos One* **12**, e0178865 (2017).
31. Zhang, H. et al. Kif18a inactivates hepatic stellate cells and alleviates liver fibrosis through the Ttc3/Akt/Mtor pathway. *Cell. Mol. Life Sci.* **81**, 96 (2024).
32. Pitt, J. M. et al. Targeting the tumor microenvironment: removing obstruction to anticancer immune responses and immunotherapy. *Ann. Oncol.* **27**, 1482–1492 (2016).

Author contributions

All the authors participated in writing the manuscript and in drawing the figures. All authors read and approved the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-91761-y.

Correspondence and requests for materials should be addressed to H.F., J.Z. or Y.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.