Integration of scRNA-Seq and Bulk RNA-Seq to analyze the heterogeneity of colorectal cancer immune cells and establish a molecular risk model

Li-Yue Sun1*, Jiao-Jiao Yang2*, Xin-Xin Zeng1, Yu-Ying Jiang1, 1, Ju Shen1, 1, Fang Wang2**, Xu-Hui Zhang2**

1Second Department of Oncology, The Affiliated Guangdong Second Provincial General Hospital of Jinan University, Guangzhou, Guangdong Province 510317, China. 2Department of Molecular Diagnostics, Sun Yat-sen University Cancer Center, Guangzhou, Guangdong Province 510317, China. 3Guangdong Medical University, Zhanjiang, Guangdong Province 524000, China.

*These authors contributed equally to this work and are co-first authors for this paper. **Corresponding to: Fang Wang, Department of Molecular Diagnostics, Sun Yat-sen University Cancer Center, No.651 Dongfeng-Dong Road, Guangzhou, Guangdong Province 510317, China. E-mail: wangfang@ysucc.org.cn; Xu-Hui Zhang, Second Department of Oncology, The Affiliated Guangdong Second Provincial General Hospital of Jinan University, No.466 Xingan-Zhong Road, Guangzhou, Guangdong Province 510317, China. E-mail: 9158863@qq.com

Abstract

Background: Colorectal cancer (CRC) is a highly heterogeneous malignant tumor that significantly impacts clinical diagnosis and treatment. Single-cell RNA sequencing is an innovative method for exploring tumor heterogeneity and understanding its role at cellular and genetic levels. Method: The colorectal cancer single-cell RNA sequencing dataset were analyzed on the immune. RNA-seq data in bulk form was utilized to assess the major genes of the immune cell subsets linked to CRC. We conducted an analysis of the abundance of immune cells in the microenvironment of CRC, and also performed weighted gene co-expression network analysis. Gene set enrichment analysis helped perform two analytical procedures of subtype groups. Furthermore, Least absolute shrinkage and selection operator regression was employed to analyse and screen for a gene signature. Finally, quantitative PCR was performed to detect the expression levels of signature genes in CRC. Results: The single-cell RNA sequencing (GSE146771) dataset was integrated to obtain 9 cell clusters. The single-sample gene set enrichment analysis showed that the related gene expression of T-cell subsets of different functional statuses could vary greatly between patients with GSE146771. Immune cell analysis of TCGA-CRC indicated an improved overall survival rate for patients with elevated T2 cell abundance. Five-gene signature (Risk Score = −0.205 × CD25C − 0.231 × GSTCD − 0.010 × KPNA2 − 0.002 × KIF15 − 0.171 × ORC1) was developed by weighted correlation network analysis, and lasso Cox regression. Then, the risk prediction efficacy of the signature was validated in four GSE datasets. Furthermore, the expression of five genes was reduced in CRC tissue by quantitative PCR. Conclusion: Five-gene signature based on CRC heterogeneity was developed as a prognosis predictor, which can serve as a potential treatment target.

Keywords: colorectal cancer; scRNA-seq; T2 cells; 5-gene signature; risk prognosis

Cancer Advances 2023;6:e23022. https://doi.org/10.53388/2023623022

Submit a manuscript: https://www.tmrjournals.com/cancer
Background

Imposing a substantial health burden worldwide, colorectal cancer (CRC) holds second place in the list of death-causing cancers [1]. Surgical treatment, along with chemotherapy, is the traditional treatment for CRC. However, a large number of CRC patients receiving traditional treatment have experienced metastases and disseminated disease resistance to treatment, eventually leading to death [2]. Recently, immunotherapies such as tumor vaccines, immune checkpoint inhibitors and small molecule therapy are widely used, and immunotherapy can benefit the prognosis of CRC patients to a certain extent, which is a promising new treatment option [3]. Many studies have demonstrated that different patients with the same cancer type have widely different tumor immune infiltrates [4]. For this reason, treatments, especially immunotherapy, have to be individualized.

Single-cell RNA sequencing (scRNA-seq) utilizes second-generation sequencing technology allowing exploration of the global gene expression profile at the individual cell level. It also functions as a promising and innovative tool in the study of immune cell heterogeneity [5]. Recently, it has been suggested that scRNA-seq promotes the cancer-fighting ability of the immune system owing to its wide range of uses [6]. For example, it enables the study of T cell heterogeneity by performing transcriptome analyses of scRNA-seq in primary human tumors and allows clarification of the dynamic relationships among T cell populations by performing analysis through the integration of transcriptomes and T cell receptors [7]. It provided new proof for antitumor immunotherapy. Liu et al. utilized scRNA-seq to uncover the differences in the immune microenvironment characteristics between primary lesions and liver metastases in CRC [8]. Therefore, immune heterogeneity-based gene analysis can serve as a promising therapeutic target and provide a robust prediction of risks associated with the disease in an accurate manner.

During this research, an array of tissue-specific clusters was constructed for the prediction of immune cell composition using the CRC scRNA-seq datasets (GSE146771) in the Gene Expression Omnibus database. ScRNA-seq data was standardized and variance-stabilized using regularized negative binomial regression through SCTransform, immune cell clusters analysis, and FindCluster function of the Seurat package. The major genes associated with immune cell subsets in CRC were analyzed using bulk RNA-seq data obtained from The Cancer Genome Atlas (TCGA) expression profile. In addition, univariate COX analysis and Lasso Cox regression were employed for establishing a gene signature. The expression of signature genes in CRC was evaluated by quantitative PCR (qPCR). Lastly, a five-gene (CDC25C, GSTCD, KPNA2, KIF15, and ORC1) signature prognostic stratification system was constructed using heterogeneity of CRC immune cells for the identification of promising immunotherapy targets and accurate evaluation of the prognostic risk.

Methodology

Extraction of data

TCGA-CRC bulk RNA-seq data, comprising 638 individuals with CRC and 56325 genes, was extracted online from TCGA (https://www.cancer.gov/ccg/research/genome-sequencing/tcga) (Table 1). CRC scRNA-seq data GSE146771, including 10 CRC patients, 13538 genes, and 43817 cells, was provided by Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) (Table 2).

Processing of single-cell RNA-seq data

The Seurat package SCTransform function was used to preprocess the data, reduce batch effects, and integrate the single-cell transcriptome dataset. SelectIntegrationFeature chose those 3000 genes which were the most variant for immune cell cluster analysis, and FindCluster function of the Seurat package (version 3.1.4) was used with a resolution value of 0.15.

Single-sample gene set enrichment analysis

The relative abundance of 22 immune cells was estimated through CIBERSORT in the TCGA-CRC bulk RNA-seq data.

Weighted correlation network analysis (WGCNA)

The “WGCNA” package helped perform the weighted correlation network analysis (WGCNA) test. To perform subsequent analysis, a value of 5 was used for β, which was the most significant parameter of this test. To identify the hub genes, we selected genes that had a module membership (MM) greater than 0.5 and a Pearson correlation coefficient of 0.1 with overall survival (OS).

Identifying the molecular subtypes through non-negative matrix factorization (NMF) algorithm

Firstly, the expression data of correlated genes were taken from the TCGA database and randomly categorized into two subgroups. This was followed by the collection of the NMF training data. Ten repetitions of the NMF method were performed with the aid of the standard “brunet” option. The NMF algorithm was applied with a range of cluster numbers (k) from two to ten. The R package NMF was used to calculate the average contour width of the common member matrix. Based on this analysis, the training samples were categorized into two subgroups.

Identification and functional analysis of differentially expressed genes (DEGs)

The limma package was employed for the identification of DEGs between both groups. Identification of differentially expressed genes was done using the mentioned criteria as follows: log2FC | ≥ 1 and FDR < 0.05.

Molecular risk model construction

A Cox model was fitted with the help of the “coxph” function from the “survival” package, and a value of survival analysis with P < 0.05 was taken as a significant value. The Least absolute shrinkage and selection operator (LASSO) Cox regression model analysis was performed using the “glmnet” package. To validate the model, K-fold cross-validation with k = 10 was employed.

Specimen collection

During surgery, CRC tumor and macroscopically normal tissue samples were collected. The samples were immediately frozen in liquid nitrogen and stored at −80°C. To fix the samples, 10% neutral formalin was used. After fixation, the samples were embedded in paraffin, and 4 μm thick sections were prepared for further analysis. The sample was taken from those patients who did not receive pre-operative chemotherapy or radiotherapy. All study subjects provided written informed consent to the research and were subjected to approval by the Guangdong Second Provincial General Hospital for sample collection, research, and analysis. Approval for this research was given by the Ethics Committee of this hospital (2021-KZ-170).

Total DNA extraction and quantitative real-time PCR

CRC tissues were purposively selected in order to be embedded in paraffin, sectioned, and stained with HE-staining. The results were interpreted under the microscope by two pathologists. And CRC tissue scrapings were collected with a surgical blade. Total DNA was extracted from 16parafin-embedded tumor sections with the FFPE DNA extraction kit (TianGen Biochemistry, Beijing). The TaqMan probe method was applied for qPCR. The primer sequences were mentioned below (“F” for Forward and “R” for Reverse sequences): CDC25C (F) 5’-3’: TTGTGACACCACACACTGC and (R) 5’–3’: CCCCTGACTGTTGAGTACA. GSTCD (F) 5’–3’: AAAGTGGAGGTAGTGCCA and (R) 5’–3’: CCCCCTCCTGGTGCAAGTACT. KPNA2 (F) 5’–3’: TTGGGACACCGTCTTGAC and (R) 5’–3’: AGCATTCACGCTGTGGG. KIF15 (F) 5’–3’: GGAGGAGTCAAGGAATGTC and (R) 5’–3’: GTTCTTACCCGCAACCCAG.
**Table 1** Data of the CRC subjects (bulk RNA-seq)

<table>
<thead>
<tr>
<th></th>
<th>TCGA</th>
<th>GSE17537</th>
<th>GSE29623</th>
<th>GSE38832</th>
<th>GSE87211</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt; 65</td>
<td>361</td>
<td>22</td>
<td>NA</td>
<td>NA</td>
<td>85</td>
</tr>
<tr>
<td>&lt;= 65</td>
<td>274</td>
<td>33</td>
<td>NA</td>
<td>NA</td>
<td>117</td>
</tr>
<tr>
<td>Male</td>
<td>337</td>
<td>26</td>
<td>40</td>
<td>NA</td>
<td>142</td>
</tr>
<tr>
<td>Female</td>
<td>298</td>
<td>29</td>
<td>25</td>
<td>NA</td>
<td>61</td>
</tr>
<tr>
<td>Male</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>NA</td>
<td>0</td>
</tr>
<tr>
<td><strong>Survival</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alive</td>
<td>487</td>
<td>35</td>
<td>40</td>
<td>83</td>
<td>169</td>
</tr>
<tr>
<td>Dead</td>
<td>126</td>
<td>20</td>
<td>25</td>
<td>9</td>
<td>28</td>
</tr>
<tr>
<td>Male</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>30</td>
<td>6</td>
</tr>
<tr>
<td><strong>Stage</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage I</td>
<td>108</td>
<td>NA</td>
<td>7</td>
<td>17</td>
<td>NA</td>
</tr>
<tr>
<td>Stage II</td>
<td>234</td>
<td>NA</td>
<td>22</td>
<td>35</td>
<td>NA</td>
</tr>
<tr>
<td>Stage III</td>
<td>183</td>
<td>NA</td>
<td>18</td>
<td>39</td>
<td>NA</td>
</tr>
<tr>
<td>Stage IV</td>
<td>89</td>
<td>NA</td>
<td>18</td>
<td>31</td>
<td>NA</td>
</tr>
<tr>
<td>Male</td>
<td>24</td>
<td>NA</td>
<td>0</td>
<td>0</td>
<td>NA</td>
</tr>
</tbody>
</table>

CRC, colorectal cancer; TCGA, The Cancer Genome Atlas.

**Table 2** Data of the CRC subjects (single-cell RNA-seq)

<table>
<thead>
<tr>
<th>Study ID</th>
<th>Sample ID</th>
<th>Sample</th>
<th>Gender</th>
<th>Age</th>
<th>Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE146771</td>
<td>P0104</td>
<td>Female</td>
<td>77</td>
<td>IIIB</td>
<td></td>
</tr>
<tr>
<td></td>
<td>P0123</td>
<td>Female</td>
<td>65</td>
<td>IIIB</td>
<td></td>
</tr>
<tr>
<td></td>
<td>P0202</td>
<td>Female</td>
<td>50</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td></td>
<td>P0305</td>
<td>Female</td>
<td>71</td>
<td>IIA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>P0323</td>
<td>Female</td>
<td>65</td>
<td>IIIB</td>
<td></td>
</tr>
<tr>
<td></td>
<td>P0408</td>
<td>Male</td>
<td>86</td>
<td>IIB</td>
<td></td>
</tr>
<tr>
<td></td>
<td>P0410</td>
<td>Female</td>
<td>66</td>
<td>IIIB</td>
<td></td>
</tr>
<tr>
<td></td>
<td>P0613</td>
<td>Male</td>
<td>89</td>
<td>IIIB</td>
<td></td>
</tr>
<tr>
<td></td>
<td>P1025</td>
<td>Male</td>
<td>40</td>
<td>II</td>
<td></td>
</tr>
<tr>
<td></td>
<td>P1026</td>
<td>Male</td>
<td>52</td>
<td>IBC</td>
<td></td>
</tr>
</tbody>
</table>

**Statistical analysis**

R software (v.3.5.3) and Prism (GraphPad7) was used to conduct all the statistical analytic procedures of this study. Comparisons were done using T-tests, Wilcoxon test and P < 0.05 indicated a significant difference.

**Results**

**Integration and clustering of scRNA-Seq data**

The heterogeneity of CRC has been characterized using the cluster GSE146771 (Table 1). For the purpose of performing integration of the single-cell transcriptome dataset, preprocessing and removing batch effects were performed in R using the Seurat package SCTransform function. Uniform manifold approximation and projection, a non-linear dimensionality reduction technique, helped confirm the patterns (Figure 1A). 9 clusters were obtained by clustering cells using the FindCluster function (Figure 1B). Immune cells were clustered based on the expression of immune cell markers. These markers include CD3D and CD3E (markers for T cells or NK cells), CD79A (marker for B cells), LYZ and CD14 (markers for myeloid cells), TPTPRC (marker for immune cells), EPCAM (marker for epithelial cells), COL1A2 (marker for fibroblasts), IL7R (marker for naive T cells), and CD8A and NKG7 (markers for CD8 + T cells and NK cells) (Figure 1C).

**Immune cell analysis**

Based on immune cell markers, a cluster analysis of T and B lymphocytes and macro cells was performed (Figure 2A). We calculated the score of each cell sample in each immune cell subpopulation according to the multiple functional states of immune cells via gene set variation analysis. Initially, the T cells were sorted and identified, followed by cluster analysis depending on the gene set variation analysis scores for every cell sample. The gene expression profiles of different T cell subsets (naive, costimulatory, regulatory, and exhausted) were measured to assess their effector functions according to the gene set expression. The expression of genes related to T cell subsets with different functional statuses varied greatly among patients with GSE146771, as shown in Figure 2B. Furthermore, the effector function of B cells was analyzed based on the expression levels of B-cell subsets related gene sets, including features such as anti-apoptosis, naive memory, cytokine production, proliferation, and germinal center gene expression (Figure 2C). The activity of tumor-infiltrating myeloid cells, specifically M2 and M1-like myeloid cells, was also evaluated (Figure 2D). The analysis indicated that the gene expression profiles of B cells and myeloid cell subsets did not show significant differences among patients with GSE146771. Therefore, the above findings implied that the result suggests that T cells with different functional states are heterogeneous among CRC patients.
ARTICLE

Cancer Advances 2023;6:e23022. https://doi.org/10.53388/2023623022

Submit a manuscript: https://www.tmrjournals.com/cancer

Figure 1 The dimension reduction of CRC scRNA-seq. (A) Color based on various subjects. (B) Label colors according to separate clusters. (C) The expression profiles of key genes. CRC, colorectal cancer.

Figure 2 GSVA enrichment analysis of immune cells. (A) Heat map of key immune cell markers. (B) T cell characterization in CRC. GSVA enrichment fractions of naive T cells, costimulatory T cells, Regulatory T cells, and exhausted T cells related gene sets. (C) B cell characterization in CRC. GSVA enrichment fraction of naive B cells, proliferative, anti-apoptotic, pro-apoptotic, cytokine, and germinal center-related gene sets. (D) Myeloid cell features in CRC. GSVA enrichment fraction of associated gene sets in M1 and M2-like myeloid cells. CRC, colorectal cancer; GSVA, Gene set variation analysis.

CIBERSORT
According to the results from the single-cell RNA-seq data and immune cell types analysis, clinical significance was analyzed, and the prognostic model was constructed using bulk data. Bulk RNA-seq data...
has several advantages, including an increased number of samples and hence more clinical information. To further evaluate the clinical significance of immune cell infiltration in CRC, the CIBERSORT algorithm was employed. This algorithm calculates the proportions of 22 immune cell types based on RNA-seq count data. In this study, the abundance of several immune cell types, including B cells, T cells, M1 macrophages, and M2 macrophages, was determined using bulk RNA-seq data from 638 TCGA-CRC subjects (Table 2). Findings of the Kaplan-Meier method-based survival analysis showed an elevated overall survival rate for patients with greater Th2 cell abundance (Figures 3B, 3C). However, this difference was not significantly different between subjects having an abundance of M2-like TAMS, M1-like TAMS, and Th1 (Supplementary Figure 1). Thus, an in-depth Th2 analysis was conducted.

**WGCNA analysis and prediction of immunotherapy**

For the purpose of further assessing Th2’s role in CRC, WGCNA analysis as per the TCGA data (56325 genes, 638 affected individuals) was carried out. The genes with median absolute deviation (MAD) of less than or equal to 0.01 were removed, yielding a total of 189 genes. In order to develop a scale-free co-expression network, a power of β = 5 was chosen as the soft threshold to identify gene signatures associated with Th2 (Figures 4A, 4B). On the whole, 20 modules were developed (Figures 4C, 4D). Among them, black modules showed the strongest association with Th2 (r = 0.73, p = 4e^-102, Figure 4E). In Figure 4F, each point on the graph represents a gene, with the x-axis representing the module membership, indicating the correlation between the gene and the module eigengene. The y-axis represents the correlation between gene expression and OS. Based on the analysis, it was found that the black module contains genes that are associated with OS. A total of 20 genes were identified within this module as being related to overall survival.

**Molecular typing according to the Th2-related genes**

The gene expression RNA-seq data of the 20 Th2-related hub genes was downloaded from the TCGA database. The TCGA samples were then divided into different subgroups using the NMF algorithm. This algorithm is commonly used for clustering and identifying distinct patterns or subgroups within a dataset based on gene expression profiles. By taking a cluster value equal to 2 as the optimal parameter, the TCGA datasets were further categorized into two sub-groups (Figure 5A). Then, the concordance matrix was constructed, and values were scaled to be within [0, 1] (Figure 5B), where the value indicated the pair to be from the different clusters, and 1 indicated the pair to be from the same cluster. The heatmap generated with the list of 20 Th2-related genes is shown in Figure 5C, and cluster 2 had a more favorable prognosis in comparison to cluster 1 (Figure 5D).

Box-plot indicated a higher Th2 proportion in cluster 2 compared to cluster 1 (p = 2.2e^-16, Wilcoxon test) (Figure 5E). Generally, poor patient outcomes were observed in cluster 1 patients. Using limma-trend analysis, DEGs between the two clusters were identified (|logFC| > 2 and adj.p.val < 0.05), obtaining 4694 DEGs. Of these 4694 genes, 737 genes were downregulated, and 3957 were overexpressed in cluster 2 (Supplementary Table 1). Furthermore, we performed the KEGG enrichment analysis with the R package ‘clusterProfiler’ for cluster 2 (Figures 5F).

**Construction of a genetic risk model**

For the purpose of facilitating further validations, 20 Th2-related genes were chosen to conduct additional analytical procedures. The prognostic model was constructed using TCGA-CRC data (Table 2).

![Figure 3](https://www.tmrjournals.com/cancer)

**Figure 3 The abundance of immune cells.** (A) The abundance of 22 immune cells constructed on RNA-seq count data. (B) Kaplan–Meier analysis of DFI in CRC patients with different Th2 cells levels. (C) Kaplan–Meier analysis of OS in CRC patients with different Th2 cells levels. CRC, colorectal cancer; DFI, disease-free interval; OS, overall survival.
Figure 4 Screening of Hub genes. (A) Network topology structure constructed with unique power values. (B) Correlation between power values and average connectivity. (C) Genes grouped into distinct modules. (D) Classification of 400 randomly selected genes into separate modules. (E) Association of different modules with Th2 cell abundance. (F) In the black module, correlation of genes with overall survival depicted as a scatter plot, with hub nodes represented as dark dots.

Figure 5 Th2-related Molecular Classification. (A) Consensus Map of NMF Clustering. (B) Sample cluster of TCGA-CRC. (C) Expression profile of genes associated with Th2 cells. (D) Kaplan-Meier survival curves of two CRC molecular subtypes. (E) Abundance of Th2 cells in two CRC molecular subtypes. (F) KEGG functional enrichment analysis of differentially expressed genes in two CRC molecular subtypes. NMF, non-negative matrix factorization; CRC, colorectal cancer; TCGA, The Cancer Genome Atlas
LASSO regression helped screen five survival-related genes (CDC25C, GSTCD, KPNA2, KIF15, and ORC1) in the CRC TCGA set. Finally, a predictive model was developed on the basis of the gene expression data and regression coefficients. Regression coefficients for each predictor in the final prediction model are recorded in Supplementary Table 2. The glmnet package was used for performing LASSO Cox analysis. This analysis helps in selecting the most relevant independent variables for predicting survival outcomes. The changing trajectories of all independent variables are shown in Figure 6A. As the lambda value increases, the number of independent variable coefficients tending to 0 also increases, indicating the shrinking effect of LASSO regularization (Figure 6A). To evaluate the performance of the models, 10-fold cross-validation was conducted. This process helps in testing all models and estimating confidence intervals for each lambda value (Figure 6B). The optimal performance of the model was observed at a lambda value of ~4. Based on the LASSO Cox regression analysis, five genes (CDC25C, GSTCD, KPNA2, KIF15, and ORC1) were selected to construct a risk model. These genes were found to have the most significant association with survival outcomes in CRC.

This was followed by the measurement of the risk score, and the distribution of the risk score (RS) was verified in the TCGA training set (Figures 6C–6E). It indicated that the lower the gene expression, the higher the risk of mortality. The median RS was standardized to have a mean of 0, and as per the median RS value of the CRC TCGA datasets, the samples were categorized into low- and high-risk groups. High-risk patients exhibited a poorer prognosis compared to the other group. (Figures 6F).

Validation of the risk model's prognostic ability

For the purpose of determining the model's robustness, three GSE datasets (GSE17537, GSE38832, GSE87211) (Figures 7A–7C) were used to calculate RS, and the RS distribution was determined, exhibiting considerably smaller high-RS samples than low-RS samples. The low expression of CDC25C, GSTCD, KPNA2, KIF15, and ORC1 was also identified as a risk factor for CRC. Lastly, KM analysis showed significant survival differences between both risk groups (P < 0.05).

The expression of signature genes in CRC tissues

Moreover, to validate the 5-gene signature's accuracy, the expression levels of signature genes (CDC25C, GSTCD, KPNA2, KIF15, and ORC1) were observed in clinical samples from cancer patients (Table 3) using qPCR analysis (Figures 8A–8E), showing that the expression of CDC25C, GSTCD, KPNA2, KIF15, and ORC1 was low in CRC tissues and high in cancerous tissues.

![Figure 6](https://www.tmrjournals.com/cancer)

**Figure 6 Construction of genetic risk model.** (A) Trajectory of individual variables: the x-axis represents the log value of the independent variable lambda, and the y-axis represents the coefficient of the independent variable. (B) Confidence interval under each lambda. (C) Risk score of the subjects, ranked from lowest to highest, is represented on the x-axis, and survival is represented on the y-axis. Green dots represent deceased subjects, and red dots represent surviving subjects. (D) Risk Score (RS) of subjects from low to high is represented on the x-axis, and standardized RS is represented on the y-axis. Red represents the high-risk group, and green represents the low-risk group. (E) Expression of 5-Gene Signature in all TCGA subjects. Higher expression of the 5-Gene Signature is represented in red, and lower expression is represented in green. (F) Survival curve of the high and low-risk groups. TCGA, The Cancer Genome Atlas; RS, Risk score.
Figure 7 Prognostic risk model validation. (A–C) GSE datasets with the Risk Score. (D–F) The results of the KM curves shown in reveal significant differences between the low and high-risk group. GSE, GEO Series.

Figure 8 5 gene expression in CRC tissues. (A–E) The qPCR results showing GSTCD, CDC25C, KPNA2, KIF15 and ORC1 expression in CRC tissues, \(^{*}P < 0.05\). qPCR, quantitative PCR; CRC, colorectal cancer.
Table 3 Data of the CRC subjects

<table>
<thead>
<tr>
<th></th>
<th>Tumor</th>
<th>Peritumoral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>9</td>
<td>36</td>
</tr>
<tr>
<td>Age</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td>&lt; 65</td>
<td>8</td>
<td>23</td>
</tr>
<tr>
<td>Type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>COAD</td>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>READ</td>
<td>5</td>
<td>22</td>
</tr>
<tr>
<td>Stage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage I</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Stage II</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>Stage III</td>
<td>5</td>
<td>16</td>
</tr>
<tr>
<td>Stage IV</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Nx</td>
<td>0</td>
<td>6</td>
</tr>
</tbody>
</table>

CRC, colorectal cancer; COAD, Colon adenocarcinoma; READ, Rectum adenocarcinoma.

Discussion

Although the prognosis of cancer treatments has improved in recent years, such as for lung cancer, breast cancer, and hematological tumors, the heterogeneity of tumor immune microenvironments limits the efficacy of immunotherapy [9–12]. There is increasing evidence that CRC displays heterogeneity, limiting the identification of effective immunotherapy targets [13]. The available literature demonstrates single-cell transcriptome technologies as powerful tools for studying tumor heterogeneity and divergent subgroups, and it is important for identifying efficient pharmaceutical targets [14, 15].

In the present research, the scRNA-seq datasets (GSE146771) were used to characterize the CRC heterogeneity. The scRNA-seq dataset obtained normalization and variance stabilization using regularized negative binomial regression of SCTRansform, revealing 9 clusters. Clustering of the immune cells was done as per the immune cell markers (CD3D and CD3E are T cell or NK cell markers; CD79A is B cell marker; LYZ and CD14 are myeloid cell markers). Furthermore, the immune-related CRC cells were identified by calculating the enrichment score by GSVA for each cell sample, which showed that the related gene expression associated with T-cell subsets with different functional landscapes can vary greatly between patients with GSE146771.

To assess the intertumoral heterogeneity in CRC, scRNA-seq datasets were evaluated. This analysis helps in understanding the variability in gene expression patterns among individual cells within a tumor. The differential interplay between tumor cells and T cells was also examined, focusing on the interactions between different immune cell subtypes. Furthermore, the proportions of 22 immune cell types were predicted using bulk RNA-seq data from the TCGA-CRC dataset. This dataset includes gene expression information from 683 CRC cancer patients and 58,385 genes. By analyzing this data, the relative abundance of different immune cell types in CRC tumors can be estimated. It was also used to find out the abundance of Th2 cells. Survival analysis indicated a better overall survival of patients with improved Th2 abundance. This is in contrast to the results reported by Salman M. Toor et al., who found that high poor prognosis score patients showed higher expression of Th2-related genes [16]. And this is consistent with that found by Anita L Ray et al., who showed that females had increased survival of Th2 populations correlated with longer survival [17].

A comprehensive assessment was conducted to check the potential role of Th2 in CRC cells, and WGCNA analysis was explored based on data collected from TCGA, indicating that the black module elements depict genes associated with the OS. Lastly, 20 hub genes were also extracted from the module. The TCGA-CRC was clustered into two groups according to the Th2-related genes, and Cox was used for the identification of 20 survival-related genes (CDC25C, GSTTCD, KPNA2, KIF15, and ORC1). The five-gene signature was $R_S = -0.205 \times \text{CDC25C} - 0.231 \times \text{GSTTCD} - 0.010 \times \text{KPNA2} - 0.002 \times \text{KIF15} - 0.171 \times \text{ORC1}$ in accordance with the lasso Cox regression analysis. To further validate the predictive risk model, the RS and distribution were calculated using all GSE datasets, showing the low expression of the five genes (CDC25C, GSTTCD, KPNA2, KIF15, and ORC1) as risk factors. In addition, qPCR analysis was conducted for verification of the gene signature’s accuracy, which showed a lowered expression level of CDC25C, GSTTCD, KPNA2, KIF15, and ORC1 in CRC tumor tissues. This demonstrated the ability of the 5-gene signature to effectively predict the risk.

Cell division cycle 25C (CDC25C) is vital for cell cycle regulation, and its expression is closely linked to tumor development and advancement [18]. Accumulating evidence has demonstrated that targeting the associated biological axes of CDC25C can suppress the proliferation of colorectal cancer cells [19, 20]. CDC25C can therefore serve as a promising pharmaceutical cancer target. Although GSTTCD is expressed in the airway cells of humans, little is known about its role [21]. It has been proposed that Karyopherin α2 (KPNA2) is expressed in cancerous colon tissue and normal colon tissue, and its high expression is linked to lower overall survival and disease-free survival (DFS) [22, 23]. Thus, KPNA2 can function as a biological target for CRC. KIF15 protein expression levels were lower in CRC tissues compared to corresponding normal tissues and were correlated with the TNM stage [24]. ORC1 has been proven to be vital for the antitumor effect in uterine cervical cancer and diffuse large B-cell lymphoma [25, 26].

Taken together, a five-gene signature based on CRC heterogeneity was developed for the first time in this study, and it can also serve as a potential treatment target and prognosis predictor. However, some limitations to this study exist, such as few samples for qPCR verification, and the mechanism underlying the interaction between the five key genes has not been investigated. Future work with larger sample sizes can focus on the validation of this model and the exploration of the molecular mechanism of the five key genes. The present study adds to the available evidence on the pharmaceutical potential of this treatment and the ability of prognostic prediction of CRC patients.

In conclusion, the scRNA-seq datasets (GSE146771) were integrated and utilized for the purpose of characterizing CRC heterogeneity, with the related gene expression of T-cell subsets of different functional statuses, which can vary greatly between patients. Collectively, this work would not only help to extend the current knowledge on tumor-infiltrated T cells but can also establish a five-gene signature incorporating genes associated with Th2 cells in the GSE-CRC data.

References


