Potential Biomarkers of Colon Adenocarcinoma Based on Weighted Gene Co-Expression Network Analysis

Zhongze Cui

Binzhou Medical College Affiliated Hospital: Binzhou Medical University Hospital <https://orcid.org/0000-0003-2145-2156>

Shuang He

Binzhou Medical College Affiliated Hospital: Binzhou Medical University Hospital

Feifei Wen

Binzhou Medical College Affiliated Hospital: Binzhou Medical University Hospital

Xiaoyang Xu

Binzhou Medical College Affiliated Hospital: Binzhou Medical University Hospital

Yangyang Li

Binzhou Medical College Affiliated Hospital: Binzhou Medical University Hospital

Lizhen Lu

Binzhou Medical College Affiliated Hospital: Binzhou Medical University Hospital

Shuhua Wu (✉ wushuhua6108@163.com)

Binzhou Medical College Affiliated Hospital: Binzhou Medical University Hospital

Research

Keywords: WGCNA, Immune-related gene, Molecular biomarkers, TCGA and GEO databases, Protein-protein interaction network, Prognostic model

Posted Date: September 17th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-871616/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Background: Colon adenocarcinoma (COAD) is one of the most common malignancies worldwide. Although a large number of studies have elucidated the aetiology of colorectal cancer, the exact mechanism of colorectal cancer development remains to be determined. To identify key modules and prognostic genes that may be involved in the occurrence and development of COAD, weighted gene coexpression network analysis (WGCNA) and differential expression analysis were performed on datasets GSE41657 and GSE74602 from the Gene Expression Omnibus (GEO) database to screen for prognostic differentially expressed genes. Gene expression profiles and clinical information were collected from The Cancer Genome Atlas (TCGA) database for verification.

Results: Through WGCNA and DEGs analysis, 439 genes in key functional modules were obtained, and 26 prognostic related genes were finally obtained through prognostic analysis: (1) We screened 5 genes (RPP40, DUSP18, PPRC1, MFSD11 and PDCD11) that have not been studied in COAD. (2) We obtained the most critical module in the occurrence and development of colon cancer and obtained one prognosis-related gene, NUP85, from the most critical module. The relationship between it and tumor immune microenvironment was verified. (3) A prognostic model comprising four coexpressed differential genes was constructed; TIMP1, PMM2, E2F3 and MORC2 were selected as the key prognosis-related genes.

Conclusions: (1) As new biomarkers, prognostic genes RPP40, DUSP18, PPRC1, MFSD11 and PDCD11 may be potential therapeutic targets for COAD, and provide new ideas for future research on the mechanism of COAD. (2) NUP85 may be an immune-related gene which was negatively correlated with CD4+ T cell and M2 macrophages that plays an important role in inhibiting the occurrence and development of colorectal adenocarcinoma. (3) A Cox proportional risk model based on gene expression can be used to predict the prognosis and survival time of patients with colon cancer.

Background

According to 2018 global cancer statistics, there were nearly 1.1 million new cases of colorectal cancer (CRC) and 550,000 CRC deaths worldwide, accounting for approximately 6% of all new cases and deaths of malignant tumours, making CRC third in morbidity and second in mortality among all cancers [1]. The main causes of death in most colon cancer patients are insidious onset, asymptomatic progression, early lymph node metastasis and poor prognosis [2]. In recent years, with the development of public databases, researchers have found through bioinformatics analysis that many genes are significantly differentially expressed during the occurrence and development of colon cancer and that the regulatory mechanism of colon cancer is complex, and the initiation and progression of colon cancer are promoted through multiple signalling pathways [3–4]. Studies have found many prognostic genes in various human cancers, and the identification of prognostic genes based on genomic data is helpful to judge cancer prognosis and understand cancer development [5–6]. Currently, screening for prognostic genes for colon cancer has been performed [7–9], but the prediction of key prognostic genes and the construction of an effective and reliable prognostic gene model for evaluating the prognosis of colon cancer patients remain to be further studied.

WGCNA, as a commonly used gene module analysis technique, has been widely applied to identify and screen for molecular markers or drug targets of complex diseases [10]. In this study, genetic changes in CRC were identified from two mRNA microarray datasets in the GEO database, and differentially expressed genes (DEGs) between CRC and normal tissues were obtained. Through WGCNA, the most significant functional modules in the process of cancer development were obtained, and the differential genes in the modules were screened out as research objects. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) were used for functional enrichment analysis. The target genes were obtained through survival analysis for further study. Potential biomarkers as new therapeutic targets were screened and a protein-protein interaction (PPI) network was constructed to analyse the associations among DEGs. We also screened coexpressed differential genes and constructed a new multigene prognostic model to identify the risk of colon cancer death. Log-rank test validation of the model was performed by using the relationships between gene expression and the prognosis and overall survival of patients with colon cancer, and the correlation of the identified genes and clinical pathologic factors was evaluated to verify the predictive value of the model.

Results

Identification of DEGs

Compared with normal tissue, the analysis of GSE41657 identified 4712 DEGs consisting of 2210 downregulated genes and 2502 upregulated genes, and the analysis of GSE74602 identified 2878 DEGs consisting of 1335 downregulated genes and 1543 upregulated genes in cancer tissue (Fig. 1A–1B).

Construction of the colon cancer coexpression network

For WGCNA analyses, all samples were included in the clusters. The soft-power threshold β was determined by the function “`sft$powerEstimate`”; $\beta = 8$ (GSE41657) and $\beta = 20$ (GSE74602) were selected for further analysis of colon cancer and normal conditions in the two GEO datasets, respectively (Fig. 1C; Fig. 1F). We set the shear height to 0.25. Then, gene modules were detected based on the TOM matrix. The analysis detected twenty-one modules in GSE41657 and five modules in GSE74602 (Fig. 1D; Fig. 1G). From the heatmap of modules and characters, it was found that the brown module (3783 genes) in GSE41657 and the black module in GSE74602 (1690 genes) had the highest degree of correlation with colon adenocarcinoma (Fig. 1E; Fig. 1H). The subsequent analyses screened the GSE41657_brown module and GSE74602 black module for seed genes.

Screening for co-altered genes

The overlapping differentially expressed genes of the most significant modules (GSE41657_DIFF, GSE41657_BROWN, GSE74602_DIFF and GSE74602_BLACK) were obtained by analysis using an online Venn graphing tool. A total of 439 genes were identified (Fig. 1I).

GO and KEGG enrichment analysis of DEGs

The R packages "clusterProfiler", "org.Hs.eg.db", "enrichplot", and "ggplot2" were used for the functional enrichment analysis of the DEGs. The results indicated that the DEGs were mainly enriched in 'mitotic nuclear division' and 'organelle fission' among BPs; 'chromosomal region' and 'kinetochore' among CCs; and 'DNA helicase activity' and 'ATPase activity' among MFs. KEGG pathway analysis revealed strong enrichment in 'cell cycle', 'RNA transport', 'oocyte meiosis', and 'cellular senescence', among other pathways (Fig. 1J-1M).

Screening prognostic genes

The correlation of gene expression and survival traits were then analysed. A total of 379 patients had complete gene expression and survival information. The overall survival (OS) was analysed for these genes using Kaplan–Meier survival plots. Briefly, the "survival" and "survminer" R packages were used to plot Kaplan–Meier curves.

Only 26 (RUVBL1, RNASEH2A, MMP3, RPP40, DUSP18, CCT6A, MORC2, NCL, IL13RA2, MELK, PPRC1, CHEK1, RRP12, EPHB2, TIMP1, MFSD11, KIF15, NUP85, NAT1, TNF, TRAP1, CDC25C, E2F3, NMB) gene from 439 differential genes were found to significantly impact survival ($p < 0.05$) (Fig. 2).

Screening for unstudied genes

Through the literature search, we found that RPP40, DUSP18, PPRC1, MFSD11 and PDCC11 gene has not yet been studied in the colonic carcinoma. Using HPA database RPP40, DUSP18, PPRC1, MFSD11 and PDCC11 in COAD protein expression levels in the organization, and comparing with the expression level of normal tissue. The results indicate that RPP40, DUSP18, PPRC1 and PDCC11 were significantly increased in tumor tissue compared with normal samples, while MFSD11 was significantly down-regulated in tumor tissue. This is consistent with the differential expression data visualized through the TCGA database (Fig. 3).

PPI network and module analysis

A PPI network of the 439 identified DEGs was constructed. Fifteen significantly enriched modules were obtained using Cytoscape software. There were 32 genes in the most critical module—the red module (Fig. 4A).

R packages were also used for the functional enrichment analysis of these key genes. The results indicated that the DEGs were mainly enriched in 'mitotic nuclear division', 'organelle fission', 'microtubule cytoskeleton organization involved in mitosis', and other functions. KEGG pathway analysis revealed strong enrichment in 'cell cycle', 'oocyte meiosis', 'RNA transport' and other pathways (Fig. 4B-4C). The functional results of the genes in the most significant module indicated that these genes were mainly enriched in processes associated with the cell cycle, cell proliferation and division.

Among the 32 genes, only NUP85 was associated with prognosis (Fig. 2). Therefore, NUP85 was selected as a key gene for subsequent analysis.

Further analysis of key genes using TCGA database

The expression of NUP85 in 398 COAD cancer tissues and 39 paracancerous tissues in TCGA database was analysed. The results showed that the expression level of NUP85 in colorectal adenocarcinoma was significantly higher than that in adjacent tissues ($P < 0.001$, Fig. 5A). The expression level of NUP85 in colorectal adenocarcinoma of the same patient was also significantly higher than that in adjacent tissues ($P < 0.001$, Fig. 5B).

NUP85 expression was further analysed by GSEA to identify activated functions and signalling pathways in colorectal adenocarcinoma. Figures were drawn for the top 10 most important functions and pathways in GO and KEGG. The GO functional enrichment analysis results showed that the NUP85 high expression group was enriched in gene sets such as "DNA repair", "M phase" and "cell cycle phase", while the low expression group was enriched in gene sets such as "basolateral plasma membrane" and "extracellular matrix". KEGG pathway enrichment showed that the samples with high NUP85 expression were enriched in the pathways 'RNA degradation', 'cell cycle', and 'nucleotide excision repair', while the low-expression group was enriched in the pathways 'Focal adhesion' and 'ECM receptor interaction' (Fig. 5C-D).

Based on the GEPIA database analysis of NUP85 and tumor-related macrophage markers, we found that no correlation was detected between NUP85 expression with marker of M1 macrophages (NOS2), CD8+(CD8A, CD8B), CD20+(MS4A1) lymphocytes ($P > 0.05$; Fig. 5E-H), and a significant positive correlation was observed between NUP85 expression and marker of CD4+(CD4) lymphocytes, M2 macrophages (CD163) ($P < 0.05$; Fig. 5I-J).

Immunohistochemical verification

Immunohistochemical test results showed that the expression of the target gene NUP85 in colorectal cancer tissues was significantly higher than that in normal mucosal tissues ($P < 0.05$; Table 1; Fig. 5K-L).

According to the NUP85 immunohistochemistry results, colorectal cancer samples were divided into a NUP85+ group and a NUP85- group. The results showed that the expression of CD4 and CD163 in the NUP85- group was higher than that in the NUP85+ group ($p < 0.05$). Spearman correlation analysis showed that the expression of NUP85 was negatively correlated with CD4 positive lymphocytes and M2-type tumour-associated macrophages (TAMs) ($p < 0.05$; Table 2; Fig. 5M-P).

Building a prognostic model

Kaplan–Meier (K–M) tests and univariate Cox analysis identified 7 prognostic differentially expressed genes from 439 differentially expressed genes with statistically significant association with survival, among which MORC2, TIMP1, PPRC1, and E2F3 were high-risk and PMM2, NAT1, and EPHB2 were low-risk (Fig. 6A-B).

After multivariate Cox analysis of the above 7 genes, the TIMP1, PMM2, E2F3 and MORC2 genes were retained, and survival curves were drawn to establish a prognostic prediction model (Fig. 2). The prognostic scoring formula was as follows: $PI = 0.193 \times \text{E2F3 expression level} - 0.253 \times \text{PMM2 expression level} + 0.099 \times \text{MORC2 expression level} + 0.003 \times \text{TIMP1 expression level}$ (Table 3).

A total of 337 patients with complete clinical data were assigned to the low-risk and high-risk groups according to the median risk score calculated by the prognostic scoring formula. The results showed that the 5-year survival rate was approximately 85% in the low-risk group and 45% in the high-risk group ($P < 0.01$). The 75% OS of the lower-risk group (5.2 years) was significantly longer than that of the higher-risk group (1.3 years) (Fig. 6C). Higher prognostic score (Fig. 6D) was associated with a greater likelihood of death at any timepoint (Fig. 6E); that is, the higher the risk score, the worse the prognosis. Fig. 6F shows the expression levels of each gene in the prognostic model in the low-risk group and the high-risk group.

Cox regression analysis of survival factors in patients with colorectal cancer and multi-indicator ROC curve were performed. Among the independent variables included in Cox regression analysis, colorectal cancer prognostic associated gene model score was a continuous variable, and the remainder were categorical variables. Univariate Cox regression model analysis showed that age, pathological stage, TNM stage and prognosis-related gene prognostic model score were risk factors affecting the overall survival of colorectal cancer patients ($P < 0.05$, Fig. 6G). Multivariate Cox regression model analysis showed that age and prognostic model score were independent factors influencing the prognosis of patients with colorectal cancer ($P < 0.05$, Fig. 6H). One-year, three-year and five-year survival rates were analysed by multi-indicator ROC curves, and the area under the curve (AUC) values of the ROC curve were 0.693, 0.743 and 0.808 (null hypothesis: real area = 0.5, $P < 0.05$, Fig. 6I), suggesting that the COAD prediction model could be accepted.

Discussion

In the present study, DEGs between COAD and noncancerous tissues were obtained from two mRNA microarray datasets. Through WGCNA, coexpression analysis was performed on the two datasets, the most significantly upregulated modules were selected from the intersection of the differentially expressed genes, and 439 coexpressed DEGs were obtained.

In order to explore the possible roles of these genes in COAD, we conducted enrichment analysis of their functions and pathways. The GO and KEGG enrichment results of 439 genes showed that these genes were mainly enriched in cell cycle, cell proliferation and division, cell senescence and other related processes. Obviously, these are all closely related to the occurrence and development of cancer.

In order to further screen for prognostic genes that may serve as new biomarkers, we analyzed them using K-M method. Through survival analysis, 26 prognostic genes were screened out from the 439 target genes, among which RPP40, DUSP18, PPRC1, MFSD11 and PDCD11 had not been studied in COAD. RPP40 has been mentioned in acute myeloid leukemia [15]. In breast cancer, DUSP18 expression was significantly increased in patients with Chemotherapy induced fatigue (CIF) [16]. MFSD11 in BREAST cancer BT-474 cells leads to decreased trastuzumab sensitivity [17]. PDCD11 positive cells from yolk sacs regulate the differentiation of zebrafish microglia through NF- κ B-Tgf β 1 pathway [18]. CircRNA CIRC-PDCD11 promotes the progression of triple negative breast cancer by enhancing aerobic glycolysis [19]. It can be seen that these genes have been used in tumor research and play an important role in tumor progression and drug resistance, which indicates that these genes are valuable for study in COAD. In this study, through differential expression analysis based on TCGA database and immunohistochemical analysis based on HPA database, we infer that RPP40, DUSP18, PPRC1 and PDCD11 were significantly increased in tumor tissue compared with normal samples, while MFSD11 was significantly down-regulated in tumor tissue. This suggests that these five genes may play an important role in tumor development. In addition, the high expression of RPP40, PPRC1 and PDCD11 in cancer predicted a poor prognosis, indicating that these three genes may promote cancer, while the high expression of DUSP18 and MFSD11 in cancer predicted a good prognosis, indicating that these two genes have the potential to become tumor suppressor genes. In conclusion, we analyzed the differential expression and prognosis of these 5 genes, and speculated that these 5 genes might play an important role in the occurrence and development of COAD, and have the potential to become new biomarkers for COAD and potential targets for treatment of COAD.

Subsequently, protein interaction network analysis was performed to identify key functional modules and key genes in 439 genes. Preliminary screening studies identified a total of 15 key modules, among which the most critical module contained 32 genes. The results showed that the main processes of 32 genes were also related to the cell cycle, cell proliferation and division.

After survival correlation analysis, only one gene, NUP85, was significant, and the prognosis of the high expression group was better. We further analysed the expression level of NUP85 in colon cancer using TCGA database, and the results showed that the expression level of NUP85 in colon adenocarcinoma was significantly higher than that in adjacent tissues, and the difference was statistically significant. The expression level of NUP85 was also significantly higher in colorectal adenocarcinoma tissue than in adjacent tissues of the same patient. We hypothesized that when cancer occurs, NUP85 increases in cancer tissue and inhibits tumor development. The function of NUP85 was further analysed by GSEA. Our results suggest that NUP85 may play a role in cell proliferation, DNA replication, and mitosis, which are associated with tumor proliferation. The enrichment of NUP85 in many pathways and functions related to the occurrence and development of cancer may provide a direction for the study of NUP85 in colon cancer.

NUP85 is also known as NUP75 and Ffront protein. According to biochemistry analysis, the expression of NUP85 in cancer is higher than that in normal tissues, and patient prognosis is better when NUP85 expression is high in cancer. Recent studies have shown that NUP85 has been used to establish a prognostic model for gastrointestinal neoplasms [20], suggesting that NUP85 has an important role to play in gastrointestinal neoplasms.

Through a literature review, we learned that NUP85 is related to CCR2-mediated monocyte chemotaxis, and combined with the CCR2 cluster formed by Ffront, it induces the activation of PI(3)K; thus, this pathway may become a therapeutic target for chronic inflammatory immune diseases associated with macrophage infiltration [21]. Studies have also shown that the intracellular chemokine signalling regulator Ffront plays a key role in the process of

macrophage promotion of tumours (lung adenocarcinoma), and it is positively correlated with tumour-associated macrophage M2 in lung adenocarcinoma [22]. M2 macrophages play a role in promoting tumour development [23–24]. Thus, NUP85 is closely associated with the immune microenvironment in cancer.

However, NUP85 and the immune microenvironment of colon cancer are rarely reported. To determine the relationship between NUP85 and immune microenvironment of colon, we analyzed their correlations using the GEPIA database. We found that NUP85 only negatively correlated with M2 macrophage marker and CD4 + T cell marker. To test this, we conducted immunohistochemical staining of NUP85 and the tumour-associated macrophage M2 marker CD163 and CD4 + T cell marker CD4 in 50 samples to explore the expression of NUP85 in colon cancer and the relationship between NUP85 and tumour-associated M2 macrophages, CD4 + T cell. The results show that NUP85 expression in cancer cells is higher than that in normal tissue and is negatively related to the presence of M2 macrophages or CD4 + T cell. The study of macrophages in colon cancer tissue mainly refers to tumor-associated macrophages (TAMs), M2 macrophages play a role in promoting tumor [25–27]. CD4 + T cell often plays a promoting role in colon cancer, and has adverse effects on the treatment of colon cancer [28]. We speculate that when NUP85 is highly expressed, a decrease in M2 macrophages occurs, inhibiting the effect of these cells on promoting cancer, CD4 + T cell and its adverse effect on treatment are decreased, leading to a better prognosis. However, our sample size was not sufficient to conclusively demonstrate this association, and the specific mechanism remains to be further studied.

In addition, the present study established a prognostic model based on differential gene co-expression. We analysed coexpressed genes in key modules of colon cancer development and used univariate Cox regression analysis, K–M tests, and multivariate Cox regression analysis to identify associated differentially expressed risk genes associated with OS. After screening, we obtained four optimal risk differential related genes—TIMP1, PMM2, E2F3 and MORC2. Many studies have proven that TIMP1 is a prognosis-related gene in colon cancer that plays an important role in the occurrence and development of colon cancer and can determine the progression and metastasis of colon cancer through the FAK-PI3K/Akt and MAPK pathways [29–31]. PMM2 is also associated with the staging and infiltration prognosis of a variety of cancers [32–34], but it has not previously been reported in colon cancer. After function enrichment, we found that its function may be related to isomerase activity and intramolecular transferase activity, which may provide insights into the mechanism by which it can be linked to colon cancer. E2F3 is a prognostic factor closely related to the proliferation and apoptosis of colorectal cancer cells. Targeting E2F3 can inhibit the proliferation and induce the apoptosis of colorectal cancer cells and induce the expression of E2F3 to promote the proliferation of colorectal cancer cells [35–38]. MORC2 is also considered a candidate gene for colon cancer, and previous studies have shown that MORC2 promotes the development of invasive colorectal cancer phenotypes by inhibiting NDRG1 [39–40].

We further constructed prognostic risk models (OS models) of genes associated with colorectal cancer and analysed the correlation between these models and clinical indicators, demonstrating that they can provide accurate prognosis for colorectal cancer patients. Subsequently, we used TCGA data to analyse the correlations of these four genes with clinical parameters.

Conclusion

In summary, in this study, WGCNA analysis and differential gene analysis were carried out on two GEO datasets. Received the 439 target genes we conducted three on the basis of analysis. Firstly, we get 26 different genes, and through the analysis of the prognosis and select 5 has not been studied genes in COAD, carries on the preliminary discussion to them, we found they could become COAD treatment of potential targets, which offers a new way for the research of COAD. Secondly, The most important elements were the identification of key gene modules through PPI analysis and subsequent immunohistochemistry to confirm the differential expression of the identified factor NUP85. Although NUP85 has been used to model the prognosis of gastrointestinal tumors, its role in COAD has not been investigated, and our study complements that. We found that NUP85 function in COAD is closely associated with tumor proliferation and progression through enrichment analysis, this provides a direction for studies of NUP85 in COAD. Moreover, this experiment proved for the first time that the role of NUP85 in colon carcinoma may be associated with tumour M2 macrophages or CD4 + T cell, as it showed a negative correlation with the infiltration of these cells, and a corresponding tumour inhibition effect. However, due to insufficient sample size, our study was inconclusive on this point, and this specific mechanism remains to be further researched. Last but not least, the establishment of a multigene prognostic model can provide more accurate prognostic evaluation and guidance than current clinicopathological indicators, reduce the waste of medical resources, guide the selection of personalized treatment options, and enable patients to benefit from clinical practice.

Methods

Downloads from the GEO database and TCGA database

Gene expression profile data with corresponding clinical data were identified and downloaded from the publicly available GEO and TCGA databases. The gene expression profiles of GSE41657 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE41657>) and GSE74602 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE74602>) were downloaded from the Gene Expression Omnibus (GEO) database. GSE41657 is an expression profile dataset based on the GPL6480 platform that contains 88 samples, including 12 normal epithelium samples, 21 low-grade dysplasia samples, 30 high-grade dysplasia samples and 25 adenocarcinoma samples. Normal and adenocarcinoma samples were selected for the study. The other gene expression profile dataset, GSE74602, is based on the GPL6104 platform and contains 60 samples, specifically, 30 normal samples and 30 primary tumour samples.

The selected TCGA (<https://portal.gdc.cancer.gov/>) colon cancer dataset contained 437 samples, including 398 tumour tissues and 39 corresponding healthy mucosal tissues, and the colon cancer clinical dataset containing data from 385 tumour patients was also downloaded. It is worth noting that in this study, we excluded special COAD samples (cystic, mucinous and serous neoplasms/complex epithelial neoplasms/complex epithelial neoplasms (NOS)) from the TCGA database to avoid interference from this disease type.

DEGs analysis

The series matrix files from GEO and TCGA were annotated with official gene symbols using the data table of the microarray platform, and then gene expression matrix files were obtained. R software was used to identify the DEGs from the GEO matrix files. $\log[\text{fold change(FC)}] > 1$ and adj. P-values < 0.05 were considered to indicate a statistically significant difference.

Co-expression network construction by WGCNA

The “WGCNA” R package was used to construct two coexpression networks for all genes in colon cancer samples and normal samples in GSE41657 and GSE74602. We removed the low-grade dysplasia samples and high-grade dysplasia samples from GSE41657. First, the RNA-seq data were filtered to reduce outliers. Second, the gene expression data were used to construct a network by selecting an appropriate weighting coefficient β (soft threshold value) to build networks of genes in the connection between the scale-free network and by using the phase relationship between genes to construct a hierarchical clustering tree. Third, the genes were classified into expression patterns based on their weighted correlation coefficients; genes with similar patterns were grouped into one module, and then the genes were classified into different modules through the gene expression pattern for the next analysis. Fourth, the correlation coefficients were applied to transform the correlation matrix into an adjacency matrix and further transform it into a topological overlap matrix (TOM) [11-12].

Finally, the Pearson correlation coefficient and P value of the module eigengene (ME) composed of each module gene and sample were calculated by the WGCNA algorithm and their P values. The relationships between different modules and clinical traits were measured according to the Pearson correlation coefficient, and the module with the highest correlation coefficient was selected for subsequent analysis. In this study, the most upregulated gene modules expressed in CRC samples from GSE41657 and GSE74602 were selected.

KEGG and GO enrichment analysis of DEGs

A Venn diagram-related website (<http://bioinformatics.psb.ugent.be/webtools/Venn/>) was used to obtain the DEGs to construct the target genes.

KEGG is a database resource for understanding advanced functions and biological systems from large-scale molecular data generated by high-throughput experimental techniques [13]. GO is a major bioinformatics tool for annotating genes and analysing their biological processes (BPs), molecular functions (MFs) and cellular components (CCs) [14]. To analyse the function of the DEGs, analysis was performed using the “clusterProfiler” “org.Hs.eg.db” “enrichplot” and “ggplot2” R packages. $p < 0.05$ was considered to indicate a statistically significant difference. The top ten BPs, MFs, CCs and KEGG pathways were visualized.

Screening for prognostic genes

The expression levels of target genes were extracted from the TCGA matrix file and combined with the clinical data. Overall survival analysis for the target genes was performed using Kaplan–Meier curves in the “survival” and “survminer” R packages by TCGA clinical database. The log-rank test was used for statistical analysis.

Screening for unstudied genes

Through literature search, unstudied genes in CRC were found. We downloaded the immunohistochemical images in normal and tumors intestinal mucosa from the HPA database (<https://www.proteinatlas.org/>), and preliminatively discussed the expression level.

PPI network construction and identification of the most critical module

The Search Tool for the Retrieval of Interacting Genes (STRING) database (Version 11.0, <http://string-db.org>) was used to predict potential interactions between candidate genes at the

protein level. A combined score of > 0.7 was considered significant. Additionally, Cytoscape software (Version 3.8.2, <http://www.cytoscape.org/>) was utilized to construct the PPI network. The Molecular Complex Detection (MCODE) app was used to analyse the PPI network modules (degree cut-off ≥ 2 , node score cut-off ≥ 0.2 , K-core ≥ 2 and max depth = 100). Functional pathway enrichment analysis and correlation analysis of genes in the most critical modules as determined by PPI were carried out in R.

Key genes were screened and analyzed

A Venn diagram-related website (<http://bioinformatics.psb.ugent.be/webtools/Venn/>) was used to obtain the key genes through the intersection of prognostic related genes and the most critical modules genes. Functional and pathway enrichment analysis of key genes was performed by GSEA. Key genes underwent correlation analysis with immune markers using the GEPIA database (<http://gepia.cancer-pku.cn/>).

Immunohistochemistry

Primary anti-NUP85 antibodies were obtained from Proteintech (WuHan, China), anti-CD4, anti-CD8, anti-CD20 and anti-CD163 antibodies from Abcam (Cambridge, MA, USA), DAB chromogenic kit and Enzyme- labeled goat anti-mouse/rabbit IgG polymer purchased from Beijing Zhongshan Jinqiao Biotechnology.

Hematoxylin and eosin-stained sections of specimens from 50 patients (2019.01-2020.12) with colon cancer were collected from the Department of Pathology, Affiliated Hospital of Binzhou Medical University Hospital, and diagnosed by two senior pathologists after a double-blind review. There were 4-6 sections of

tumor tissue and normal tissue from each case, and the corresponding paraffin tissue with more lymphocyte infiltration sections was selected for continuous sections.

According to the EnVision immunohistochemistry method, tissue sections were dewaxed and rehydrated, subjected to antigen retrieval under high pressure and heat, and treated with 3% H₂O₂ to eliminate endogenous peroxidase activity. Then, the sections were incubated with primary antibody at the dilutions recommended by the manufacturers (Affinity Biosciences and Abcam) in block overnight at 4 °C. After PBS washing, the sections were incubated with secondary antibody in block at 37 °C for 30 min, followed by application of the DAB chromogen kit, haematoxylin redyeing, differentiation, dehydration, clearing, and then mounting with neutral rubber seal. As suggested by the antibody manufacturers, known positive tissue was used as the positive control, and PBS used instead of the primary antibody was the negative control.

Two researchers who were unaware of the clinicopathological statuses of the specimens scored each section separately. NUP85, CD4, CD8, CD20 and CD163 expression in somatic cells of cancer tissues was counted: CD4, CD8, CD20 and CD163 were mainly expressed in the cell membrane or cytoplasm, NUP85 were mainly expressed in the cell membrane, cytoplasm or nucleus, and the presence of brown-yellow diffuse staining or brown-yellow particles was scored as positive. Ten independent fields were observed under a 400x microscope, and the 5 fields with the largest number of positive cells were selected to determine the experimental results according to the intensity of cell staining and the proportion of positive cells: 0 for no staining, 1 for light yellow, 2 for yellow-brown, and 3 for brown; 0 for ≤5% positive cells, 1 for 6%~25%, 2 for 26%~50%, 3 for 51%~75%, and 4 for > 75%. The two scores were multiplied to produce the final score, and a final score of ≥4 points was considered positive.

Building a prognostic model

Clinical data were extracted from TCGA colorectal adenocarcinoma data. R packages were used to screen for prognostic genes that affect overall patient survival. Kaplan–Meier survival curves and univariate Cox regression models were used to screen for prognostic gene status ($P < 0.05$). Multivariate Cox regression analysis was used to establish the prediction model and calculate the risk score (RS). Patients were then divided into high-risk and low-risk groups with the median score as the cut-off. Kaplan–Meier survival curves were used to analyse the difference in survival between the high-risk group and the low-risk group. Prognostic score and clinicopathological factors such as TNM stage, pathological stage, age (≤65 years vs. >65 years), sex and TNM stage were included in univariate and multivariate Cox regression models to analyse the factors affecting the survival of patients with COAD. The receiver operating characteristic (ROC) curve was used to evaluate the accuracy of the prognostic prediction model.

Statistical analysis

R software (v.3.6.3) and SPSS 25.0 software were used for statistical analysis. Enumeration data were analyzed by the χ^2 test, the independent samples t test was used for comparisons between two groups, and one-way analysis of variance was used for comparisons between multiple groups. The correlation of protein expression was analyzed by the Spearman's rank correlation coefficient, with $P < 0.05$ was considered statistically significant.

Abbreviations

CRC: Colon cancer; COAD: Colon adenocarcinoma; WGCNA: Weighted gene co-expression network analysis; TCGA: The Cancer Genome Atlas; GEO: Gene Expression Omnibus; GO: Gene Ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes. PPI: protein-protein interaction; K-M: Kaplan-Meier; OS: overall survival; ROC: Receiver operating characteristic; DEGs: differentially expressed genes; AUC: Area under the curve; TAM: tumor-associated macrophage.

Declarations

Supplementary information

No.

Acknowledgements

We appreciate The Cancer Genome Atlas and Gene Expression Omnibus for the open data.

Authors' contributions

ZC and SW conceived and designed the study. ZC analyzed public databases of colorectal cancer patients. SH performed immunohistochemical analysis of colorectal tumor microarrays and SW and YL graded the results as a specialized pathologist. ZC and FW managed the preparation of the manuscript. SH and LL proofread the manuscript. All authors read and approved the final manuscript.

Funding

National Natural Science Foundation of China (NO.81772637).

Availability of data and materials

All analyzed data are included in this published article. The original data are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

The experiments were approved by the Ethics Committee of the Binzhou Medical University Hospital. Written informed consent was obtained from all patients. The experiments were carried out following the Declaration of Helsinki. No animal experiment is applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

Binzhou medical university Hospital, Binzhou, Shandong.

References

1. Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J]. *CA Cancer J Clin*, 2018, 68(6): 394-424.
2. Banskota S, Dahal S, Kwon E, et al. B-Catenin gene promoter hypermethylation by reactive oxygen species correlates with the migratory and invasive potentials of colon cancer cells [J]. *Cell Oncol (Dordr)* 2018;41 (5) : 569-580
3. Bhatlekar S, Fields JZ, Boman BM. HOX genes and their role in the development of human cancers [J]. *J Mol Med* 2014;92(8) : 811-823
4. Hong SN. Genetic and epigenetic alterations of colorectal cancer [J]. *Intest Res* 2018;16(3) : 327
5. McDonald A, Hata A, MacLean M, et al. Micro RNA and vascular remodelling in acute vascular injury and pulmonary vascular remodelling [J]. *Cardiovasc Res* 2012;93(4) : 594-604
6. Davis-Dusenbery BN, Wu C, Hata A. Micromanaging vascular smooth muscle cell differentiation and phenotypic modulation [J]. *Arterioscler Thromb Vasc Biol* 2011;31(11) : 2370-2377
7. Xu G, Zhang M, Zhu H, Xu J. A 15-gene signature for prediction of colon cancer recurrence and prognosis based on SVM. *Gene*. 2017;604:33-40. doi:10.1016/j.gene.2016.12.016
8. Sun D, Chen J, Liu L, et al. Establishment of a 12-gene expression signature to predict colon cancer prognosis. *PeerJ*. 2018;6:e4942. Published 2018 Jun 14. doi:10.7717/peerj.4942
9. Zuo S, Dai G, Ren X. Identification of a 6-gene signature predicting prognosis for colorectal cancer. *Cancer Cell Int*. 2019;19:6. Published 2019 Jan 5. doi:10.1186/s12935-018-0724-7
10. Jia R, Zhao H, Jia M, et al. Identification of co-expression modules and potential biomarkers of breast cancer by WGCNA [J]. *Gene*, 2020, 750: 144757.
11. Li A, Horvath S. Network neighborhood analysis with the multi-node topological overlap measure[J]. *BIOINFORMATICS*. 2007;23(2):222–31.
12. Yip AM, Horvath S. Gene network interconnectedness and the generalized topological overlap measure[J]. *BMC BIOINFORMATICS*. 2007;8:22.
13. Kanehisa M, Furumichi M, Tanabe M, et al: KEGG: New perspectives on genomes, pathways, diseases and drugs. 2017, *Nucleic Acids Res* 45: D353-D361.
14. Ashburner M, Ball CA, Blake JA, et al: Gene ontology: Tool for the unification of biology. The gene ontology consortium. 2000, *Nat Genet* 25: 25-29.
15. Aasebø E, Berven FS, Hovland R, et al. The Progression of Acute Myeloid Leukemia from First Diagnosis to Chemoresistant Relapse: A Comparison of Proteomic and Phosphoproteomic Profiles. *Cancers (Basel)*. 2020;12(6):1466. Published 2020 Jun 4. doi:10.3390/cancers12061466
16. de Alcântara BBR, Cruz FM, Fonseca FLA, et al. Chemotherapy-induced fatigue is associated with changes in gene expression in the peripheral blood mononuclear cell fraction of patients with locoregional breast cancer. *Support Care Cancer*. 2019;27(7):2479-2486. doi:10.1007/s00520-018-4519-0
17. Gao Y, Wu N, Wang S, Yang X, Wang X, Xu B. Concurrent mutations associated with trastuzumab-resistance revealed by single cell sequencing. *Breast Cancer Res Treat*. 2021;187(3):613-624. doi:10.1007/s10549-021-06237-0
18. Yang R, Zhan M, Guo M, et al. Yolk sac-derived Pcd11-positive cells modulate zebrafish microglia differentiation through the NF-κB-Tgfb1 pathway. *Cell Death Differ*. 2021;28(1):170-183. doi:10.1038/s41418-020-0591-3
19. Xing Z, Wang R, Wang X, et al. CircRNA circ-PDCD11 promotes triple-negative breast cancer progression via enhancing aerobic glycolysis. *Cell Death Discov*. 2021;7(1):218. Published 2021 Aug 21. doi:10.1038/s41420-021-00604-y
20. Xia R, Tang H, Shen J, et al. Prognostic value of a novel glycolysis-related gene expression signature for gastrointestinal cancer in the Asian population. *Cancer Cell Int*. 2021;21(1):154. Published 2021 Mar 4. doi:10.1186/s12935-021-01857-4
21. Terashima Y, Onai N, Murai M, et al. Pivotal function for cytoplasmic protein FROUNT in CCR2-mediated monocyte chemotaxis. *Nat Immunol*. 2005;6(8):827-835. doi:10.1038/ni1222
22. Terashima Y, Toda E, Itakura M, et al. Targeting FROUNT with disulfiram suppresses macrophage accumulation and its tumor-promoting properties. *Nat Commun*. 2020;11(1):609. Published 2020 Jan 30. doi:10.1038/s41467-020-14338-5
23. Locati M, Curtale G, Mantovani A. Diversity, Mechanisms, and Significance of Macrophage Plasticity[J]. *Annu Rev Pathol*. 2020;15:123-147. doi:10.1146/annurev-pathmechdis-012418-012718
24. Najafi M, Hashemi Goradel N, Farhood B, et al. Macrophage polarity in cancer: A review[J]. *J Cell Biochem*. 2019;120(3):2756-2765. doi:10.1002/jcb.27646

25. Locati M, Curtale G, Mantovani A. Diversity, Mechanisms, and Significance of Macrophage Plasticity[J]. Annu Rev Pathol. 2020;15:123-147. doi:10.1146/annurev-pathmechdis-012418-012718
26. Isidro RA, Appleyard CB. Colonic macrophage polarization in homeostasis, inflammation, and cancer[J]. Am J Physiol Gastrointest Liver Physiol. 2016;311(1):G59-G73. doi:10.1152/ajpgi.00123.2016
27. Najafi M, Hashemi Goradel N, Farhood B, et al. Macrophage polarity in cancer: A review[J]. J Cell Biochem. 2019;120(3):2756-2765. doi:10.1002/jcb.27646
28. Toor SM, Murshed K, Al-Dhaheeri M, Khawar M, Abu Nada M, Elkord E. Immune Checkpoints in Circulating and Tumor-Infiltrating CD4+ T Cell Subsets in Colorectal Cancer Patients. Front Immunol. 2019;10:2936. Published 2019 Dec 17. doi:10.3389/fimmu.2019.02936
29. Huang R, Wang K, Gao L, Gao W. TIMP1 Is A Potential Key Gene Associated With The Pathogenesis And Prognosis Of Ulcerative Colitis-Associated Colorectal Cancer. Onco Targets Ther. 2019;12:8895-8904. Published 2019 Oct 30. doi:10.2147/OTT.S222608
30. Song G, Xu S, Zhang H, et al. TIMP1 is a prognostic marker for the progression and metastasis of colon cancer through FAK-PI3K/AKT and MAPK pathway. J Exp Clin Cancer Res. 2016;35(1):148. Published 2016 Sep 20. doi:10.1186/s13046-016-0427-7
31. Chen L, Lu D, Sun K, et al. Identification of biomarkers associated with diagnosis and prognosis of colorectal cancer patients based on integrated bioinformatics analysis. Gene. 2019;692:119-125. doi:10.1016/j.gene.2019.01.001
32. Yamada Y, Arai T, Sugawara S, et al. Impact of novel oncogenic pathways regulated by antitumor miR-451a in renal cell carcinoma. Cancer Sci. 2018;109(4):1239-1253. doi:10.1111/cas.13526
33. Yang C, Ren J, Li B, et al. Identification of clinical tumor stages related mRNAs and miRNAs in cervical squamous cell carcinoma. Pathol Res Pract. 2018;214(10):1638-1647. doi:10.1016/j.prp.2018.07.035
34. Davalieva K, Kiprijanovska S, Broussard C, Petrusevska G, Efremov GD. Proteomic analysis of infiltrating ductal carcinoma tissues by coupled 2-D DIGE/MS/MS analysis. Mol Biol (Mosk). 2012;46(3):469-480.
35. Chang SW, Yue J, Wang BC, Zhang XL. miR-503 inhibits cell proliferation and induces apoptosis in colorectal cancer cells by targeting E2F3. Int J Clin Exp Pathol. 2015;8(10):12853-12860. Published 2015 Oct 1.
36. Wu S, Wu F, Jiang Z. Identification of hub genes, key miRNAs and potential molecular mechanisms of colorectal cancer. Oncol Rep. 2017;38(4):2043-2050. doi:10.3892/or.2017.5930
37. Yang B, Du K, Yang C, et al. CircPRMT5 circular RNA promotes proliferation of colorectal cancer through sponging miR-377 to induce E2F3 expression. J Cell Mol Med. 2020;24(6):3431-3437. doi:10.1111/jcmm.15019
38. Fang Y, Gu X, Li Z, et al. miR-449b inhibits the proliferation of SW1116 colon cancer stem cells through downregulation of CCND1 and E2F3 expression. Oncol Rep. 2013;30(1):399-406. doi:10.3892/or.2013.2465
39. Liu J, Shao Y, He Y, et al. MORC2 promotes development of an aggressive colorectal cancer phenotype through inhibition of NDRG1. Cancer Sci. 2019;110(1):135-146. doi:10.1111/cas.13863
40. Tuupanen S, Hänninen UA, Kondelin J, et al. Identification of 33 candidate oncogenes by screening for base-specific mutations. Br J Cancer. 2014;111(8):1657-1662. doi:10.1038/bjc.2014.429

Tables

Tab.1 The expression of NUP85 in colorectal cancer and normal [n (%)]

	n	NUP85
Normal	50	10%20.0%
Tumor	50	27%54.0%*

*P<0.05

Tab.2 Correlation between NUP85 and immune marker in colorectal cancer

		n	CD4		r	p	CD163		r	P
			+	-			+	-		
NUP85	+	29	9	20	-0.446	0.001*	12	17	-0.364	0.009*
	-	21	16	5			17	4		

*P<0.05

Tab.3 The prediction model of gene prognosis was established by multivariate Cox analysis

Gene	coef	HR	HR.95L	HR.95H	pvalue
E2F3	0.139	1.149	1.052	1.255	0.002
PMM2	-0.253	0.776	0.623	0.967	0.024
MORC2	0.099	1.104	1.007	1.210	0.035
TIMP1	0.003	1.003	1.001	1.006	0.001

Figures

Figure 1

WGCNA analysis of GSE41657 and GSE74602 dataset. a Differential genes in GSE41657. b Differential genes in GSE74602. c Determination of soft threshold (β) for weighted gene co-expression network analysis. d Gene cluster dendrogram. e Heat map of correlation between clinical features and module features. f-h The figures type are the same as dataset GSE41657. i Co-expression DEGs of the most significant modules in two datasets. j-k Go barplotPval and bubble of co-expression differential genes. l-m KEGG barplotPval and bubble of co-expression differential genes.

Figure 2

Kaplan-Meier survival curves of 26 prognostic genes.

Figure 3
Expression levels of genes in clinical tissue samples. Representative images of immunohistochemistry staining of five unstudied genes in COAD from the Human Protein Atlas (HPA) database.

Figure 4
PPI analysis and target gene screening of co-expressed genes. a PPI network of 439 DEGs. Different colors represent different modules, and blue is the other genes. The red module is the most important module, including 32 genes. b GO enrichment analysis of the most important module genes. c KEGG enrichment analysis of the most important module genes.

Figure 5

Investigating the function of NUP85 and verifying the relationship between NUP85 and the immune microenvironment. a-b Scatter plot and pair plot of gene expression differences of key genes. c GO enrichment analysis of NUP85. d KEGG enrichment analysis of NUP85. e-j NUP85 is associated with tumor-associated macrophage markers. k The expression of NUP85 in normal tissues. l The expression of NUP85 in tumor tissues. m The expression of CD4 in CD4+T cell. n The expression of CD8 in CD8+T cell. o The expression of CD20 in CD20+ lymphocyte. p The expression of CD163 in M2 TAM.

Figure 6

Prognostic model.a Differential genes screened by significant.The red is the high-risk gene, and the green is the low-risk gene.b Differential genes screened by univariate Cox and K-M.c Kaplan-Meier plotter of the high-risk and low-risk COAD patients in the training group.d Risk score distribution of COAD patients with different risks in the training set and verification set. e Scatterplots of COAD patients with different survival status in the training set and verification set. f The heatmap of risk genes in COAD between high-risk and low-risk patients in the training and verification sets.g Univariate COX was independent of post-analysis.h Multivariate COX was independent of post-analysis.i Time-dependent ROC curves for predicting one-year, three-year, and five-year survival.

Supplementary Files

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

- [GraphicalAbstract1.jpg](#)