Colon Cancer Classification and Prognosis Prediction Based on Genomics Multi-features

Jiasheng Xu
Nanchang University Second Affiliated Hospital
Kaili Liao
Nanchang University Second Affiliated Hospital
Chengfeng Wu
Nanchang University Second Affiliated Hospital
Qijun Yang
Nanchang University
Hongping Wan
Nanchang University Second Affiliated Hospital
Xiaozhong Wang (xiaozhongwangncu@163.com)
the Second Affiliated Hospital of nanchang university

Research

Keywords: genome, multiple characteristics, colon cancer, molecular subtype, prognosis prediction

Posted Date: September 14th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-837390/v1

Abstract

Background: To classify colon cancer and predict the prognosis of patients with multiple characteristics of the genome.

Methods: We used the mRNA expression profile data and mutation maf files of colon cancer patients in the TCGA database to calculate the TMB value of patients. Combined with CNV, MSI, and corresponding clinical information, the patients were clustered by the "K-means" method to identify different molecular subtypes of colon cancer. Comparing the differences of prognosis, and immune cell infiltration, and other indicators among patients in each subgroup, we used COX and lasso regression analysis to screen out the prognosis difference genes among subgroups and construct the prognosis prediction model. We used the external data set to verify the model, and carried out the hierarchical analysis of the model to compare the immune infiltration of patients in the high and low-risk groups. And detected the expression differences of core genes in tumor tissues of patients with different clinical stages by qPCR and immunohistochemistry.

Results: We successfully calculated the TMB value and divided the patients into three subgroups. The prognosis of the second subgroup was significantly different from the other two groups. The mmunoinfiltration analysis showed that the expression of NK.cells.resting increased in cluster1 and cluster 3, and the expression of T.cells.CD4.memory.resting increased in cluster3. By analyzing the differences among subgroups, we screened out eight core genes related to prognoses, such as HYAL1, SPINK4, EREG, and successfully constructed a patient prognosis evaluation model. The test results of the external data set shows that the model can accurately predict the prognosis of patients; Compared with risk factors such as TNM stage and age, the risk score of the model has higher evaluation efficiency. The experimental results confirmed that the differential expression of eight core genes was basically consistent with the model evaluation results.

Conclusion: Colon cancer patients were further divided into three subtypes by using genomic multi-features, and eight-core genes related to prognosis were screened out and the prognosis evaluation model was successfully constructed. With external data and experiments, it verified that the model had good evaluation efficiency.

Background

Colon cancer is the third most common cancer in the world, which seriously threatens human health [1–2]. With the progress of surgical techniques and the emergence of new therapeutic drugs, the 5-year survival rate of colon cancer has gradually increased, but colon cancer is still one of the important causes of cancer-related deaths. In recent years, the mortality of colon cancer has dropped significantly which mainly due to the progress of adjuvant therapy technology. The choice of adjuvant therapy is mainly based on TNM staging of the tumor, which is the most valuable staging method for judging prognosis at present, and the colorectal staging is mainly based on the pathology and anatomy of the tumor [1]. However, TNM staging can not effectively distinguish the differences in tumor biological behavior caused by tumor heterogeneity, which leads to the lack of accuracy of adjuvant therapy based on TNM staging.

Colon cancer is highly heterogeneous, and its biological behavior has wide genetic and epigenetic differences among different individuals or different lesions within the same individual, which can further lead to different prognosis and response to treatment. Because of the differences in biological behaviors of tumors, the prognosis and response to adjuvant therapy of colon cancer patients in the same TNM stage are quite different. Some studies have pointed out that there are differences in anatomical features, histological features, prognosis, and tumor biological behavior between left colon cancer and right colon cancer. According to the National Comprehensive cancer network Guidelines, patients with metastatic unresectable colon cancer have different therapeutic effects on cetuximab due to the mutation status of the RAS gene. Because of the clinical heterogeneity of colon cancer, it is always being a difficult problem to provide accurate and effective treatment for individual differences of patients. Classifying colon cancer at the molecular level, and understanding the molecular change characteristics of each subtype and its corresponding tumor biological behavior characteristics have important guiding significance for the treatment of colon cancer. In the era of precise treatment, there is an urgent need for a method to classify colon cancer according to tumor heterogeneity and classify tumors at the molecular biological level.

The occurrence of tumors is closely related to the accumulation of gene mutation. In 1914, CALKINS et al. [3] first discovered that abnormal chromosome distribution may be related to the malignant tumor during cancer cell division, and then people began to explore the relationship between abnormal genetic material and tumor occurrence. Tumor genomes carry a variety of mutations, including single nucleotide variations, structural rearrangement, and copy number variations, CNVs). According to the number of gene mutations per million bases, we can get the tumor mutation load value. Studies have confirmed that patients can benefit from immunotherapy when TMB (Tumor Mutation Burden) is greater than 10. In addition, the change of gene expression and the mutation of microsatellite instability genes in colon cancer are closely related to the choice of treatment methods and prognosis of patients. Therefore, the comprehensive summary analysis of these features in the genome of colon cancer patients is conducive to the classification of colon cancer, so as to further stratify and accurately treat colon cancer patients at the molecular level and improve the prognosis of patients.

With the accumulation of biological data and the maturity of data mining technology, the application of advanced calculation methods is helpful to build a multi-factor comprehensive scoring system and solve the above problems. Therefore, this study integrates the mRNA expression profile data, mutation maf file, CNV information, MSI information, and corresponding clinical information of colon cancer patients, and calculates the TMB of patients. The patients were clustered by multi-feature indexes of the genome, and different molecular subtypes of colon cancer were identified, and the patients were divided into different subgroups. Comparing the differences of prognosis, immune cell infiltration, and other indicators among patients in each subgroup, further screening the differential genes among subgroups with significant differences in prognosis by regression analysis and constructing a prognosis prediction

model. The external data set was used to verify the model, and the hierarchical analysis of the model was carried out. The difference in immune infiltration and tumor purity of patients in high and low-risk groups were calculated, and established the Nomogram model to evaluate the prognosis of patients.

Materials And Methods

2.1 Object of study

We downloaded the mRNA expression profile data, mutated maf files, CNV information, and corresponding clinical information of patients with colon cancer (COAD) from the database of The Cancer Genome Atlas (tcga, https://tcga-data.nci.nih.gov/tcga/). There are 379 samples containing CNV information, mutation information, MSI information, and mRNA matrix information, among which 359 samples have complete survival information. In addition, we also downloaded the data set numbered GSE17536 from the Gene Expression Omnibus (GEO database, https://www.ncbi.nlm.nih.gov/geo/) database. the data set contains 177 colon cancer patients, of which 172 patients contain complete survival information, and the samples were analyzed using Affymetrix human genome u133 plus 2.0 array platform

2.2 Calculation and Cluster Analysis of Tumor Mutation Load (TMB)

We use the "maftools" function package in R language to calculate the tumor mutation load (TMB) of each sample according to the maf file. We use the "factoextra" function package of the R language, based on the "K-mean" method, and integrate CNV, TMB, and MSI data to cluster 379 samples. At the same time, the principal component analysis is carried out on the samples, and the results of clustering analysis are compared with those of subcomponent analysis to see if the results are consistent. For the obtained subgroups, do survival analysis between each cluster, and judge whether the differences among subgroups are significant. Compare CNV, TMB, and MSI among each subgroup, and visualize the results.

2.3 Immune differences between subtypes

Cibersort R package was used to analyze the immune cell infiltration of each subgroup, and the difference of immune cell infiltration between each group was compared. The difference of expression content of each immune cell in each group was compared and analyzed, and the results were displayed visually. The expressions of CTLA4, PDL1, LAG3, TIGIT, IDO1, TDO2, and other classical immune checkpoints in each subgroup were compared and analyzed.

2.4 Screening and enrichment analysis of different genes among subtypes

According to the difference of survival analysis between groups, we divided the groups into two groups: good prognosis and poor prognosis. Then, the limma[PMID: 25605792] function package of r language (version3.5.2, the same below) is used to analyze the difference of gene expression between the two subgroups, and the absolute value of Log2FC is greater than 1 and FDR $\leq$ 0.05 is used as the standard to screen the differentially expressed genes. Cluster profiler R package was used for enrichment analysis, and online analysis tools of string (HTTPS: / / www.string-db. org /) and metascape (http: / / metascape. Org /) were used for protein interaction analysis.

2.5 Regression analysis to construct prognosis model

Univariate Cox regression analysis was performed on colon cancer samples based on the expression value of differential genes, and the differential expression genes significantly related to the prognosis of colon cancer were screened with $P < 0.05$ as the threshold. Then LASSO Cox regression analysis was carried out with glmnet package of r language [PMID: 20808728] to further select differentially expressed genes related to the prognosis of colon cancer. And use the screened differentially expressed genes to calculate the Risk Score of each sample according to the following formula:

2.6 Coefi is the risk coefficient of each factor calculated by the LASSO-Cox model, Xi is the expression value of each factor, and in this study, it refers to the expression value of mRNA. Then, r package survival, survminer, and bilateral log-rank test were used to determine the optimal cutoff value of the Risk score. according to the cutoff value, patients were divided into the Low-Risk group and the High-Risk group.

2.7 Verification of model by external data set

The time-dependent receiver operating characteristics (ROC) curve of the model was drawn by the r language survivalROC package [PMID: 10877287]. TCGA and GSE17536 data sets are used as verification sets to verify the model. the model is used to calculate the Risk scores of patients in the data sets, and the survminer package is used to find the best cutoff value so that patients can be divided into high and low-risk groups. the r language survival package is used to estimate the overall survival rate of different groups based on the Kaplan-Meier method, and the log-rank test is used to test the significance of survival rate difference between different groups. A Multivariate Cox regression model was used to analyze whether the Risk Score can predict the survival of colon cancer patients independently of other factors.

2.8 Hierarchical analysis of the model

The model was analyzed hierarchically to explore the relationship between age, gender, clinical stage, and the prognosis of patients. According to clinical stage, gender, age over 65, and other criteria, patients were divided into two groups by using this model, and the survival rate of patients with different clinical characteristics was analyzed.

2.9 Difference of immune cell infiltration between high and low-risk groups and calculation of tumor purity

We use the software CIBERSORT [PMID: 25822800] to calculate the relative proportion of 22 immune cells in each cancer sample. CICERSORT software will use the deconvolution algorithm to characterize the composition of immune infiltrating cells with preset 547 barcode genes according to the gene expression matrix. The sum of all estimated immune cell types in each sample is equal to 1. The estimate function package of r language [PMID: 24113773] was used to calculate the tumor purity of each cancer sample.

2.10 Establishing Nomogram prognosis prediction model

Nomogram is widely used to predict the prognosis of cancer. To predict the survival probability of patients in 1 year, 3 years, and 5 years, based on all independent prognostic factors determined by multivariate Cox regression analysis, we use R language rms package to establish nomogram, draw the calibration curve of the nomogram, and observe the relationship between nomogram prediction probability and actual incidence rate.

2.11 Clinical sample verification

100 cases of patients with stage I and stage III colon cancer(50 cases at each stage) were screened out from the First Affiliated Hospital of Nanchang University from June 2017 to June 2019. Inclusion criteria: (1) Pathologically confirmed (2) New cases diagnosed by this hospital for the first time. Exclusion criteria: (1) with other malignant tumors or diabetes, hypertension Systemic diseases such as cardiovascular and cerebrovascular diseases; (2) patients who have received radiotherapy, chemotherapy, or other anti-neoplastic drugs before surgery

Tumor tissue and normal specimens(Tissues more than 5 cm from the outermost periphery of tumor tissues are regarded as normal tissues)from patients surgically were collected. GAPDH was used as the internal reference, the expression of selected RNA was detected by qPCR. Primers were shown in Table 1. Using the expression level of the GAPDH as the standard value "1", the relative expression levels of 8 RNAs in cancer tissues of a patient with stage I and III were calculated and draw statistical charts. The expression of prognostic-related genes in tumor tissues of patients with I and III stages of colon cancer was detected by immunohistochemistry.

Table 1
qPCR primers for 8 core genes related to prognosis

Gene	HYAL1	SPINK4	EREG	VWDE
Forward Primer	TATGGCCCAAGGCTTTAGGG	CAGTGGGTAATCGCCCTGG	GTGATTCCATCATGTATCCCAGG	ACTGCTGTCTCTTTCAAATCCC
Reverse Primer	GCTACCACATCGAAGACACTGA	CACAGATGGGCATTCTTGAGAAA	GCCATTCATGTCAGAGCTACACT	CCGCACAGTAGCCCATACAT
Gene	ITLN1	ZBTB7C	KLK12	B3GNT6
Forward Primer	ACGTGCCCAATAAGTCCCC	CGCCACCCTTCCCTAATGAC	CTGTGTGTTCTTGGGCTCAG	GTGCGCCGCCTCTTTCTATT
Reverse Primer	CCGTTGTCAGTCCAACACTTTC	GATAGGCACCGTAGTCGTTCT	CTGCCACGGCTGTGAGTTA	CCAGCCAGTCGAGCAAGTG

2.12 Statistical analysis

Kaplan-Meier method was used to estimate the overall survival rate of different groups, and log-rank was used to test the significant difference of survival rate among different groups Wilcoxon signed rank-sum test was used to compare the infiltration differences of immune cells in different groups, with $p < 0.05$ as the significant threshold. R software, version number v3.5.2, is used for statistical analysis.

Result

3.1 Integration and clustering of 1.TMB, MSI, and CNV information

We downloaded CNV, mutation maf file, and MSI information of COAD in TCGA database, and the corresponding clinical information. There were 379 samples with CNV information, mutation information, MSI information, and mRNA matrix information, of which 359 had complete survival information. According to the maf file, the TMB (Fig. 1A, Supplemented Table 1) of each sample is calculated first, and then three kinds of data are integrated to cluster 379 samples using the "factoextra" function package. samples can be clustered into three categories (Supplemented Table 2). PCA analysis shows that the samples can be well divided into three categories (Fig. 1B), and PCA results are visualized in three dimensions (Fig. 1C). Comparing PCA results with cluster analysis results, it is found that the classification results are consistent, which shows that our clustering is more accurate. The survival analysis among the three clusters shows that there are significant differences among the subgroups (Fig. 1D). A comparative analysis of TMB and MSI among the three types of samples shows that the TMB of Cluster3 is significantly higher than that of the other two types of samples, and the proportion of MSI-H samples in Cluster3 is also significantly higher than that of the other two types of samples (Fig. 1E-F).

Table 2
The best cutoff value of 8 prognostic gene

Gene	Cutpoint	Statistic
HYAL1	6.247927513	2.575825733
SPINK4	8.243173983	3.735550973
EREG	6.022367813	2.601903964
VWDE	4.247927513	3.885960058
ITLN1	8.592457037	3.467487375
ZBTB7C	9.317412614	2.644323788
KLK12	3.169925001	2.730170447
B3GNT6	6.285402219	3.632614081

3.2 Analysis of characteristic differences and immune infiltration among subgroups

The mutations among the three subgroups were analyzed, and the results were shown in Fig. 2A, Fig. 2B and Fig. 2C respectively. Immunoinfiltration analysis was carried out on the samples of three subtypes, and the display diagram of immune infiltration cells among the three types of samples is shown in Fig. 3A-C. Furthermore, the infiltration content of each immune cell in three groups of samples was compared. The results showed that compared with cluster2, cluster3 expression content of Macrophages.M0 in Cluster3 decreased, the expression content of Macrophages.M1 increased, and the expression content of Macrophages.M2 in cluster 1 decreased. Compared with cluster3, the expression of NK. Cells. activated in cluster3,cluster1 decreased, the expression of Plasma. cells increased, and the expression of t. cells. folliculous. helper decreased in cluster1. Compared with cluster 2, the expression level of NK. cells. restoring is higher in cluster1 and cluster 3, and the expression level of t. cells. CD4. memory. restoring is higher in cluster3 (Fig. 4A-D). Comparing and analyzing the expression of several important immune checkpoints in three groups of samples, the results showed that compared with cluster3, CTLA4 expression content of CTLA4, PDL1, LAG3, TIGIT, and IDO1 genes in cluster1 and cluster2 samples decreased, and the difference was statistically significant. The expression of TDO2 had no significant difference among the three groups (Fig. 5A-F).

3.3 Screening and enrichment analysis results of differential genes

According to the analysis of the difference of prognosis among the three subgroups, we found that the prognosis of patients with cluster1 and 3 was worse than that of patients with cluster2, and the difference was statistically significant; There is no significant difference in prognosis between cluster1 and cluster3 patients. Therefore, we combined the patients of cluster1 and cluster3 and analyzed the differential genes with those of cluster2. Comparing the gene expression differences between Cluster1 + Cluster3 samples and Cluster2 samples, we got 128 differentially expressed genes, including 51 up-regulated genes and 77 down-regulated genes (Fig. 5G). GO enrichment analysis is carried out on the differential genes, The results showed that the differential genes were mainly related to anion transport, Negative Regulation of Cell Proliferation, Multicellular Organic Homeostasis, Humoral Immunization Response, Organic anion transport, positive regulation of defense response, epigenetic cell differentiation and tissue homeostasis are related to biological activities (Fig. 5H-I). KEGG enrichment analysis showed that the differential genes were mainly enriched in alpha-Linolenic acid metabolism, linolenic acid metabolism, Fat digestion and absorption, Ether lipid metabolism, Arachidonic acid metabolism, Wnt signaling pathway, and etherlipid metabolism (Fig. 5J). The protein-protein interaction analysis results are shown in Fig. 5K-M. The results show that these differential genes can be divided into four interaction modules. CCK、GPR143、TAC1、NPFFR1、GNG4、MUC5A、MUC5B、MUC6、DRD2、KY6G6D and other genes are the core genes and interact most closely with other targets.

Comparing the difference in the expression of differential genes between the two groups, it was found that the expression of differential genes between Cluster1 + Cluster3 samples and Cluster2 samples was significantly different (Fig. 6A).

3.4 Construction and verification of prognosis model

Univariate TCGA regression analysis was carried out with 128 differential gene expression values as continuous variables, and the Hazard ratio(HR) of each gene was calculated. Eight genes were obtained by screening with P value < 0.05 as the threshold, and the survival analysis of these eight genes was carried out (Fig. 6B-Fig. 6I). The best cutoff value is shown in Table 2. COX regression analysis showed that the HR value of 7 of these 8 genes was less than 1, which was beneficial to the prognosis. There is one gene whose HR value is greater than 1, which is a dangerous gene and unfavorable to prognosis (Fig. 7A). Then, the selected eight genes were analyzed by LASSO regression. according to the lambda values corresponding to different gene numbers in LASSO analysis, we determined that the optimal number of genes was eight (Fig. 7B, lambda value was the smallest), and the eight genes were HYAL1, SPINK4, EREG, VWDE, ITLN1, ZBTB7C, KLK12, B3GNT6.

According to the expression level of each gene and the regression coefficient of LASSO Cox regression analysis, a risk-scoring model for predicting the survival of patients is established,

RiskScore=(-0.0725*HYAL1)+(-0.0151*SPINK4)+(-0.1859*EREG)+(0.1787*VWDE)+(-0.0226*ITLN1)+(-0.0926*ZBTB7 C)+(-0.0231*KLK12)+(-0.0000407*B3GNT6)。 We calculated the risk score of each patient and divided the samples of TCGA data set and GEO verification set into high-risk group

and low-risk group according to the median. Survival analysis found that in TCGA and GEO data sets, the overall survival of high-risk colon cancer samples was worse than that of low-risk patients (Fig. 7C and Fig. 7D). It shows that the risk model can effectively predict the prognosis of colon cancer patients in both data sets. Generally speaking, the results show that the Risk Score calculated by the evaluation model constructed by HYAL1, SPINK4, EREG, VWDE, ITLN1, ZBTB7C, KLK12, and B3GNT6 can better predict the prognosis of colon cancer patients.

3.5 Risk Score is an independent prognostic marker of colon cancer

To further explore the prognostic value of Risk Score in colon cancer samples with different clinicopathological factors (including age, TNM Stage and gender), we regrouped colon cancer patients according to these factors and performed Kaplan-Meier survival analysis. It was found that among patients with Stage I + StageII and StageIII + StageIV, patients with low risk score had better prognosis than those with high risk score (Fig. 7E-F). Among patients aged > 67 and ≤ 67, patients with low risk score had better prognosis than those with high risk score (Fig. 7G-H). In female and male patients, the overall survival rate of was significantly lower in the high-risk group than in the low-risk group (Fig. 7I-J). The difference was statistically significant. These results indicate that the Risk Score can be used as an independent index to predict the prognosis of colon cancer patients.

Then, we included age, MSI classification, TNM Stage, gender and Risk Score for multivariate Cox regression analysis to determine whether the Risk Score is an independent prognostic indicator. The result is shown in Fig. 8A. We found that the Risk Score, TNM Stage and age were still significantly correlated with overall survival, and the samples with high risk score had a higher death risk and were adverse prognostic factors (HR = 2.63, 95%CI:1.83–3.8, P< 0.001).

3.6.Nomogram model can better predict the prognosis and survival of patients

The Nomogram model (Fig. 8B) was successfully constructed by using three independent prognostic factors: age, TNM Stage and Risk Score. For each patient, draw up three lines to determine the Points obtained from each factor in Nomogram. The sum of these Points is located on the "Total Points" axis, and then a line is drawn down from the “Total Points” axis to determine the survival probability of colon cancer patients for 1, 3 and 5 years. The corrected curve in the calibration chart is close to the ideal curve (a 45-degree line passing through the origin of the coordinate axis with slope 1). The result shows that the predicted result of the model is basically consistent with the actual result (Fig. 8C-E).

3.7. Immune status of colon cancer patients in high and low risk groups

CIBERSORT method and LM22 feature matrix were used to estimate the difference of immune infiltration among 22 kinds of immune cells in colon cancer patients with high and low risk groups, and the result was visualized (Fig. 9A-B). The result shows that there was a significant difference in the infiltration ratio of T cell CD4 memory resting immune cells between high and low risk groups, and its expression level was significantly higher in low risk groups which suggesting that the expression of T cell CD4 memory resting is related to the improvement of prognosis.

3.8.qPCR and immunohistochemical verification

There was no significant difference in baseline data between patients with Stage I and III (Table 3). Results of real-time quantitative PCR detection of selected 8 genes listed in Fig. 9C. The expression of HYAL1、SPINK4、EREG、ITLN1、ZBTB7C、KLK12、B3GNT6 were lower in tumor tissues of Stage III patients than in Stage I. Compared with Stage I patients, the expression of VWDE was higher in Stage III patients. Through immunohistochemical detection, we found that compared with patients with Stage I colon cancer, VWDE was expressed in higher levels in the tumor tissues of patients with Stage III colon cancer; Compared with patients with Stage I colon cancer, HYAL1、SPINK4、ITLN1、KLK12 and B3GNT6 were lower in tumor tissues of patients with Stage III colon cancer (Fig. 10). The experimental verification results were basically consistent with the previous data analysis results.

Table 3
Comparison of baseline data of patients with stage I and stage III colon cancer

Group/Character	age	male	female	Pathological type
stage I	63.5 ± 2.5	26	24	Adenocarcinoma
stage III	65.1 ± 1.5	28	22	Adenocarcinoma
P value	> 0.05	> 0.05		

Discussion

As a common tumor, the incidence of colon cancer is high and increasing year by year. Actively exploring the prevention, treatment and prognosis evaluation of colon cancer is of great significance to reduce its morbidity and mortality. In recent years, with the development of tumor molecular biology, researchers have gradually realized that tumor is also a genetic disease, and its occurrence and development are the result of long-term interaction between genes and environment [4]. The evaluation of tumor risk factors and genetic predisposing factors helps to further clarify the mechanism of cancer and contribute to clinical treatment. Because of the high heterogeneity among tumor patients, the traditional pathological and clinical staging can no longer better predict the prognosis and guide the treatment for patients. In order to predict the prognosis of individual patients accurately, it is necessary to stratify patients in more detail and precision at the genomic level, and divide them into different tumor subtypes, so as to carry out more targeted anti-tumor treatment. It is of great significance to find out the types of cancer subtypes accurately, whether it is the understanding of cancer, the treatment of cancer or the practical clinical application.

Researchers found that the prognosis of patients with the same cancer type can be quite different, and even patients of the same pathological type have significant differences in prognosis [5–7]. Through NGS sequencing and related research on a large number of tumor tissue samples from colon cancer patients, We found that the prognosis is related to patients' TMB, MSI, CNV and other genomic characteristics [8–15]. Taieb et al[8] investigated the mutations frequence and prognostic value of BRAF and KRAS in MSI and MSS Stage III Colon Cancer.They confirmed significant differences in mutation frequency of BRAF and KRAS and prognosis of patients in MSI-H and MSS. Chouhan et al[12] found the interaction between BRAF mutation and microsatellite instability (MSI) status were very important in determining survival outcomes after adjuvant 5FU based chemotherapy in Stage III colon cancer.Researchers had confirmed the correction between TMB and prognosis in patients with colorectal cancer treated with adjuvant Fluoropyrimidine and Oxaliplatin. Studies also found that TMB was associated with the treatment outcomes in patients with colorectal cancer[13–14]. Wang et al[15] studied the genome-wide expression profiling-based copy number variations and colorectal cancer risk in Chinese.Their study suggested that the amplified copy number of HLA-DQB1 is associated with lower risk of colorectal cancer and able to induce the apoptosis of colon cancer cells.

Generally speaking, patients with TMB greater than 10 can benefit from immunotherapy, thus improving the prognosis. MSI mutation also plays an important role in guiding the selection of clinical chemotherapy drugs. A number of evidences also showed that the differences of patients' prognosis can be analyzed more accurately from the perspective of genomic differences, and colon cancer patients can be further divided into different hypotypes in detail at the genetic level, so as to explore the potential mechanisms and predict the prognosis of patients.

Therefore, we decided to analyze the genome sequencing data of colon cancer patients and calculate the TMB value of each patient; Furthermore, combined with CNV, MSI mutation and other genomic characteristic data indicators for comprehensive analysis, so as to use these genomic characteristics to subgroup colon cancer patients at the gene level.We analyzed the differences of prognosis and immune cell infiltration between the three groups wuth different genomic characteristics. We found that compared with the patients in the second subgroup, the prognosis of first and third one was worse, and the difference was statistically significant. However, the prognosis curves of subgroup 1 and subgroup 3 were close, and there was no significant difference in prognosis between the two groups. Therefore, when using COX and LASSO regression analysis, we combined the patients in subgroup 1 and subgroup 3 first, and then compared them with those in subgroup 2 to screen the differentially expressed genes related to prognosis. We eventualky identified eight core genes related to prognosis: HYAL1, SPINK4, EREG, VWDE, ITLN1, ZBTB7C, KLK12, B3GNT6, and used them to construct a prognostic evaluation model for colon cancer patients.

Jin et al [16] found that HYAL1 and HYAL2 played a suppressive role in the metastasis of colorectal cancer. The expression of hyaluronan synthase2 or hyaluronidase1 differentially could affect the growth rate of transplantable colon carcinoma cell tumors [17]. Wang et al [18] confirmed that down-regulated SPINK4 was associated with poor survival in colorectal cancer. Another study demonstrated [19] that the serum SPINK4 level increased in CRC and was associated with the location and distant metastasis of CRC. They thought SPINK4 had a high diagnostic value in CRC but was not associated with the survival of CRC patients. Previous studies found [20] miR-215-5p-EREG/TYMS axis in colon cancer cells to produce resistance to 5-FU. Study showed that high EREG expression was predictive of better outcomes in rectal cancer patients receiving neoadjuvant concurrent chemoradiotherapy [21]. Qu et al revealed the activation of EGFR was depended on demethylation of the EREG promoter by integrated genomic analysis of colorectal cancer progression [22].

Human Intelectin-1 (ITLN-1) is a novel identified galactose-binding lectin that is expressed in the colonic goblet cells. Study has confirmed that the aberrant ITLN-1 expression in gastric cancer is correlated with clinicopathological features and may be a useful prognostic factor for predicting the outcomes of gastric cancer patients [23]. Yousef et al reported that the expression of KLK12 was down-regulated at the mRNA level in breast cancer tissues and was up-regulated by steroid hormones in breast and prostate cancer cell lines. This gene may be involved in the pathogenesis and/or progression of certain cancer types and may find applicability as a novel cancer biomarker [24].

We independently verified the model by using GEO and TCGA datasets to confirm its accuracy and credibility. Then we analyzed the relationship between age, sex, clinical stage, model risk score and the prognosis. The results showed that besides TNM staging and age, the risk score of the model was also significantly correlated with the overall survival time of colon cancer patients, which further affirmed the effectiveness of our model in evaluating the prognosis of patients. By analyzing and comparing the immune cell infiltration of patients in high and low risk groups divided by the model, it was found that the expression of T cell CD4 memory resting decreased in the high risk group. This suggests that this cell may be related to the prognosis of patients. Previously, Lin et al's research [25] found that activated T(reg) cells secreted significantly lower levels of effector cytokines (interleukin-2, tumor necrosis factor-α and interferon-γ) than did resting T(reg) cells and nonsuppressive cells upon ex vivo stimulation. Activated, but not resting, T(reg) cells in cancer tissue correlated with tumor metastases. They confirmed that activated T(reg) cells are a distinct subgroup with effector memory phenotype and fully functional regulatory activity against human colonic cancer immunity. This showed that during the development of colon cancer, the infiltration content of activated T cell CD4 gradually increased with the disease progression, and correspondingly, the content of another subtype T cell CD4 memory resting gradually decreased with the disease progression. This is consistent with our research results.

For the first time, our research explores the correlation between multiple genomic characteristics and the prognosis of colon cancer patients from the perspective of genome, which brings fresh ideas to researchers. In the past, many studies used the expression data of gene transcriptome to construct the prognosis model and evaluate the prognosis of colon cancer patients, but the expression of some transcriptome genes is rarely detected in the current clinical treatment. Therefore, it is difficult for other researchers to obtain more data about the expression of these genes to verify the validity of the model or to expand the sample size for further study. Compared with gene transcriptome sequencing in tumor tissues, the more routine clinical examination for colon cancer patients is to detect MSI mutation in tumor tissues by immunohistochemistry and gene mutation in patients by NGS, so as to evaluate CNV and TMB values of patients. Therefore, compared with previous studies and prognosis models, we not only provide a more accurate prognostic evaluation method, but also enable other researchers to use existing data to expand the research sample size and further optimize patient grouping and model

construction. Of course, our research also has a few shortcomings. For example, we did not use our own MSI, TMB and CNV data of colon cancer patients to increase the sample size or verify the grouping. Since not all patients complete the NGS sequencing in our department, we can only obtain more abundant data through cooperation in future research. To verify the expression of model genes, we used patients with different clinical stages to represent patients with different prognosis, and detected the expression level of each gene in tumor tissues of patients with different stages. However, there are a small number of stage Ⅰcolon cancer patients with poor prognosis. According to the experience of our center, this part of stage I colon cancer patients with poor prognosis may account for 1–3% of all stage I colon cancer patients. Since the sample size we used to verify the conclusion is not very large, a small amount of result deviation may occur, but we think this is statistically allowable.

Conclusion

Our study successfully calculated the TMB value of colon cancer patients by using mRNA expression profile data and mutation information, and identified three subtypes by combining MSI and CNV genomic characteristic data. Through the analysis of each subgroup, we successfully screened out 8 core genes related to prognosis and constructed a prognosis prediction model. Proved by external data sets and experimental verification, our model can accurately evaluate the prognosis of colon cancer patients.

Abbreviations

maf：Mutation Annotation Format; TCGA：The Cancer Genome Atlas; TMB：Tumor Mutation Burden; CNV：Copy Number Variation; MSI: Microsatellite instability; qPCR：quantitative Polymerase Chain Reaction; DCs: Dendritic cells; TNM：Tumor Node Metastasis; GEO：Gene Expression Omnibus; ROC： Receiver Operating Characteristics; COAD：Colon adenocarcinoma; PCA：Principal Component Analysis; NK cells：Natural Killer cells

Declarations

Acknowledgements:

None

Consent for publication:

Not applicable

Availability of data and materials:

All datas and materials are available. Please contact us to access if it is needed.

Ethical approval and consent to participate:

No animals or humans were involved in this study. This study was carried out in accordance with the Declaration of Helsinki.

Authors' contributions:

JSX: research design and drafting the manuscript; KLL，CFW and QJY: literature search; HPW and XZW: review and revision of the manuscript and writing guidance. All authors approved this manuscript.

Acknowledgements:

The authors thank Xiaozhong Wang for his kindly supports during this study.

Funding:

None

Disclosure statement:

Competing interests: There are no conflicts of interest in this study.

References

1. Benson AB, Venook AP, Al-Hawary MM, Cederquist L, Chen YJ, Ciombor KK, Cohen S, Cooper HS, Deming D, Engstrom PF, Garrido-Laguna I, Grem JL, Grothey A,Hochster HS, Hoffe S, Hunt S, Kamel A, Kirilcuk N, Krishnamurthi S, Messersmith WA, Meyerhardt J, Miller ED, Mulcahy MF, Murphy JD, Nurkin S, Saltz L, Sharma S,Shibata D, Skibber JM, Sofocleous CT, Stoffel EM, Stotsky-Himelfarb E, Willett CG, Wuthrick E, Gregory KM, Freedman-Cass DA. NCCN Guidelines Insights: Colon Cancer, Version 2.2018. J Natl Compr Canc Netw. 2018 Apr;16(4):359-369.
2. Siegel RL, Miller KD, Fedewa SA. Colorectal cancer statistics, 2017[J]. Ca Cancer J Clin, 2017, 67 (3) :177-193.
3. CALKINS G N.Zur frage der entstehung maligner tumoren[J].Sci-ence,1914,40(1041):857-859.

4. Figueiredo J.C., Hsu L., Hutter C.M., Lin Y., Campbell P.T.,Baron J.A., Berndt S.I., Jiao S., Casey G., Fortini B., Chan A.T., Cotterchio M., Lemire M., Gallinger S., Harrison T.A., Le Marchand L., Newcomb P.A., Slattery M.L., Caan B.J., Carlson C.S., Zanke B.W., Rosse S.A., Brenner H., Giovan[1]nucci E.L., Wu K., Chang-Claude J., Chanock S.J., Curtis K.R., Duggan D., Gong J., Haile R.W., Hayes R.B.,Hoffmeister M., Hopper J.L., Jenkins M.A., Kolonel L.N., Qu C., Rudolph A., Schoen R.E., Schumacher F.R., Semi[1]nara D., Stelling D.L., Thibodeau S.N., Thornquist M., War[1]nick G.S., Henderson B.E., Ulrich C.M., Gauderman W.J.,Potter J.D., White E., Peters U., CCFR, and GECCO, 2014,Genome-wide diet-gene interaction analyses for risk of col[1]orectal cancer, PLoS Genetics, 10(4): e1004228.
5. Benedix F, Kube R, Meyer F, Schmidt U, Gastinger I, Lippert H; Colon/Rectum Carcinomas (Primary Tumor) Study Group. Comparison of 17,641 patients with right- and left-sided colon cancer: differences in epidemiology, perioperative course, histology, and survival. Dis Colon Rectum. 2010 Jan;53(1):57-64.
6. Klaver CEL, Kappen TM, Borstlap WAA, Bemelman WA, Tanis PJ. Laparoscopic surgery for T4 colon cancer: a systematic review and meta-analysis. Surg Endosc. 2017 Dec;31(12):4902-4912.
7. Yahagi M, Okabayashi K, Hasegawa H, Tsuruta M, Kitagawa Y. The Worse Prognosis of Right-Sided Compared with Left-Sided Colon Cancers: a Systematic Review and Meta-analysis. J Gastrointest Surg. 2016 Mar;20(3):648-55.
8. Taieb J, Le Malicot K, Shi Q, Penault-Llorca F, Bouché O, Tabernero J, MiniE, Goldberg RM, Folprecht G, Luc Van Laethem J, Sargent DJ, Alberts SR, EmileJF, Laurent Puig P, Sinicrope FA. Prognostic Value of BRAF and KRAS Mutations in MSI and MSS Stage III Colon Cancer. J Natl Cancer Inst. 2016 Dec 31;109(5):djw272.
9. Zaanan A, Taieb J. Valeur prédictive et pronostique du phénotype MSI dans lecancer du colon non métastatique : qui et comment traiter ? [Predictive and prognostic value of MSI phenotype in adjuvant colon cancer: Who and how to treat?. Bull Cancer. 2019 Feb;106(2):129-136. French.
10. Beypinar I, Demir H, Araz M, Baykara M, Aykan NF. The View of Turkish Oncologists Regarding MSI Status and Tumor Localization in Stage II and III Colon Cancer. J Gastrointest Cancer. 2020 Nov 6.
11. Yadav VK, Huang YJ, George TA, Wei PL, Sumitra MR, Ho CL, Chang TH, Wu ATH,Huang HS. Preclinical Evaluation of the Novel Small-Molecule MSI-N1014 for Treating Drug-Resistant Colon Cancer via the LGR5/β-catenin/miR-142-3p Network and Reducing Cancer-Associated Fibroblast Transformation. Cancers (Basel). 2020 Jun 16;12(6):1590.
12. Chouhan H, Sammour T, Thomas ML, Moore JW. The interaction between BRAF mutation and microsatellite instability (MSI) status in determining survival outcomes after adjuvant 5FU based chemotherapy in stage III colon cancer. J Surg Oncol. 2018 Dec;118(8):1311-1317.
13. Lee DW, Han SW, Bae JM, Jang H, Han H, Kim H, Bang D, Jeong SY, Park KJ, KangGH, Kim TY. Tumor Mutation Burden and Prognosis in Patients with Colorectal Cancer Treated with Adjuvant Fluoropyrimidine and Oxaliplatin. Clin Cancer Res.2019 Oct 15;25(20):6141-6147.
14. Pai SG, Carneiro BA, Chae YK, Costa RL, Kalyan A, Shah HA, Helenowski I, Rademaker AW, Mahalingam D, Giles FJ. Correlation of tumor mutational burden and treatment outcomes in patients with colorectal cancer. J Gastrointest Oncol.2017 Oct;8(5):858-866.
15. Wang K, Yu X, Jiang H, Huang J, Wang H, Jiang H, Wei S, Liu L. Genome-wide expression profiling-based copy number variations and colorectal cancer risk in Chinese. Mol Carcinog. 2019 Jul;58(7):1324-1333.
16. Jin Z, Zhang G, Liu Y, He Y, Yang C, Du Y, Gao F. The suppressive role of HYAL1 and HYAL2 in the metastasis of colorectal cancer. J Gastroenterol Hepatol.2019 Oct;34(10):1766-1776.
17. Jacobson A, Rahmanian M, Rubin K, Heldin P. Expression of hyaluronan synthase2 or hyaluronidase 1 differentially affect the growth rate of transplantable colon carcinoma cell tumors. Int J Cancer. 2002 Nov 20;102(3):212-9.
18. Xie M, Li K, Li J, Lu D, Hu B. Association and diagnostic value of serum SPINK4 in colorectal cancer. PeerJ. 2019 Apr 4;7:e6679.
19. Wang X, Yu Q, Ghareeb WM, Zhang Y, Lu X, Huang Y, Huang S, Sun Y, Lin J, Liu J, Chi P. Downregulated SPINK4 is associated with poor survival in colorectal cancer. BMC Cancer. 2019 Dec 30;19(1):1258.
20. Chen S, Yue T, Huang Z, Zhu J, Bu D, Wang X, Pan Y, Liu Y, Wang P. Inhibition of hydrogen sulfide synthesis reverses acquired resistance to 5-FU through miR-215-5p-EREG/TYMS axis in colon cancer cells. Cancer Lett. 2019 Dec 1;466:49-60.
21. Lin CY, Hsieh PL, Chou CL, Yang CC, Lee SW, Tian YF, Shiue YL, Li WS. High EREG Expression Is Predictive of Better Outcomes in Rectal Cancer Patients Receiving Neoadjuvant Concurrent Chemoradiotherapy. Oncology. 2020;98(8):549-557.
22. Qu X, Sandmann T, Frierson H Jr, Fu L, Fuentes E, Walter K, Okrah K, Rumpel C, Moskaluk C, Lu S, Wang Y, Bourgon R, Penuel E, Pirzkall A, Amler L, Lackner MR, Tabernero J, Hampton GM, Kabbarah O. Integrated genomic analysis of colorectal cancer progression reveals activation of EGFR through demethylation of the EREG promoter. Oncogene. 2016 Dec 15;35(50):6403-6415.
23. Zheng L, Weng M, Qi M, Qi T, Tong L, Hou X, Tong Q. Aberrant expression of intelectin-1 in gastric cancer: its relationship with clinicopathological features and prognosis. J Cancer Res Clin Oncol. 2012 Jan;138(1):163-72.
24. Yousef GM, Magklara A, Diamandis EP. KLK12 is a novel serine protease and a new member of the human kallikrein gene family-differential expression in breast cancer. Genomics. 2000 Nov 1;69(3):331-41.
25. Lin YC, Mahalingam J, Chiang JM, Su PJ, Chu YY, Lai HY, Fang JH, Huang CT,Chiu CT, Lin CY. Activated but not resting regulatory T cells accumulated in tumor microenvironment and correlated with tumor progression in patients with colorectal cancer. Int J Cancer. 2013 Mar 15;132(6):1341-50.

Figures

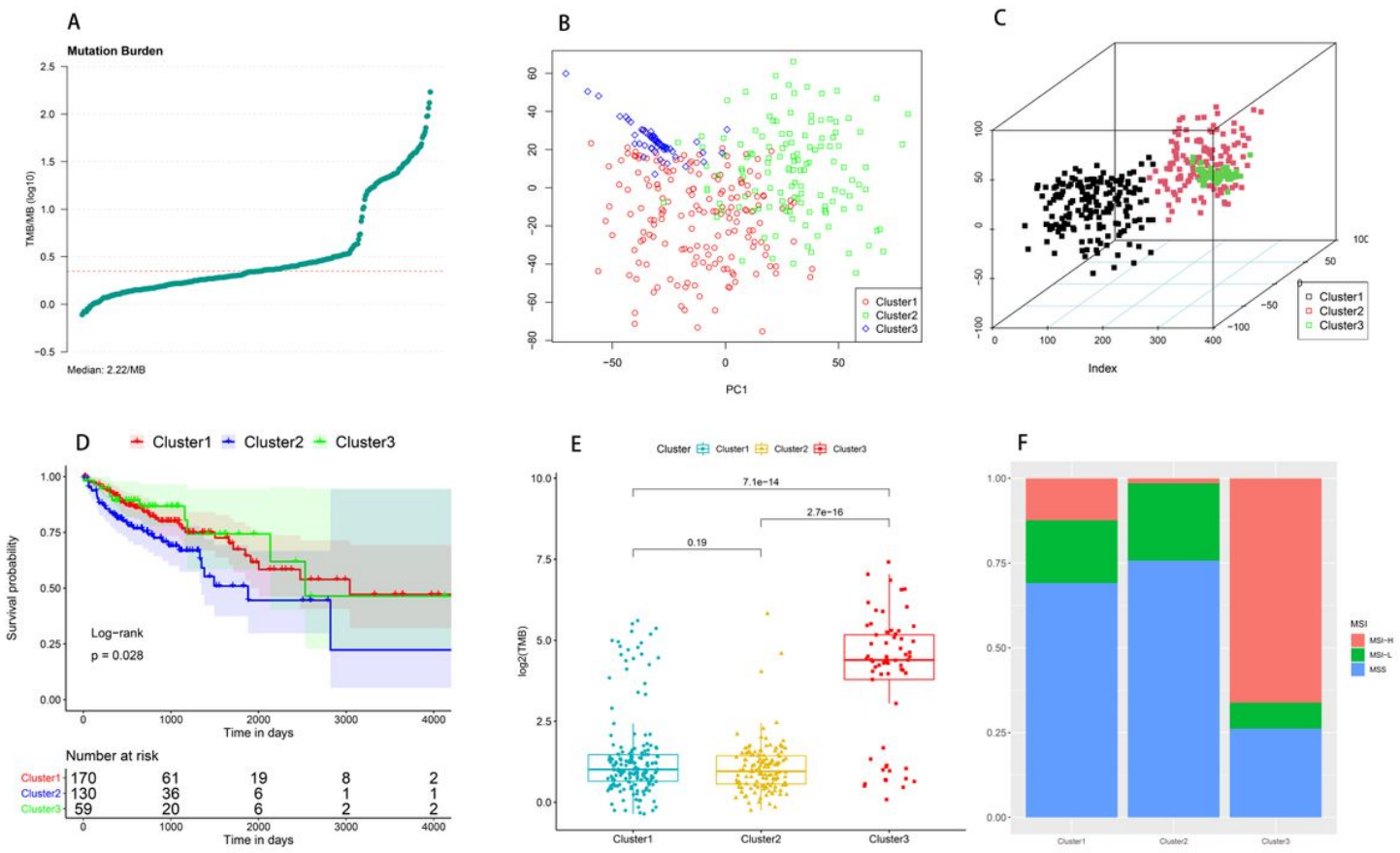

Figure 1

1A: Results of calculation of tumor mutation burden value in colon cancer samples; 1B: PCA analysis results of all samples; 1C: 3D visualization of PCA analysis results; 1D: Survival analysis results of three clusters obtained by clustering; 1E: TMB comparative analysis results among three types of samples; 1F: MSI comparative analysis results among three types of samples

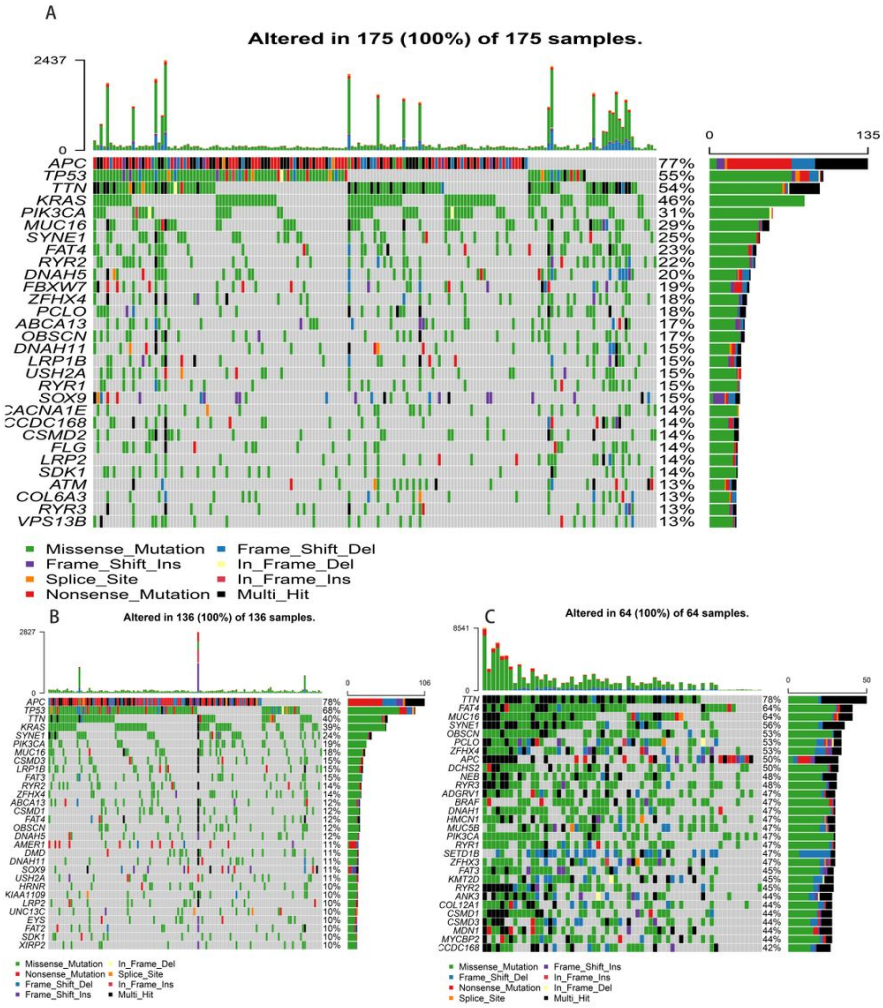

Figure 2

2A: Mutation analysis results of Cluster 1 samples; 2B: Mutation analysis results of Cluster 2 samples; 2C: Mutation analysis results of Cluster 3 samples

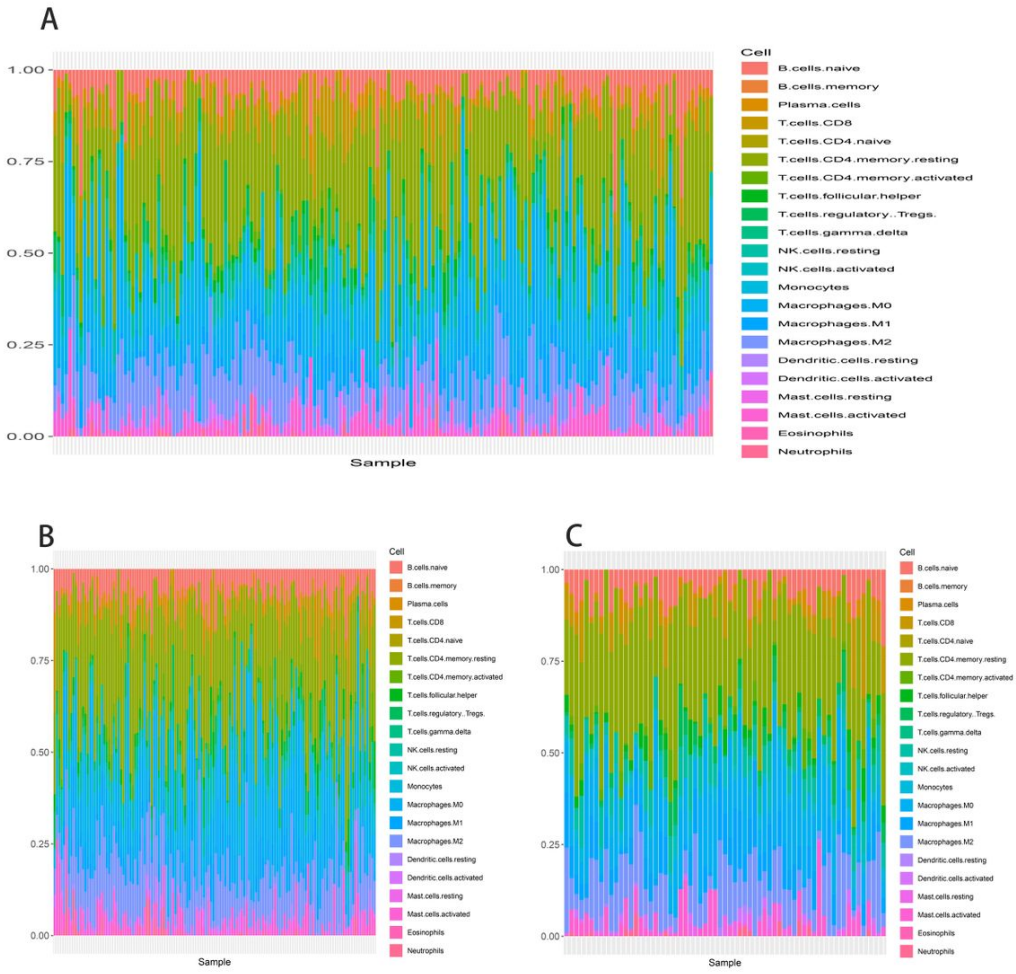

Figure 3

3A: Immune cell infiltration analysis results of Cluster 1 samples; 3B: Immune cell infiltration analysis results of Cluster 2 samples; 3C: Immune cell infiltration analysis results of Cluster 3 samples

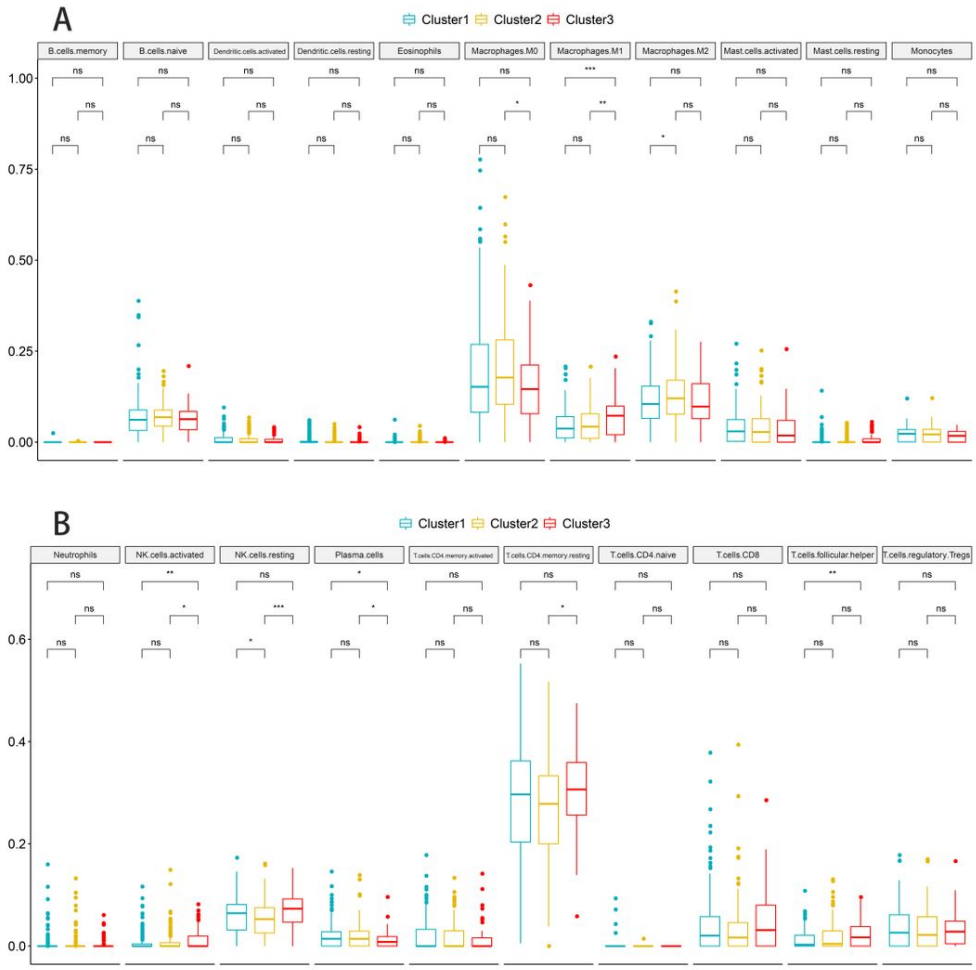

Figure 4

4A-B: Comparison results of infiltration content of 22 kinds of immune cells in three groups of cluster samples

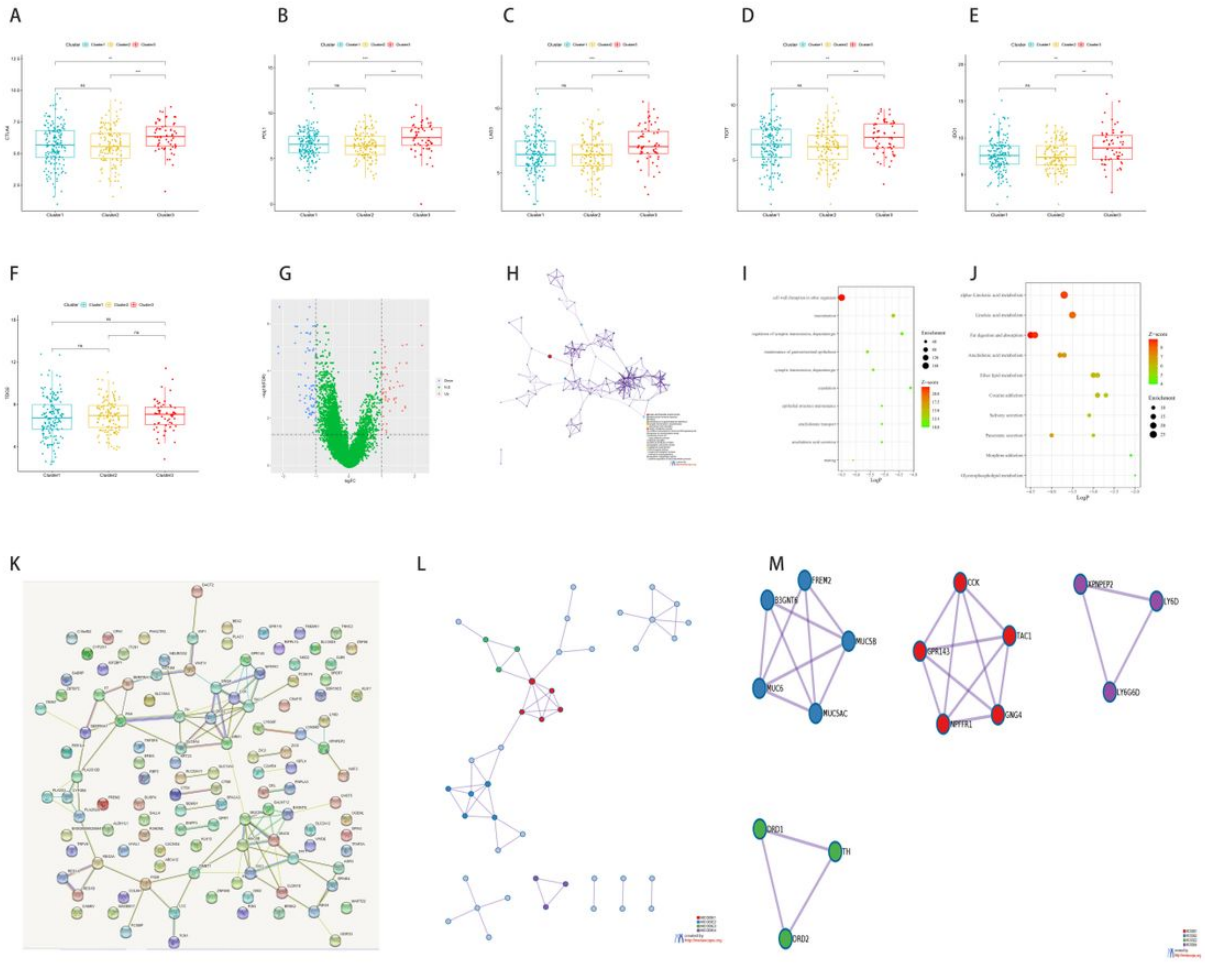

Figure 5

5A-F: Comparative analysis results of CTLA4, PDL1, LAG3, TIGIT, IDO1, TDO2 and other immune checkpoint genes in three groups of samples; G: Differentially expressed genes between Cluster 1+Cluster 3 samples and Cluster 2 samples; H: The network diagram of GO analysis results of differential genes between groups; I: Bubble chart of GO analysis results of differential genes between groups; J: Bubble chart of KEGG analysis results of differential genes between groups; K: Protein interaction analysis results of differential genes between groups; L: Topological network of protein interaction analysis results of differential genes; M: The core module obtained from the interaction analysis of differential gene protein.

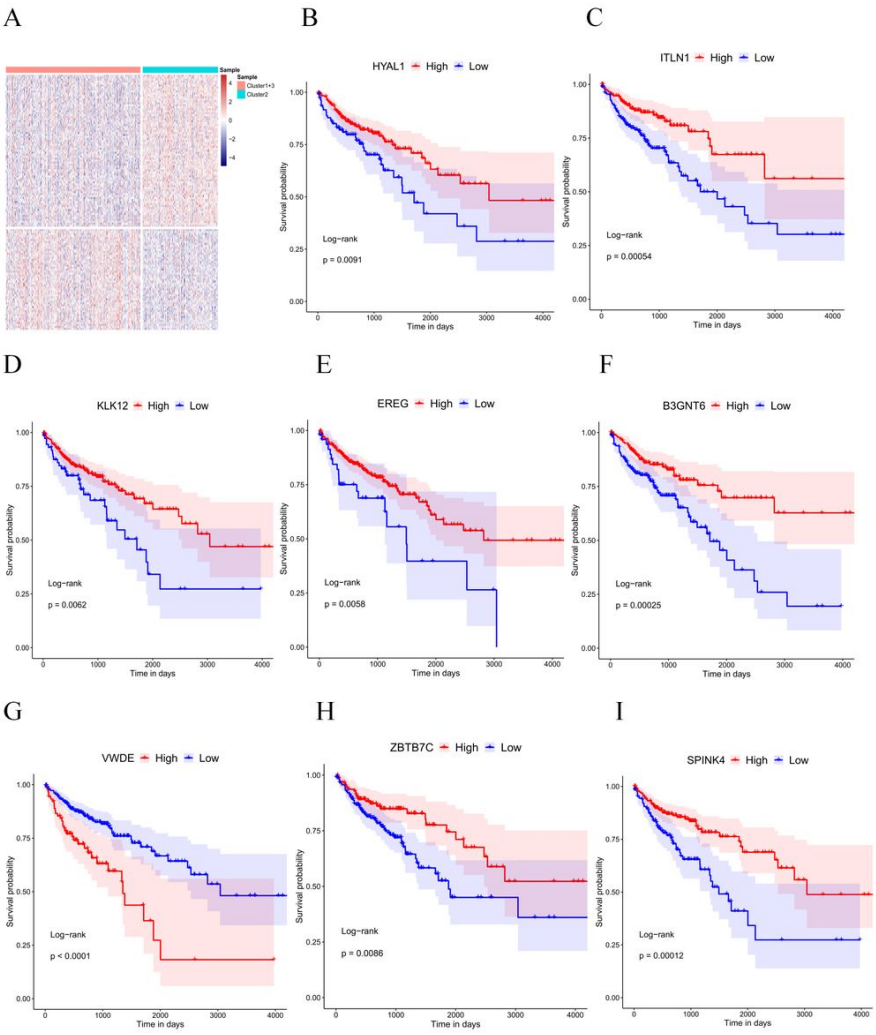

Figure 6

6A: Heat map display of the difference in expression of differential genes between the two groups; B-I: Survival analysis results of eight prognostic related genes

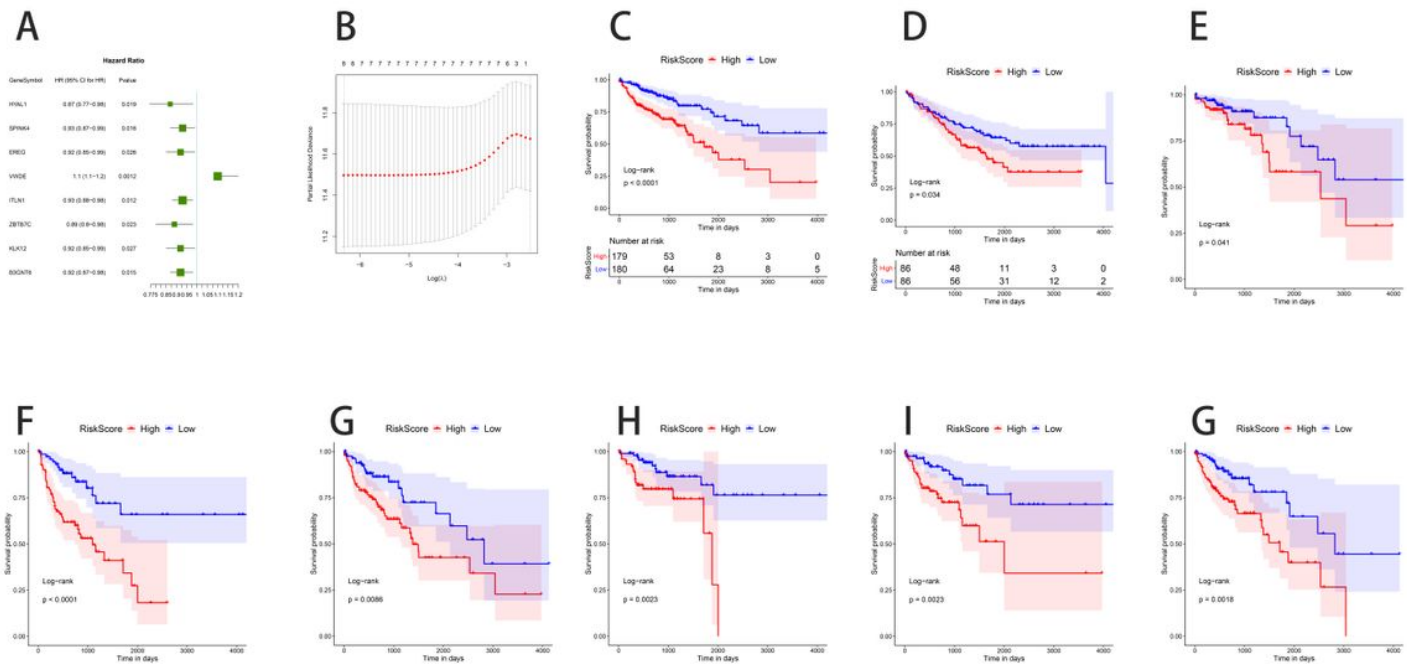

Figure 7

7A: COX regression analysis results of 8 prognosis-related genes; B: Lasso regression analysis results of 8 prognosis-related genes; C: The verification result of survival analysis of the high and low risk patients in the TCGA patient samples; D: The verification result of survival analysis of the high and low risk patients in the patient samples in GSE17536 data set; E: Comparison of survival analysis results between high and low risk groups divided by Stage I+StageII patients; F: Comparison of survival analysis results between high and low risk groups divided by Stage III+StageIV patients; G: Comparison of survival analysis results between high and low risk groups divided by patients over 67 years old; H: Comparison of survival analysis results between high and low risk groups divided by patients aged ≤67 years old; I: Comparison of survival analysis results between high and low risk groups divided by female patients; J: Comparison of survival analysis results between high and low risk groups of male patients.

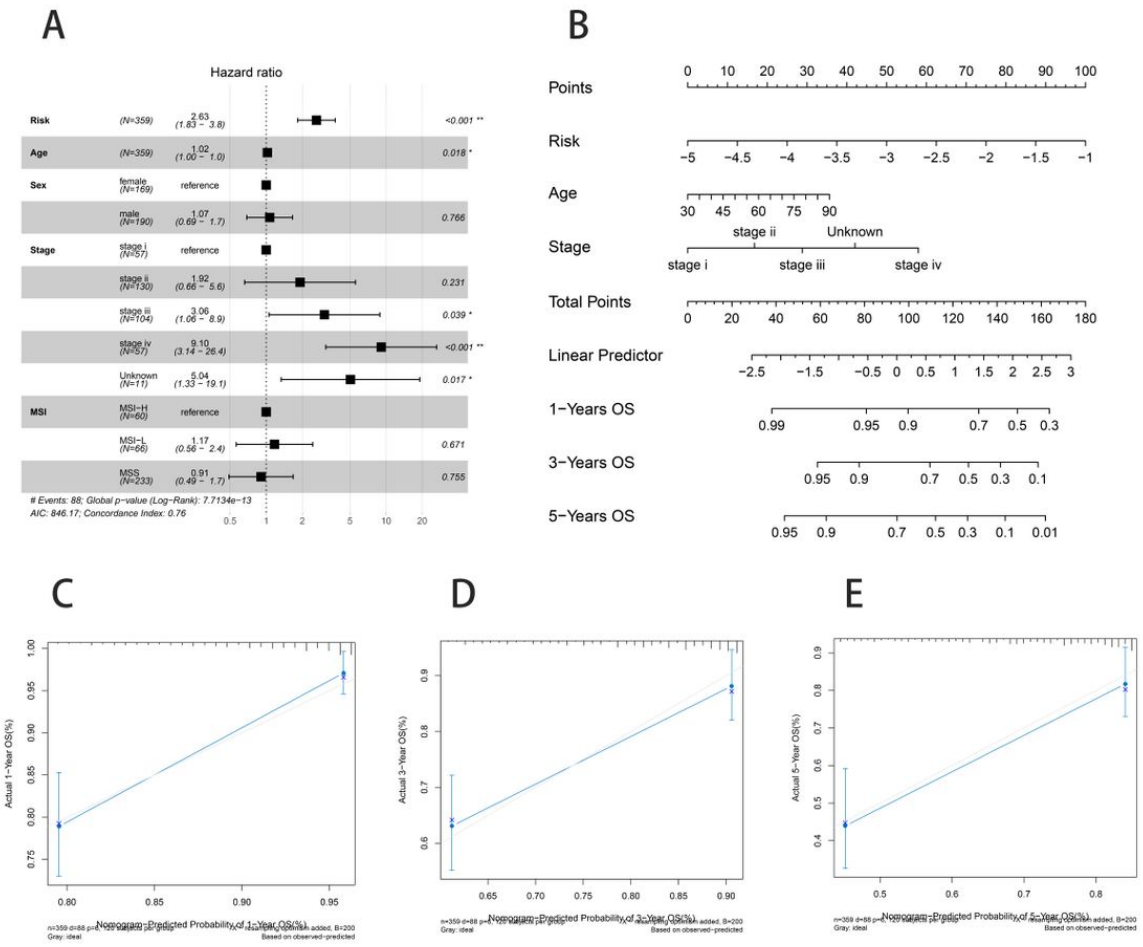

Figure 8

8A: Multivariate Cox regression analysis result of age, MSI classification, TNM Stage, gender and Risk Score; B: Nomogram model constructed by three independent prognostic factors: age, TNM Stage and Risk Score; C: Comparison of 1-year survival rate between model evaluation and actual colon cancer patients; D: Comparison of 3-year survival rate between model evaluation and actual colon cancer patients; E: Comparison of 5-year survival rate between model evaluation and actual colon cancer patients;

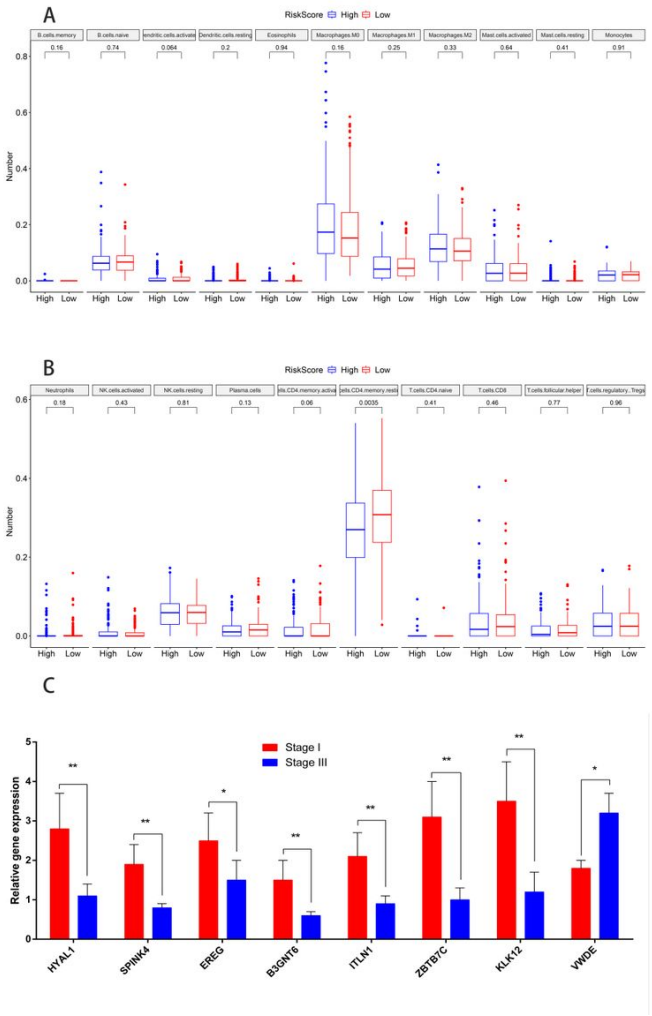

Figure 9

9A-B: Analysis of the difference of immune cell infiltration of 22 kinds of immune cells between high and low risk groups of colon cancer patients; 9C: qPCR verified the expression results of 8 prognostic genes in tumor tissues of Stage I and Stage III patients.

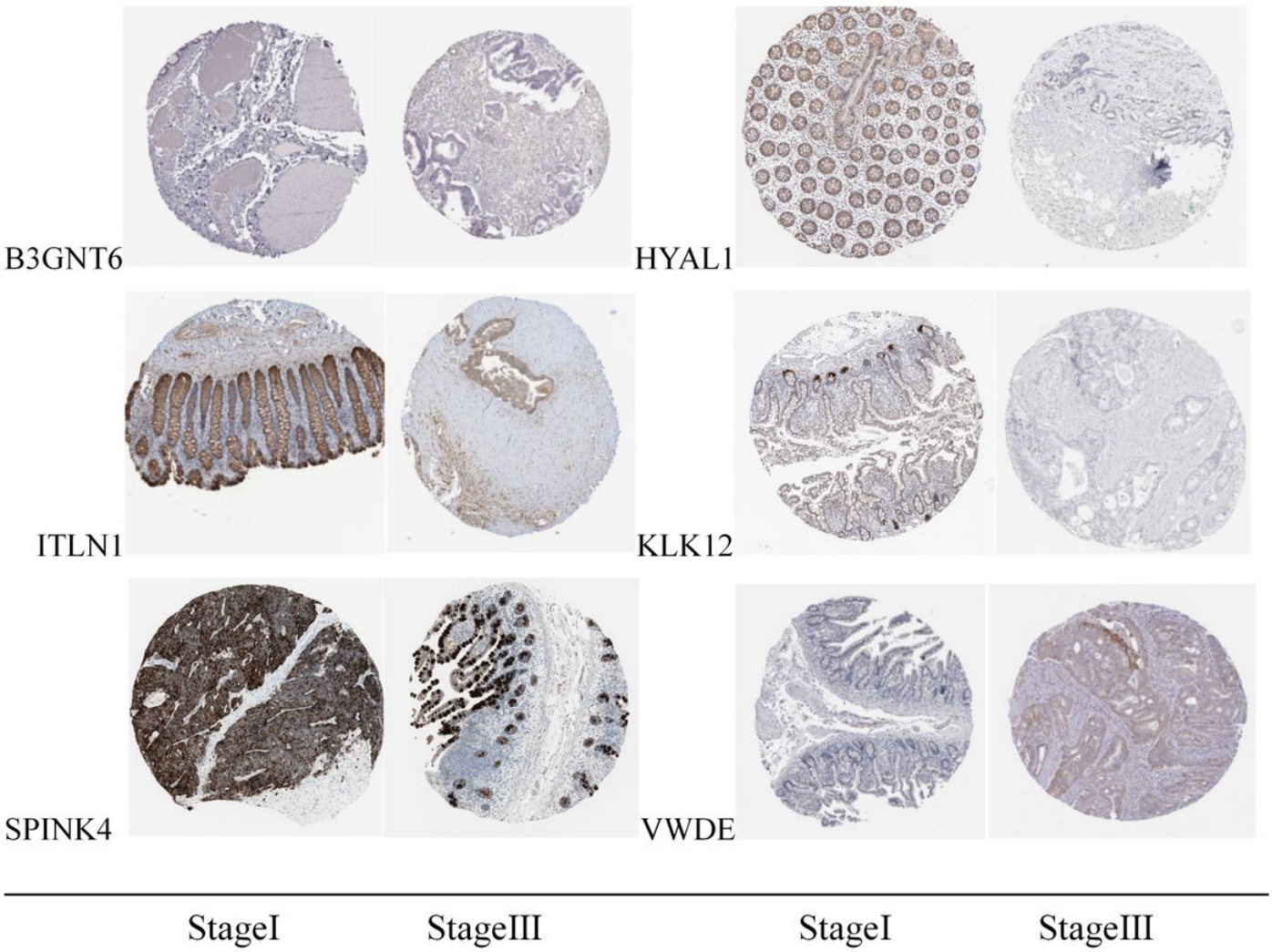

Figure 10

Immunohistochemical results of differential expression of 8 prognostic genes in tumor tissues of Stage I and Stage III colon cancer patients

Supplementary Files

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

- supplementedTable1tmb.csv
- supplementedTable23lei.csv