Bioinformatics analysis of prognostic value genes in colorectal cancer microenvironment

Taohua Yue  
Beida Hospital

Jing Zhu  
Beida Hospital

Xin Wang  
Beida Hospital

Yisheng Pan  
Beida Hospital

Yucun Liu  
Beida Hospital

Shanwen Chen  
Beida Hospital

pengyuan wang (pengyuan_wang@bjmu.edu.cn)  
Beida Hospital

Research article

Keywords: TCGA, microenvironment, colorectal cancer, immune scores, stromal scores

Posted Date: October 18th, 2019

DOI: https://doi.org/10.21203/rs.2.16191/v1

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

Colorectal cancer (CRC) is one of the most deadly gastrointestinal malignancies. The openness of the Cancer Genome Atlas (TCGA) allows us to perform correlation analysis between large-scale transcriptome data and overall survival (OS) of multiple malignancies. Previous literature reports that the infiltration of immune cells and stromal cells in the tumor microenvironment (TME) significantly associate with the prognosis of cancers. Based on the ESTIMATE algorithm, the immune and stromal components in TME can be quantified by immune and stromal scores. To determine the effects of immune and stromal cell associated genes on CRC prognosis, we divided the CRC cases into high- and low-groups based on the immune/stromal scores and identified 999 differentially expressed genes (DEGs). Heatmaps, functional enrichment analysis and protein-protein interaction (PPIs) networks further indicated that 999 DEGs mainly participated in stromal composition and immune response. Finally, we obtained 56 genes that were significantly associated with CRC prognosis from 999 DEGs and identified the PPIs networks. The role of 41 genes in CRC has been reported in previous literature, and the other 15 genes have never been reported. Therefore, we found 15 novel TME genes associated with CRC prognosis waiting for more researches.

Background

Colorectal cancer (CRC) was one of the most fatal gastrointestinal malignancy with a mean incidence rate of 6.1%, and mortality rate of 9.2% in both sexes combined reported on GLOBOCAN 2018 produced by the International Agency for Research on Cancer [1]. The Cancer Genome Atlas (TCGA) had established a genome-wide gene expression profile that increased the understanding of the impact of tumor genetic composition on clinical outcomes [2, 3]. In the TCGA database, the disease type of CRC was mainly classified into 6 subtypes: (1) adenomas and adenocarcinomas, (2) cystic, mucinous and serous neoplasms, (3) complex epithelial neoplasms, (4) epithelial neoplasms, NOS, (5) squamous cell neoplasms, (6) mature B-cell lymphomas. In terms of the number of samples, our study incorporated the former 4 subtypes. Previous large-scale clinical trials had found that women with BRCA1 or BRCA2 mutations had a high risk of breast and ovarian cancer, and women with BRCA1 mutations had an increased risk of early-onset CRC [4]. With these advances, gene expression profiling had increasingly been accepted as clinical diagnostic criteria.

Intrinsic genes in tumor cells determined the fate of CRC [5, 6]. The environment where tumor cells reside was called the tumor microenvironment (TME), which consisted of borders, blood vessels, lymph vessels, extracellular matrix, stromal cells, immune/inflammatory cells, secreted proteins, RNAs and small organelles. It was reported that the TME also significantly affected the gene expression of tumor cells, thereby affecting clinical outcomes [7–9]. In the TME, immune cells and stromal composition were 2 major types of non-tumor components. Feasible algorithms, ESTIMATE (Estimation of STromal and Immune cells in MAInfiant Tumor tissues using Expression data), could predict the tumor purity of tumor specimens [7, 10]. In this algorithm, the authors calculated the immune and stromal scores by analyzing the specific gene expression of immune cells and stromal cells, and indirectly quantified the proportion of
non-tumor cells. The ESTIMATE algorithm had been applied to glioblastoma [11], and head and neck squamous cell carcinoma [12] and was firstly applied into the CRC to determine the role of TME related genes in CRC prognosis. We extracted TME-related 56 genes that could be used to predict the prognosis of patients with CRC. The role of 15 genes in CRC had never been reported before and the research prospect of TME-related genes was very meaningful.

Methods

Database

Clinical data including gender, age, TNM stage, survival, etc. and level 3 transcriptome profiling (level 3 data) of CRC patients were downloaded from the TCGA Database (https://portal.gdc.cancer.gov/) [13]. The ESTIMATE algorithm was applied to calculate immune scores and stromal scores [7].

Extraction of differentially expressed genes (DEGs)

Data processing was performed by using R software package limma (http://www.bioconductor.org/packages/release/bioc/html/limma.html) [14], the wilcox test and the intersection of the R software. Fold change > 1 and adj. p < 0.05 were set as the cutoffs to filter for DEGs.

Heatmaps and venn programs

Based on the Pearson distance measurement method and the average linkage method, the heatmaps were drawn as follow. Heatmaps and Venn programs were made using the R software package pheatmap (https://cran.r-project.org/web/packages/pheatmap/) and VennDiagram (https://cran.r-project.org/web/packages/VennDiagram/), respectively.

Construction of PPIs network

The protein-protein interactions (PPIs) network was obtained from the public website STRING (https://string-db.org/) [15] and analyzed by cytoscape software [16]. To determine the most significant PPIs among 999 DEGs, we set the minimum required interaction score to 0.95. To simplify the display, we hided disconnected nodes in the network. K-core was set to 5 in cytoscape to highlight the point. The degree of connectivity for each node of the network was calculated. “Molecular Complex Detection” (MCODE) [17] was used to find clusters based on topology to locate densely connected regions. The minimum required interaction score was set as 0.4 in the building of PPIs network among 56 prognostic genes.

Function enrichment analysis of DEGs
Functional enrichment analysis of DEGs was performed by the open database DAVID 6.8 (a database for annotation, visualization and integration discovery) (https://david.ncifcrf.gov/) [18] that identifies the GO (Gene Ontology) categories (biological process (BP), molecular function (MF) or cellular component (CC)) and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway. Figures were made by using R software packages DOSE (http://www.bioconductor.org/packages/release/bioc/html/DOSE.html), clusterProfiler (http://www.bioconductor.org/packages/release/bioc/html/clusterProfiler.html) and enrichplot (http://www.bioconductor.org/packages/release/bioc/html/enrichplot.html). A false discovery rate (FDR) <0.05 was used as the cutoff value.

Kaplan-Meier survival analysis

Kaplan-Meier plots were produced to interpret the correlation between CRC OS and immune scores, stromal scores, and gene expression levels of DEGs, respectively. Figures drawing were performed by using the R software package survival. This significance was tested by log-rank test [19].

Results

1. Immune scores were significantly associated with CRC clinical stages and metastasis

We downloaded the transcriptome profiling and corresponding clinical information of 544 CRC patients on the TCGA database. Among them, 253 (46.5%) patients were female, 288 (52.9%) cases were male, 3 (0.6%) patients were of unknown gender. There were 449 (82.5%) cases of colon adenocarcinoma and 95 (17.5%) rectal adenocarcinoma. The number of patients in stages I, II, III, and IV was 96, 210, 149, and 78, respectively. 472 cases of adenomas and adenocarcinomas, 69 cases of Cystic, Mucinous and Serous Neoplasms were included in our study. Based on ESTIMATE algorithm, immune scores were ranged from −967.3 to 2999.3, and stromal scores were distributed between −2204 to 1969.1, respectively. The average immune scores of stage IV patients ranked the lowest of all stages and the difference was significant (P = 0.014). Similarly, the average immune scores of metastatic CRC cases were significantly lower than that of CRC patients without metastasis (P = 0.001). However, there was no difference in stromal scores, whether in CRC staging or metastasis, and the average stromal score had an increasing trend with the increase of stages (Fig.1A). This indicated that the average immune score was meaningful in the staging and metastasis of CRC.

To determine the association of immune scores and stromal scores with prognosis, we divided 540 CRC (excluding cases without complete prognostic information) cases into upper and lower halves (high and low) based on medians. Kaplan-Meier survival curves showed that the overall survival (OS) of high immune score group was not higher than that of the low scoring group (P = 0.982). Similarly, patients with lower stromal scores showed longer OS compared to patients with higher stromal scores (P = 0.347), although they were not statistically significant (Fig.1B).
2. Comparison of transcriptome profiling between stromal scores and immune scores in CRC

To reveal TME-related genes that were closely associated with the development and prognosis of CRC, we compared the transcriptome profiling between immune cells and/or stromal cells and constructed mRNA matrix based on 540 CRC cases and further drew heatmaps and Venn diagrams aiming to highlight DEGs. Heatmaps showed these different gene expression profiles based on high vs. low stromal score/immune scores groups. For the stromal scores, 1589 genes were upregulated and 11 genes downregulated in the high score compared with the low score group (log fold change (FC) >1, false discovery rate (FDR) < 0.05) (Fig.2A). For the immune scores, 1194 genes were upregulated and 29 genes downregulated in the high score (log FC >1, FDR < 0.05) (Fig.2B). In addition, Venn diagrams indicated 993 genes were up-regulated in both high scores groups (Fig.2C), and 6 genes were down-regulated in both low scores groups (Fig.2D). In order to accurately identify TME potential genes that played a key role, we performed a subsequent analysis of these 999 genes that were commonly differentially expressed in both the immune scores and the stromal scores, whether upregulated or downregulated.

3. Protein-protein interactions among 999 DEGs.

To better interpret the interactions between these identified DEGs, we used the free STRING website (https://string-db.org/) to explore protein-protein interactions (PPIs) networks. In our study, the network consisted of 8 modules, including 307 nodes and 606 edges. We chose the top 3 important modules for further investigative. We named these modules CXCL11, COL6A3 and SERPING1 modules for convenience, respectively. CXCL11 is the top module including 10 nodes and 44 edges. The names of these nodes are CXCL11, CCL13, CCL19, CCL21, CCR2, CCR5, CXCL10 and CXCL13 (Fig.3A). The second COL6A3 module containing 7 nodes and 21 edges, the names of these nodes are COL6A3, COL1A1, COL1A2, COL3A1, COL5A2, COL6A1 and COL6A2 (Fig.3B). In the SERPING1 modules, 15 edges involving 6 nodes were formed in the network. This list of nodes is as follows SERPING1, C1QA, C1QB, C1QC, C1R and C1S (Fig.3C). In the PPIs network, the top 3 nodes with the highest number of adjacent nodes were CXCL8, IL10 and FN1 (fibronectin 1) (Fig.3D). PPIs network analysis focused on the immune and stromal aspects of CRC, further validating the accuracy of our TME analysis.

4. Functional enrichment analysis of 999 DEGs.

To summarize the potential functions of DEGs in the TME of CRC, we performed functional enrichment analysis of up-regulated 993 genes and down-regulated 6 genes together. Part of gene ontology (GO) analysis results were shown in table 1. Top GO terms identified included regulation of leukocyte activation and T cell activation, extracellular matrix, receptor ligand activity and receptor regulator activity (Fig.4A), consistent with the subject of our research target TME.
In order to explore the specific signal pathways involved in CRC on immune and stromal aspects, we further performed Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis of the total of 999 genes. Part of KEGG analysis results were shown in table 2. The top 3 signal pathways were cytokine-cytokine receptor interaction, chemokine signaling pathway and cell adhesion molecules (Fig.4B), all of which were mainly associated with immune response and stromal composition.

5. Correlation between expression of individual DEGs and OS of CRC

To investigate the potential roles of individual DEGs in OS, we performed survival analysis on individual DEGs and plotted Kaplan-Meier survival curves. Of the 999 genes differentially expressed, 56 had a significant effect on the OS of CRC patients. Among them, 44 genes significantly worsened the prognosis (Fig.5A) while 12 genes significantly improved the prognosis of CRC (Fig.5B) (p < 0.05, DEGs significantly associated with CRC prognosis were shown in table 3). Of the 56 genes, 15 genes had never been studied in CRC previously based on literature search. More research on these TME related genes needed to be done.

6. PPIs network analysis of 56 prognostic TME-related genes

Based on the String website, we found that PTGS2 and ADAMTS4 played the central role in linking prognostic TME-related genes together (Fig.6). More research could be done around PTGS2 and ADAMTS4.

Discussion

In this study, we sought to identify TME-related genes that impacted the OS of CRC patients. In particular, by comparing transcriptome profiling in 540 CRC cases with high/low TME scores, we obtained 999 DEGs both in immune scores and stromal scores and displayed in the form of heatmaps and venn programs. In addition, we constructed 3 major PPIs modules by clustering, all of which were related to stromal components/immune responses. Top relevant nodes in these 3 modules, including CXCL11, COL6A3 and SERPING1, and the three nodes with the most adjacent nodes containing CXCL8, IL10, FN1, were associated with proliferation, invasion, angiogenesis, distant metastasis and prognosis in CRC [20–28]. GO terminology and KEGG pathway analysis further showed function enrichment of these 999 genes. This was consistent with previous research findings that the function of stromal composition and immune responses were interrelated in the construction of the TME [29–32].

Next, we performed prognostic analysis and identified 56 genes significantly associated with the CRC prognosis. Among them, 41 genes (genes in non-bold in table 3) reported previously, were meaningful in the pathogenesis of CRC or in predicting OS, indicating that TCGA database analysis had essential references for determining the prognosis of CRC patients from the genetic level. The remaining 15 genes had not been associated with CRC prognosis previously, now all of which could serve as potential
prognostic markers for CRC. These 15 genes included SPOCD1, ADAMTSL4, CD1B, NALCN, TBX1, CLECL1, KRT16, PODN, IGLON5, P4HA3, SLC6A16, CHRDL1, TREML1, RFX8 and ANKRD1 (genes in bold in table 3). The flow chart of this study was shown in figure 7.

PPIs network on these 56 genes displayed that PTGS2 and ADAMTS4 were in the central position. Among the 19 genes involved in building the PPIs network, roles and mechanisms of TBX1, ADAMTSL4 and ANKRD1 needed to be studied. TBX1 encoded a T-box transcription factor and extinction of TBX1 expression in the third pharyngeal pouch was a prerequisite for thymus organogenesis [33]. In addition, it played a potential role as a tumor suppressor in thyroid cancer [34]. ADAMTSL4 was immune-related biomarker for primary glioblastoma multiforme [35]. ANKRD1 was a potential molecular target to enhance sensitivity of ovarian cancer to chemotherapy [36]. It could be seen that they were related to the development of malignant tumor.

Significant advances had been made in the analysis of the association between OS and intrinsic gene expression of cancer cells. However, the study of TME was relatively backward. Many studies had been conducted in colon cancer cell lines, tumor-bearing mice, or a small number of tumor samples from CRC patients. Based on the complexity of the CRC microenvironment, we need a comprehensive analysis of large sample data. Fortunately, the development of the TCGA database and its free access for the public had promoted TME related research [37].

The interaction of CRC and its TME severely affected the evolution of the tumor, and subsequently impacted patient staging, recurrence, drug resistance, and OS [38, 39]. Previous reports had focused on how intrinsic genes in tumor cells shaped the TME and affected survival. Our research focused on the genetic characteristics of the microenvironment surrounding the tumor cells, which in turn affected the occurrence and development of CRC, thereby influencing the OS of patients. Our results could provide part of data explaining the impact of the CRC TME on patient survival from the genetic level.

In conclusion, through the ESTIMATE algorithm for immune and stromal scores, we analyzed the functions, signaling pathways, and prognosis of these DEGs associated with the TME. Focusing on the “soil” of tumor cell, previously neglected genes had the potential to become novel CRC biomarkers. Finally, more molecular biology research related to the TME DEGs needed to be carried out, in order to obtain a more comprehensive interpretation of the interaction of the TME and tumor cell itself.

**Conclusion**

In conclusion, there were 56 prognostic-related genes in the colorectal cancer microenvironment, and among them 15 unknown functional genes needed to be studied.

**Abbreviations**

CRC: Colorectal cancer;
TCGA: The Cancer Genome Atlas;
OS: Overall survival;
TME: Tumor microenvironment;
ESTIMATE: Estimation of STromal and Immune cells in MAignant Tumor tissues using Expression data;
DEGs: Differentially expressed genes;
PPIs: Protein-protein interactions;
MCODE: Molecular Complex Detection;
GO: Gene Ontology; BP: biological process; MF: molecular function; CC: cellular component;
KEGG: Kyoto Encyclopedia of Genes and Genomes;

**Declarations**

**Ethics approval and consent to participate**

Not applicable.

**Consent for publication**

Not applicable.

**Availability of data and materials**

Data analyzed during the current study could be obtained from corresponding authors at the reasonable request.

**Competing interests**

The authors declared that they had no competing interests.

**Funding**

This work was supported by the National Natural Science Foundation of China (No. 81770522). The funder had no role in the design of the study, data collection or analysis, interpretation of data or preparation of manuscript.
Authors’ contributions

YL, PW, and TY, experimental design. TY and SC: bioinformatics analysis, writing and editing. YL and PW: sources of funds. JZ, XW, YP, YL and PW: supervision and manuscript modification. All authors had read and approved this manuscript.

Acknowledgements

The author would like to thank maintainers of TCGA database, string website and DAVID website for their contribution to research.

Authors’ information

Taohua Yue, Email: taohua_yue@bjmu.edu.cn.

Jing Zhu, Email: pwjingzhu@163.com.

Xin Wang, Email: wangxin_guo@hotmail.com.

Yisheng Pan, Email: yisheng.pan@pkufh.com.

Yucun Liu, Email: yucunliu@bjmu.edu.cn.

Shanwen Chen, Phone: 15001357973, Email: shanwen@bjmu.edu.cn.

Pengyuan Wang, Phone: 1881132155, Email: pengyuan_wang@bjmu.edu.cn.

References


Tables

Due to technical limitations, tables are only available as a download in the supplemental files section

Figures
Figure 1

The correlations between immune scores/stromal scores and clinical stage, metastasis and OS of CRC. (A), Immune and stromal scores for CRC clinical staging and metastasis. Box-plot showed that CRC clinical stage and metastasis were significantly associated with immune scores (n=544, p < 0.05); However, the stromal scores were weakly correlated with these two points. (B), based on the corresponding median scores of immune scores and stromal scores, CRC cases were divided into high
scores group and low scores group. As shown by the Kaplan-Meier survival curve, the OS of the high immune scores group was longer than that of the low immune scores group; The OS of the low stromal scores group was longer than that of the high stromal scores group. However, both were not statistically significant (p >0.05).

Figure 2
Transcriptome profiling in CRC immune scores and stromal scores, respectively. Red, green and black represented higher expressed genes, lower expressed genes and genes with the same expression level, respectively. (A) The difference of the lower half (low score) stromal scores DEGs was significant ($p < 0.05$, fold change $> 1$). (B) The difference of the upper half (high score) of the immune scores DEGs was significant ($p < 0.05$, fold change $> 1$). Venn diagrams showed the number of DEGs that were generally up-regulated (C) or down-regulated (D) in the stromal and immune scores groups.
Top 3 PPIs networks of CXCL11, COL6A3, and SERPING1 modules. In cluster networks CXCL11 (A), COL6A3 (B), and SERPING1 (C) modules, the color and size of the nodes in the PPI network represented the score of the MCODE. Deep color and large nodes meant high MCODE scores. (D) Number of adjacent nodes per node in the PPIs network of the top 30 nodes.
Functional enrichment analysis of 999 DEGs. The top 10 GO terminology in aspect of (A) Biological processes (BP), cellular components (CC), and molecular functions (MF). (B) The top 30 KEGG pathways.

Figure 5

56 DEGs significantly associated with prognosis. Correlation of 999 DEGs with OS. Based on corresponding medians, individual gene expression was divided into high and low groups. red line and blue line represented higher and lower gene expression, respectively. 56 genes significantly associated with OS as shown by the Kaplan-Meier survival curve ($p<0.05$). Among them, 44 genes (A) worsened disease while 12 genes (B) improved prognosis.
Figure 6

PPIs networks of 56 genes associated prognosis. Of the 56 TME-related prognostic genes, 19 were involved in the construction of a PPIs network, in which PTGS2 and ADAMTS4 were at the central position, and green represented the TME-associated prognosis gene with unknown function in colorectal cancer.
Figure 7

Flow chart of our study.

Supplementary Files

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