Establishing and validating a prognostic model for colorectal cancer patients by integrating data from single-cell RNA sequencing and bulk RNA sequencing

Qiujin Huang  
Guizhou Medical University

Dengwei You  
the Affiliated Hospital of Guizhou Medical University

Zhi Huang  
the Affiliated Hospital of Guizhou Medical University

Zhaohui Yin  
the Guizhou Moutai Hospital

Xuya Zhao (✉️ zhaoxuya7@sina.com)  
the Affiliated Hospital of Guizhou Medical University

Research Article

Keywords: scRNA-seq, prognosis, prognostic model, colorectal adenocarcinoma

Posted Date: January 2nd, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3787497/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

Colorectal adenocarcinoma (COAD) is a major global cause of mortality. While conventional RNA sequencing (RNA-seq) has been used to study its prognostic indicators, it lacks precision in identifying cellular alterations. This study aimed to develop a predictive framework for COAD by integrating scRNA-seq with conventional RNA-seq.

Methods

This study acquired primary RNA sequencing data from The Cancer Genome Atlas (TCGA) database and single-cell RNA sequencing data on colorectal adenocarcinoma (COAD) from the Gene Expression Omnibus (GEO) database. The t-SNE method reduced dimensionality and identified clusters. Additionally, Weighted Gene Correlation Network Analysis (WGCNA) identified crucial modules and genes with differential expression (DEGs). Cox regression analysis was utilized to construct the prognostic model and explore mutation profiles and immune statuses across different risk groups.

Results

Integration of scRNA-seq data from four samples revealed 15 distinct clusters covering 8 cell types. Differential analysis identified important cell types, including B cells (Naïve and Plasma cells), Endothelial cells, Epithelial cells, Monocytes, Natural Killer (NK) cells, Smooth muscle cells, and T cells (CD8+). Subsequently, a prognostic model was built using 28 genes showing differential expression, with four DEGs displaying a significant correlation with higher risk scores, poorer survival outcomes, and increased APC mutation rates. Various prognostic and immune characteristics were observed within this context.

Conclusion

Integrating 10x scRNA-seq and bulk RNA-seq data, we established a prognostic framework for colorectal adenocarcinoma (COAD) in this study. Additionally, we identified two distinct groups, each displaying different prognoses and immune features.

MATERIALS AND METHODS

Raw Data Acquisition

Using 10x scRNA-seq, we examined two COAD samples (T1 and T2) and two corresponding normal tissue samples (N1 and N2) to analyze colon adenocarcinoma. These data, sourced from the GSE161277 series, comprised 3190, 4438, 2882, and 4097 cells for each respective sample[5]. Additionally, we acquired bulk RNA sequencing data, mutation profiles, and clinicopathological details for COAD from the TCGA (The Cancer Genome Atlas) database.
scRNA-Seq Data Processing and Analysis

To begin with, the 10x scRNA-seq data was converted into a Seurat object using the 'Seurat' package in R software during the processing of the data. This was followed by a quality control step, where raw counts were assessed. To eliminate cells of poor quality, the calculation of the ratio between mitochondrial or ribosomal genes was performed. Lastly, to identify the top 2000 genes with the most variability, we employed the "FindVariableFeatures" function after quality control. In the processing of 10× scRNA-seq data, the fourth step involved conducting principal component analysis (PCA) on the selected 2000 genes[6]. This was followed by using t-SNE for reducing dimensions and identifying clusters. To identify important marker genes in various clusters, we used the 'Find All Markers' feature as the fifth step. This involved setting a log2 [Foldchange (FC)] threshold of 0.35 and a minimum percentage (min. pct) of 0.15. Finally, the sixth step involved the use of R software's "SingleR" package for the annotation of clusters to distinguish various cell types. Then, let us conduct a differential analysis after annotation using the same parameters. We conducted functional enrichment analysis on the key cell types identified using the "ReactomeGSA" package in R software. As part of the enrichment process, the "analyze_sc_clusters" function was used, followed by the "pathways" function to extract the results. The 'monocle' package in R was utilized for cell trajectory and pseudo-time analysis, employing the 'DDRTree' technique to reduce dimensions. We then utilized the "BEAM" statistical method to assess gene contributions during cell development, focusing on the top 200 genes for visualization purposes. To analyze cell-cell communication and visualize networks, we utilized the R software packages 'CellChat' and 'patchwork'.

Differentially Expressed Genes Identification and Functional Enrichment

Analysis

We performed differential expression analysis using the R software package 'limma' to identify genes that were expressed differently in the TCGA cohort. The criteria set for filtering included an absolute log2 fold change (log2FC) greater than 1.2 and a false discovery rate (FDR) below 0.05. A volcano plot was then created to display the DEGs' distribution. Afterward, we employed the R software's 'clusterProfiler' package to perform analyses on the Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO). The purpose of these analyses was to investigate the pathways and biological processes that are most significantly enriched and linked to the DEGs.

Analysis of gene correlation networks using weights

The "WGCNA" package in R was used to identify key hub genes among differentially expressed genes (DEGs) using the Weighted Gene Correlation Network Analysis (WGCNA) method. WGCNA involves a two-pronged approach: clustering gene expressions and correlating them with phenotypic traits. This analysis is primarily comprised of four stages: calculating gene correlation coefficients, identifying gene modules, constructing a co-expression network, and correlating these modules with specific traits. During the network construction, we chose a soft thresholding power of 12. The modules were then visualized in a dendrogram following clustering. To identify the most important DEGs in the development of colorectal
adenocarcinoma (COAD), a heatmap of module traits was created using correlation coefficients and p-values. Finally, we focused on the overlapping genes between the marker genes and the DEGs identified in the WGCNA for in-depth analysis.

**Construction of the prognostic risk model**

The risk score for each patient was determined using the following equation:

\[ \text{RiskScore} = \sum \beta_i \times \text{EXP}i \]

In this context, ‘i’ denotes the gene expression level, and beta represents the coefficient corresponding to the receptor-ligand pair derived from multivariate Cox regression analysis\[7\]. Patients were divided into high-risk and low-risk categories based on a predetermined threshold, specifically the middle value. Survival curves were generated using the Kaplan-Meier method for prognostic analysis. Furthermore, the log-rank test was utilized to assess the statistical significance of the noted disparities.

**Comparison of clinical significance and mutation patterns in high-risk and low-risk populations**

We then investigated the correlation between the risk score and the clinicopathological characteristics of patients within the TCGA group. Cox regression analysis was conducted using the "survcomp" package in R software to assess if the risk score serves as an independent prognostic indicator for colorectal adenocarcinoma (COAD) patients. At the same time, the forest plot package in R was utilized to generate forest plots that depict the findings from both univariate and multivariate Cox regression analyses. Furthermore, to examine the distinct genetic mutation patterns in the groups at high and low risk, we utilized the 'oncoplot' feature in the 'maftools' package of the R software to create waterfall plots.

**Comparison of immune cell infiltration and immune function status between high and Low-Risk Groups**

Afterward, the cibersort technique was employed to assess the levels of immune cell infiltration and the activity in pathways related to the immune system. The Wilcoxon rank-sum test was used to evaluate the statistical disparities between the high-risk and low-risk groups. Furthermore, the R 'ESTIMATE' package was employed to evaluate the level of immune infiltration in colorectal cancer patients categorized into high and low-risk cohorts. For visualization, the "ggplot2" package in R was employed.

**INTRODUCTION**

Colorectal cancer (CRC), a leading malignancy globally, originates from the epithelial cells of the colon or rectum. Benign polyps often develop into invasive cancers, with the development of colorectal cancer (CRC) being influenced by a combination of genetic, environmental, and lifestyle factors[1]. Risk factors that are acknowledged include family background, eating patterns, age, as well as specific hereditary conditions such as Lynch syndrome and familial adenomatous polyposis[2]. The pathogenesis of CRC typically follows a progression from normal mucosa to adenomatous polyps, leading eventually to
carcinoma. This progression is often paralleled by a series of genetic and epigenetic alterations, including mutations in the APC, KRAS, and TP53 genes. The understanding of these molecular pathways has been instrumental in the development of targeted therapies and personalized treatment approaches.

Diagnostically, CRC screening is crucial and includes methods such as fecal occult blood tests, colonoscopy, and imaging techniques. Early detection greatly improves prognosis, emphasizing the importance of regular screening, especially in high-risk groups.

Therapeutically, CRC treatment is stage-dependent and can involve a combination of surgical resection, chemotherapy, radiotherapy, and increasingly, targeted biological therapies and immunotherapies. Surgical removal of the tumor is the primary treatment modality for localized disease. In contrast, advanced stages often require a multimodal approach, including systemic therapies. The molecular comprehension of CRC has made significant progress lately, resulting in the creation of targeted medications. These drugs, including EGFR and VEGF inhibitors, have demonstrated potential in enhancing outcomes, especially in cases of metastasis. Immunotherapy, especially checkpoint inhibitors, has also emerged as an effective treatment option in certain subtypes of CRC, notably those with high microsatellite instability[3]. The advent of single-cell analysis combined with transcriptomics represents a groundbreaking advancement in cancer research, offering unprecedented insights into the complex biology of tumors. This integrative approach provides a detailed landscape of tumor heterogeneity, enabling researchers to dissect the intricate cellular composition of tumors and the dynamic interactions within the tumor microenvironment.

Single-cell transcriptomics allows for the characterization of individual cells within a tumor, revealing variations in gene expression that are obscured in bulk analyses. This granularity is crucial in cancer, where heterogeneity among cancer cells can drive disease progression, metastasis, and treatment resistance. By profiling individual cells, researchers can identify distinct subpopulations, including rare cancer stem cells or immune cells, which play pivotal roles in cancer biology but are often undetected in bulk analyses. Furthermore, the use of single-cell analysis helps to decipher the intricacies of the tumor microenvironment, encompassing diverse cell populations including immune cells, fibroblasts, and endothelial cells. Understanding the interactions among these cells and the cancer cells is essential for comprehending tumor progression and response to therapies. For instance, the immune landscape within a tumor can predict response to immunotherapies, and single-cell approaches can identify immunosuppressive mechanisms employed by tumors[4].

The future of colorectal cancer (CRC) research is poised for transformation through the application of single-cell analysis. This innovative technique promises to unravel the complex heterogeneity of CRC, providing a more nuanced understanding of the disease at the cellular level. By dissecting the intricate interactions between diverse cell types within the tumor microenvironment, single-cell analysis will offer insights into the mechanisms driving CRC progression, metastasis, and treatment resistance.

RESULTS
scRNA-Seq and Cell Typing of Normal and Colon adenocarcinoma Samples

The single-cell data were derived from two patients with colorectal cancer and two normal individuals, totaling 14,607 cells sourced from the GSE161277 dataset. Post-quality control, the results of these cells are detailed in Figure S1. After conducting principal component analysis (PCA) and t-SNE analysis, we successfully classified the cells into 15 separate clusters. The "SingleR" package was then used to annotate these defined cell clusters. The subgroups of cells that were identified consisted of B cells (Naive and Plasma cells), Endothelial cells, Epithelial cells, Monocytes, Natural Killer (NK) cells, Smooth muscle cells, and T cells (CD8+). Functional enrichment analysis using ReactomeGSA[8] revealed that the cell types identified are primarily engaged in processes such as the binding of Ficolins to repetitive carbohydrate structures on the surface of target cells, COX reactions, and serotonin metabolism. We analyzed the cell path and estimated time for the three cell types that were significantly identified using the 'monocle' software (Fig. 1). Additionally, the probability of cell-cell communication was calculated to examine the cell-cell communication network[9]. The deduction of this network was extended by utilizing particular pathways and interactions between ligands and receptors. Our analysis highlighted the pivotal role of the COLLAGEN signaling pathway in the cell-cell communication network (Fig. 2).

Detection of Genes with Altered Expression in Bulk RNA-Seq Data

Differential expression analysis yielded 8,800 genes as differentially expressed genes (DEGs), with 5,012 being up-regulated and 3,788 down-regulated (Fig. 3). The analysis of Gene Ontology (GO) revealed a notable enrichment of these differentially expressed genes (DEGs) in biological processes such as the formation of ribonucleoprotein complexes and the development of ribosomes. Based on the analysis from the Kyoto Encyclopedia of Genes and Genomes (KEGG), the enrichment was mainly observed in pathways associated with the cell cycle and infection caused by the Human T-cell leukemia virus 1. Following that, an analysis called Weighted Gene Correlation Network Analysis (WGCNA) was performed to identify differentially expressed genes (DEGs) linked to the growth and advancement of colorectal adenocarcinoma (COAD). During the co-expression network's construction, a soft thresholding power of 12 was noted. The pink module emerged as significantly correlated with COAD development, as evidenced by its correlation coefficient and p-value. Finally, a group of 28 genes, representing both marker genes and genes from the WGCNA module, were selected for the creation of an expression matrix to facilitate further investigation (Fig. 4).

Prognostic Model Construction

Initially, we used univariate Cox regression analysis to identify 12 genes with potential prognostic value in our analysis of colorectal adenocarcinoma (COAD) within the TCGA cohort. Subsequently, through LASSO regression analysis, we refined this list to four key DEGs for our final risk model[10]. CD177, CLCA4, CLDN23, and SMPDL3A were identified as independent prognostic DEGs using multivariate Cox
analysis. The risk score formula used was: risk score = (expression level of CD177 * -0.176) + (CLCA4 * -0.142) + (CLDN23 * -0.592) + (SMPDL3A * -0.804). Afterward, the patients were categorized into groups of high and low risk, depending on the risk score value that represented the middle value. The analysis of survival showed that individuals in the high-risk category experienced a lower overall survival (OS) in comparison to those in the low-risk category (Fig. 5).

**Comparison of clinical significance and mutation patterns in high-risk and low-risk populations**

Following this, our inquiry concentrated on the connection between risk scores and clinicopathological (Fig. 6) characteristics, uncovering a link between elevated risk scores and T-Stage condition. The additional examination was performed using both univariate and multivariate Cox regression to evaluate whether risk type could predict prognosis independently in COAD patients, compared to other standard clinicopathological factors. Our findings indicated that risk type indeed acts as an independent prognostic indicator for these patients (Fig. 7). Afterward, a cascade diagram was generated to analyze the particular genetic mutation patterns in the high- and low-risk categories. In both groups, the genes APC, KRAS, and NEB exhibited the highest frequencies of mutation. Significantly, the high-risk group exhibited a greater frequency of mutations in the APC gene in contrast to the low-risk group (Fig. 8).

**Comparing the immune function of high-risk and low-risk groups**

We used the 'estimate' package in R to evaluate patients with colorectal cancer and categorize them into high- and low-risk groups in our predictive model. The high-risk group exhibited a significantly lower immune score in comparison to the low-risk group. Furthermore, we evaluated the presence of 21 different immune cell types in these patients using Cibersort. According to our research, the high-risk group exhibited reduced levels of four types of immune cells, namely Plasma cells, M2 Macrophages, Activated Dendritic cells, and Resting Mast cells, in comparison to the low-risk group[11]. In contrast, the high-risk group showed higher levels of expression in Resting NK cells and M0 Macrophages compared to the low-risk group (Fig. 9).

**Discussion**

It's crucial to emphasize the significant implications this approach has for both tumor treatment and research. Firstly, single-cell analysis in COAD allows for an unprecedented resolution in understanding tumor heterogeneity. Colorectal tumors, like many cancers, are not uniform; they consist of a diverse array of cell types, each with unique genetic expressions. By dissecting these variations at a single-cell level, we gain a deeper insight into the molecular mechanisms driving tumor progression and resistance to therapy. This is particularly valuable in identifying subpopulations of cancer cells that might be resistant to current treatments or are more aggressive, thereby guiding more effective therapeutic strategies. Secondly, the integration of differential gene expression analysis enhances our ability to identify key driver genes and pathways in COAD. By correlating specific gene expression patterns with clinical
outcomes, such as patient survival or response to treatment, we can pinpoint potential therapeutic targets. For example, discovering overexpressed genes in malignant tumor cells can result in the creation of focused treatments targeting these particular genes. Moreover, the use of single-cell sequencing in COAD can also aid in the identification of new biomarkers for early detection and prognosis. As we better understand the gene expression profiles of early-stage tumors and their evolution into more advanced stages, we can develop diagnostic tools that detect COAD at an earlier, more treatable stage. Additionally, the analysis of the tumor microenvironment at a single-cell level provides insights into the complex interactions between cancer cells and immune cells. Understanding these interactions is key to developing and optimizing immunotherapies for COAD. For instance, identifying immune evasion mechanisms employed by tumor cells can guide the development of immune checkpoint inhibitors or other immunomodulatory treatments. In summary, the integration of differential gene expression with single-cell analysis in COAD is a powerful approach that promises to revolutionize our understanding of the disease. It holds the potential to identify novel therapeutic targets, enhance the efficacy of existing treatments, contribute to the development of personalized medicine, and ultimately improve patient outcomes in colorectal cancer. Prognostic models play a crucial role in predicting patient outcomes and guiding clinical decision-making. By integrating various clinical and molecular data, including genetic, epigenetic, and expression profiles, these models can predict disease progression, patient survival, and response to therapy with greater accuracy. This is especially critical in cancers where the heterogeneity of the tumor and its microenvironment significantly influence the disease course and treatment response. With personalized prognostic models, clinicians can tailor treatment plans to individual patients, potentially improving outcomes and minimizing unnecessary treatments. In terms of immune infiltration, its study has revolutionized the field of oncology. The tumor microenvironment, comprising various immune cells, influences tumor progression, metastasis, and response to therapy. By analyzing the types and states of immune cells within tumors, researchers can gain insights into the tumor's behavior and its interaction with the host's immune system. This understanding is crucial for the development of effective immunotherapies, such as checkpoint inhibitors, which have shown remarkable success in treating certain types of cancer.

Furthermore, the study of immune infiltration helps in identifying new therapeutic targets and understanding mechanisms of drug resistance. For example, tumors that evolve mechanisms to evade immune detection might be resistant to certain therapies, prompting the need for alternative strategies.

In conclusion, the development of sophisticated prognostic models and the in-depth study of immune infiltration are indispensable in the realm of tumor research. These approaches facilitate a more nuanced understanding of cancer biology, aid in the creation of more effective treatment strategies, and are integral to the future of personalized oncology.

Conclusion

By merging information from 10× scRNA-seq and bulk RNA-seq, a predictive framework was constructed to forecast the progression of colorectal adenocarcinoma (COAD) in this study. Within this study, we
identified two distinct patient cohorts, each displaying different prognoses and immune characteristics. Notably, a higher risk score, indicative of worse survival outcomes, was also linked to a higher incidence of APC mutations. The potential of this model to function as a valuable biomarker for categorizing the risk of COAD patients and forecasting their reaction to treatments is significant. Future prospective studies are essential to confirm and expand upon these initial findings.

**Declarations**

**Data availability**

The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

**Funding**

This work was supported by science and technology fund project of Guizhou Provincial Health Commission NO. gzwkj2021-179

**Ethical approval**

Not applicable.

**Consent to participate**

Not applicable.

**Consent for publication**

Not applicable.

**Conflict of interest**

The authors declare no conflicts of interest.

**Author information**

Department of anus and intestine surgery, Guizhou Medical University, Guiyang 550002, P.R. China;

Qiujin Huang

Department of Interventional Radiology, the Affiliated Hospital of Guizhou Medical University, Guiyang 550002, P.R. China

Dengwei You
Department of Interventional Radiology, the Affiliated Hospital of Guizhou Medical University, Guiyang 550002, P.R. China

Zhi Huang

Department of anus and intestine surgery the GuiZhou Moutai Hospital, Zunyi 5190100, P.R. China

Zhaohui Yin

Department of Interventional Radiology, Affiliated Cancer Hospital of Guizhou Medical University, Guiyang; 550002, P.R. China

Xuya Zhao

Contributions

QJH designs the study, implements the code, analyzes the results and writes the paper. DWY and ZH implements the code and analyze the results. All authors read and approved the final version. ZHY writes the paper. XYZ designs the study, analyzes the results and writes the paper. All authors read and approved the final manuscripts.

References


Figures
Figure 1

(A-D) Annotation of distinct clusters and identification of cell types in colorectal adenocarcinoma (COAD) using 10x Genomics single-cell RNA sequencing (scRNA-seq) data.
Figure 2

(A) Pseudo temporal analysis. (B) Functional enrichment analysis for the identified hub cell types using the “ReactomeGSA” package.
Figure 3

Analysis of intercellular communication networks and identification of differentially expressed genes in the TCGA cohort. (A-B) Cell-cell communication network. (C) DEGs in TCGA cohort. (D-E) GO and KEGG enrichment analysis of the identified DEGs.
Figure 4

Identification of hub DEGs that participate in COAD development through WGCNA.
Figure 5

Prognostic model establishment for patients with COAD.
Figure 6

The association between risk score and prevalent clinicopathological features in colorectal adenocarcinoma (COAD).
Waterfall charts provide a summary of the gene mutation profiles in groups with high and low-risk.
Figure 8

The correlation between risk score and its implications in clinical practice.
Figure 9

Estimation of immune cell infiltration scores in low- and high-risk groups using Cibersort and estimate analysis.