Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Integrative Transcriptomic Analysis Reveals Multicellular Dysregulation and Spatial Heterogeneity in the Colorectal Cancer Microenvironment

Yan Lin

linyan@gxmu.edu.cn

Guangxi Medical University Cancer Hospital

Wenfeng Luo

Guangxi Medical University Cancer Hospital

Xiaoqing Li

Guangxi Medical University Cancer Hospital

Jingting Su

Guangxi Medical University Cancer Hospital

Yanying Pan

Guangxi Medical University Cancer Hospital

Guohai Yang

Guangxi Medical University Cancer Hospital

Shanshan Luo

Guangxi Medical University Cancer Hospital

Yubin Huang

The Second Affiliated Hospital of Guangxi Medical University

Research Article

Keywords: Single cell RNA-sequencing, Spatial transcriptomics, Colorectal cancer, Prognosis, Biomarkers

Posted Date: September 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10693359/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Tumor heterogeneity impedes personalized therapy for colorectal cancer (CRC), while the complex cellular interactions within the tumor microenvironment critically influence disease progression through mechanisms not yet fully elucidated.

Methods

We integrated multi-omics data including bulk RNA-seq, scRNA-seq, and spatial transcriptomics (ST) from CRC and normal tissues. Bulk RNA-seq datas were subjected to differential expression analysis, functional enrichment, and screened prognostic feature genes by machine learning-based prognostic modeling and survival analysis. Consensus high-expression risk genes and low-expression protective genes at the intersection of feature genes and epithelial cell DEGs were designated as important genes. scRNA-seq data were used for cluster analysis, trajectory analysis, cell-cell communication and high-dimensional weighted gene co-expression network analysis. ST data were deconvoluted and analyzed for spatial heterogeneity and copy number variations analysis.

Results

A total of 5,127 overlapping DEGs were identified, revealing key dysregulated pathways: cell cycle, DNA replication, mismatch repair, base excision repair, and p53 signaling. We explored an eight-gene prognostic signature (ADH5, FOXD1, GDI1, PSMF1, PTP4A2, RFXANK, RNF113A, SSNA1) for patient stratification. scRNA-seq delineated functional subtypes of T cells and cancer-associated fibroblasts (CAFs), revealing extensive interaction networks. And ten co-expression network modules were identified associated with CAFs. ST analysis spatially suggested that malignant epithelial cells may act as signaling hubs, actively remodeling an immunosuppressive, pro-tumor microenvironment.

Conclusion

This integrative multi-omics analysis characterizes the multicellular dysregulation in the CRC microenvironment, identifies an exploratory prognostic gene signature, and highlights candidate functional modules in CAFs. These findings provide a resource for further investigation of CRC microenvironment biology and potential biomarker development.

1. Introduction

Colorectal cancer (CRC) is the third most common malignancy and second leading cause of cancer-related deaths worldwide (1). The five-year survival rate for patients with early stage CRC exceeds 90%,

whereas that for patients with advanced disease presenting with metastases and complications is below 10% (2). Although personalized therapies reduce CRC mortality, tumor heterogeneity remains a barrier to effective clinical management (3, 4). Such heterogeneity spans multi-omic layers from the genomic to the transcriptomic level and extends to cellular interactions within the tumor microenvironment (TME) (5). Therefore, deciphering the complexity of the molecular and cellular crosstalk in the CRC TME may offer novel insights and therapeutic avenues for CRC prevention and improved patient outcomes.

The TME, a central focus in oncology research, influences treatment outcomes through immune infiltrates (6, 7). Within the tumor ecosystem, diverse cell populations form dynamic interaction networks that shape microenvironmental homeostasis (8–10). The TME is primarily composed of diverse cell types, including immune cells, fibroblasts, vascular endothelial cells, and noncellular components (11). Cellular components within the TME directly orchestrate malignant tumor progression, organ-specific metastasis, and immunotherapy outcomes, wherein immune-infiltrating cells occupy a pivotal hub position in the TME (12). Studies have reported multiple cell types within the local TME in breast cancer, renal cell carcinoma, and CRC (13–15). Furthermore, memory B/T cell accumulation in the TME determines the clinical stage of CRC and reflects the efficacy of the antitumor immune response (16). Consequently, elucidating the cellular composition within the CRC TME may enhance the precision of immunotherapy.

A recent study by Zhou et al. integrated single-cell and spatial transcriptomics to characterize MUC1-mediated cellular crosstalk in mucinous colorectal adenocarcinoma using the GSE236696 and GSE236697 datasets(17). Building upon this foundation, the present study extends the analysis by: (i) constructing and assessing a machine learning-based prognostic gene signature using TCGA and GSE39582 bulk RNA-seq cohorts, (ii) identifying CAF-associated co-expression modules through hdWGCNA, and (iii) performing in silico gene perturbation analysis to infer potential functional consequences of GDI1 modulation in CAFs. Through these analyses, we aimed to further characterize multicellular dysregulation in the CRC microenvironment and explore clinically relevant biomarkers.

2. Methods

2.1 Data collection

This study included one 10x Genomics spatial transcriptomic dataset (GSE236697), one 10x Genomics scRNA-seq dataset (GSE236696), and an Affymetrix microarray gene-expression dataset (GSE39582) from the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo/)(Table 1). GSE236696 comprises paired mucinous colorectal adenocarcinoma and adjacent normal colon tissue samples from six patients profiled by scRNA-seq. GSE236697 comprises spatial transcriptomic profiles of mucinous colorectal adenocarcinoma and adjacent normal tissues from one patient (17). Notably, these datasets are restricted to the mucinous colorectal adenocarcinoma subtype, limiting the generalizability of our findings to non-mucinous CRC. GSE39582 contains gene-expression profiles and survival data from 566 primary colon cancer samples and 19 non-tumor samples generated on the GPL570 platform (18).

Transcriptomic and survival data for 471 colon adenocarcinoma (COAD) tumor samples and 41 normal samples, as well as 166 rectum adenocarcinoma (READ) tumor samples and 10 normal samples, were obtained from The Cancer Genome Atlas COAD (TCGA-COAD) and READ (TCGA-READ) projects, respectively (19).

Table 1
Information of the data set in this study

Deposited Data	Species	Platform	Identifier
single-cell RNA-sequencing dataset	Homo sapiens	10x Genomics	GEO: GSE236696
Spatial Transcriptomics dataset	Homo sapiens	10x Genomics	GEO: GSE236697
bulk RNA-seq dataset	Homo sapiens	GPL570	GEO: GSE39582
bulk RNA-seq dataset	Homo sapiens	Illumina Hiseq	The Cancer Genome Atlas database

2.2 Differential expression analysis and enrichment analysis in transcriptome data

The ComBat_Seq function of the sva package was used to merge and debatch COAD and READ samples. Subsequently, principal component analysis (PCA) was employed to perform dimensionality reduction on the TCGA datasets (COAD and READ). We performed a differential expression analysis using the DESeq2 package, comparing 637 tumor and 51 normal samples, to identify CRC-related DEGs in TCGA cohort. The limma package (20) was used to perform a differential analysis of the GSE39582 expression profile. Significance thresholds for DEGs were set at an adjusted $p < 0.05$ and an absolute average log2 fold change ($|avg_log2FC| > 0.1$.

We integrated the TCGA and GSE39582 datasets to extract overlapping DEGs and identify their functional roles in CRC. Functional enrichment analysis was performed using ClusterProfiler (21) for gene ontology (GO), including cellular components (CC), biological processes (BP), molecular functions (MF), and KEGG pathways. Gene set enrichment analysis (GSEA) was performed on overlapping DEGs using the GSEABase package (http://www.bioconductor.org/packages/release/bioc/html/GSEABase.html), with pathways ranked by normalized enrichment score (NES). The top five significantly activated and inhibited pathways were selected and visualized using GseaVis to highlight the conserved functional signatures. The significance threshold for GO and pathways was set at adjusted $p < 0.05$.

2.3 Constructing prognostic models by integrative machine learning methods

We performed Kaplan–Meier survival analysis with log-rank testing was performed on tumor samples from the GSE39582 and TCGA cohorts to identify prognosis-associated genes ($p < 0.05$). Subsequently, we constructed a prognostic model using consensus high- and low-expression protective genes that showed significant survival associations in both datasets to investigate the prognostic value of CRC feature genes. Multiple machine learning approaches were applied for prognostic modeling, including random survival forest (RSF), elastic network, Lasso, Ridge, stepwise Cox, CoxBoost, partial least squares regression for Cox, supervised principal components, generalized boosted regression modeling, survival support vector machine, CForest, xgBoost, and oblique RSF. The TCGA cohort served as the training set, whereas the GSE39582, COAD, and READ cohorts were used as testing sets. The C-index and area under the curve (AUC) were calculated across the four datasets, and the highest C-index and AUC were considered the best-performing prognostic model (22). An AUC value > 0.60 was considered indicative of acceptable discriminatory performance.

2.4 Assessments of prognostic genes

We used a multi-tiered gene screening strategy to identify prognosis-related genes in CRC. We first obtained the intersecting genes between DEGs of the major cell types in CRC and epithelial cells in the tumor and normal groups. Next, consensus high-expression risk genes and low-expression protective genes of common genes between feature genes and intersecting genes were considered important genes. Thereafter, a waterfall plot was used to visualize the single nucleotide polymorphisms (SNPs) of important genes from the 583 samples in TCGA dataset.

According to the important genes, patients were divided into high- and low-risk groups, and survival differences were determined based on KM curves. Important genes and 3- and 5-year survival probabilities were constructed to predict CRC prognosis using the rms R package. In addition, a risk-scoring model was built based on the median risk score, and CRC samples were divided into high- and low-risk groups in TCGA cohort. Correlations among the important genes and differences in their expression levels between the CRC and normal groups were further evaluated.

2.5 scRNA-seq data processing and construction of cell atlas

First, low-quality cells in the GSE236696 expression profile were filtered out using the Seurat R package (23). Additionally, contaminated cells were removed using the decontX package, and double-cell artifacts were eliminated using the DoubletFinder package. After quality control, 62,288 cells from the GSE236696 dataset were selected for subsequent analysis.

Harmony was used to remove batch effects from the samples and further performed dimensionality reduction using UMAP in the Seurat package (23). The resulting cell clusters were visualized on a cellular atlas in a loop map using the plot1cell package. Cell-type annotation was performed using known markers. Subsequent proportional composition analysis revealed significant proportional differences in these cell types between the tumor and normal groups using the ggplot2 package.

Nonnegative matrix factorization (NMF) offers an effective approach to dimensionality reduction, yielding biologically interpretable outputs (24). Here, we applied this approach to T cells and fibroblasts using the NMF package, and subsequently visualized the results by UMAP in Seurat package (23). The MiloR package was employed for the differential abundance analysis of subtypes, identifying those showing significant proportional differences. Additionally, the corrplot function from the corrplot package was used to compute correlations between these subtypes.

2.6 Differential expression analysis and enrichment analyses of scRNA-seq

We applied the FindAllMarkers function (25) to extract DEGs with exclusive expression patterns for each cell type in CRC and define the specific signatures of the cell types. Subsequently, the FindAllMarkers function was used to identify the DEGs of epithelial cells between CRC and normal samples to compare the changes in major cell types between the normal and diseased states. Statistical significance was defined as an adjusted $p < 0.05$ and an $|avg_log2FC| > 0.1$.BPs and KEGG pathway enrichment analyses were conducted using the clusterProfiler package (21) to characterize the specific functions of major cell types in CRC.

2.7 Verifying potential tumour cell clusters using InferCNV

We used the InferCNV package to explore large-scale chromosomal CNVs to identify malignant cells in patients with CRC.

2.8 Cell-cell communication analysis

We used the R software package CellChat (26) to analyze ligand-receptor pairs across major cell types and identify cell state-specific signaling networks to explore how these signaling molecule networks can achieve coordinated regulation among various cell types.

2.9 Pseudotime analysis

We first analyzed the differentiation states of T cell and fibroblast subtypes using the CytoTRACE package to identify cells in the least differentiated state. Based on the differentiation scores predicted by CytoTRACE, we selected cells with the lowest differentiation state as the starting point for pseudotime analysis. T cell subtypes and fibroblast subtypes were analyzed in separate pseudotime trajectories using the monocle3 package. The T cell trajectory contained both CD4 + and CD8 + lineages and the inferred ordering reflects transcriptional similarity. Genes meeting $q < 0.05$ and Moran's index > 0 were screened, and the top 20 genes ranked by Moran's absolute value were displayed in a heat map.

2.10 Expression of important genes and pathways

The plot_density function of the Nebulosa package was used to generate heat maps of gene expression for the important genes in the two subtypes and the CRC cell atlas to investigate the functional activity regions of T cells and fibroblasts in the CRC microenvironment. The AddModuleScore algorithm was

used to calculate gene set activity scores to quantify the activity of key pathways, thereby elucidating how microenvironmental positioning regulates cellular functions.

2.11 Virtual gene perturbation analysis

scTenifoldKnk employs a machine learning workflow to simulate gene knockout on scRNA-seq data and assess its functional impact (27). To elucidate the downstream consequences of GDI1, we established a single-cell gene regulatory network from CAF expression data, conducted in silico knockout and overexpression of GDI1, quantified the resulting perturbation impact, and thereby screened for differentially regulated genes (DRGs). These in silico perturbations do not establish causality. We then displayed the DRGs affected by GDI1 perturbation in a regulatory network, applying a significance threshold of $p < 0.05$.

2.12 hdWGCNA

The R package hdWGCNA (28) was utilized to construct gene co-expression networks and identify CAF-associated modules. Based on the selected soft-thresholding power, we constructed an adjacency matrix and a topological overlap matrix (TOM). Hierarchical clustering was then performed to generate a dendrogram, from which co-expression gene modules were identified. UMAP plots further visualized the correlation between module genes. The volcano plot explored the differences between each module based on CRC and control samples. Finally, Enrichr tools were used to characterize the GO enrichment within each module (29).

2.13 Processing and deconvolution of ST data

Spots with unique molecular identifier counts below 500 or above 45,000 were filtered from both samples to exclude outliers to assess the quality of the ST data (30). Harmony (31) was used to remove batch effects from the ST data, and the Seurat package was used to reduce the UMAP dimensions.

The 10x Genomics of ST data lacked single-cell spatial resolution, and the gene expression of each spot constituted different cell types of varying numbers. Therefore, we used SPOTlight for spatial deconvolution analysis to integrate single-cell transcriptome data into ST data and show the proportion of cell types in each spatial spot.

2.14 Analyses of ST data

We conducted CNV analysis of CRC tissues using the SPATA2 package based on ST data to analyze the spatial genomic heterogeneity in tumor tissues. CellChat v2 (32) was used to map spatial cellular communication networks to infer the interactions between cell types. Subsequently, key pathways were quantified using the AddModuleScore function and their expression patterns were visualized using SpatialFeaturePlot from the Seurat package. Finally, the spatial feature plot from Seurat was used to visualize the ST expression profiles of important genes in CRC tissues.

2.15 Statistical analysis

Data processing and visualization were performed using R version 4.3.0. Survival analysis with log-rank testing was used to determine statistical significance. Furthermore, the statistical significance threshold for DEGs, GO, and KEGG pathways was set at an adjusted $p < 0.05$.

3. Results

3.1 The core dysregulation pathway in CRC molecular heterogeneity

To characterize the transcriptional landscape of CRC, we performed differential expression analysis across four transcriptomic datasets: TCGA-COAD, TCGA-READ, GSE39582, and the merged TCGA cohort, and its individual components COAD and READ. Following batch correction of COAD and READ, PCA confirmed integration quality (Fig. 2A). Each cohort showed substantial transcriptional dysregulation, yielding 11,463, 12,200, 8,958, and 12,469 DEGs, respectively (Fig. 2B). Intersection across all four datasets identified a core set of 2,973 upregulated and 2,154 downregulated overlapping DEGs (Fig. 2C). Enrichment analysis of these 5,127 genes revealed enrichment in cell cycle, DNA replication, focal adhesion, chromosomal regions, ATP hydrolysis activity, and kinase regulator activity (Fig. 2D). Five KEGG pathway gene sets were prominently enriched: cell cycle, DNA replication, mismatch repair, base excision repair, and p53 signaling (Fig. 2E). GSEA independently confirmed that cell cycle- and DNA replication-related gene sets were among the most activated, while drug metabolism and lipolysis regulatory pathways were suppressed (Fig. 2F). Together, these results define cell cycle, DNA replication, and DNA repair as the core transcriptional programs dysregulated in CRC.

3.2 Global single-cell transcriptome map of CRC and normal tissue samples

To resolve the cellular composition of the CRC microenvironment, we analyzed scRNA-seq data from six paired tumor and adjacent normal tissue samples (GSE236696). After quality filtering, 62,288 cells were retained and clustered into 25 transcriptionally distinct groups, which were annotated as nine major cell types using canonical markers: B cells (n = 14,174), T cells (n = 24,126), NK cells (n = 2,038), mast cells (n = 228), myeloid cells (n = 7,272), fibroblasts (n = 7,289), epithelial cells (n = 2,544), endothelial cells (n = 3,921), and enteric glial cells (n = 696) (Fig. 3A–B). Comparison of cellular proportions between tumor and normal tissues revealed a shift in composition: T cells declined from 47.5% to 29.2%, while myeloid cells expanded from 8.8% to 23.1%, fibroblasts increased from 6.9% to 14.5%, and endothelial cells rose from 3.7% to 10.8% (Fig. 3C). Marker gene expression confirmed the identity of each cell type (Fig. 3D). These compositional changes suggest a remodeling of the CRC microenvironment characterized by immune cell depletion and stromal expansion.

3.3 Signaling dysregulation reveals functional heterogeneity across all cell types in CRC

To assess the functional state of each cell population, we performed differential expression analysis within all nine cell types, identifying 23,755 cell-type-specific DEGs (Fig. 3E). An additional 4,656 DEGs were identified in epithelial cells between tumor and normal tissues (Fig. 4A), and the top 10 DEGs per cell type revealed marked transcellular transcriptional heterogeneity (Fig. 4B). GO enrichment of each cell type's DEGs highlighted type-specific biological programs: T cells were enriched in T cell activation and lymphocyte differentiation; myeloid cells in innate immune response; fibroblasts in extracellular matrix organization; B cells in antigen receptor signaling (Fig. 4C). At the pathway level, the MAPK pathway and the p53 signaling gene set were broadly enriched across all nine cell types (Fig. 4D). TP53 is frequently inactivated in CRC, and enrichment of the p53 gene set may reflect proliferation-associated expression of cell cycle and DNA repair genes within this gene set. To distinguish malignant from non-malignant cells, we performed copy number variation (CNV) inference using T cells, B cells, fibroblasts, myeloid cells, and NK cells as normal references. Epithelial cells displayed prominent CNV signals, particularly on chromosomes 1 and 12, confirming their malignant identity (Fig. 4E). Scoring of the five core pathways across the cell atlas confirmed broad pathway activity across all cell types (Fig. 4F).

3.4 Screening of key genes related to CRC prognosis and construction of prognostic score model

To identify genes with prognostic relevance, we applied Kaplan–Meier survival analysis to tumor samples from the GSE39582 and TCGA cohorts. KM pre-selection was performed on both datasets prior to model training. Among twelve machine learning algorithms evaluated, the Gradient Boosting Machine combined with random survival forest (GBM + RSF) achieved the highest C-index across the four datasets and was selected for feature gene identification, yielding 73 candidate feature genes (Fig. 5A). Across the nine evaluated algorithm combinations, C-index values ranged from 0.40 to 0.97, with the highest performance (C-index > 0.90) observed in TCGA and GSE39582, and lower values (C-index 0.50–0.89) in the smaller COAD and READ subsets. Intersecting these feature genes with epithelial cell DEGs and filtering for consensus high-expression risk and low-expression protective genes identified eight important genes: ADH5, FOXD1, GDI1, PSMF1, PTP4A2, RFXANK, RNF113A, and SSNA1. Mutational profiling of TCGA samples revealed low-frequency mutations (1%) in ADH5, GDI1, PSMF1, PTP4A2, and RNF113A, while FOXD1, RFXANK, and SSNA1 showed no detectable mutations (Fig. 5B). In univariate Kaplan–Meier analysis, all eight genes were significantly associated with overall survival (Fig. 5C): ADH5 ($p = 0.022$, HR = 0.88) and PTP4A2 ($p = 0.025$, HR = 0.88) were associated with favorable prognosis, whereas FOXD1 ($p = 0.006$, HR = 1.54), GDI1 ($p = 0.001$, HR = 1.59), PSMF1 ($p = 0.020$, HR = 1.49), RFXANK ($p = 0.044$, HR = 1.44), RNF113A ($p = 0.036$, HR = 1.40), and SSNA1 ($p = 0.047$, HR = 1.43) were associated with worse survival. A nomogram integrating all eight genes was constructed to predict 3- and 5-year survival probabilities (Fig. 5D), and calibration curves showed acceptable agreement between predicted and observed outcomes (Fig. 5E). Stratification by median risk score separated patients into high- and low-risk groups with distinguishable survival trajectories (Fig. 5F). Gene-gene correlation analysis (Fig. 5G) and differential expression between tumor and normal tissues (Fig. 5H) confirmed that

the six risk-associated genes were upregulated in tumors, while ADH5 and PTP4A2 were downregulated, consistent with their protective associations.

3.5 Coordinated reprogramming of T cells and cancer-associated fibroblasts (CAF) in CRC microenvironment

Given the prominent compositional changes observed in T cells and fibroblasts, we performed sub-clustering analysis of these populations alongside endothelial and B cells to refine their functional taxonomy (Fig. 6A). T cells were resolved into six subtypes across seven clusters: Tem/Temra cytotoxic T cells (n = 11,920), Effector T cells (n = 5,112), CD8 + T cells (n = 3,635), Tem/Trm cytotoxic T cells (n = 1,447), Th17 cells (n = 1,378), and follicular helper T cells (TFH, n = 634). CAFs were resolved into five subtypes: inflammatory CAF (iCAF), mesenchymal CAF (mesCAF), myofibroblastic CAF (myCAF), antigen-presenting CAF (apCAF), and vascular CAF (vCAF) (Fig. 6B). Differential abundance analysis revealed that Tem/Trm cytotoxic T cells and CD8 + T cells were significantly depleted in tumor tissue, whereas mesCAFs and iCAFs were significantly expanded and myCAFs were reduced (Fig. 6C). Correlation analysis among all subtypes showed coordinated abundance patterns within each lineage (Fig. 6D). Activity scores for the five core pathways demonstrated elevated pathway engagement in both T cell and CAF subtypes, most prominently in mesCAFs and myCAFs (Fig. 6E). Among the eight prognostic genes, GDI1 showed the strongest risk association (HR = 1.59, p = 0.001) and was highly expressed in CAFs. As an exploratory analysis, in silico gene perturbation was performed: simulated GDI1 knockout in CAFs was associated with altered expression of 14 genes including VEGFA and HSPB1, while simulated overexpression was associated with changes in 83 genes including TIMP2 and CD63 (Fig. 9A–B). These computational predictions do not establish a functional role for GDI1 and require experimental validation.

3.6 Coordinated T cell-CAF crosstalk in the CRC microenvironment

Having established the transcriptional profiles of T cell and CAF subtypes, we next asked how these populations interact within the tumor microenvironment. To explore intercellular communication, we applied CellChat to infer ligand-receptor interactions among T cell and CAF subtypes. Effector T cells served as the dominant communication hub within the T cell compartment, interacting with all five other T cell subtypes, while among CAFs, mesCAFs and myCAFs showed the strongest communication intensity (Fig. 7A). A total of 24 significant ligand-receptor pairs were identified among T cell subtypes, with CLEC- and MHC-I-mediated pathways exhibiting the highest bidirectional signaling intensity. Among CAF subtypes, 112 significant pairs were detected, with MIF, APP, and COLLAGEN showing the highest communication flux (Fig. 7B; Figure S1). To infer transcriptional trajectories, we applied CytoTRACE and pseudotime analysis. Among T cells, TFH cells exhibited the highest CytoTRACE scores, suggesting a less differentiated transcriptional state compared to effector and cytotoxic populations. Among CAFs, mesCAFs showed the highest scores (Fig. 7C). Pseudotime analysis organized T cell and CAF subtypes

along separate trajectories, with TFH positioned near the origin of the T cell trajectory and mesCAFs near the origin of the CAF trajectory (Fig. 7D). The eight prognostic genes showed heterogeneous expression distribution across T cell and CAF subtypes as well as the broader CRC cell atlas (Fig. 8A–B), providing a view of their spatial distribution within the tumor ecosystem.

3.7 Co-expression network analysis reveals immunomodulatory CAF subtypes in CRC

To further characterize the transcriptional organization of CAFs, we performed high-dimensional weighted gene co-expression network analysis (hdWGCNA). A soft-thresholding power of 6 was selected to achieve scale-free network topology (Fig. 9C). Hierarchical clustering identified ten co-expression modules, designated M1 through M10 (Fig. 9D), which were visualized as a UMAP embedding of the gene co-expression network (Fig. 9E). Comparison between CRC and normal samples revealed that modules M2, M5, and M6 were significantly upregulated in CRC, while module M4 was downregulated (Fig. 9F). GO enrichment analysis of each module revealed distinct functional repertoires: M1 was enriched in cell migration and signal transduction; M2 in focal adhesion and cell junction organization; M3 in extracellular vesicle biology and respiratory chain components; M4 in translation and ribosome biogenesis; M5 in DNA replication; M6 in RNA processing; M7 in chromatin organization; M8 in antigen processing and presentation via MHC class II, suggesting a CAF subset with T cell-activating potential; M9 in extracellular matrix disassembly and metallopeptidase activity; and M10 in extracellular structure organization and angiogenesis (Fig. 9G). These modules highlight functionally divergent CAF states within the CRC microenvironment.

3.8 ST deciphers the pro-tumor ecosystem orchestrated by malignant epithelial modules in CRC

To assess the spatial organization of the CRC microenvironment, we analyzed ST data from tumor (GSM7573205) and adjacent normal (GSM8286350) tissue sections from a single patient with mucinous colorectal adenocarcinoma. UMAP clustering of 4,978 spots resolved nine spatial clusters, which were grouped into four modules: immune (n = 2,055), epithelial (n = 455), fibroblast (n = 823), and myeloid (n = 1,645) (Fig. 10A). Deconvolution of ST spots using scRNA-seq as a reference mapped cell-type proportions to each spatial location (Figure S2A). The spatial distribution of T cells and fibroblasts showed distinct patterns between tumor and normal tissue (Fig. 10B; Figure S2B). CellChat analysis at the spatial level revealed prominent intercellular interactions between the epithelial module and the immune, myeloid, and fibroblast modules in the tumor, suggesting that malignant epithelial cells may function as signaling hubs (Fig. 10C). CNV analysis of spatial transcriptomic data confirmed enrichment of chromosomal aberrations on chr1 and chr12 in the epithelial and myeloid modules (Fig. 10D). Among the five core pathway gene sets, the p53 signaling gene set showed spatial heterogeneity with localized alterations in tumor regions (Fig. 10E). The eight prognostic genes displayed heterogeneous spatial expression across the tissue (Figure S2C). These spatial findings are derived from a single patient and at 55 μm resolution.

4. Discussion

In this study, we integrated bulk RNA-seq, single-cell transcriptomics, and ST data to characterize multicellular dysregulation in the CRC tumor microenvironment and explore a prognostic gene signature using machine learning-based feature selection. Our analysis provides a descriptive framework spanning molecular, cellular, and spatial dimensions, although all conclusions derive from computational inference.

Integrative analysis of multiple large-scale datasets revealed a transcriptional dysregulation landscape in CRC. Enrichment analysis and GSEA of overlapping DEGs consistently identified the cell cycle, DNA replication, mismatch repair, and base excision repair pathways as core dysregulated molecular events. The p53 signaling gene set was also enriched; however, TP53 is among the most frequently mutated genes in CRC, and the observed enrichment of the p53 gene set may reflect proliferation-driven expression of cell cycle and DNA repair component genes within this gene set. Subsequently, scRNA-seq analysis revealed that the dysregulation of core pathways is a universal feature across all major cell types within the tumor microenvironment, pointing to a field effect of comprehensive carcinogenic reprogramming. This global dysregulation was further substantiated by ST data, which pinpointed highly active p53 pathway signals, specifically within the core tumor region. Loss of p53 accelerates the malignant progression of CRC (33). Dysfunction of the mismatch repair and base excision repair pathways promotes genomic instability and elevates mutation rates, ultimately driving more aggressive tumor behavior and worse clinical outcomes in CRC (34). Furthermore, inactivation of the core regulatory mechanism of the cell cycle and continuous abnormal activation of the DNA replication process jointly drive the proliferation of tumor cells, which is the core mechanism of the occurrence and development of CRC (35). Collectively, these findings identify the coordinated dysregulation of core pathways involved in cell proliferation and genomic stability, such as the cell cycle, DNA replication, and p53 signaling, may be as the central drivers of CRC development.

We constructed a single-cell transcriptomic atlas for CRC, which revealed a characteristic immunosuppressive landscape within the CRC TME. Specifically, we observed the exhaustion of T and B cells and significant expansion of CAFs, myeloid cells, and endothelial cells. Studies have shown that the presence of T cells within CRC tumors serves as a prognostic biomarker for improved outcomes, underscoring the therapeutic potential of harnessing immune editing processes (36). Conversely, CAFs are highly abundant, linked to poor prognosis (37) and interact with myeloid cells to drive malignancy (38). Similarly, the prevalence of tip-like endothelial cells correlates with poor survival in CRC and other cancers (39). Thus, the interplay between functionally impaired T cells and activated CAFs has emerged as a central node in fostering immunosuppression. Consequently, targeting this co-regulatory axis may present a novel therapeutic strategy for restoring immune control and inhibiting stromal-supported tumor progression.

Based on these findings, we identified eight key genes (ADH5, FOXD1, GDI1, PSMF1, PTP4A2, RFXANK, RNF113A, and SSNA1) and developed an exploratory prognostic model. Among them, six genes,

including FOXD1 and GDI1, were classified as high-risk factors, whereas ADH5 and PTP4A2 showed potential protective effects. GDI1 encodes a small GTPase dissociation inhibitor that regulates GTPase membrane cycling and thereby influences proliferation, migration, and adhesion (40). Specifically, overexpression of GDI1 correlates with enhanced malignant progression and unfavorable prognosis in CRC, highlighting its potential as a prognostic biomarker and target for therapeutic development (41). FOXD1 promotes tumor growth and chemoresistance in non-small-cell lung cancer (42), whereas its knockdown inhibits the proliferation, migration, and invasion of CRC cells, highlighting its context-dependent roles in cancer progression (43). PSMF1 (44) and RNF113A (45) are associated with CRC progression. ADH5, a member of the alcohol dehydrogenase family, participates in inflammation, immune regulation, and cancer development, and its dysregulation may disrupt cellular homeostasis and contribute to disease pathogenesis (46, 47). Although a high expression of PTP4A2 (48) and SSNA1 (49) has been linked to poor prognosis in hepatocellular carcinoma, their roles in CRC remain poorly characterized. Thus, multiple genes in CRC play dual roles in promoting and inhibiting cancer, and their functional analysis has opened up new avenues for the development of novel prognostic biomarkers and targeted therapy strategies.

We further investigated the cellular ecology and interactions between the two most heterogeneous and influential populations in the TME— T cells and CAFs. Our analysis revealed significant functional subtype differentiation and state heterogeneity in both groups, further confirming the high complexity of the TME at the cellular and molecular levels (50). Notably, $CD8^+$ T cells exhibit functionally depleted phenotypes with significantly suppressed effector functions and cytotoxicity (51). Concurrently, the proportion of iCAF markedly increases, actively reshaping the inflammatory microenvironment of tumors through the secretion of specific cytokines and chemokines (52). These results suggest that T cell dysfunction and iCAF activation may be regulated bidirectionally to exacerbate immunosuppression. Together, these findings indicate that targeting these key cellular populations can effectively reverse the inhibitory state of the TME, providing a new theoretical foundation and potential therapeutic targets for enhancing anti-tumor immunity and improving CRC treatment strategies.

In silico gene perturbation analysis predicted that simulated GDI1 knockout was associated with altered expression of VEGFA and HSPB1, whereas simulated overexpression was associated with changes in TIMP2 and CD63. VEGFA is an endothelial mitogen that drives angiogenesis via promoting endothelial cell proliferation, migration, and vascular permeability (53, 54). VEGFA upregulated is a robust predictor of unfavorable prognosis in CRC, with strong links to late-stage disease, aggressive metastasis, and reduced patient survival (55). HSPB1 polymorphism (rs2070804) is associated with primary tumor depth but not with invasiveness, recurrence, or metastasis in CRC (56). TIMP2 is a transcriptional repressor that plays a critical role in proliferation, differentiation, matrix remodeling, and vascular development (57), and has been implicated in CRC (58). CD63, a member of the tetraspanin family, has been independently associated with adverse clinicopathological features and poor prognosis in CRC, suggesting its potential as a novel prognostic biomarker (59). GDI1, acting as a Rab GTPase dissociation inhibitor, was found to indirectly regulate downstream DRGs via Rab-dependent transcriptional signaling. Functionally, GDI1 overexpression in CRC CAFs produces stronger pathogenic effects than knockout, and

its downstream DRGs are enriched in angiogenesis and matrix remodeling pathways associated with poor prognosis, supporting a functional role for GDI1 in CAF malignant transformation.

Furthermore, hdWGCNA delineated ten CAF–related modules, with module 8 and module 10 representing two functionally divergent subtypes characterized by opposing immunomodulatory activities. This modular framework advances our understanding of CAF heterogeneity in CRC, and pinpoints candidate mediators of CAF–T cell interactions. Future studies incorporating functional assays and spatial profiling are required to translate these findings into prognostic and therapeutic insights.

ST analysis revealed that malignant epithelial cells act as central signaling molecules within the CRC microenvironment, engaging extensively with immune, myeloid, and fibroblast modules to construct a pro-tumor ecosystem (17). Notably, CNV events were observed in epithelial cells in ST and scRNA-seq data, which are associated with the malignant transformation of tumor cells (60). These findings highlight the spatially organized multicellular interactions driving CRC progression and provide a foundation for the development of niche-targeting therapeutic strategies.

This study has several important limitations. First, the ST data are derived from a single patient (one tumor and one adjacent normal sample), which limits the generalizability of spatial conclusions. Second, all cell-cell communication, trajectory inference, and gene perturbation analyses are computational and have not been experimentally validated. Third, the extensive overlap in pathway enrichment across all cell types, including populations with very low cell counts (e.g., mast cells, n = 228), may reflect ambient RNA contamination or imbalanced cell numbers; patient-level pseudobulk analysis would strengthen these findings. Fourth, the prognostic gene signature was developed and tested using the same datasets on which KM pre-selection was performed, and has not been validated in an independent external cohort. Although all eight genes showed significance in univariate survival analysis (Fig. 5C), multivariate analyses adjusting for tumor stage, MSI status, or CMS subtype have not been performed. Fifth, CAF subtypes were defined by NMF decomposition without validation against established CRC CAF classification systems. Sixth, the datasets are restricted to the mucinous colorectal adenocarcinoma subtype, limiting applicability to non-mucinous CRC. Finally, the pseudotime analysis placed CD4 + and CD8 + T cells on a single trajectory, and the inferred ordering reflects transcriptional similarity. Future studies should incorporate larger, multi-patient ST cohorts, independent prognostic validation datasets, multivariate Cox modeling, and orthogonal experimental approaches to test the hypotheses generated by this analysis.

In conclusion, this integrative multi-omics analysis provides a descriptive characterization of multicellular dysregulation in the CRC tumor microenvironment. We identified cell cycle and DNA replication as prominently dysregulated pathways, with the p53 signaling gene set showing transcriptional alterations consistent with TP53 inactivation. An eight-gene prognostic signature was explored using machine learning-based feature selection. In silico perturbation analysis generated hypotheses about potential GDI1 downstream effectors in CAFs, and ST analysis from a single patient provided preliminary spatial context. All findings are computational in nature and require experimental

validation. This study provides a resource and a set of testable hypotheses for future investigation of the CRC microenvironment and biomarker development.

Abbreviations

CAFs
cancer-associated fibroblasts
CNV
copy number variation
COAD
colon adenocarcinoma
CRC
colorectal cancer
DEGs
differentially expressed genes
EMT
epithelial-mesenchymal transition
GSEA
gene set enrichment analysis
hdWGCNA
high-dimensional weighted gene co-expression network analysis
iCAF
inflammatory cancer-associated fibroblast
MAPK
mitogen-activated protein kinase
MCA
mucinous colorectal adenocarcinoma
mesCAF
mesenchymal cancer-associated fibroblast
MHC-I
major histocompatibility complex class I
myCAF
myofibroblastic cancer-associated fibroblast
NK
natural killer
NSCLC
non-small cell lung cancer
PCA
principal component analysis
READ

rectum adenocarcinoma
scRNA-seq
single-cell RNA sequencing
SNPs
single nucleotide polymorphisms
ST
spatial transcriptomics
TCGA
The Cancer Genome Atlas
TME
tumor microenvironment
UMAP
uniform manifold approximation and projection.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Guangxi Natural Science Foundation(2024GXNSFAA010401),the “Empowerment” Action Plan (Guangxi Key Research and Development Program)[Project Task Contract No.:Guike FN2600640166],Guangxi Medical and health key discipline construction project,Guangxi Medical and health key cultivation discipline construction project,Guangxi Key Laboratory of Extremely Weak Magnetic Field in Cancer Medicine,University Engineering Research Center of Oncolytic and Nanosystem Development,Guangxi Medical University First-class Disciplines Innovation-Driven Talent Program.

Author Contribution

All authors contributed to the study conception and design. Y.L., W.L., and X.L. conceived and designed the experiments;Y.L.,W.L.,X.L.,J.S.,Y.P.,and G.Y. collected, analyzed and interpreted the data; Y.L., W.L., and X.L. wrote original draft; S.L.,Y.L., and Y.H. reviewed and edited draft. All authors have read and agreed to the published version of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable

Data Availability

All datasets analyzed in this study are publicly available. The single-cell RNA sequencing dataset GSE236696 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE236696), the spatial transcriptomics dataset GSE236697 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE236697), and the microarray gene-expression dataset GSE39582 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE39582) were obtained from the National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO). RNA sequencing, clinical and survival, and somatic mutation data were obtained from The Cancer Genome Atlas colon adenocarcinoma (TCGA-COAD; https://portal.gdc.cancer.gov/projects/TCGA-COAD) and rectum adenocarcinoma (TCGA-READ; https://portal.gdc.cancer.gov/projects/TCGA-READ) projects through the National Cancer Institute Genomic Data Commons (NCI GDC) Data Portal. No new primary sequencing data were generated in this study.

References

1. Biller LH, Schrag D. Diagnosis and Treatment of Metastatic Colorectal Cancer: A Review. JAMA. 2021;325(7):669–85.
2. Siegel RL, Miller KD, Goding Sauer A, Fedewa SA, Butterly LF, Anderson JC, et al. Colorectal cancer statistics, 2020. CA Cancer J Clin. 2020;70(3):145–64.
3. Wang W, Kandimalla R, Huang H, Zhu L, Li Y, Gao F, et al. Molecular subtyping of colorectal cancer: Recent progress, new challenges and emerging opportunities. Semin Cancer Biol. 2019;55:37–52.
4. Okita A, Takahashi S, Ouchi K, Inoue M, Watanabe M, Endo M, et al. Consensus molecular subtypes classification of colorectal cancer as a predictive factor for chemotherapeutic efficacy against metastatic colorectal cancer. Oncotarget. 2018;9(27):18698–711.
5. Chan DKH, Buczacki SJA. Tumour heterogeneity and evolutionary dynamics in colorectal cancer. Oncogenesis. 2021;10(7):53.
6. Combes AJ, Samad B, Tsui J, Chew NW, Yan P, Reeder GC, et al. Discovering dominant tumor immune archetypes in a pan-cancer census. Cell. 2022;185(1):184–e20319.

7. Cheng S, Li Z, Gao R, Xing B, Gao Y, Yang Y, et al. A pan-cancer single-cell transcriptional atlas of tumor infiltrating myeloid cells. Cell. 2021;184(3):792–809. e23.
8. Wang D, Liu B, Zhang Z. Accelerating the understanding of cancer biology through the lens of genomics. Cell. 2023;186(8):1755–71.
9. Ren X, Zhang L, Zhang Y, Li Z, Siemers N, Zhang Z. Insights Gained from Single-Cell Analysis of Immune Cells in the Tumor Microenvironment. Annu Rev Immunol. 2021;39:583–609.
10. Luca BA, Steen CB, Matusiak M, Azizi A, Varma S, Zhu C, et al. Atlas of clinically distinct cell states and ecosystems across human solid tumors. Cell. 2021;184(21):5482–e9628.
11. Hinshaw DC, Shevde LA. The Tumor Microenvironment Innately Modulates Cancer Progression. Cancer Res. 2019;79(18):4557–66.
12. Schulz M, Salamero-Boix A, Niesel K, Alekseeva T, Sevenich L. Microenvironmental Regulation of Tumor Progression and Therapeutic Response in Brain Metastasis. Front Immunol. 2019;10:1713.
13. Wang XQ, Danenberg E, Huang CS, Egle D, Callari M, Bermejo B, et al. Spatial predictors of immunotherapy response in triple-negative breast cancer. Nature. 2023;621(7980):868–76.
14. Bi K, He MX, Bakouny Z, Kanodia A, Napolitano S, Wu J, et al. Tumor and immune reprogramming during immunotherapy in advanced renal cell carcinoma. Cancer Cell. 2021;39(5):649–61. e5.
15. Wu T, Zhang X, Liu X, Cai X, Shen T, Pan D, et al. Single-cell sequencing reveals the immune microenvironment landscape related to anti-PD-1 resistance in metastatic colorectal cancer with high microsatellite instability. BMC Med. 2023;21(1):161.
16. Picard E, Verschoor CP, Ma GW, Pawelec G. Relationships Between Immune Landscapes, Genetic Subtypes and Responses to Immunotherapy in Colorectal Cancer. Front Immunol. 2020;11:369.
17. Zhou H, Shen Y, Zheng G, Zhang B, Wang A, Zhang J, et al. Integrating single-cell and spatial analysis reveals MUC1-mediated cellular crosstalk in mucinous colorectal adenocarcinoma. Clin Transl Med. 2024;14(5):e1701.
18. Marisa L, de Reynies A, Duval A, Selves J, Gaub MP, Vescovo L, et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS Med. 2013;10(5):e1001453.
19. Tomczak K, Czerwinska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn). 2015;19(1A):A68–77.
20. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
21. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16(5):284–7.
22. Liu Z, Liu L, Weng S, Guo C, Dang Q, Xu H, et al. Machine learning-based integration develops an immune-derived lncRNA signature for improving outcomes in colorectal cancer. Nat Commun. 2022;13(1):816.

23. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019;177(7):1888–902. e21.
24. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999;401(6755):788–91.
25. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36(5):411–20.
26. Jin S, Guerrero-Juarez CF, Zhang L, Chang I, Ramos R, Kuan CH, et al. Inference and analysis of cell-cell communication using CellChat. Nat Commun. 2021;12(1):1088.
27. Osorio D, Zhong Y, Li G, Xu Q, Yang Y, Tian Y, et al. scTenifoldKnk: An efficient virtual knockout tool for gene function predictions via single-cell gene regulatory network perturbation. Patterns (N Y). 2022;3(3):100434.
28. Morabito S, Reese F, Rahimzadeh N, Miyoshi E, Swarup V. hdWGCNA identifies co-expression networks in high-dimensional transcriptomics data. Cell Rep Methods. 2023;3(6):100498.
29. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44(W1):W90–7.
30. Valdeolivas A, Amberg B, Giroud N, Richardson M, Galvez EJC, Badillo S, et al. Profiling the heterogeneity of colorectal cancer consensus molecular subtypes using spatial transcriptomics. NPJ Precis Oncol. 2024;8(1):10.
31. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16(12):1289–96.
32. Jin S, Plikus MV, Nie Q. CellChat for systematic analysis of cell-cell communication from single-cell transcriptomics. Nat Protoc. 2025;20(1):180–219.
33. Herranz-Montoya I, Angulo-Aguado M, Perna C, Zagorac S, Garcia-Jimeno L, Park S, Djouder N. p53 protein degradation redefines the initiation mechanisms and drives transitional mutations in colorectal cancer. Nat Commun. 2025;16(1):3934.
34. Leguisamo NM, Gloria HC, Kalil AN, Martins TV, Azambuja DB, Meira LB, Saffi J. Base excision repair imbalance in colorectal cancer has prognostic value and modulates response to chemotherapy. Oncotarget. 2017;8(33):54199–214.
35. Yu H, Jiang W, Chen G, Yang F, Zhao X, Ji Z, et al. Impact of Colon-Specific DNA Methylation-Regulated Gene Modules on Colorectal Cancer Patient Survival. Med Sci Monit. 2019;25:3549–57.
36. Yang W, Liu S, Mao M, Gong Y, Li X, Lei T, et al. T-cell infiltration and its regulatory mechanisms in cancers: insights at single-cell resolution. J Exp Clin Cancer Res. 2024;43(1):38.
37. Deng L, Jiang N, Zeng J, Wang Y, Cui H. The Versatile Roles of Cancer-Associated Fibroblasts in Colorectal Cancer and Therapeutic Implications. Front Cell Dev Biol. 2021;9:733270.
38. Peng Z, Ren Z, Tong Z, Zhu Y, Zhu Y, Hu K. Interactions between MFAP5 + fibroblasts and tumor-infiltrating myeloid cells shape the malignant microenvironment of colorectal cancer. J Transl Med.

2023;21(1):405.
39. Zhang J, Lu T, Lu S, Ma S, Han D, Zhang K, et al. Single-cell analysis of multiple cancer types reveals differences in endothelial cells between tumors and normal tissues. Comput Struct Biotechnol J. 2023;21:665–76.
40. Muller MP, Goody RS. Molecular control of Rab activity by GEFs, GAPs and GDI. Small GTPases. 2018;9(1–2):5–21.
41. Xie X, Lin H, Zhang X, Song P, He X, Zhong J, Shi J. Overexpression of GDP dissociation inhibitor 1 gene associates with the invasiveness and poor outcomes of colorectal cancer. Bioengineered. 2021;12(1):5595–606.
42. Li D, Fan S, Yu F, Zhu X, Song Y, Ye M, et al. FOXD1 Promotes Cell Growth and Metastasis by Activation of Vimentin in NSCLC. Cell Physiol Biochem. 2018;51(6):2716–31.
43. Pan F, Li M, Chen W. FOXD1 predicts prognosis of colorectal cancer patients and promotes colorectal cancer progression via the ERK 1/2 pathway. Am J Transl Res. 2018;10(5):1522–30.
44. Chang Z, Liu X, Zhao W, Xu Y. Identification and Characterization of the Copy Number Dosage-Sensitive Genes in Colorectal Cancer. Mol Ther Methods Clin Dev. 2020;18:501–10.
45. Chen M, Song X, Lin H, Cheng W. Prognostic value of RNF113A shows a correlation with immune infiltrates in colorectal cancer. Am J Transl Res. 2024;16(4):1281–94.
46. Chida K, Oshi M, Roy AM, Sato T, Endo I, Takabe K. Pancreatic ductal adenocarcinoma with a high expression of alcohol dehydrogenase 1B is associated with less aggressive features and a favorable prognosis. Am J Cancer Res. 2023;13(8):3638–49.
47. Lee DW, Ji YB, Song CM, Kim JK, Lee SH, Tae K. Impact of Alcohol Dehydrogenase 7 Polymorphism and Alcohol Consumption on Risk of Head and Neck Squamous Cell Carcinoma: A Korean Case-Control Study. J Clin Med. 2023;12(14).
48. Wu Y, Peng H, Chen G, Tu Y, Yu X. Bulk and single-cell RNA sequencing identify prognostic signatures related to FGFBP2(+) NK cell in hepatocellular carcinoma. PeerJ. 2025;13:e19337.
49. Wu LW, Hu X. SSNA1 Promotes Hepatocellular Carcinoma Metastasis Via STAT3/EMT Induction. Anticancer Res. 2023;43(8):3479–86.
50. Gao Y, Luo C, Yang H, Xie Q, He H, Li J, Miao J. Enhanced efficacy of dual chimeric antigen receptor-T cells targeting programmed death-ligand 1 and cancer-associated fibroblasts in colorectal cancer in vitro. Cytojournal. 2025;22:29.
51. Liu Y, Zhang Q, Xing B, Luo N, Gao R, Yu K, et al. Immune phenotypic linkage between colorectal cancer and liver metastasis. Cancer Cell. 2022;40(4):424–37. e5.
52. Shin S, Kim HH, Kim JW, Rim D, An C, Chung YJ, Lee SH. Cellular responses to neoadjuvant FOLFOX6-bevacizumab treatment in colorectal cancers analyzed by single-cell transcriptome analysis. Pathol Res Pract. 2024;263:155681.
53. Leung DW, Cachianes G, Kuang WJ, Goeddel DV, Ferrara N. Vascular endothelial growth factor is a secreted angiogenic mitogen. Science. 1989;246(4935):1306–9.

54. Tischer E, Gospodarowicz D, Mitchell R, Silva M, Schilling J, Lau K, et al. Vascular endothelial growth factor: a new member of the platelet-derived growth factor gene family. Biochem Biophys Res Commun. 1989;165(3):1198–206.
55. Eswarappa SM, Potdar AA, Koch WJ, Fan Y, Vasu K, Lindner D, et al. Programmed translational readthrough generates antiangiogenic VEGF-Ax. Cell. 2014;157(7):1605–18.
56. Hung CS, Huang CY, Hsu YW, Makondi PT, Chang WC, Chang YJ, et al. HSPB1 rs2070804 polymorphism is associated with the depth of primary tumor. J Cell Biochem. 2020;121(1):63–9.
57. Peeney D, Liu Y, Lazaroff C, Gurung S, Stetler-Stevenson WG. Unravelling the distinct biological functions and potential therapeutic applications of TIMP2 in cancer. Carcinogenesis. 2022;43(5):405–18.
58. Sun J, Zhao J, Jiang F, Wang L, Xiao Q, Han F, et al. Identification of novel protein biomarkers and drug targets for colorectal cancer by integrating human plasma proteome with genome. Genome Med. 2023;15(1):75.
59. Kaprio T, Hagstrom J, Andersson LC, Haglund C. Tetraspanin CD63 independently predicts poor prognosis in colorectal cancer. Histol Histopathol. 2020;35(8):887–92.
60. Sha Z, Gao Q, Wang L, An N, Wu Y, Wei D, et al. Investigating the Cell Origin and Liver Metastasis Factors of Colorectal Cancer by Single-Cell Transcriptome Analysis. Onco Targets Ther. 2024;17:345–58.

Figures

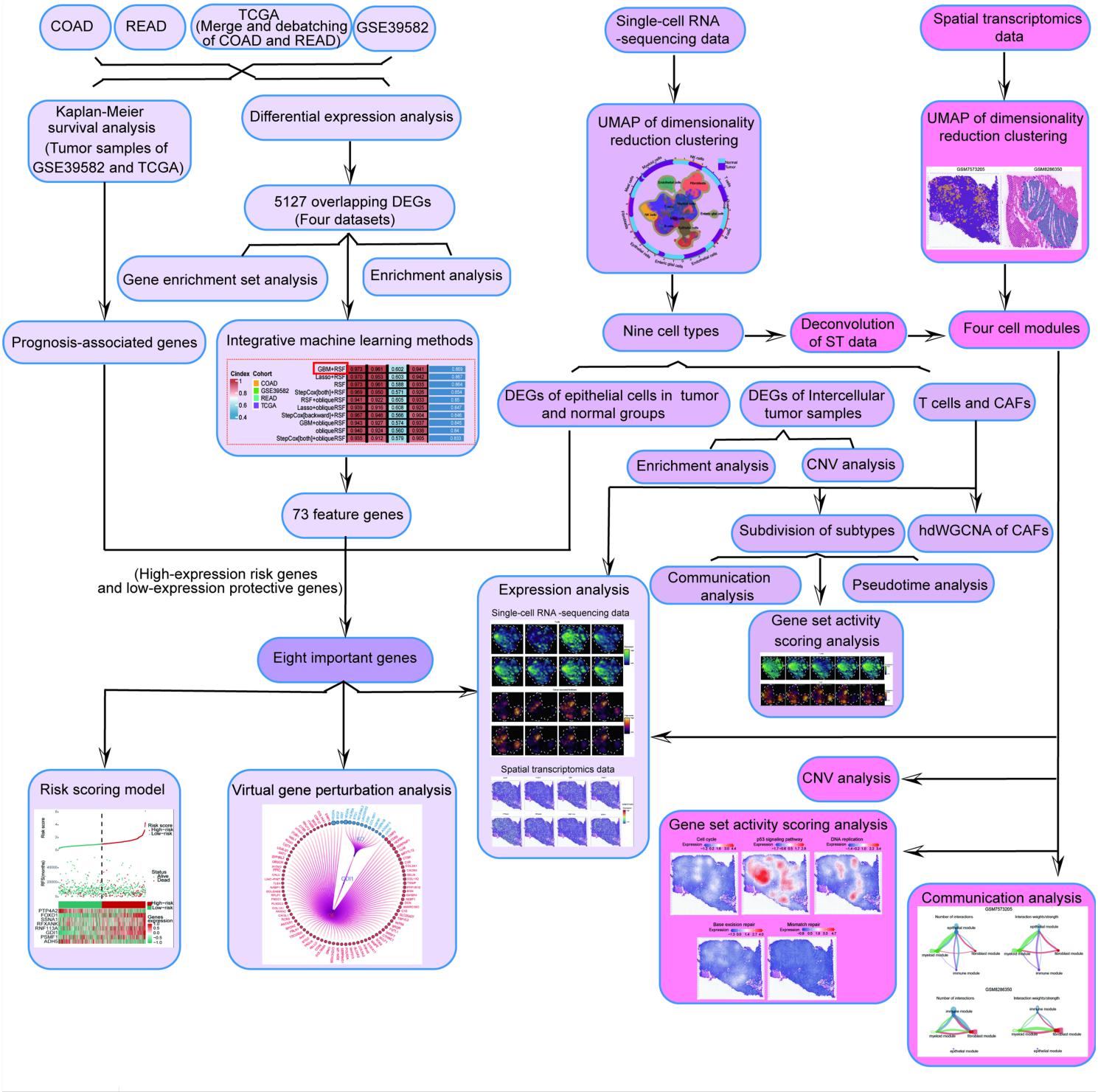


Figure 1

Study workflow

Abbreviations: CAF, cancer associate fibroblast; CNV, chromosomal copy number variations; COAD, colon adenocarcinoma; DEGs, differentially expressed genes; hdWGCNA, high-dimensional weighted gene co-expression network analysis; READ, rectum adenocarcinoma; UMAP, uniform manifold approximation and projection; TCGA, The Cancer Genome Atlas database

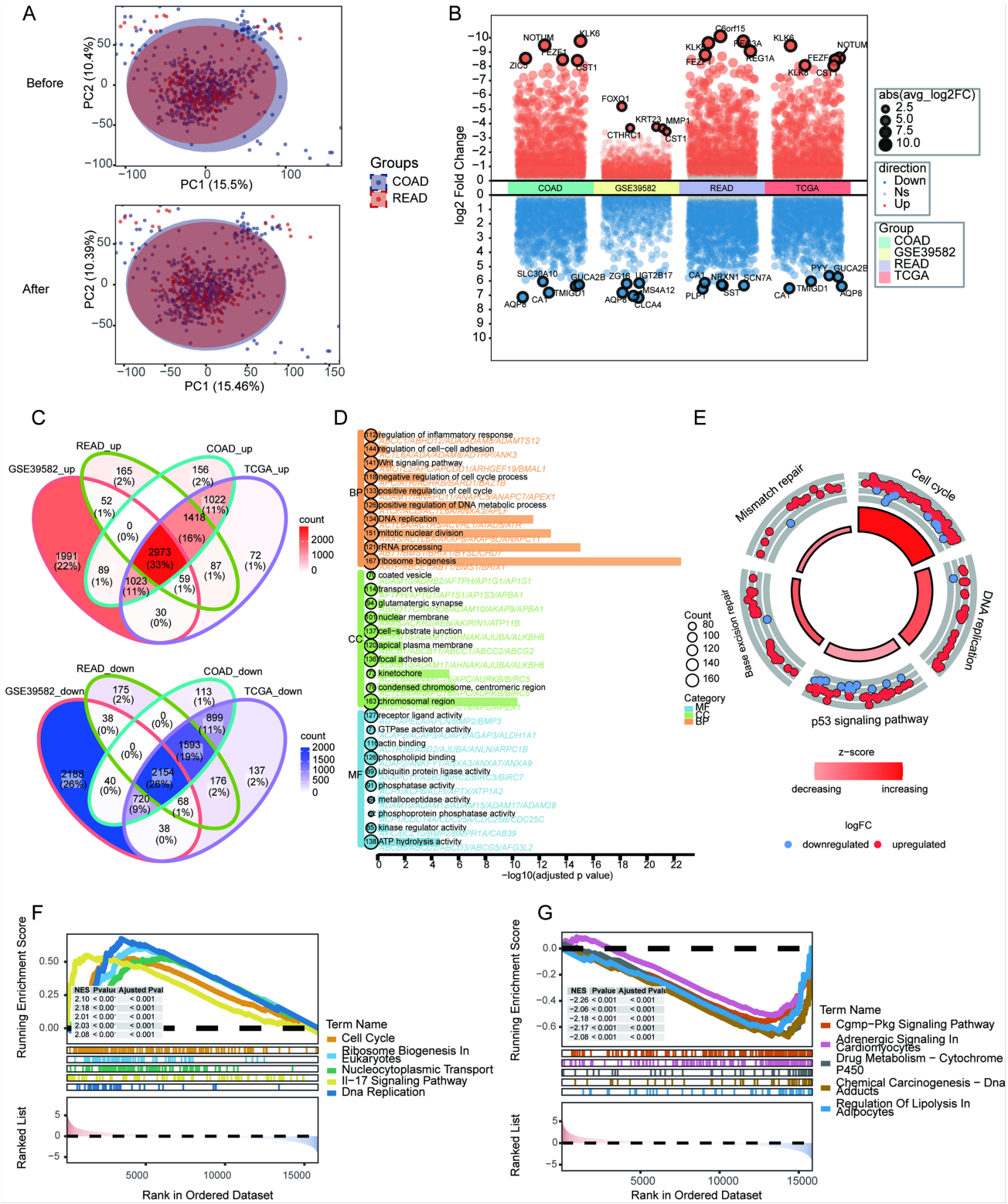


Figure 2

Enrichment analyses were performed on the overlapping DEGs in samples from GSE39582 and COAD and READ samples from TCGA

(**A**) Principal component analysis between COAD and READ samples from TCGA. (**B**) Differential expression analysis of different colorectal cancer data sets. (**C**) Up and down regulated overlapping

differentially expressed genes between GSE39582, COAD and READ samples from TCGA. (**D**) Gene ontology analysis. (**E**) Kyoto Encyclopedia of Genes and Genomes analysis. (**F-G**) Gene set enrichment analysis was performed on the overlapping DEGs

Abbreviations: BP, biological process; CC, cellular component; COAD, colon adenocarcinoma; differentially expressed genes, DEGs; MF, molecular function; NES, normalized enrichment score; READ, rectal adenocarcinoma; TCGA, The Cancer Genome Atlas database

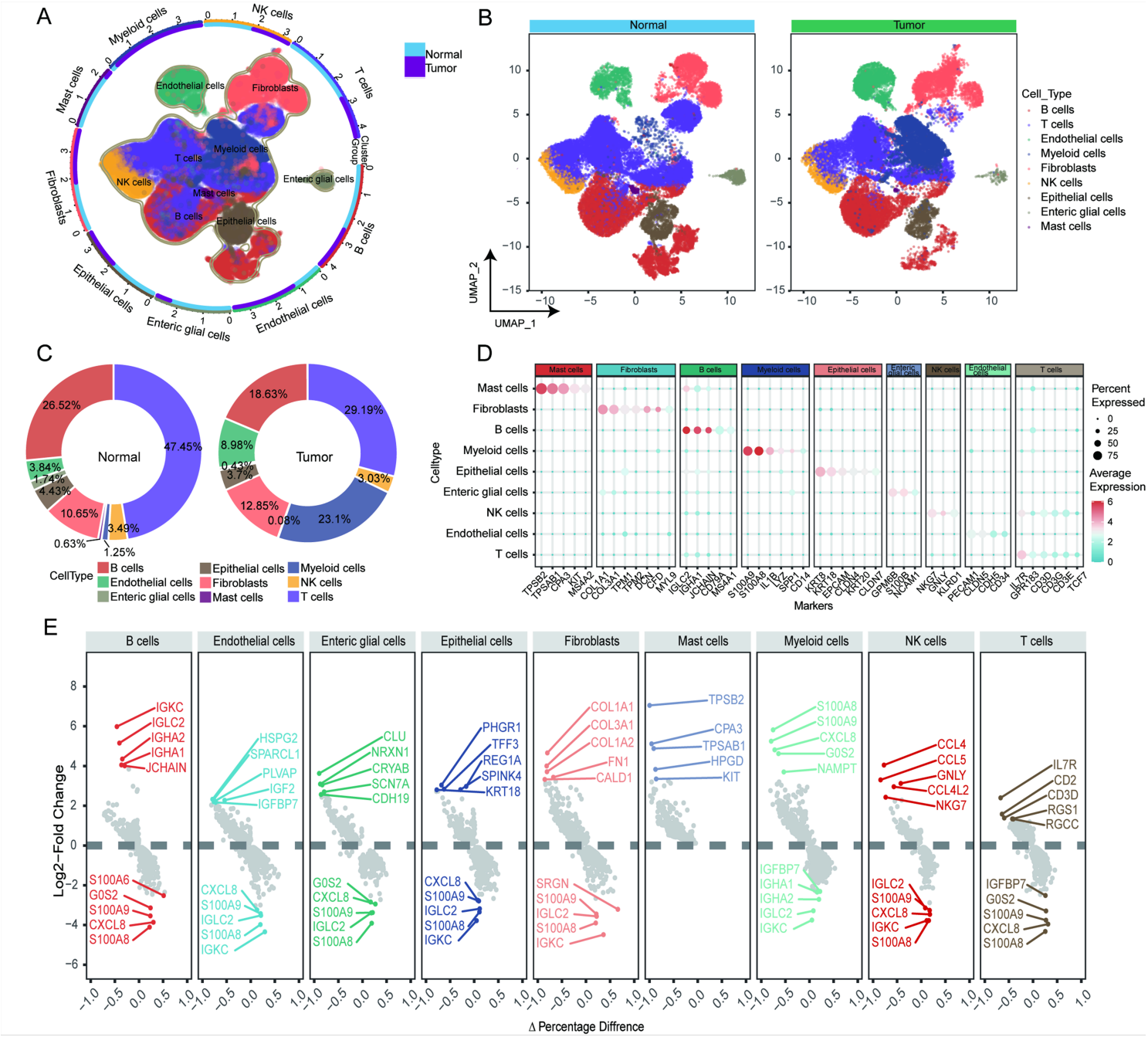


Figure 3

Landscape of cell types in CRC and normal tissues

(**A**) UMAP plot of nine major cell types of CRC. (**B**) The distribution of nine cell types in CRC and normal groups. (**C**) The proportions of cell types are shown in normal and CRC circular diagram. (**D**) Expression of known marker genes in different cell types. The size of each bubble represents the proportion of expression, whereas the color (Gradating from blue to red) indicates the expression level. (**E**) Differentially expressed genes between CRC and normal samples in different cell types.

Abbreviations: colorectal cancer, CRC; NK, natural killer cell; UMAP, Uniform Manifold Approximation and Projection

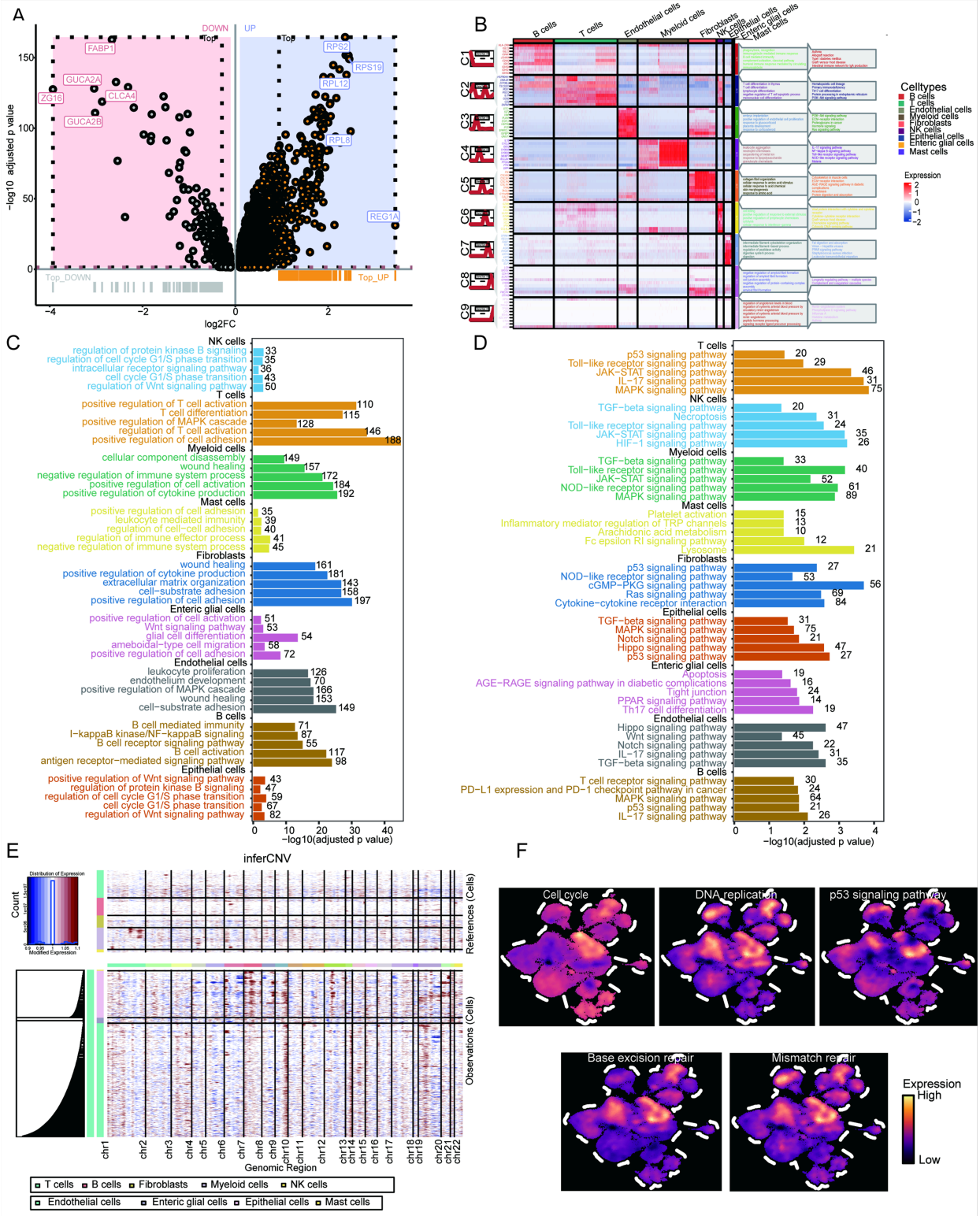


Figure 4

Cell type specific molecular landscapes in CRC

(**A**) Differential expression analysis of epithelial cells between CRC and normal groups. (**B**) Heatmap shows expression of the top 10 genes in the nine cell types. (**C**) Biological processes of the nine cell types. (**D**) Kyoto Encyclopedia of Genes and Genomes analysis of the nine cell types. (**E**) Heatmap

showing chromosomal copy number variations patterns in nine major cell types across CRC samples. (**F**) Expression of the five key pathways. Yellow indicates the high expression, and black indicates the low expression.

Abbreviations: colorectal cancer, CRC

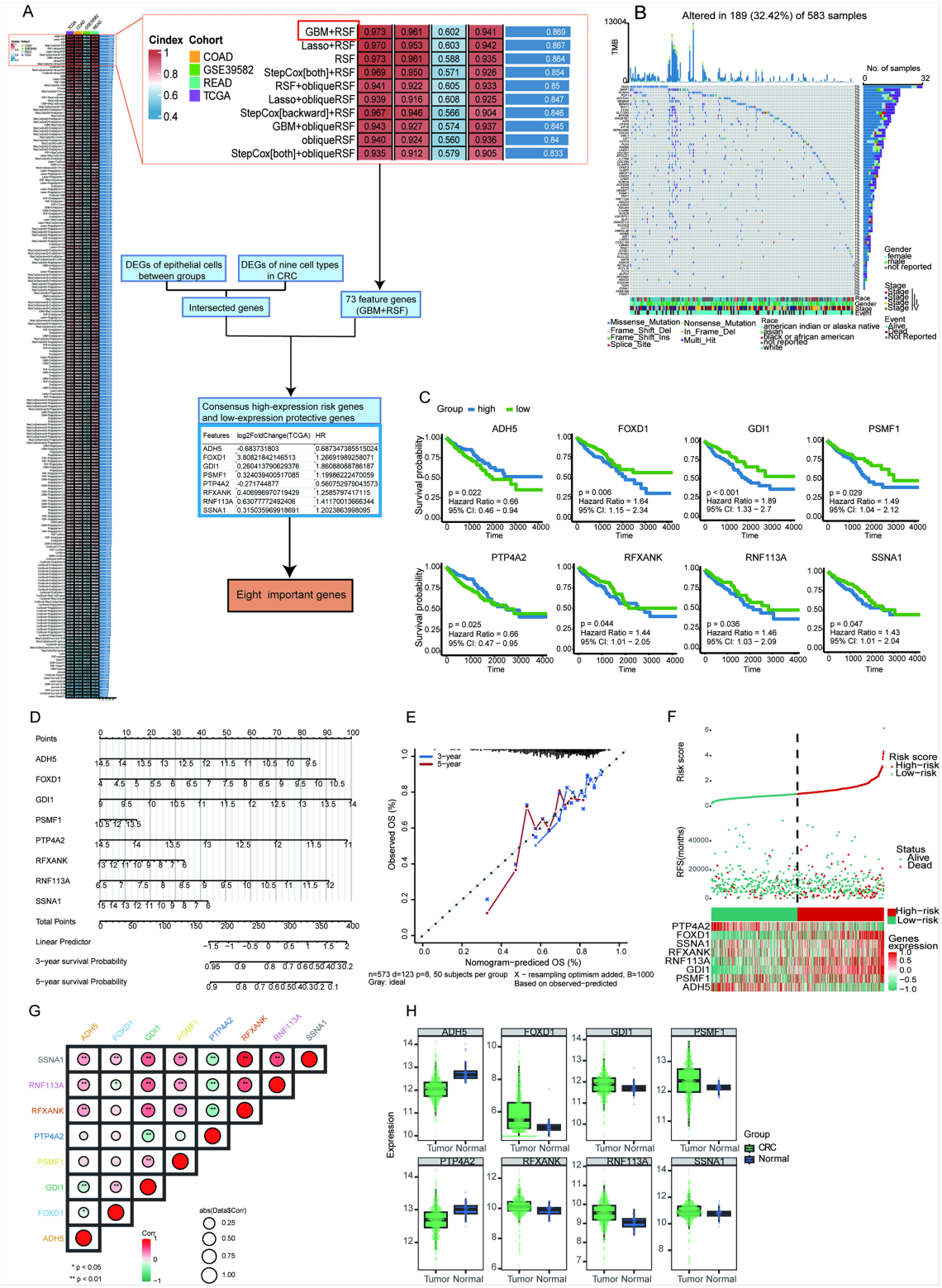


Figure 5

Eight key genes associated with prognosis were identified in CRC

(A) Identification of eight important genes. (B) Single nucleotide polymorphism analysis of 583 CRC samples. (C) Kaplan–Meier plot of eight important genes in the low- or high-risk group. (D) Nomogram the eight important genes. (E) The calibration curve of the constructed Nomogram of 3-year, and 5-year survival. (F) The risk score for each patient with CRC, their corresponding survival outcomes, and a heatmap of eight important genes stratified by high- and low-risk groups. (G) Correlation between the eight important genes. Red shows a positive correlation, green shows a negative correlation. * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$ (H) Expression levels of eight important genes in tumor (Green) and normal (Blue) groups.

Abbreviations: COAD, colon adenocarcinoma; CRC, colorectal cancer; DEGs, differentially expressed genes; READ, rectum adenocarcinoma; TCGA, The Cancer Genome Atlas database

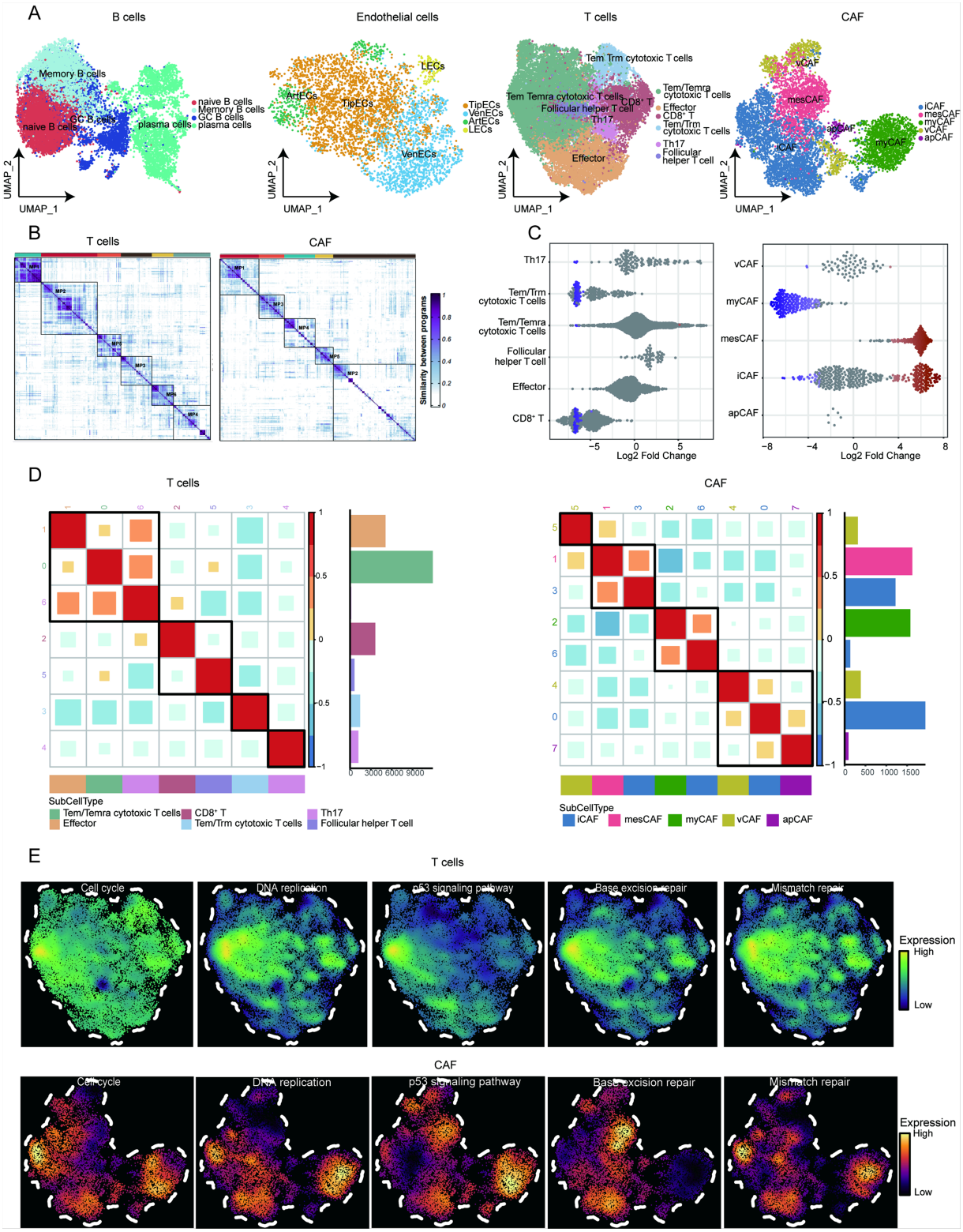


Figure 6

Cellular landscape and expression of five key pathways in cell subtypes of colorectal cancer

(**A**) The distribution of four cell subtypes of B and endothelial cells, six cell subtypes of T cells and five cell subtypes of CAFs by UMAP plot. (**B**) Non-negative matrix factorization of T cells and CAFs. (**C**) Differential expression of cell subtypes. Red indicates high expression, and purple indicates low

expression. (**D**) Correlation analysis between cell subtypes. Red indicates positive correlation, and blue indicates negative correlation. (**E**) Gene set activity scoring analysis was quantified the activity of five key pathways in T cells and CAFs.

Abbreviations: CAFs, cancer-associated fibroblasts; EC, endothelial cells; UMAP, Uniform Manifold Approximation and Projection

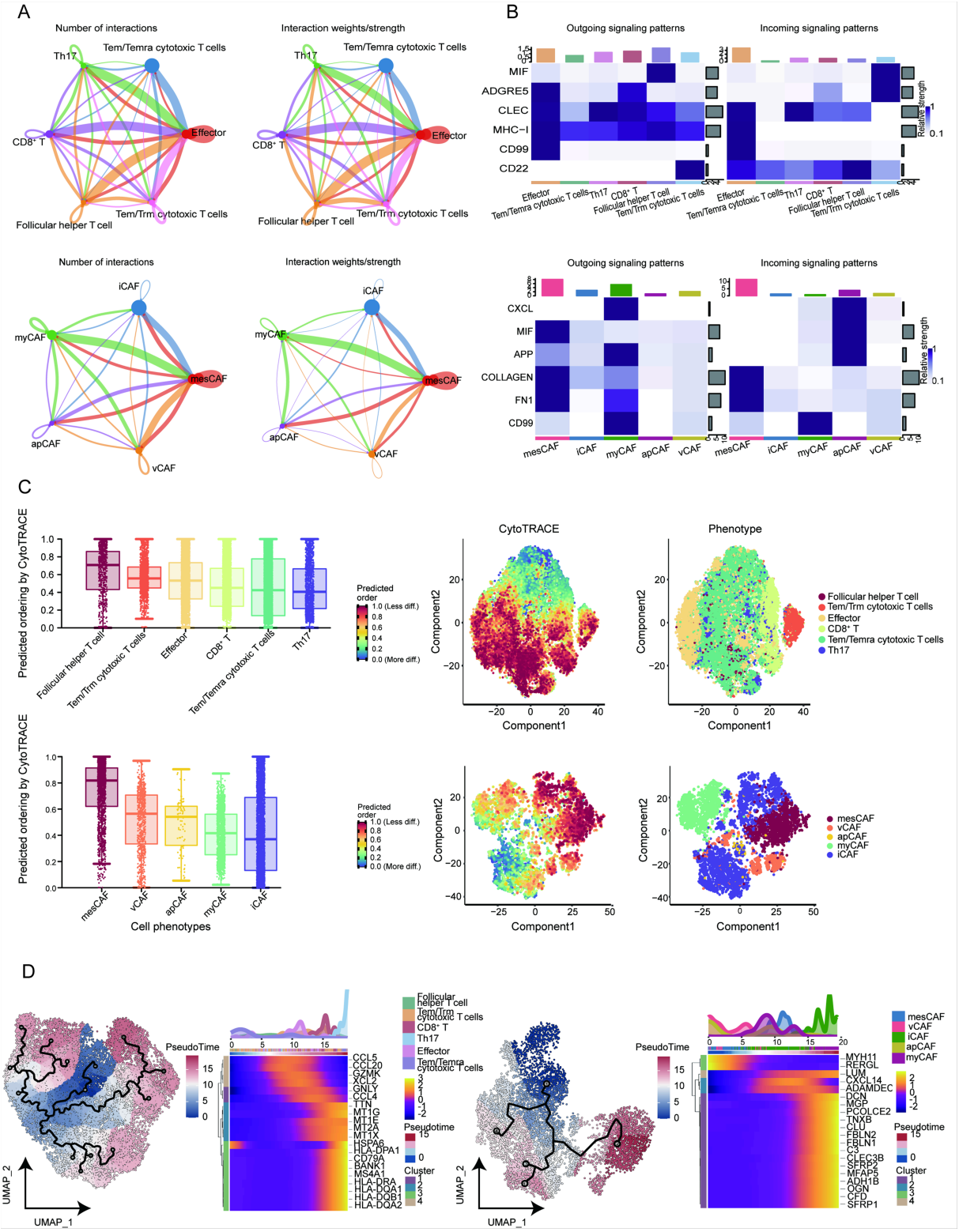


Figure 7

Cell communication and pseudotime trajectory analyses of T cells and CAFs

(**A**) Communication between cell subtypes. The thicker the line, the stronger the correlation. (**B**) Heatmaps of the overall (comprising both outgoing and incoming) signaling flows of each cell subtype mediated by individual signaling axes in CRC sample. (**C**) Cell subtypes of T cells and CAFs with the differentiation state. (**D**) Pseudotime trajectory analyses of T cell and CAF subtypes, and the top 20 genes according to Moran's absolute value were displayed in a heat map.

Abbreviations: CAF, cancer-associated fibroblast; CRC, colorectal cancer; UMAP, Uniform Manifold Approximation and Projection

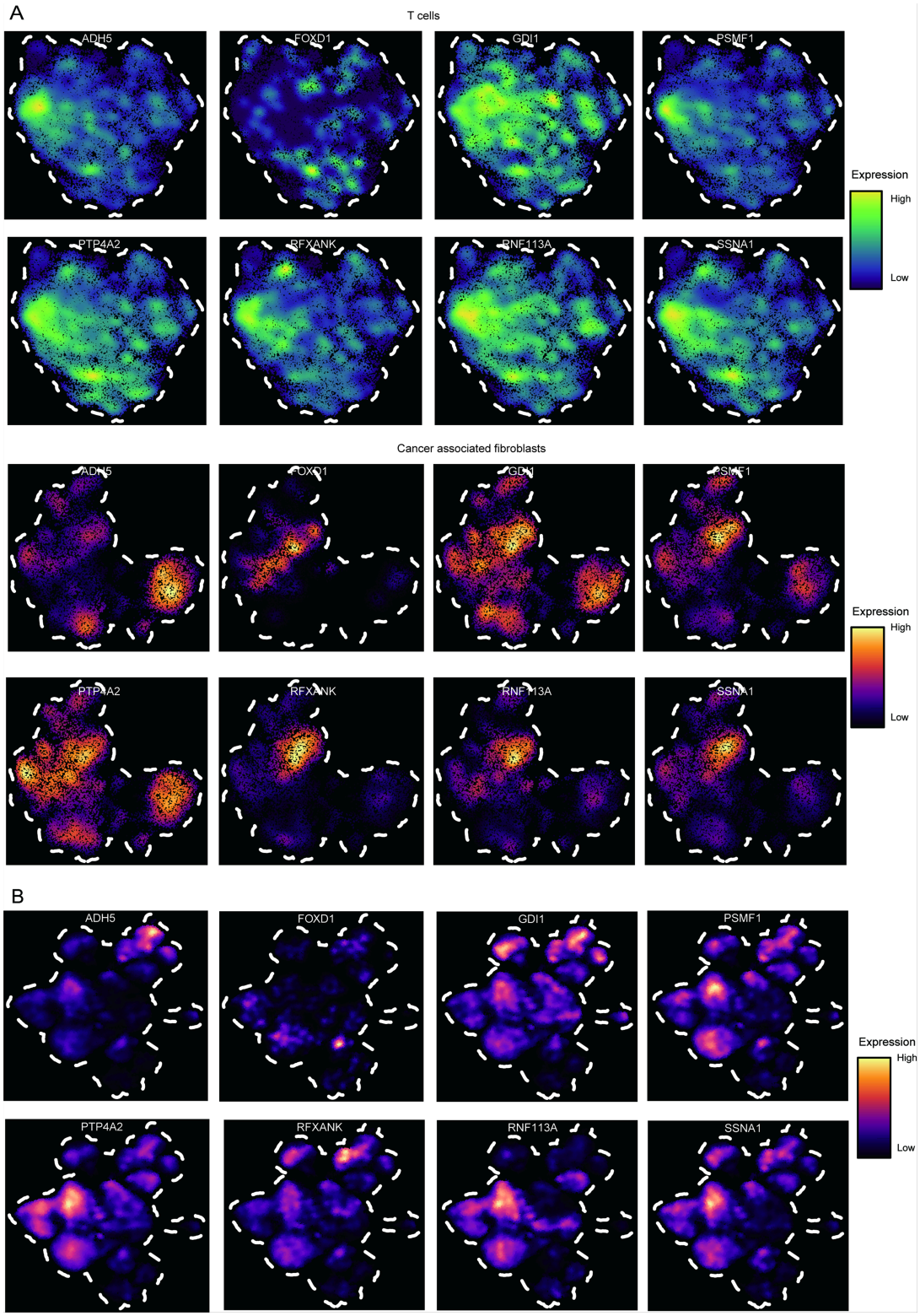


Figure 8

Expression of eight key genes in different cell atlas

Yellow indicates high expression, black indicates low expression.

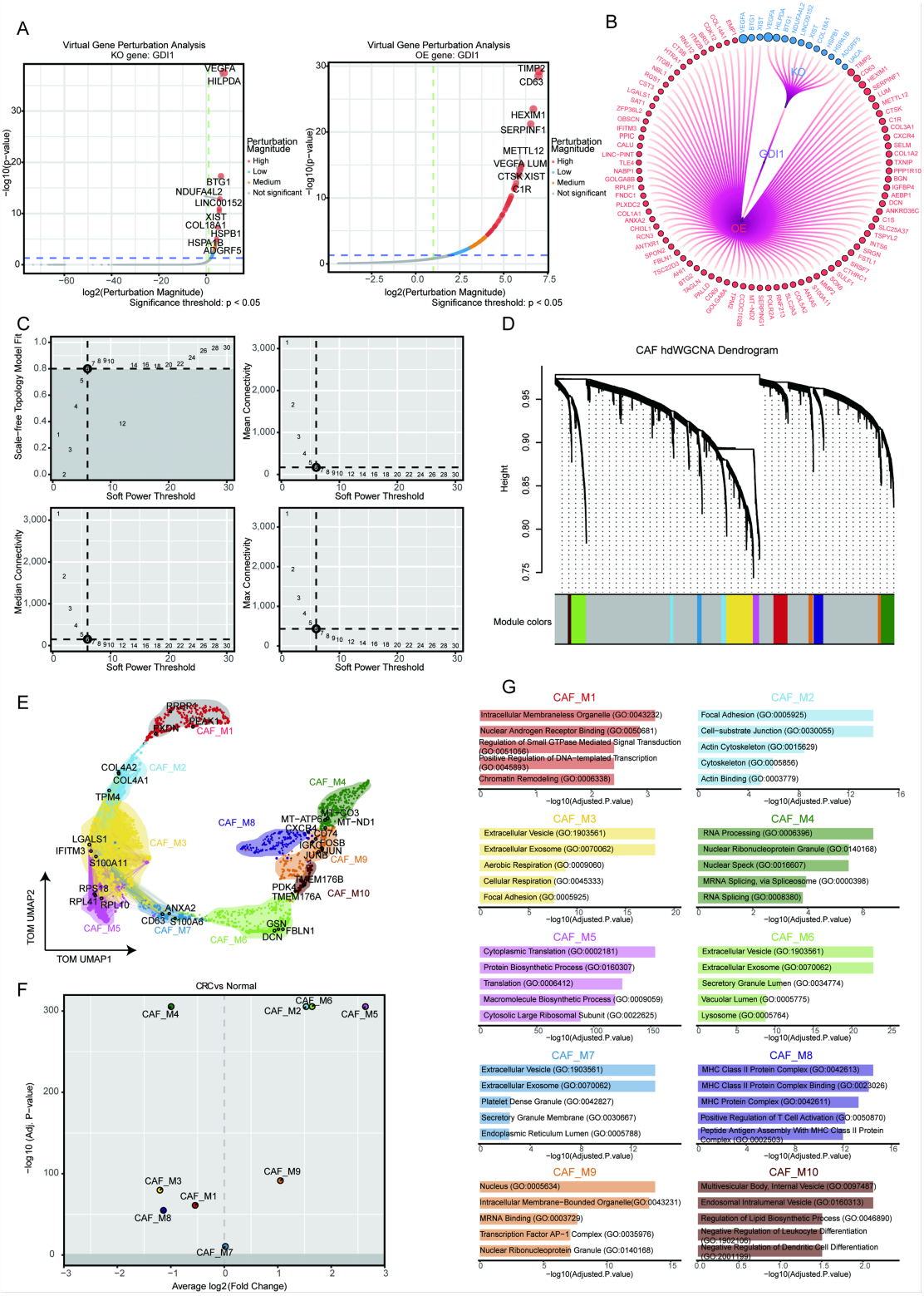


Figure 9

Virtual gene perturbation analysis and high-dimensional weighted gene co-expression network analysis of cancer-associated fibroblast (CAF). (A) Virtual gene perturbation analysis of GDI1. (B) Network chart showed the differentially regulated genes of GDI1. (C) Soft-thresholding power selection for scale-free network. (D) Dendrogram of ten modules. (E) Uniform Manifold Approximation and Projection of the CAF co-expression network. Nodes represent genes, colored by module, sized by kME. (F) Volcano plot

visualizing the differences among the ten modules in the CRC and normal groups. (G) Bar plots showing Gene Ontology enrichment results for co-expression modules. CRC, hdWGCNA, high-dimensional weighted gene co-expression network analysis.

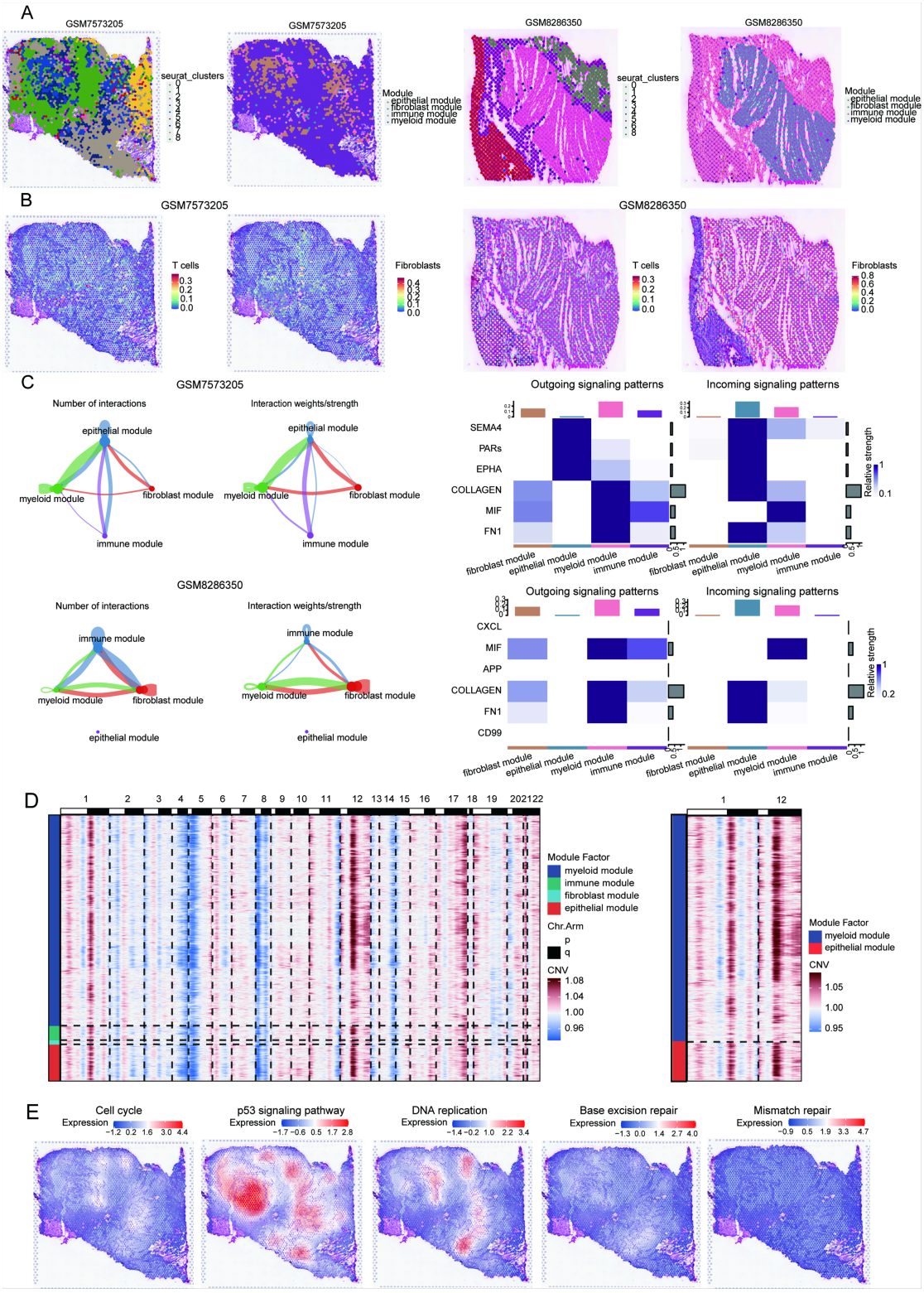


Figure 10

Cell communication and chromosomal copy number variations analyses of spatial transcriptomics data

(**A**) Nine cell clusters were divided into four modules. (**B**) Expression of T cells and CAFs in GSM7573205 and GSM8286350. (**C**) Cell communication analysis between the four modules, and heatmaps of the outgoing and incoming signaling flows of each cell subtypes mediated by individual signaling axes. (**D**) Heat map shows the chromosomal copy number variations patterns in four modules. (**E**) Expression of pathways in spatial transcriptomics data.

Supplementary Files

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

- FS1.tif
- FS2.tif