Immune Characteristics-Related Typing of Colorectal Cancer and the Establishment of 14-Gene Prognostic Model

Jianxing Ma
Lanzhou University Second Hospital

Chen Wang (chenwang@lzu.edu.cn)
Lanzhou University Second Hospital

Research Article

**Keywords:** Immune, colorectal cancer, TME, NMF

**Posted Date:** November 30th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-1093264/v1

**License:** © This work is licensed under a Creative Commons Attribution 4.0 International License.  [Read Full License](https://creativecommons.org/licenses/by/4.0)
Abstract

This study is to establish NMF (nonnegative matrix factorization) typing related to the tumor microenvironment (TME) of colorectal cancer (CRC) and to construct a gene model related to prognosis to be able to more accurately estimate the prognosis of CRC patients. NMF algorithm was used to classify samples merged clinical data of differentially expressed genes (DEGs) of TCGA that are related to the TME shared in The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) datasets, and survival differences between subtype groups were compared. By using createData Partition command, TCGA database samples were randomly divided into train group and test group. Then the univariate Cox analysis, Lasso regression and multivariate Cox regression models were used to obtain risk model formula, which is used to score the samples in the train group, test group and GEO database, and to divide the samples of each group into high-risk and low-risk groups, according to the median score of the train group. After that, the model was validated. Patients with CRC were divided into 2, 3, 5 subtypes respectively. The comparison of patients with overall survival (OS) and progression-free survival (PFS) showed that the method of typing with the rank set to 5 was the most statistically significant ($p=0.007, p<0.001$, respectively). Moreover, the model constructed containing 14 immune-related genes (PPARGC1A, CXCL11, PCOLCE2, GABRD, TRAF5, FOXD1, NXP4, ALPK3, KCNJ11, NPR1, F2RL2, CD36, CCNF, DUSP14) can be used as an independent prognostic factor, which is superior to some previous models in terms of patient prognosis. The 5-type typing of CRC patients and the 14 immune-related genes model constructed by us can accurately estimate the prognosis of patients with CRC.

Introduction

Colorectal cancer (CRC) is one of the most common cancers with high morbidity and high mortality in the world. Worldwide, there are approximately 1.8 million new cases of CRC and 900,000 CRC-related deaths each year, accounting for about 10% of all diagnosed cancers and cancer-related deaths, ranking the fourth among 36 cancers. So far, how to diagnose CRC at an early stage and accurately estimate the prognosis of patients still need further research and optimization. In recent years, cancer immunotherapy has become an important means of cancer treatment and one of the hot spots in the field of cancer research. However, the complex composition and heterogeneity of tumor microenvironment (TME) pose a huge challenge to tumor immunotherapy. CRC is one of the tumors with poor effect in immunotherapy. One problem that needs us to think is to how to perform the immune-related typing of patients with CRC and use immune-related genes to construct prognostic models, so that patients with different subtypes and different risk conditions can be given reasonable treatments and their survival can be assessed accurately. In order to solve this related problem, many researchers have made efforts and put forward some innovative viewpoints. People such as Jun Jun Wang, Xiao Wang and others have previously constructed genetic models for the prognosis of CRC, which are of high value in the evaluation of the prognosis and recurrence of CRC. However, there is no one has compared the accuracy of evaluation, statistical differences among their respective models. Even some scholars also used the NMF algorithm to classify CRC samples in TCGA, most of them used methods with the rank value of 2 or 3, and they did not compare the statistical differences in OS and PFS of patients caused by typing with different rank values.

In view of the ability to better classify patients with CRC, a total of 540 CRC samples that meet the requirements from the TCGA (TCGA-COAD, TCGA-READ) and GEO (GSE39582, GPL570) database were screened by us, with the use of the NMF (method=brunet) algorithm to gain the typing, and the relative optimal typing method with a rank value of 5 was determined by comparing the impact of the three typing methods with a rank value of 2, 3, and 5 respectively on the OS and PFS of patients in each subtype. This typing method was then compared with others. In addition, with the purpose to predict the survival of patients with CRC, a risk score model containing 14 immune-related genes from the TCGA database (TCGA-COAD, TCGA-READ) and GEO database (GSE39582, GPL570) has also been established. In order to obtain a relatively more accurate risk score model, the immune-related risk score model constructed by us was compared with the models of some scholars. The results showed that the risk score model we constructed was superior to some scholars' models in terms of patient prognosis, and the results of univariate and multivariate independent prognostic analysis showed that the constructed immune-related risk score model could be used as an independent prognostic factor.

Methods And Materials

Data acquisition. The CRC-related gene expression data obtained from the TCGA (https://tcga-data.nci.nih.gov) database contains 56461 genes and their expression levels from 544 original samples, and the clinical data from the TCGA database includes 548 original samples and their clinical traits. While the CRC expression data obtained from the GEO (http://www.ncbi.nlm.nih.gov/geo/) database (GSE39582, GPL570) contains 21654 genes and their expression levels, and the clinical data from the GEO database contains 585 clinical samples and their clinical traits. By using $|\text{LogFC}|>1.5$ and fdr<0.05 as filter conditions, DEGs were found in TCGA, and the 559 TCGA DEGs related to TME shared in the two databases of TCGA and GEO were obtained by the method of intersection. Finally, the samples of these DEGs were respectively intersected and merged with the clinical samples of the TCGA and GEO databases, and 540 and 579 samples were obtained respectively for subsequent typing and modeling. The genes of part of the CRC prognostic risk models constructed by the predecessors were obtained from their articles.

NMF algorithm immune-related typing. The univariate Cox survival analysis was performed on genes of 540 samples in TCGA, and survival related differential genes were obtained with p value <0.01 as the filter condition. The NMF command (rank=2:10, method=brunet, NRUN=10, seed=123456) was used to classify the samples of differential genes obtained by us (use “survival” package, “NMF” package), so that the graph of Cophenetic typing parameter and 9 heatmaps were obtained, and the optimal rank value was determined by comparing the graphical results and subsequent analysis on
OS and PFS in patients with CRC for typing. There are three criteria for the selection of rank value according to the graphic results: firstly, select the point at the front of the maximum slope segment in the Cophenetic graph; secondly, the color within the group is required to be dark red, while the color between groups is required to be dark blue in the typed heatmap; thirdly, the number of samples in each group is not very different. By comparison with previous pan-cancer typing methods, differences between our typing and others could be obtained, and Sanki plots were used for visualization.

**MCPcounter immune cell infiltration analysis.** To explore whether there were differences in infiltration content between subtypes in our typing method, the immune infiltration analysis was performed. The MCPcounter estimate command (use "MCPcounter" package and "limma" package) was used to calculate the immune cell infiltration through the expression data to obtain the content of immune cell infiltration in each sample. Finally, the differences in infiltration of different immune cells between subtypes were analyzed. The infiltration differences of ten immune cells, including CD8 T cells, Endothelial cells, Fibroblasts, Monocytic lineage, Myeloid dendritic cells, Neutrophils, NK cells, Cytotoxic lymphocytes, B lineage, and T cells were analyzed between subtype groups.

**Construction of prognostic risk model.** The 540 samples containing expression and clinical data in TCGA were randomly divided into the train group and test group by using the `createDataPartition` (p=0.7) command (use "survival", "caret", "glmnet", "survminer", and "timeROC" packages). The univariate Cox survival analysis was performed on the genes in the train group, and p<0.05 was used as the filter condition to obtain the gene expression levels with significant differences. Then, the Lasso regression model was constructed by taking the expression level of single-factor significantly different genes as x-axis as well as the survival time and survival state of the samples as y-axis. After that, the genes expression levels corresponding to the minimum error point was obtained through cross-validation, and the model genes and corresponding coefficients were obtained by constructing the multivariate Cox model. The formula of the model was as follow: the expression of A gene x A gene coefficient + the expression of B gene x B gene coefficient +... the expression of N gene x N gene coefficient. The model formula was applied to score the samples from the train group, test group and GEO database (579 samples), and the median risk score of samples in the train group was used for risk grouping (above the median value is high, and below the median value is low). The survival analysis was then performed on the high-risk and low-risk samples in each group, and the ROC curve analysis was used to test the accuracy of the model.

**Decision Curve Analysis (DCA) and ROC curve analysis.** To compare the prognostic ability and accuracy of the risk score model, nomogram$^{23}$, and various clinical traits, DCA and ROC curve analysis were performed. The "survival", "survminer", "timeROC", and "ggDCA" packages were used to perform the Cox survival analysis on the prognostic risk score model, nomogram, and risk scores of the clinical traits, and then DCA was performed for all factors to obtain the optimal prognostic factor. Then predictTime was set to 1 to conduct the ROC curve analysis of each factor and the value of Area Under The Curve (AUC) can be obtained later, and the accuracy of prediction of each factor was judged by comparing the AUC value.

**Gene Set Enrichment Analysis (GSEA).** To obtain the active functions and pathways in the high-risk and low-risk groups, the GSEA was conducted. With p value <0.05 as the filtering condition (use "limma", "org.Hs.eg.db", "clusterProfiler", and "enrichplot" packages), the GSEA was performed on 540 samples in the TCGA database. Active functions and pathways in high-risk and low-risk groups could be obtained respectively through the GO, KEGG annotated gene set files and the expression matrix, and the results were finally visualized.

**Results**

**Research flow chart.** The data of TCGA database and GEO database were screened for subsequent NMF typing and the construction of prognostic risk model. For a purpose of illustration, a flowchart for the main steps of this research (Fig. 1) has been plotted.

**Comparison of NMF immune-related typing.** In order to get the optimal grouping method, the rank value was set as 2:10 to get the cophenetic graph and 9 heatmaps. With the choice of three selection conditions: the point at the front of the segment with the largest slope, the deep red within the group and the deep blue between the groups in the heatmap, the number of samples in each group should be as close as possible. By analyzing and comparing the three types with the rank value set to 2, 3, and 5 respectively, the optimal typing method was determined. After comparing the OS and PFS of the samples in the subtypes of each typing method, it was found that the typing method with the rank value of 5 was the most statistically significant (p=0.007, p<0.001, Fig. 2A-J). The three immune-related typing methods with the rank severally set to 2, 3, and 5 were compared with the pan-cancer typing methods of the predecessors, and it was found that the differences between the typing method with the rank set to 5 and others was the most obvious among the three types (Fig. 2K-M).

**Results of the MCPcounter immune cell infiltration.** The MCPcounter immune cell infiltration analysis was performed on CD8 T cells, Endothelial cells, Fibroblasts, Monocytic lineage, Myeloid dendritic cells, Neutrophils, NK cells, Cytotoxic lymphocytes, B lineage, and T cells, and the differences between subtypes were obtained (Fig. 3). The results showed that Monocytic lineage gradually increased in the subtype groups from c1 to c5, and the differences were statistically significant. Between them, the c1 and c5 subtypes were the most statistically significant. In terms of Fibroblasts, except that there was no significant difference between subtypes c1 and c3, there were significant differences between the other subtypes, and its overall trend of infiltration gradually increased from subtypes c1 to c5. The infiltration content of Myeloid dendritic cells, Cytotoxic lymphocytes, Neutrophils, NK cells and B lineage gradually increased from subtype c1 to the other subtypes respectively, and it was statistically significant. However, the trend of infiltration content of CD8 T cells and T cells was not obvious between subtypes.

**Results of the construction of the prognostic risk model**
Model formula and survival analysis. Through the univariate Cox analysis of the samples in the train group, the construction of Lasso regression model (Fig. 4A), the multivariate Cox regression model, and optimization of the model, 14 immune-related model genes (PPARGC1A, CXCL11, PCOLCE2, GABRD, TRAF5, FOXD1, NXPH4, ALPK3, KCNJ11, NRP1, F2RL2, CD36, CCNF, DUSP14) and the expression levels of them were obtained. The model formula was the sum of the product of each model gene expression and the corresponding coefficient. The specific formula was: the expression of PPARGC1A × (0.92348824433332) + the expression of CXCL11 × (-0.204281207401165) + the expression of PCOLCE2 × (0.47681870221687) + the expression of TRAF5 × (0.328217688439647) + the expression of FOXD1 × (0.328311268298874) + the expression of NXPH4 × (0.302207021939709) + the expression of ALPK3 × (0.2956614563205) + the expression of KCNJ11 × (0.539567484564305) + the expression of NRP1 × (-1.19788016724308) + the expression of F2RL2 × (-0.289185498901046) + the expression of CD36 × (0.6844747316158) + the expression of CCNF × (-0.71540050393579) + the expression of DUSP14 × (0.35138433759811). Among them, 5 genes (PPARGC1A, CXCL11, NRP1, F2RL2, CCNF) were protective genes with coefficients less than 0, while the remaining 9 genes were risk genes with coefficients greater than 0.

The model formula was used to score the samples of the train group, the test group and the GEO group respectively, and the samples of each group were divided into high and low-risk groups according to the median value of the score in the train group. Survival analysis were performed on the high and low-risk groups of the train group, test group, all samples of TCGA group, and GEO group, and the survival differences between the high and low-risk groups were obtained then (Fig. 4C-F). Moreover, the accuracy of this risk model for 1-year, 3-year, and 5-year prognosis was tested by using the ROC curve analysis (Fig. 4G-J). The results showed significant differences in survival among risk groups (p<0.001). According to whether the AUC value is greater than 0.65, the accuracy of the prediction was judged, and the results showed that the prediction effect of this risk model was good.

Cox Survival analysis of clinical traits by group. According to the clinical traits, 540 patients in TCGA were divided into groups, and the Cox survival analysis between the high-risk and low-risk groups of each group was carried out (Fig. 5A-J). The results showed that except no statistical difference in the survival of patients in the high-risk and low-risk groups in the T1-T2 group, others were significantly different (p<0.001, p=0.005), which suggested that the survival of CRC patients who invade the submucosa and muscularis propria was not associated with the risk.

Nomogram scoring forecast. With the use of clinical traits (gender, age, stage, T, N) and risk score, a nomogram model was built, and the model was visualized to obtain the graph of scoring method and calibration curve (Fig. 5K-L). According to the nomogram, the total score of the patient's clinical traits and risk score could be obtained, and the total score could be used to predict the 1-year, 3-year, and 5-year survival rate (Fig. 5K). In addition, the calibration curve showed that the curves of 1, 3, and 5 years were closer to the gray line segment, which indicated that the nomogram was accurate in predicting the survival rate of patients with CRC (Fig. 5L).

Decision Curve Analysis (DCA). The DCA was used to analyze the ability of nomogram, risk model, and clinical traits to predict survival of patients with CRC. The results showed that the nomogram has the strongest predictive ability, followed by the risk model, while the gender was the worst (Fig. 5M). In addition, the ROC curve analysis results also showed that the nomogram has the highest prediction accuracy, followed by the risk model, while gender was the lowest (Fig. 5N). The reason why the nomogram has the strongest predictive ability was that it combines multiple factors including the risk model and clinical traits.

Independent prognostic analysis. The univariate Cox independent prognostic analysis of patients’ age, gender, stage and risk scores showed that age, stage and risk scores were statistically significantly prognostic factors (p<0.05; Table 1). The multivariate independent prognostic analysis was performed on the statistically significant prognostic factors of univariate Cox analysis. Then the results showed that the p-values of age, stage, and risk scores were still less than 0.05 (Table 1), indicating that the constructed risk score model could be used as an independent prognostic factor.

Table1. univariate and multivariate Cox independent prognostic analysis

| Univariable analysis Multivariable analysis |
|--------------------------------------------|
| Variables | HR 95% CI of HR | P value | Variables | HR 95% CI of HR | P value |
|-----------|----------------|---------|-----------|----------------|---------|
| lower     | higher         |         | lower     | higher         |         |
| Age       | 1.036565199    | 1.017917911 | 0.000105584 | 1.051213616 | 1.071405676 | 2.67E-07 |
| Gender    | 1.052620671    | 0.713634557 | 1.552629297 | 0.795937211 | 1.716676975 | 2.93E-11 |
| Stage     | 2.052653076    | 1.644863825 | 2.561315483 | 1.95E-10 | 1.037517749 | 1.025523397 | 1.049652386 | 2.151630571 | 1.023236217 | 1.044455031 | 2.20E-10 |

Comparison with other prognostic risk models. By plotting survival curves for each model, it was found that the curves of the high and low-risk groups distinguished by the model constructed by us were more separated (p<0.001; Fig. 6A-F), indicating that the model could better assess the level of patient risk. The AUC values of 1-year, 3-year, and 5-year were 0.768, 0.758, and 0.801 respectively, which suggested that our model has a higher prognostic accuracy (Fig. 6G-L). The higher value of the C-index analysis, the better the prognostic effect of the model. As shown in the figure, the score of the risk
model constructed by us was 0.744, and the prognostic effect was relatively best (Fig. 6M). What is more, the RMS graph showed that our risk model has good accuracy, 95%HR: 1.035 (1.025-1.045), p<0.001, which was statistically significant (Fig. 6N).

**Results of GSEA.** The GSEA was used to obtain functions and pathways that were active in the high and low-risk groups. The results of the GSEA functional analysis showed that comification, epidermal cell differentiation, and epidermis development were active in the high-risk group, while in the low-risk group, spliceosomal tri-snRNP complex assembly, DNA packaging complex, protein DNA complex, Sm-like protein family complex, small nuclear ribonucleoprotein complex were the top 5 functionally active (Fig. 7A-B). Moreover, the results of GSEA pathway analysis showed that complement and coagulation cascades, ECM-receptor-interaction, PPAR signaling pathway, ribosome, Wnt-signaling pathway were the top 5 active pathways in the high-risk group, while there was no obvious active pathway in the low-risk group (Fig. 7C-D).

**Distribution of risk scores for clinical groups.** Through using each clinical trait (Age, Gender, T, N, M, and Stage), the samples were divided into groups respectively, and the distribution of risk scores in each group was analyzed (Fig. 7E-J). The results showed that there was no significant difference in the distribution of risk scores in each group after grouping by age (≤65 and >65, p>0.05) and gender (Male and Female, p>0.05). After grouping by T, it could be found that there was no significant difference between T1 and T2 (p>0.05) as well as between T2 and T3 (p>0.05), which indicated that there was no significant difference in the risk between patients with CRC invading the submucosa and invading the muscularis propria, and there was also no significant difference in the risk between patients with CRC invading the muscularis propria and subserosal. According to N grouping, there was no significant difference between N0 and N1 (p>0.05), but the differences between them and N2 were significant (p<0.001), which indicated that there was no significant difference in risk when the number of lymph node metastases in patients with CRC does not exceed 3, while exceeds 3, the risk increased significantly. In general, after grouping by M and Stage, the average risk score of each group gradually increased.

**Genes and immune cells associated with the risk model.** The correlation analysis was used to obtain genes and immune cells related to the prognostic risk score model constructed by us. The analysis results showed that genes including PDCD1, SERPINE1, CCD80, etc., had a significant positive correlation with the model (p<0.05; Fig. 7K), and immune cells such as T cells, B Lineage, CD8 T cells, Endothelial cells, Monocytic lineage had a significant positive correlation with the model (p<0.05; Fig. 7L). The results of Tumor Mutation Burden (TMB) correlation analysis showed that TMB was significantly positively correlated with Cytotoxic lymphocytes, Monocytic lineage, and Myeloid dendritic cells, but had no obvious correlation with riskscore (Fig. 7M).

**Discussion**

The tumor microenvironment (TME) is a component of cancer. It is a complex ecosystem that supports tumor growth, metastasis, while weakening immune surveillance, thereby affecting the disease progression and prognosis of patients with tumors. The TME contains stromal cells and immune cells, which accumulate at different stages of tumor development. Among them, the immune cells mainly include macrophages, natural killer (NK) cells, dendritic cells (DC), and lymphocytes, which are essential for the early invasion of tumors. The stromal cells mainly include endothelial cells, fibroblasts, mesenchymal cells, etc. How to classify patients with CRC more accurately and select appropriate treatment methods according to different subtypes is very important. At the molecular level, it is worth thinking about how to find immune-related targets and construct immune-related molecular prognostic models to improve the therapeutic effect and more accurately assess the prognosis of patients. In order to solve this problem, many scholars have made affirmative efforts and found many effective methods. Hu and other researchers used the NMF algorithm with the rank value set to 3 to classify patients with CRC, but they did not compare the differences in different typing methods when the rank value was at 2, 3, and 5 respectively. Some scholars built risk models to predict the survival of patients with CRC, but no one compared the accuracy of their respective models.

CRC is one of the tumors with poor effect in immunotherapy. How to classify patients with CRC more accurately and select appropriate treatment methods according to different subtypes is very important. At the molecular level, it is worth thinking about how to find immune-related targets and construct immune-related molecular prognostic models to improve the therapeutic effect and more accurately assess the prognosis of patients. In order to solve this problem, many scholars have made affirmative efforts and found many effective methods. Hu and other researchers used the NMF algorithm with the rank value set to 3 to classify patients with CRC, but they did not compare the differences in different typing methods when the rank value was at 2, 3, and 5 respectively. Some scholars built risk models to predict the survival of patients with CRC, but no one compared the accuracy of their respective models.

In this study, the NMF algorithm was used to perform an immune-related typing on the CRC samples of the TCGA DEGs shared by TCGA and GEO. The differences in OS and PFS of samples in subtypes with the rank value set to 2, 3, 5 respectively were compared, and it was found that the typing method with the rank value of 5 was the most statistically significant. Through the analysis of immune cell infiltration between subtypes, it was found that the infiltration content of Monocytic lineage was statistically different between the subtype groups, and the Fibroblasts had significant differences between the other subtypes except for the C1 and C3 subtypes. The overall trend of infiltrating content of them was gradually increasing from subtypes C1-C5, which suggested that our typing method could distinguish CRC patients with different infiltration content of Monocytic lineage and Fibroblasts. In the model of 14 immune-related genes constructed by us (PPARGC1A, CXCL11, PCOLCE2, GABRD, TRAF5, FOXD1, NXP4H, ALPK3, KCNJ11, NPR1, F2RL2, CD36, CCNF, DUSP14), there were 5 genes (PPARGC1A, CXCL11, NPR1, F2RL2, CCNF) had the coefficients less than 0, which were protective genes, and the remaining 9 genes (PCOLCE2, GABRD, TRAF5, FOXD1, NXP4H, ALPK3, KCNJ11, CD36, DUSP14) had the coefficients greater than 0, which were risk genes. Additionally, univariate and multivariate independent prognostic analysis showed that the immune-related risk model constructed by us could become an independent prognostic factor. Through the DCA and ROC curve analysis and comparison with models constructed by some predecessors, it was shown that the risk model constructed by us had a high prognostic accuracy. In terms of function, the results of GSEA suggested that comification, epidermal cell differentiation, and epidermis development, etc., were active in the high-risk group, while functions such as spliceosomal tri-snRNP complex assembly were active in the low-risk group, which suggested that functions such as comification might promote the progression of CRC, while spliceosomal tri-snRNP complex assembly might inhibit the progression of CRC. In terms of pathways, the results of GSEA showed that pathways such...
as complement and coagulation cascades as well as ECM-receptor-interaction were active in the high-risk group, suggesting that the activation of these pathways might promote the development of CRC. From the risk score distribution of clinical groups, whether the tumor has metastasized, whether the number of lymph node metastasis exceeds 3, and whether the tumor penetrates the visceral peritoneum, etc., were associated with the risk score of patients with CRC, while age and gender were not associated with the risk score. In addition, there was no significant difference in the risk scores of patients with CRC between Stage I and Stage II as well as between Stage III and Stage IV. Moreover, PDCD1, SERPINE1, CCDC80, etc., were significantly positively correlated with the model (p<0.05), indicating that these genes might promote the progression of CRC and increase the risk of patients with CRC. Besides, T cells, B Lineage, CD8 T cells, Endothelial cells, Monocytic lineage, etc., had a significant positive correlation with the model (p<0.05), suggesting that the infiltration of these cells might promote the development of CRC. The specific expression, mechanism of action and specific pathways of the 14 model genes in CRC need to be further verified and researched in vitro and in vivo. The correlation of genes and immune cells related to the risk model with CRC also need to be verified in vivo and in vitro.

**Conclusion**

In conclusion, the method of typing by using the NMF algorithm with the rank value set to 5 was more suitable for patients with CRC, and the 14 immune-related genes prognostic risk model constructed by us could more accurately assess the prognosis of patients with CRC, which could be used as an independent prognostic factor.

**Declarations**

**Consent for publication**

All individuals involved in this study provided consent for publication.

**Data availability**

The clinical and expression data of this study were all from TCGA and GEO databases as well as the patients’ information in the database was obtained after patients’ knowledge and consent. We can provide raw data as supplementary files.

**Competing interests**

The author(s) declare no competing interests.

**References**

1. Rabeneck, L., Han, M. C. & Senore, C. International Perspective on the Burden of Colorectal Cancer and Public Health Effects, 158, 447–452. https://doi.org/10.1053/j.gastro.2019.10.007 (2020).
2. Dekker, E., Tanis, P. J., Vleugels, J. L. A., Kasi, P. M. & Wallace, M. B. Colorectal cancer. THE LANCET, 394, 1467–1480. https://doi.org/10.1016/S0140-6736(19)32319-0 (2019).
3. Wong, S. & Yu, J. Gut microbiota in colorectal cancer: Mechanisms of action and clinical applications. NATURE REVIEWS GASTROENTEROLOGY & HEPATOLOGY, 16, 690–704. https://doi.org/10.1038/s41575-019-0209-8 (2019).
4. Bray, F. et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA-Cancer J Clin, 68 (6), 394–424. https://doi.org/10.3322/caac.21492 (2018).
5. Ahmedin, J. et al. Global cancer statistics. CA: A Cancer journal for clinician, 61, 69–90 (2011).
6. Marwan, G. F. & Metastatic Colorectal, C. (2015) Current State and Future Directions. JOURNAL OF CLINICAL ONCOLOGY, 33, 1809-1824, doi:10.1200/JCO.2014.59.7633 (2015).
7. Li, J., Luo, Y. & Pu, K. Electromagnetic Nanomedicines for Combinational Cancer Immunotherapy. ANGEWANDTE CHEMIE-INTERNATIONAL EDITION, 60, 23 https://doi.org/10.1002/anie.202008386 (2021).
8. Tan, S., Li, D. & Zhu, X. Cancer immunotherapy Pros, cons and beyond. BIOMEDICINE & PHARMACOTHERAPY, 124, https://doi.org/10.1016/j.biopharma.2020.109821 (2020).
9. Poeta, V. M., Massara, M., Capucetti, A. & Bonecchi, R. Chemokines and Chemokine Receptors: New Targets for Cancer Immunotherapy. FRONTIERS IN IMMUNOLOGY, 10, 1664–3224. https://doi.org/10.3389/fimmu.2019.00379 (2019).
10. Ciardiello, D. et al. Immunotherapy of colorectal cancer: Challenges for therapeutic efficacy. Cancer Treat Rev, 76, 22–23 https://doi.org/10.1016/j.ctrv.2019.04.003 (2019).
11. Wang, J. et al. A novel prognostic signature of immune-related genes for patients with colorectal cancer. J Cell Mol Med, 24, 8491–8504 https://doi.org/10.1111/jcmm.15443 (2020).
12. Wang, X. & Li, T. Development of a 15-gene signature for predicting prognosis in advanced colorectal cancer. BIOENGINEERED, 11 (1), 165–174 https://doi.org/10.1080/21655979.2020.1718459 (2020).
13. Chena, L. et al. Identification of biomarkers associated with diagnosis and prognosis of colorectal cancer patients based on integrated bioinformatics analysis. Gene, 692, 119–125 https://doi.org/10.1016/j.gene.2019.01.001 (2019).
14. Lu, H. et al. A Hallmark-Based Six-Gene Expression Signature to Assess Colorectal Cancer and Its Recurrence Risk. *GENETIC TESTING AND MOLECULAR BIOMARKERS*, 23 (8), 557–564 https://doi.org/10.1089/gtmb.2018.0332 (2019).

15. Tian, X. et al. Recurrence-associated gene signature optimizes recurrence-free survival prediction of colorectal cancer. *Mol. Oncol*, 11, 1544–1560 https://doi.org/10.1002/1878-0261.12117 (2017).

16. Wu, B. et al. Development of an Immune Infiltration-Related Eight-Gene Prognostic Signature in Colorectal Cancer Microenvironment. *BioMed Research International*, 2020, 43, doi:10.1155/2020/2719739 (2020).

17. Wei, R. et al. Integrative Analysis of Biomarkers Through Machine Learning Identifies Stemness Features in Colorectal Cancer. *Front. Cell Dev. Biol*, https://doi.org/10.3389/fcell.2021.724860 (2021).

18. Li, X. et al. Identification and Validation of a Six Immune-Related Genes Signature for Predicting Prognosis in Patients With Stage II Colorectal Cancer. *Front. Genet*, 12, https://doi.org/10.3389/fgene.2021.666003 (2021).

19. Wang, Y. et al. Identification of prognostic immune-related gene signature associated with tumor microenvironment of colorectal cancer. *BMC Cancer*, 21, 905 https://doi.org/10.1186/s12885-021-08629-3 (2021).

20. Jazayeri, N. & Sajedi, H. Breast cancer diagnosis based on genomic data and extreme learning machine. *SN APPLIED SCIENCES*, 2, 3, https://doi.org/10.1007/s42452-019-1789-1 (2020).

21. Yu, N. et al. Graph Regularized Robust Non-negative Matrix Factorization for Clustering and Selecting Differentially Expressed Genes. *IEEE INTERNATIONAL CONFERENCE ON BIOINFORMATICS AND BIOMEDICINE*. 1752-1756 (2017).

22. Dienstmann, R. et al. Relative contribution of clinicopathological variables, genomic markers, transcriptomic subtyping and microenvironment features for outcome prediction in stage II/III colorectal cancer. *Annals of oncology*, 30, 1622–1629 https://doi.org/10.1093/annonc/mdz287 (2019).

23. Balachandran, V. P., Gonen, M. & Smith, J. J. Nomograms in oncology: more than meets the eye. *LANCET ONCOLOGY*, 16, E173–E180 https://doi.org/10.1016/S1470-2045(14)71116-7 (2015).

24. Vickers, A. J. & Holland, F. Decision curve analysis to evaluate the clinical benefit of prediction models. *SPINE JOURNAL*, 21, 1643–1648 https://doi.org/10.1016/j.spinee.2021.02.024 (2021).

25. Sande, S. Z. et al. Statistical inference for decision curve analysis, with applications to cataract diagnosis. *STATISTICS IN MEDICINE*, 39, 2980–3002 https://doi.org/10.1002/sim.8588 (2020).

26. Kumar, R. et al. Receiver Operating Characteristic (ROC) Curve for Medical Researchers. *INDIAN PEDIATRICS*, 48, 277–287 https://doi.org/10.1007/s13312-011-0055-4 (2011).

27. Subramanian, A. et al. GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Gene expression*, 23, 3251–3253 https://doi.org/10.1093/bioinformatics/btm369 (2007).

28. Zito, A. et al. Gene Set Enrichment Analysis of Interaction Networks Weighted by Node Centrality. *Front. Genet*, 12, https://doi.org/10.3389/fgene.2021.577623 (2021).

29. Pitt, J. M. et al. Targeting the tumor microenvironment: removing obstruction to anticancer immune responses and immunotherapy. *ANNALS OF ONCOLOGY*, 27, 1482–1492 https://doi.org/10.1093/annonc/mdw168 (2016).

30. Roma, R. C. et al. Targeting Tumor Microenvironment for Cancer Therapy. *INTERNATIONAL JOURNAL OF MOLECULAR SCIENCES*, 20, https://doi.org/10.3390/ijms20040840 (2019).

31. Gallo, G. et al. Therapeutic Targets and Tumor Microenvironment in Colorectal Cancer. *J Clin Med*, 10 (11), 2295 https://doi.org/10.3390/jm20040840 (2021).

32. Wu, X. et al. Construction of a New Tumor Immunity-Related Signature to Assess and Classify the Prognostic Risk of Colorectal Cancer. *INTERNATIONAL JOURNAL OF GENERAL MEDICINE*, 14, 6661–6676 https://doi.org/10.2147/IJGM.S325511 (2021).

33. Qian, Y. et al. Prognostic Cancer Gene Expression Signatures: Current Status and Challenges. *CELLS*, 10 (3), https://doi.org/10.3390/cells10030648 (2021).

34. Ahluwalia, P. et al. Identification and Clinical Validation of a Novel 4 Gene-Signature with Prognostic Utility in Colorectal Cancer. *INTERNATIONAL JOURNAL OF MOLECULAR SCIENCES*, 20 (15), https://doi.org/10.3390/ijms20153818 (2019).

35. Ahluwalia, P. K. et al. Exploring prognostic and immunological impact of a novel 4 gene signature in colorectal cancer patients. *CANCER RESEARCH*, 80 (16), https://doi.org/10.1158/1538-7445.AM2020-4334 (2020).

36. Bai, J. et al. Identification of prognostic immune-related signature predicting the overall survival for colorectal cancer. *EUROPEAN REVIEW FOR MEDICAL AND PHARMACOLOGICAL SCIENCES*, 24, 1134–1141 https://doi.org/10.26355/eurrev_202002_20164 (2020).

37. Jiang, C. et al. Identification of a novel 15-gene expression signature predicting overall survival of human colorectal cancer. *CLINICAL AND TRANSLATIONAL MEDICINE*, 10 (8), https://doi.org/10.26355/eurrev_202002_20164 (2020).

38. Zuo, S. et al. Identification of a 6-gene signature predicting prognosis for colorectal cancer. *CANCER CELL INTERNATIONAL*, 19, 6, https://doi.org/10.1186/s12935-018-0724-7 (2019).

39. Qian, Y. et al. Prognostic Risk Model of Immune-Related Genes in Colorectal Cancer. *FRONTIERS IN GENETICS*, 12, https://doi.org/10.3389/fgene.2021.619611 (2021).
40. Li, M. *et al.* Identification and validation of an immune prognostic signature in colorectal cancer. *INTERNATIONAL IMMUNOPHARMACOLOGY, 88*, https://doi.org/doi:10.1016/j.intimp.2020.106868 (2020).

**Figures**

**Figure 1**

Flow chart of the main steps of this study. This study mainly consists of three parts: the blue part - mainly for data screening, the green part - mainly for the typing of NMF and the typing comparison, and the red part - mainly for the construction, analysis and comparison of prognostic risk models.
Figure 2

A-J. Comparison of typing methods with the rank value set to 2, 3 and 5 respectively: A. The cophenetic graph with the rank value set to 2:10 showed that the 5-6 and 3-4 segment have a larger slope, and combining the results of 9 heatmaps, the typing methods with rank value set to 2, 3 and 5 respectively were selected for comparison; B-D. The heatmap with the rank value set to 2, the OS survival analysis between the 2 subtypes, and the PFS survival analysis between the 2 subtypes; E-G. The heatmap with the rank value set to 3, the OS survival analysis among the 3 subtypes, and the PFS survival analysis among the 3 subtypes; H-J. The heatmap with the rank value set to 5, the OS survival analysis among the 5 subtypes, and the PFS survival analysis among the 5 subtypes. K-M. Comparison between each typing method and the predecessor's pan-cancer typing (the right side is the predecessor's typing): K. The comparison between typing with the rank set to 2 and others; L. The comparison between typing with the rank set to 3 and others; M. the comparison with the rank set to 5 and others.
Figure 3

Results of the MCPcounter immune cell infiltration analysis. Differences of immune cells of CD8 T cells, Endothelial cells, Fibroblasts, Monocytic lineage, Myeloid dendritic cells, Neutrophils, NK cells, Cytotoxic lymphocytes, B lineage, and T cells shown in the different subtypes.
Figure 4

A-B. Graphs of Lasso regression and cross-validation: A. The change trajectory of each variable, where abscissa represents the logarithmic value of the independent variable lambda, and the ordinate represents the coefficient of the independent variable; B. The Cross-validation graph, where the confidence intervals for different log(λ) were shown. C-J. The survival analysis and ROC curve analysis between the high and low-risk groups in the train group, the test group, all the TCGA samples group, and the GEO group: C-F. The survival curve of train group, test group, all TCGA samples group, and GEO group in sequence; G-J. The ROC curve of train group, test group, all TCGA samples group, and GEO group in sequence. The results showed that after grouping according to the risk model, there were significant differences in the survival of patients between the high and low-risk groups. Additionally, the AUC value in the ROC curve showed that the prediction effect of the risk score model was good.
Figure 5

A-J. Survival analysis between high and low-risk groups after clinical traits grouping. The patients were grouped according to gender, T, N, M, and Stage, and the Cox survival analysis was performed on the high and low-risk groups in each group. K-L. Nomogram and calibration curve: K. The Nomogram. It could be used to score each clinical trait and high or low-risk to obtain the total score, and according to the total score, the survival rate of patients at 1, 3, and 5 years could be found; L. The calibration curve. The closer each calibration curve is to the gray line in the figure, the higher the accuracy of the prediction is. M-N. Decision Curve Analysis (DCA) and ROC curve analysis: M. DCA. In the figure, "All" represented the curve assuming that all samples were positive, "None" represented the curve assuming that all samples were negative, and "Risk" represented the curve of the risk model. The further away from the purple curve, the stronger the ability of survival prediction; N. ROC curve analysis. "Risk" represented the curve of the risk model. The larger the Area Under The Curve (AUC), the higher the accuracy of survival prediction.
Figure 6

A-L. Survival analysis and ROC curve analysis of each prognostic risk model: A-F. The survival analysis grouped by our risk model as well as risk models of Huiqi, Jun, Libo, Xiaolong, and Xiao respectively; G-L. The ROC curve analysis performed on our risk model as well as risk models of Huiqi, Jun, Libo, Xiaolong, and Xiao respectively. M-N. C-index analysis results of respective risk models and the RMS graph: M. C-index analysis of the respective risk models. The higher the score, the higher the prognostic accuracy of the model; N. The RMS graph of the respective risk models. The results showed that the prognostic accuracy of our risk model was high, which was statistically significant (p<0.001).
Figure 7

A-D. GSEA result graph: A. Related functions that were active in the high-risk group; B. Related functions that were active in the low-risk group; C. Related pathways that were active in the high-risk group; D. No obvious active pathway in the low-risk group. E-J. Distribution of samples risk scores after grouping according to each clinical trait: The analysis for the distribution of risk scores of samples in each group after grouping by Age, Gender, T, N, M, and Stage respectively. K-M. Genes, immune cells related to the risk model and correlation analysis of TMB: K. The genes related to the risk model. The genes with * in the first row were related to the risk model (p<0.05). Red indicated a positive correlation, while blue indicated a negative correlation; L. The immune cells associated with the risk model; M. The TMB correlation analysis. Red indicated a positive correlation, while green indicated a negative correlation.