frontiers in Oncology

ORIGINAL RESEARCH
published: 28 October 2020
doi: 10.3389/fonc.2020.573295

Differential Expression Analysis Revealing CLCA1 to Be a Prognostic and Diagnostic Biomarker for Colorectal Cancer

*Fang-Ze Wei, Shi-Wen Mei, Zhi-Jie Wang, Jia-Nan Chen, Hai-Yu Shen, Fu-Qiang Zhao, Juan Li, Zheng Liu and Qian Liu**

Department of Colorectal Surgery, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union College, Beijing, China

OPEN ACCESS

Edited by:
Stephen J. Pandol,
Cedars Sinai Medical Center,
United States

Reviewed by:
Deguang Kong,
University of Colorado Denver,
United States
Xiaojie Wang,
Fujian Medical University, China
Liping Sun,
The First Affiliated Hospital of China
Medical University, China

***Correspondence:**
Qian Liu
fcwpumch@163.com

Specialty section:
This article was submitted to
Gastrointestinal Cancers,
a section of the journal
Frontiers in Oncology

Received: 16 June 2020
Accepted: 02 October 2020
Published: 28 October 2020

Citation:
Wei F-Z, Mei S-W, Wang Z-J, Chen J-N, Shen H-Y, Zhao F-Q, Li J, Liu Z and Liu Q (2020) Differential Expression Analysis Revealing CLCA1 to Be a Prognostic and Diagnostic Biomarker for Colorectal Cancer. Front. Oncol. 10:573295. doi: 10.3389/fonc.2020.573295

Colorectal cancer (CRC) is a common malignant tumor of the digestive tract and lacks specific diagnostic markers. In this study, we utilized 10 public datasets from the NCBI Gene Expression Omnibus (NCBI-GEO) database to identify a set of significantly differentially expressed genes (DEGs) between tumor and control samples and WGCNA (Weighted Gene Co-Expression Network Analysis) to construct gene co-expression networks incorporating the DEGs from The Cancer Genome Atlas (TCGA) and then identify genes shared between the GEO datasets and key modules. Then, these genes were screened *via* MCC to identify 20 hub genes. We utilized regression analyses to develop a prognostic model and utilized the random forest method to validate. All hub genes had good diagnostic value for CRC, but only CLCA1 was related to prognosis. Thus, we explored the potential biological value of CLCA1. The results of gene set enrichment analysis (GSEA) and immune infiltration analysis showed that CLCA1 was closely related to tumor metabolism and immune invasion of CRC. These analysis results revealed that CLCA1 may be a candidate diagnostic and prognostic biomarker for CRC.

Keywords: CRC, diagnostic, biomarker, prognostic model, database

INTRODUCTION

Colorectal cancer (CRC) is the third most common cancer and the second most deadly cancer worldwide (1). The incidence and mortality of CRC continue to increase because of the lack of diagnostic biomarkers and inadequate understanding of the molecular mechanism (2). Detection and monitoring of CRC occurrence and progression are dependent on a combination of radiologic examinations and serum biomarker measurements (3); however, these methods have some limitations. In some cases, the levels of biomarkers do not change. In other diseases, the levels of biomarkers can change (4, 5). In addition, some patients do not undergo colonoscopy because of the discomfort of this procedure (6). In the past few decades, advanced gene microarray and high-throughput sequencing technologies have been used to explore novel gene expression, treatment targets, and pathogenesis in CRC (7).

Robust rank aggregation (RRA) has been utilized in various recent cancer studies to overcome the limitations of substantial interstudy variability and the different statistical analysis methods used with different technological platforms (8, 9). In our study, we used RRA to analyze 10 microarray datasets from the Gene Expression Omnibus (GEO) database and explored data from The Cancer Genome Atlas (TCGA) through WGCNA to identify differentially expressed genes (DEGs). Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses were used to explore the potential functions of these DEGs. We utilized regression analyses and the random forest method to develop and validate the prognostic model. Among the genes included in the model, we used MCC to calculate the top 20 hub genes. We explored biological functions through GO and KEGG analyses and utilized ROC curves to explore diagnostic value. We also explored the relationship between them. Based on the 20 hub genes, we utilized Kaplan-Meier (K-M) analysis to explore relationships with prognosis, and only CLCA1 had a close relationship with prognosis. We continued to explore the potential biological value of CLCA1. In addition, we utilized the online tool TISIDB and R packages to explore the functions of these genes in immunity and performed gene set enrichment analysis (GSEA) to investigate their potential functions in CRC.

MATERIALS AND METHODS

Gene Expression Datasets

All microarray datasets were downloaded from the TCGA and GEO databases. The RNA sequencing data were downloaded from the TCGA database (https://portal.gdc.cancer.gov/), which contained 41 control tissues and 482 CRC tissues with clinical data. Other datasets that satisfied the following criteria were downloaded from GEO (http://www.ncbi.nlm.nih.gov/geo/): 1) Gene expression data in the microarray datasets included data for both control tissues and CRC tissues, and 2) each microarray contained a minimum of 5 tumor and control tissues. According to the above criteria, 10 GEO datasets were incorporated in this study: GSE9348 (10), GSE44076 (11), GSE4183 (12), GSE20916 (13), GSE37364 (14), GSE44861 (15), GSE81558 (16), GSE22598 (17), GSE113513, and GSE110224 (18).

Identification of Significant DEGs in CRC Samples

We downloaded the series matrix files from GEO and screened them with the R package "limma" for normalization and DEG identification. Then, the RRA method was utilized to integrate the results of these 10 datasets to identify the most significantly upregulated and downregulated genes (**Supplementary file 1**). Genes with an adjusted P value of <0.05 were considered significantly differentially expressed. For TCGA database analysis, we first separated mRNA and lncRNA data and used the R package "edgeR" (19) to identify DEGs. The following criteria were used to select DEGs: $|\log(\mathrm{foldchange})|>2$ and P value<0.01 (**Supplementary file 2**). After obtaining the 2 sets of DEGs, we used the R package "WGCNA" to identify clinical trait-related modules (20). We used the online tool "VENN" (http://bioinformatics.psb.ugent.be/webtools/Venn/) to generate a Venn diagram to identify genes shared between the key modules from the TCGA and GEO datasets (21). We ultimately obtained 129 DEGs.

GO and KEGG Functional Enrichment Analyses

We conducted GO enrichment analysis using the online tool Database for Annotation, Visualization, and Integrated Discovery (DAVID; https://david.ncifcrf.gov/) (22) and the R packages "digest" and "GOplot"; an adjusted P value of <0.05 was considered statistically significant. For KEGG pathway analyses, we used the R packages "clusterprofiler" (23), "org.Hs.eg.db", "enrichplot", and "ggplot2", with an adjusted P value of <0.05 considered statistically significant. Both GO enrichment and KEGG analysis results were visualized using the R package "GOplot".

Hub Genes from the DEG Network

We utilized the online STRING database (https://string-db.org/cgi/input.pl/) to explore connections among the DEGs and visualized these connections by constructing a PPI network with Cytoscape software (version 3.6.1) (24). We utilized cytoHubba MCC to calculate the top 20 hub genes. We analyzed relationships between the 20 hub genes using the R package "psych".

Development and Validation of the Prognostic Model

We utilized R (version 3.6.1) to generate a matrix that included the clinical information and DEG expression. We used Cox regression analysis to build the prognostic model using the R package "survival" (25, 26) and online tool "SangerBox". Then, we utilized the R package "randomForest" to validate the prognostic model through risk score and calculate the accuracy, rrror rate, sensitivity and precision from a confusion matrix. The prognostic model was based on the TCGA database.

Diagnostic and Prognostic Value of the Hub Genes

We utilized SPSS to explore the diagnostic value of the genes for CRC and K-M analysis to determine the prognostic value. We validated the differential expression levels between control tissue and tumor tissue with the R packages "limma" and "beeswarm"

Abbreviations: Act_B, Activated B cell; Act_CD4, Activated CD4 T cell; Act_CD8, Activated CD8 T cell; ACT_DC, Activated dendritic cell; CD56bright, CD56bright natural killer cell; CD56dim, CD56dim natural killer cell; DEG, differentially expressed gene; GEO, Gene Expression Omnibus; GO, Gene Ontology; GSEA, Gene Set Enrichment Analysis; iDC, Immature dendritic cell; Imm_B, Immature B cell; KEGG, Kyoto Encyclopedia of Genes and Genomes; Mast, Mast cell; MCC, Maximal Clique Centrality; MDSC, Myeloid-derived suppressor cell; Mem_B, Memory B cell; NK, Natural killer cell; NKT, Natural killer T cell; pDC, Plasmacytoid dendritic cell; ROC, Receiver Operating Characteristic; RRA, robust rank aggregation; Tcm_CD4, Central memory CD4 T cell; Tcm_CD8, Central memory CD8 T cell; TCGA, the Cancer Genome Atlas; Tem_CD4, Effector memory CD4 T cell; Tem_CD8, Effector memory CD8 T cell; Tfh, T follicular helper cell; Tgd, Gamma delta T cell; Th1, Type 1 T helper cell; Th17, Type 17 T helper cell; Th2, Type 2 T helper cell; TNM, Tumor Node Metastasis; Treg, Regulatory T cell; WGCNA, Weighted Gene Co-Expression Network Analysis.

utilized GSE44076. We utilized the Wilcoxon and Kruskal-Wallis tests to explore the relationship between gene expression and clinical features in the TCGA-COAD and TCGA-READ datasets.

Analysis of the Association of Hub Gene Expression With Tumor-Infiltrating Immune Cell Infiltration

We utilized TISIDB (http://cis.hku.hk/TISIDB/) to explore the relationship between the expression of genes and infiltration of tumor-infiltrating immune cells, including $CD4^+$ T cells, $CD8^+$ T cells, B cells, neutrophils, monocytes, eosinophils, mast cells, DCs, NKT cells, NK cells, MDSCs, and CD56 cells (27, 28). TISIDB is an online tool that includes genomic, transcriptomic and clinical data for 30 cancer types from the TCGA database.

GSEA of Hub Genes

We utilized GSEA, which was downloaded from (https://www.gsea-msigdb.org/gsea/msigdb), to explore the functions of the hub genes. We performed GSEA of the hub genes with the R package "clusterprofiler" (29) in data downloaded from the TCGA-COAD and TCGA-READ datasets and divided 482 samples into two groups: high expression and low expression. We utilized "c2.cp.kegg.v6.2.symbols.gmt" for analysis and to select the top 5 genes. Then, we used the R packages "plyr", "ggplot2", "grid", and "gridExtra" to integrate different significant pathways into a single diagram.

Validation of Protein Expression and Prognostic Value of CLCA1

We utilized GEO online tools PROGgene online database (http://genomics.jefferson.edu/proggene/), The Human Protein Atlas (https://www.proteinatlas.org/), and Kaplan-Meier Plotter (http://kmplot.com/analysis/) to explore the protein expression and prognostic value of CLCA1. The Human Protein Atlas is the online database which provides the distribution of human proteins in tissues and cells, and immunohistochemical techniques are used to examine the distribution and expression of each protein in 48 normal tissues and 20 tumor tissues. Kaplan-Meier Plotter is the online database which including the data from GEO, EGA and TCGA.

RESULTS

Identification of the Significantly Differentially Expressed Genes in the Datasets

Figure 1 shows the workflow of our study. We downloaded CRC samples from the TCGA-COAD and TCGA-READ datasets and identified DEGs between the control tissues and tumor tissues. In total, 2097 genes in the TCGA-COAD dataset and 2887 genes in the TCGA-READ dataset were differentially expressed between tumor and control tissues. The volcano plots of these genes are shown in **Figure 2**. According to the selection criteria for the GEO data, we selected 10 eligible CRC datasets for exploration. The characteristics of all datasets are shown in **Table 1**. RRA analysis of the GEO datasets identified 212 significantly downregulated and 136 significantly upregulated genes. The top 20 downregulated and top 20 upregulated genes are shown in a heatmap (**Figure 3**).

WGCNA and Identification of DEGs

To identify the key modules most associated with CRC clinical traits, we performed WGCNA on the significant genes in the TCGA-COAD and TCGA-READ datasets (**Figures 4A–E**). Clinical information such as age, TNM grade, and survival time was retrieved from TCGA. By setting a soft-thresholding power of 5 (scale free R^2 = 0.89), we eventually identified 5 modules. From the heatmap of module-trait correlations, we found that the bule module was the most highly correlated with clinical traits, especially the futime (P=5.2e-10; **Figure 4F**). The

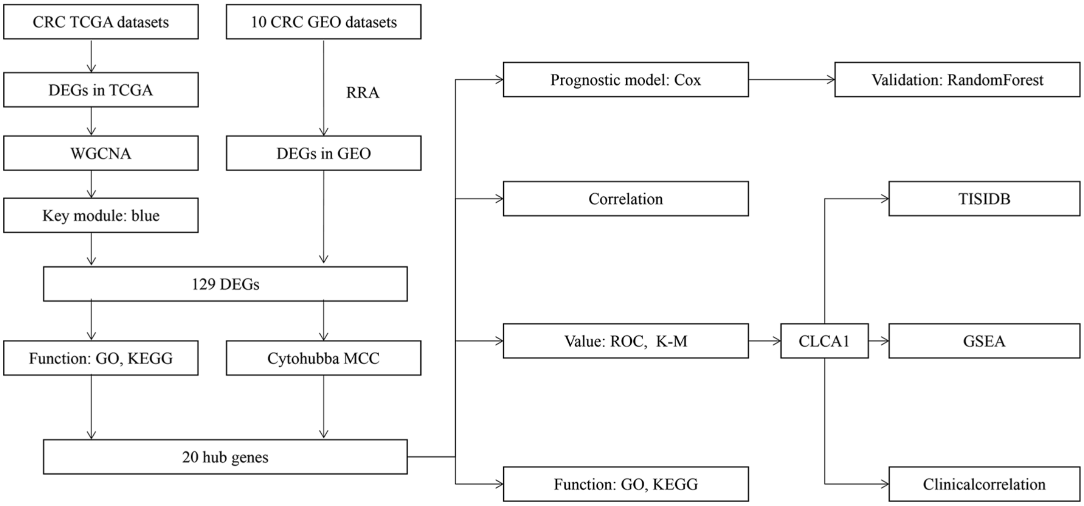


FIGURE 1 | Analysis workflow of this study.

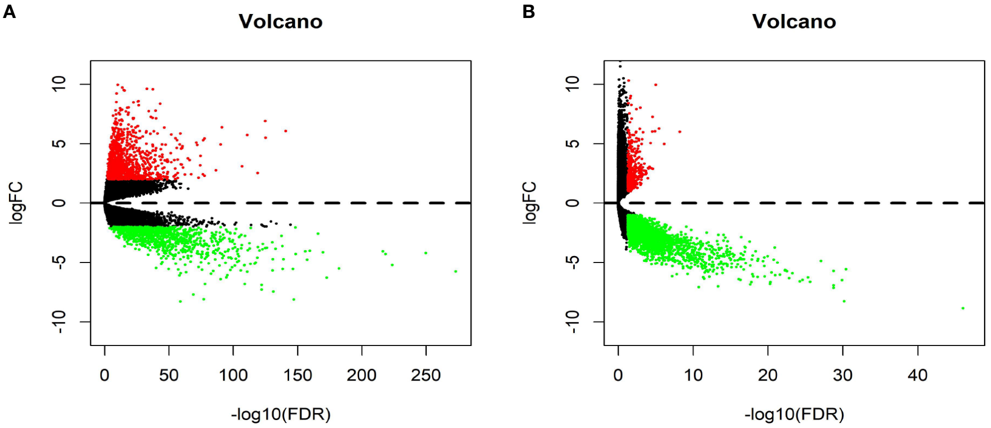


FIGURE 2 | Volcano plot of The Cancer Genome Atlas (TCGA) data. **(A)** Volcano plot depicting the differential expression and distribution of TCGA-COAD data. **(B)** Volcano plot depicting the differential expression and distribution of TCGA-READ data.

blue module contained a total of 299 genes, as shown in **Figure 4F**. We combined the genes from the blue module and the RRA analysis and used Venn diagrams to identify significantly DEGs common to the 2 datasets, as shown in **Figure 5A**. The 129 DEGs were visualized using STRING and Cytoscape software, and CytoHubba MCC was used to calculate the top 20 hub genes, as shown in **Figure 5B**.

GO and KEGG Enrichment Analyses of DEGs

We used DAVID to explore the main 3 categories of GO enrichment: biological process (BP), cellular component (CC), and molecular function (MF). In the BP category, we explored bicarbonate transport (P=1.62E-08), one-carbon metabolic process (P=1.57E-06), negative regulation of growth (P=6.97E-06), cellular response to zinc ion (P=6.97E-06) and regulation of intracellular pH (P=9.70E-05) **(Figure 5C)**. In the CC category, we identified plasma membrane (P=7.41E-04), anchored component of membrane (P=0.001050845), integral component of plasma membrane (P=0.007269043), extracellular space (P=0.009771536) and integral component of membrane (P=0.013058288) **(Figure 5D)**. In the MF category, we explored carbonate dehydratase activity (P=1.81E-06), hormone activity (P=4.20E-04), zinc ion binding (P=7.40E-04), chloride channel activity (P=4.76E-04) and transporter activity (P=0.002384793) (**Figure 5E**). For KEGG pathway analysis, we explored the top 5 pathways that satisfied the criteria of pFilter<0.05 and adjPfilter<1: nitrogen metabolism, pentose and glucuronate interconversions, retinol metabolism, ascorbate and aldarate metabolism, and steroid hormone biosynthesis (**Figure 5F**).

TABLE 1 | Characteristics of the datasets.

Dataset	No. of Normal	No. of Tumor	Platform ID	No. of Row Perl Platforms
TCGA-COAD	39	398	RNAseq	17557
TCGA-READ	2	84	RNAseq	17418
GSE9348	12	70	GPL570	54,675
GSE44076	148	98	GPL13667	49,386
GSE4183	15	38	GPL570	54,675
GSE20916	109	36	GPL570	54,675
GSE37364	67	27	GPL570	54,675
GSE44861	55	56	GPL3921	22,277
GSE81558	9	42	GPL15207	49,395
GSE22598	17	17	GPL570	54,675
GSE113513	14	14	GPL15207	49,395
GSE110224	17	17	GPL570	54,675

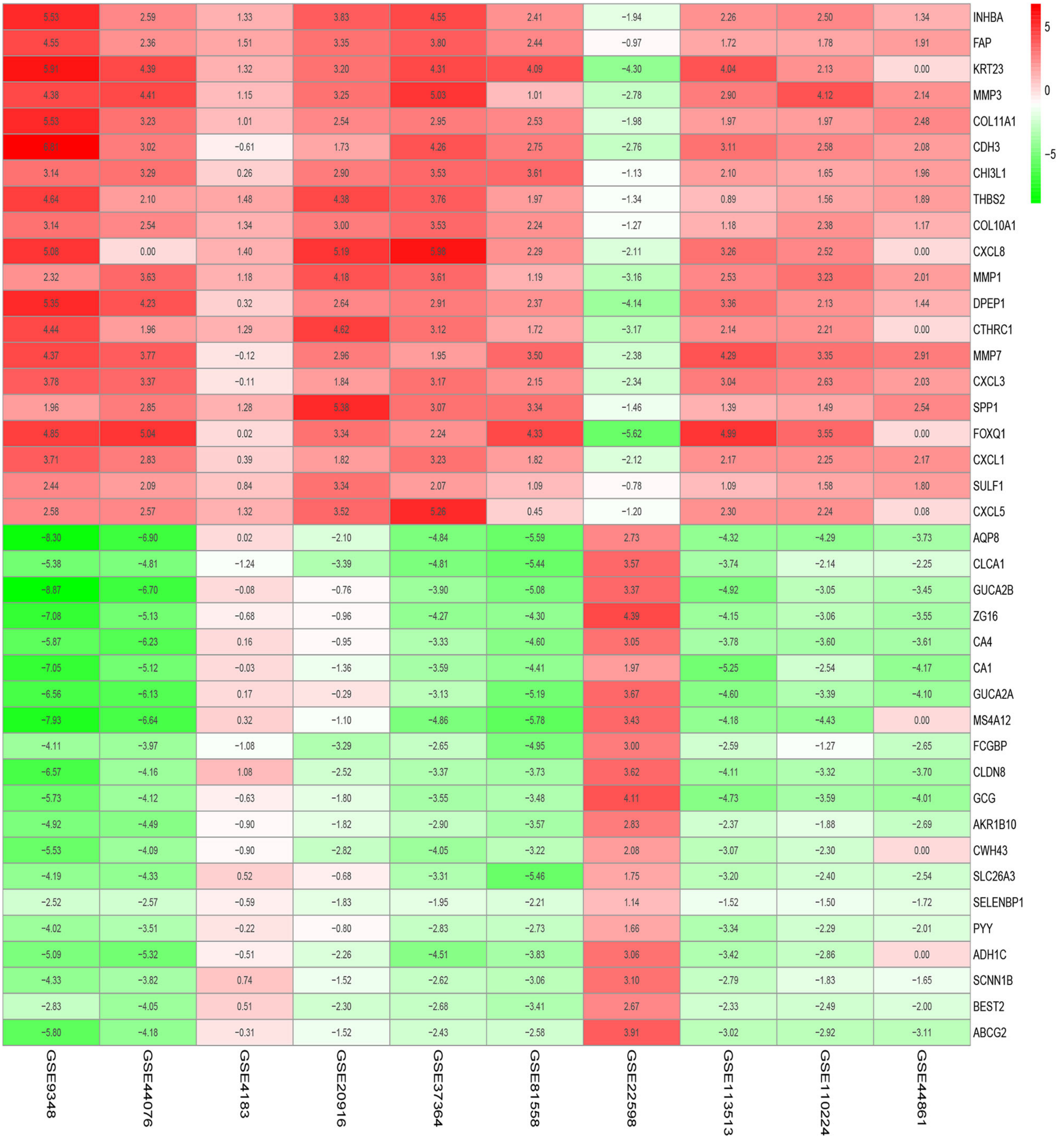
TCGA-COAD, TCGA-Colon adenocarcinoma; TCGA-READ, TCGA-Rectum adenocarcinoma; GSE, Gene Expression Omnibus Series; GPL, Gene Expression Omnibus Platform.

Biological Value of the Hub Genes

Through CytoHubba MCC calculation, we obtained 20 hub genes. The 20 hub genes, which are shown in **Figure 6A**, were also closely related to each other. We utilized SPSS to explore their diagnostic value. ROC curve analysis showed that these 20 genes have high diagnostic value for CRC: CLCA1 AUC= 0.959, TMIGD1 AUC= 0.998, SLC30A10 AUC= 0.993, MT1F AUC= 0.933, MT1M AUC= 0.975, MT1G AUC= 0.944, MT1H AUC= 0.947, MT1E AUC=0.943, GUCA2B AUC=0.991, GUCA2A AUC=0.99, SLC26A3 AUC= 0.989, CLCA4 AUC= 0.984, MS4A12 AUC= 0.978, SI AUC= 0.94, SLC9A2 AUC= 0.959, GCG AUC= 0.992, PYY AUC= 0.993, SST AUC= 0.992, SLC4A4 AUC= 0.997, and SLC16A9 AUC= 0.903 (**Figure 6B**). We also explored the prognostic value of the hub genes, and only CLCA1 was closely related to survival time (**Figure 6C**); the other genes are shown in **Supplementary Figure 1**. To further explore the functions of the hub genes, we conducted GO and KEGG analyses. The most significant GO terms for BPs, CCs, and MFs, as well as KEGG pathways, are shown in **Figures 6D, E**.

Development and Validation of a Prognostic Model Based on the Hub Genes

We utilized Cox proportional hazards regression analysis of the survival-related genes to develop the prognostic model



	GSE9348	GSE44076	GSE4183	GSE20916	GSE37364	GSE81558	GSE22598	GSE113513	GSE110224	GSE44861
INHBA	5.53	2.59	1.33	3.83	4.55	2.41	−1.94	2.26	2.50	1.34
FAP	4.55	2.36	1.51	3.35	3.80	2.44	−0.97	1.72	1.78	1.91
KRT23	5.91	4.39	1.32	3.20	4.31	4.09	−4.30	4.04	2.13	0.00
MMP3	4.38	4.41	1.15	3.25	5.03	1.01	−2.78	2.90	4.12	2.14
COL11A1	5.53	3.23	1.01	2.54	2.95	2.53	−1.98	1.97	1.97	2.48
CDH3	6.81	3.02	−0.61	1.73	4.26	2.75	−2.76	3.11	2.58	2.08
CHI3L1	3.14	3.29	0.26	2.90	3.53	3.61	−1.13	2.10	1.65	1.96
THBS2	4.64	2.10	1.48	4.38	3.76	1.97	−1.34	0.89	1.56	1.89
COL10A1	3.14	2.54	1.34	3.00	3.53	2.24	−1.27	1.18	2.38	1.17
CXCL8	5.08	0.00	1.40	5.19	5.98	2.29	−2.11	3.26	2.52	0.00
MMP1	2.32	3.63	1.18	4.18	3.61	1.19	−3.16	2.53	3.23	2.01
DPEP1	5.35	4.23	0.32	2.64	2.91	2.37	−4.14	3.36	2.13	1.44
CTHRC1	4.44	1.96	1.29	4.62	3.12	1.72	−3.17	2.14	2.21	0.00
MMP7	4.37	3.77	−0.12	2.96	1.95	3.50	−2.38	4.29	3.35	2.91
CXCL3	3.78	3.37	−0.11	1.84	3.17	2.15	−2.34	3.04	2.63	2.03
SPP1	1.96	2.85	1.28	5.38	3.07	3.34	−1.46	1.39	1.49	2.54
FOXQ1	4.85	5.04	0.02	3.34	2.24	4.33	−5.62	4.99	3.55	0.00
CXCL1	3.71	2.83	0.39	1.82	3.23	1.82	−2.12	2.17	2.25	2.17
SULF1	2.44	2.09	0.84	3.34	2.07	1.09	−0.78	1.09	1.58	1.80
CXCL5	2.58	2.57	1.32	3.52	5.26	0.45	−1.20	2.30	2.24	0.08
AQP8	−8.30	−6.90	0.02	−2.10	−4.84	−5.59	2.73	−4.32	−4.29	−3.73
CLCA1	−5.38	−4.81	−1.24	−3.39	−4.81	−5.44	3.57	−3.74	−2.14	−2.25
GUCA2B	−8.87	−6.70	−0.08	−0.76	−3.90	−5.08	3.37	−4.92	−3.05	−3.45
ZG16	−7.08	−5.13	−0.68	−0.96	−4.27	−4.30	4.39	−4.15	−3.06	−3.55
CA4	−5.87	−6.23	0.16	−0.95	−3.33	−4.60	3.05	−3.78	−3.60	−3.61
CA1	−7.05	−5.12	−0.03	−1.36	−3.59	−4.41	1.97	−5.25	−2.54	−4.17
GUCA2A	−6.56	−6.13	0.17	−0.29	−3.13	−5.19	3.67	−4.60	−3.39	−4.10
MS4A12	−7.93	−6.64	0.32	−1.10	−4.86	−5.78	3.43	−4.18	−4.43	0.00
FCGBP	−4.11	−3.97	−1.08	−3.29	−2.65	−4.95	3.00	−2.59	−1.27	−2.65
CLDN8	−6.57	−4.16	1.08	−2.52	−3.37	−3.73	3.62	−4.11	−3.32	−3.70
GCG	−5.73	−4.12	−0.63	−1.80	−3.55	−3.48	4.11	−4.73	−3.59	−4.01
AKR1B10	−4.92	−4.49	−0.90	−1.82	−2.90	−3.57	2.83	−2.37	−1.88	−2.69
CWH43	−5.53	−4.09	−0.90	−2.82	−4.05	−3.22	2.08	−3.07	−2.30	0.00
SLC26A3	−4.19	−4.33	0.52	−0.68	−3.31	−5.46	1.75	−3.20	−2.40	−2.54
SELENBP1	−2.52	−2.57	−0.59	−1.83	−1.95	−2.21	1.14	−1.52	−1.50	−1.72
PYY	−4.02	−3.51	−0.22	−0.80	−2.83	−2.73	1.66	−3.34	−2.29	−2.01
ADH1C	−5.09	−5.32	−0.51	−2.26	−4.51	−3.83	3.06	−3.42	−2.86	0.00
SCNN1B	−4.33	−3.82	0.74	−1.52	−2.62	−3.06	3.10	−2.79	−1.83	−1.65
BEST2	−2.83	−4.05	0.51	−2.30	−2.68	−3.41	2.67	−2.33	−2.49	−2.00
ABCG2	−5.80	−4.18	−0.31	−1.52	−2.43	−2.58	3.91	−3.02	−2.92	−3.11

FIGURE 3 | Significantly differentially expressed genes in Gene Expression Omnibus (GEO) datasets by robust rank aggregation (RRA) analysis. These heatmaps show the top 20 downregulated and top 20 upregulated genes. Each column indicates one dataset, and each row indicates one gene. Green indicates downregulation, and red indicates upregulation. The numbers in the heatmap indicate the logarithmic fold changes in the expression of each gene in the dataset.

(**Figures 7A–C**). According to the prognostic risk score value, CRC patients were divided into a low-risk and a high-risk group. The risk score distribution was analyzed and is shown in **Figure 7A**. The risk scores reflected the 1-year, 3-year and 5-year survival rates of CRC patients. The AUCs for 1-year, 3-year and 5-year survival are shown in **Figure 7B**. K-M curves were used to show the relationship of the risk score with overall survival (OS) in the low-risk and high-risk groups and verified that a low risk score had a stronger positive association with OS (P=0.0079; **Figure 7C**).

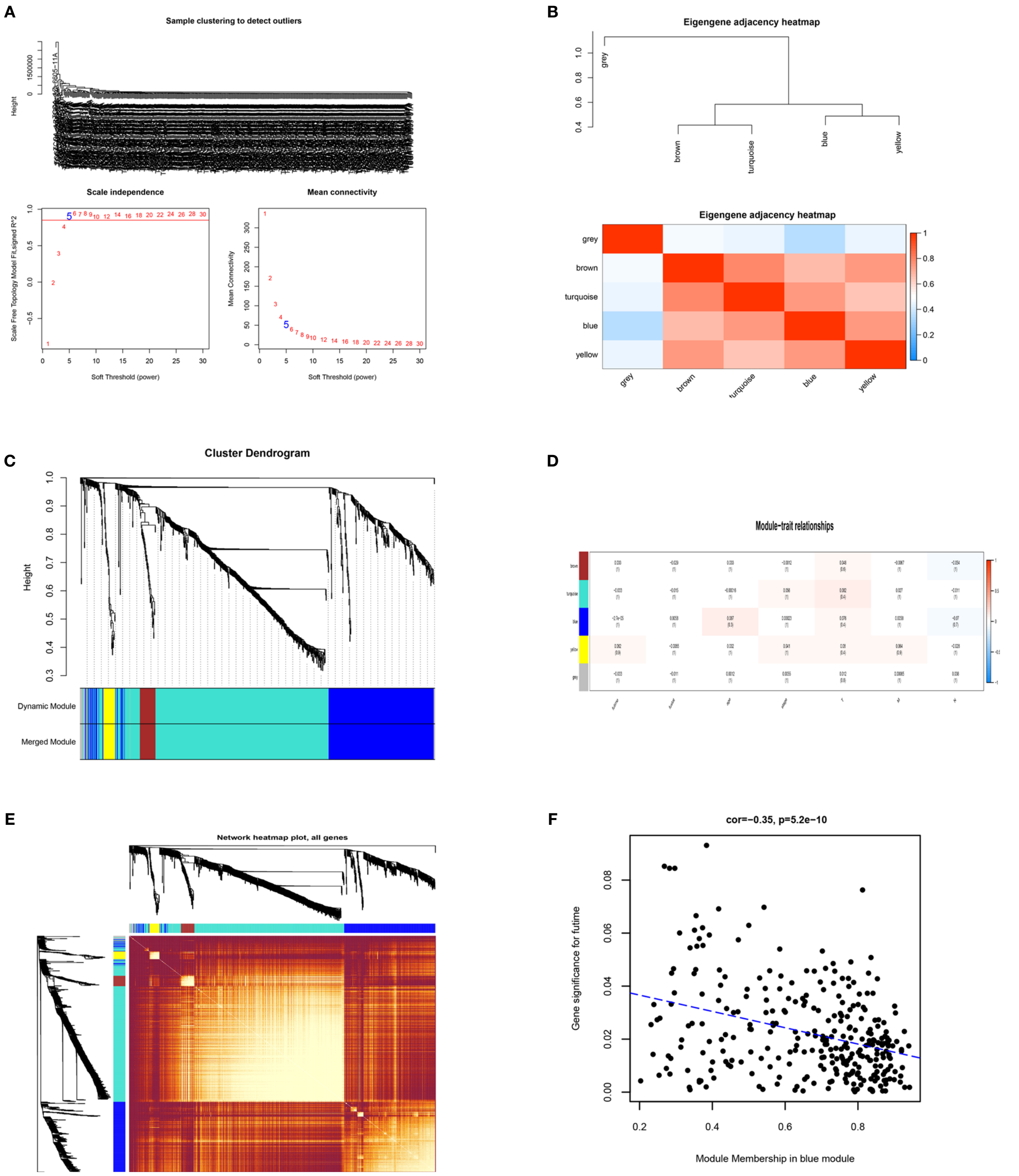


FIGURE 4 | Identification of key modules correlated with clinical traits in the The Cancer Genome Atlas (TCGA)-COAD and TCGA-READ datasets through Weighted Gene Co-Expression Network Analysis (WGCNA). **(A)** Clustering dendrograms of genes. Analysis of the scale-free fit index (left) and the mean connectivity (right) for various soft-thresholding powers. **(B)** Clustering of module eigengenes and a heatmap of adjacent eigengenes. **(C)** Dendrogram of all differentially expressed genes (DEGs) clustered from TCGA based on a dissimilarity measure. **(D)** Heatmap of the correlation between module eigengenes and clinical traits of colorectal cancer (CRC). Each cell contains the correlation coefficient and P value. **(E)** TOM network heatmap of all genes. **(F)** Scatter plot of module eigengenes in the blue module. The genes in the blue module with significance in futime.

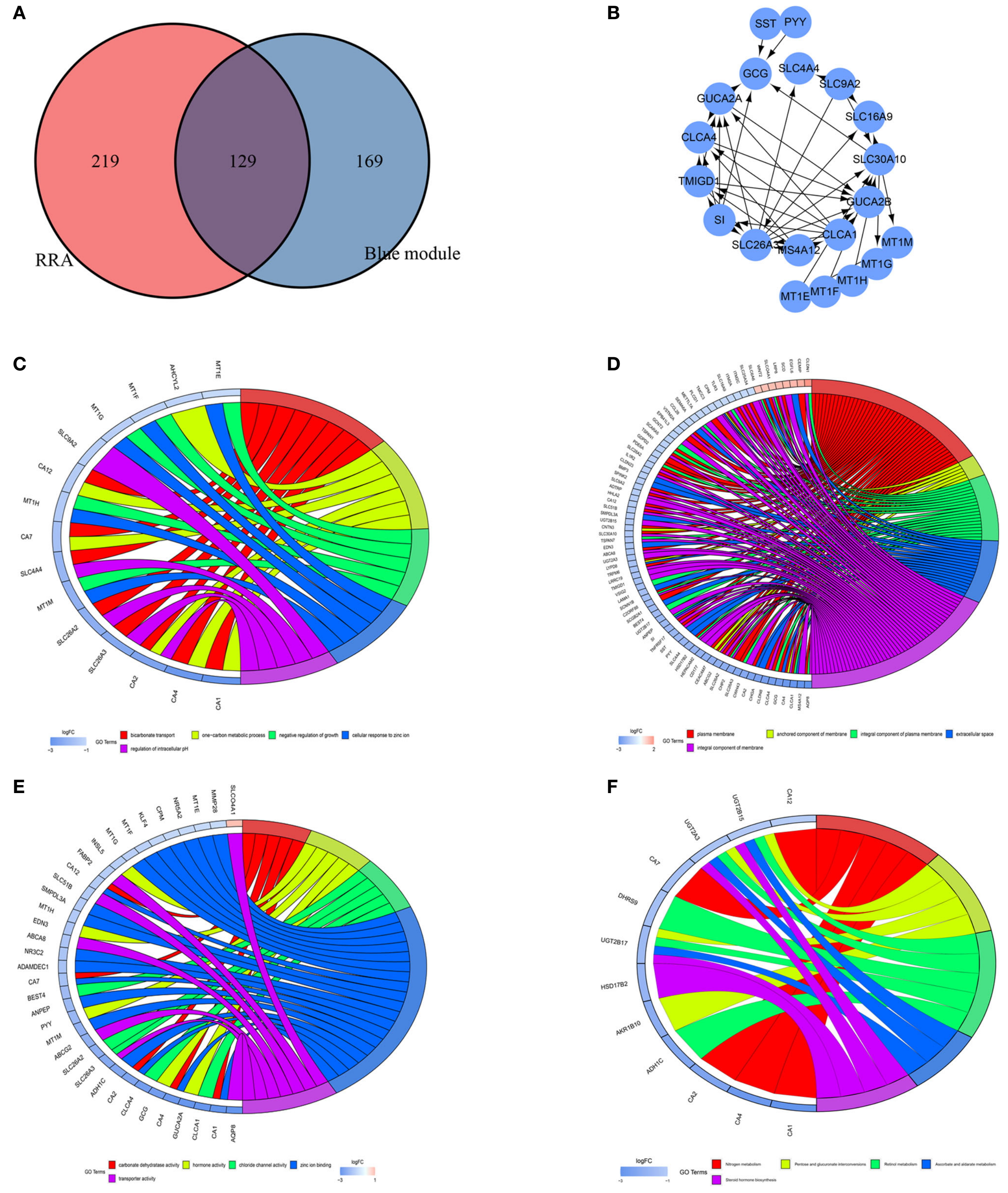


FIGURE 5 | Identification of differentially expressed genes (DEGs) and hub genes. Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses of the DEGs. **(A)** Venn diagram showing the numbers of DEGs. Red indicates blue module genes in Weighted Gene Co-Expression Network Analysis (WGCNA) from the TCGA-COAD and TCGA-READ datasets. Blue indicates significantly differentially expressed genes in robust rank aggregation (RRA) analysis of Gene Expression Omnibus (GEO) datasets. **(B)** PPI network of 20 hub genes. **(C)** Chord plot depicting the relationships between the genes and GO biological process (BP) terms. **(D)** Chord plot depicting the relationships between the genes and GO cellular component (CC) terms. **(E)** Chord plot depicting the relationships between the genes and GO molecular function (MF) terms. **(F)** Chord plots depicting the functions of the genes in KEGG pathways.

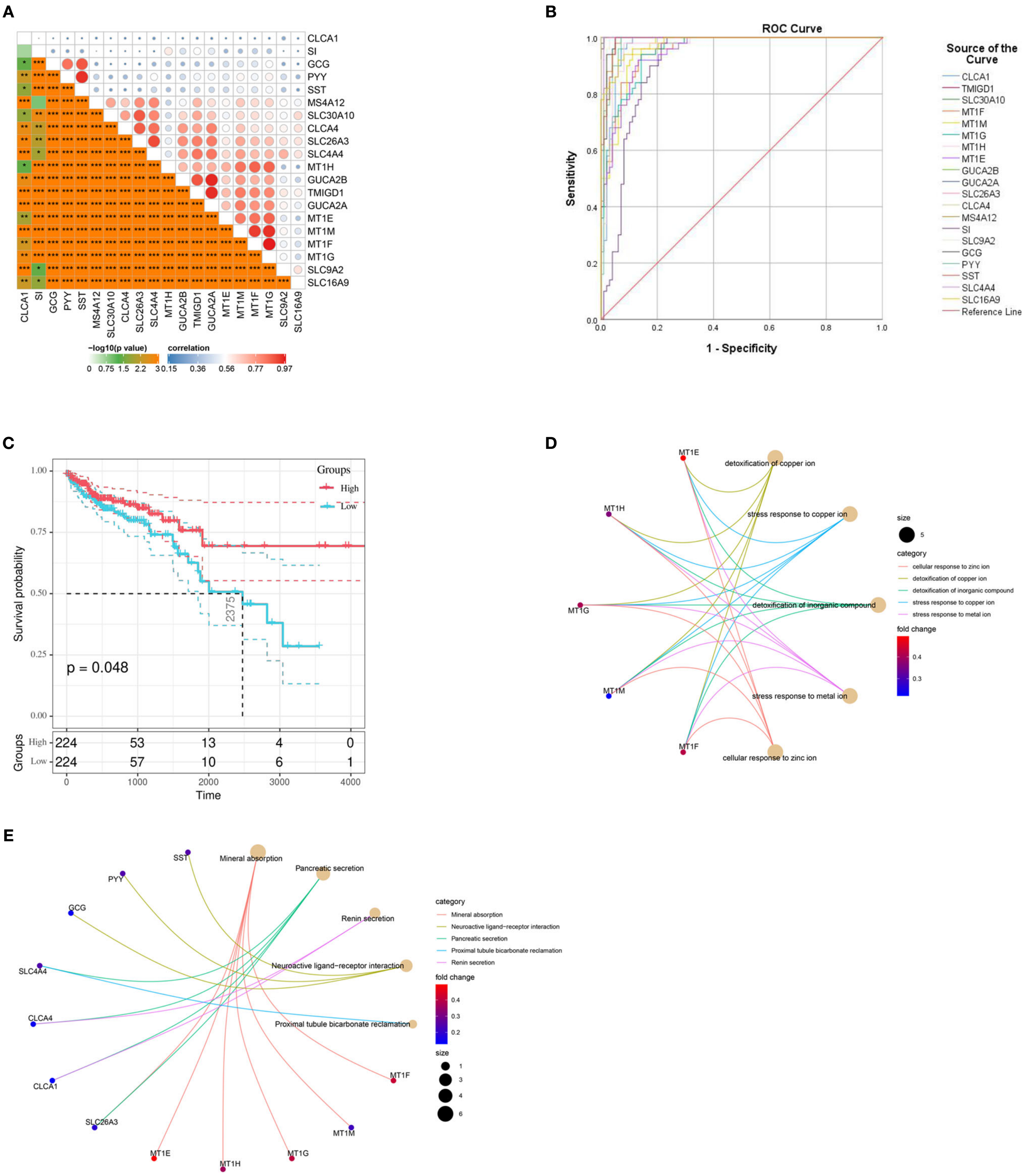


FIGURE 6 | Different values of hub genes and ROC curves of the diagnostic model. **(A)** Hub genes show strong associations with each other. **(B)** ROC curves for the hub genes. **(C)** K-M plot for CLCA1. **(D)** Circo plot depicting the relationships between the hub genes and gene ontology (GO) terms. **(E)** Circo plots depicting the functions of the hub genes in Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. K-M, Kaplan-Meier.

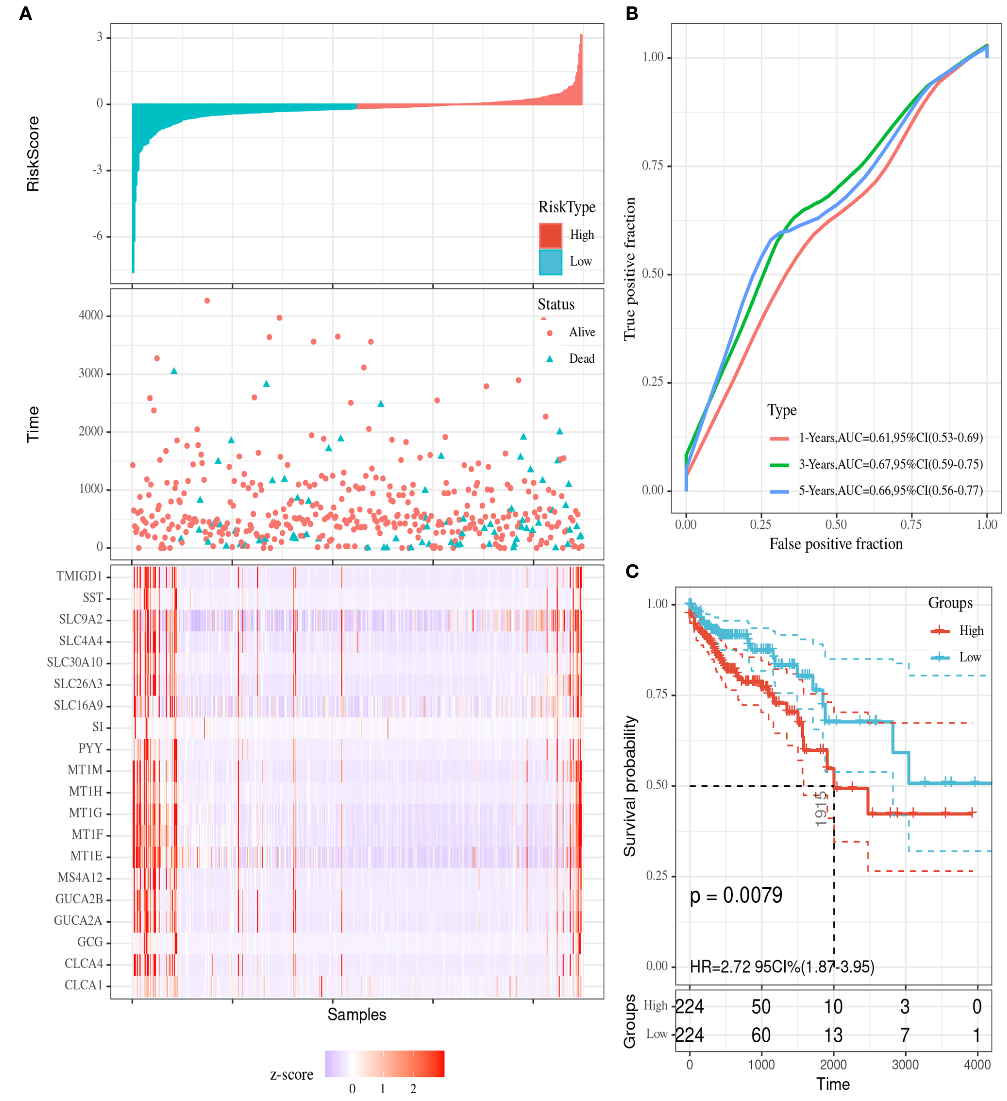


FIGURE 7 | Visualization of the prognostic model. **(A)** The risk score distribution in colorectal cancer (CRC) patients. **(B)** ROC curves for the 1-year, 3-year and 5-year survival rates of CRC patients. **(C)** K-M OS curves for the low-risk and high-risk groups.

We utilized randomForest to validate the prognostic model. The training group contained 50 died and 263 living patients, and the validation group contained 23 died and 112 living patients. From the confusion matrix, we obtained the following values: accuracy = 79.3%%, error rate = 20.7%, sensitivity = 85%, and precision = 91.1%.

Assessment of the Clinical Significance of the Hub Genes

Among the 20 hub genes, CLCA1 was associated with survival time. We explored correlations between gene expression levels (**Figures 8A, B**) and clinical features (**Figures 8C–F**). CLCA1 was downregulated during CRC, and no differences were identified in its expression across different stages and TNM grades. The persistently downregulated expression of CLCA1 underscores its diagnostic effectiveness.

GSEA for Hub Genes

We performed GSEA to investigate the potential functions of CLCA1 in CRC in the TCGA-COAD and TCGA-READ datasets (**Figure 8G**). The top 5 upregulated pathways in which CLCA1 was enriched included "ascorbate and aldarate metabolism", "butanoate metabolism", "fatty acid metabolism", "starch and sucrose metabolism", and "valine, leucine, and isoleucine degradation".

Relationship of Hub Genes with Immune Infiltration of CLCA1

We utilized TISIDB to explore the relationship between hub gene expression levels and lymphocyte levels in colon and rectal cancer. CLCA1 exhibited no relationship or only a weak relationship with immune infiltration (**Figure 9**). CLCA1

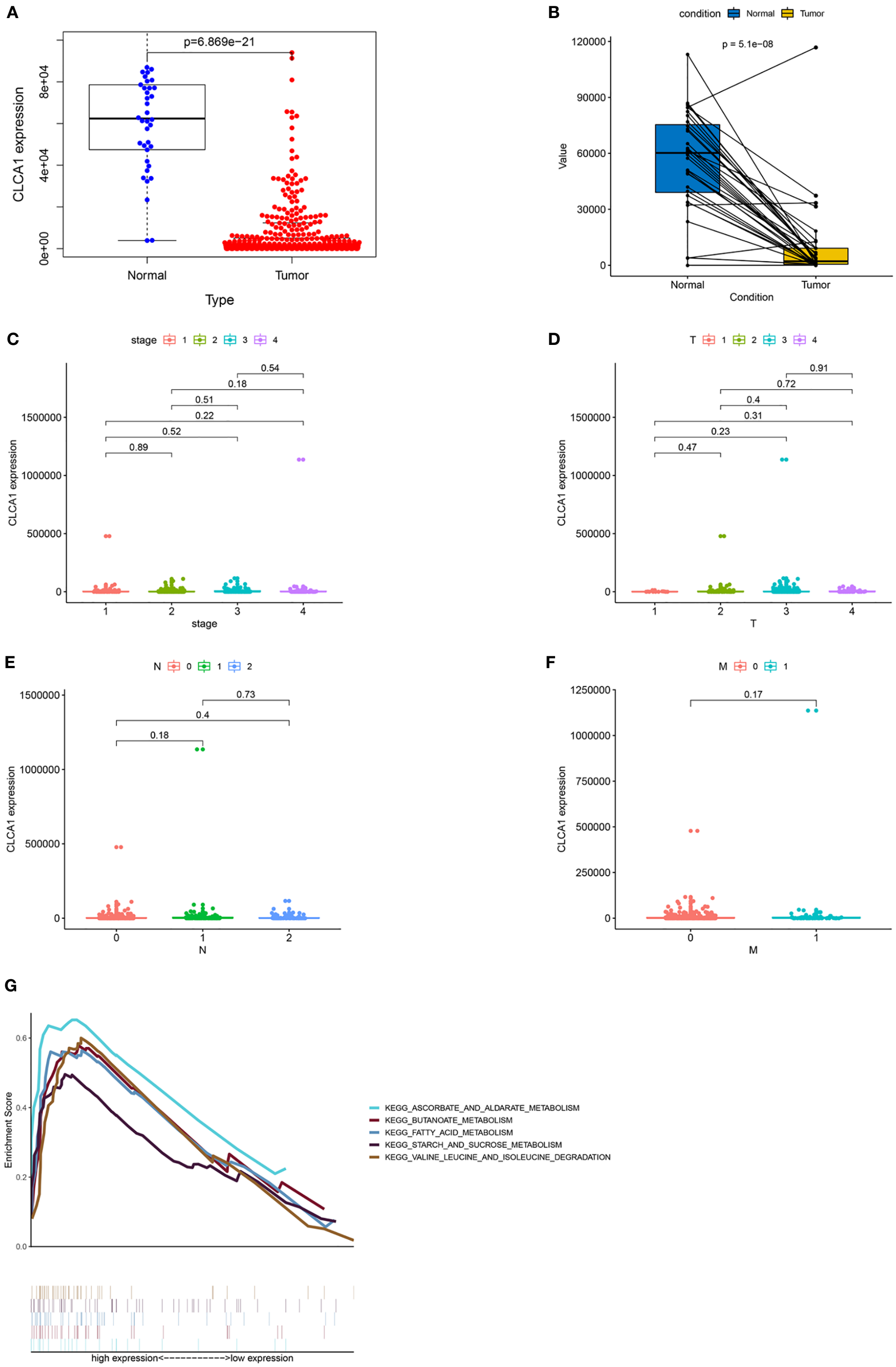


FIGURE 8 | Visualization of correlations between CLCA1 expression levels and clinical features. GSEA of CLCA1. **(A)** Differences in CLCA1 expression between control tissues and CRC tissues. **(B)** Differential expression of CLCA1 between tumor tissues and adjacent tissues. **(C)** Differences in CLCA1 expression between different clinical stages. **(D)** Differences in CLCA1 expression between different T stages. **(E)** Differences in CLCA1 expression between different N stages. **(F)** Differences in CLCA1 expression between different M stages. T, tumor; N, regional lymph node; M, metastasis. **(G)** GSEA for CLCA1. GSEA, Gene set enrichment analysis.

expression was closely correlated with Th17 (rho=0.379, P=3.71e-19) levels in colon cancer and with Act-B (rho=0.419, P=2.46e-08), ImmB (rho=0.365, P=1.5e-06), neutrophil (rho=0.414, P=3.68e-08), and Th17 (rho=0.517, P<2.2e-16) levels in rectal cancer.

Validation of Protein Expression and Prognostic Value of CLCA1

We utilized GEO datasets to valdiate the CLCA1,P value was >0.05 (**Supplementary Figure 2**). We also utilized The Human Protein Atlas and Kaplan-Meier Plotter database to validate the

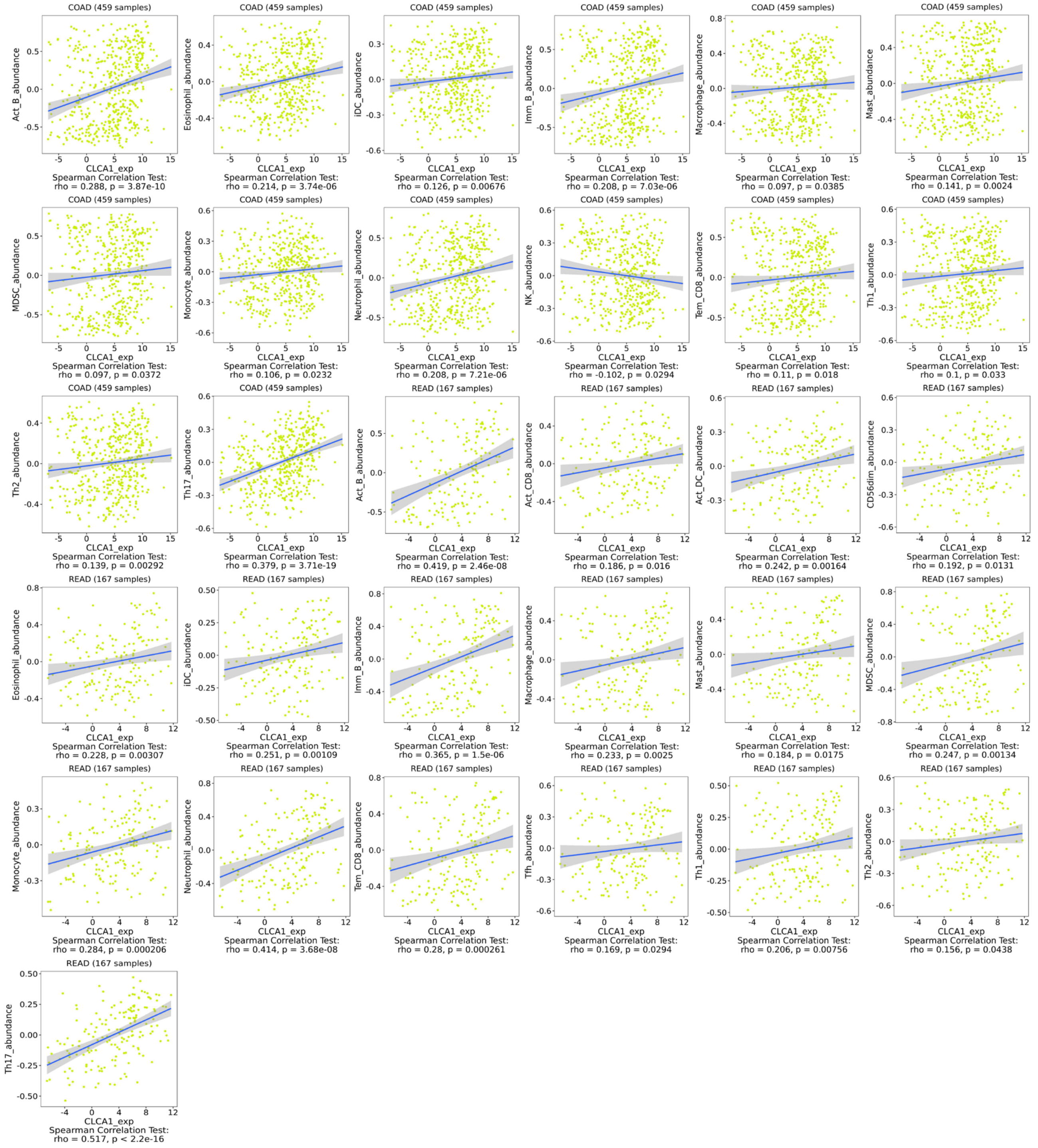


FIGURE 9 | Relationship with immune infiltration. CLCA1 in colon cancer and rectal cancer.

protein expression and prognostic value of CLCA1 (**Figure 10**). The protein expression in normal colon tissue was significantly higher than that in colon cancer tissue, and the same was true in the rectum. Both of the two databases showed the good prognostic value for CRC, the P value of The human protein atlas is <0.001 and KM plotter database is 0.0064.

DISCUSSION

CRC is the most frequently diagnosed gastrointestinal cancer (30), and the current colonoscopic diagnosis of CRC has limitations (31); therefore, identifying a significant biomarker for CRC is necessary. Gene microarrays were utilized to discover novel biomarkers or therapeutic targets for CRC (32). To our knowledge, our work is the first to combine RRA analysis of GEO datasets with WGCNA of TCGA datasets to explore the significant genes associated with CRC. Some diseases of the intestinal tract, such as intestinal polyps (33) and inflammatory bowel disease (34), can have symptoms similar to those of CRC and can also develop into cancer. To explore potential DEGs between tumor tissue and noncancerous tissue, we compared normal tissue, normal matched tissue and paratumor tissue with CRC tissue as control tissue. We integrated 10 datasets from GEO, TCGA-COAD and TCGA-READ and identified robust DEGs, such as SST (35), SLC26A3 (36), and SLC4A4 (37), which have been reported to be diagnostic biomarkers or therapeutic targets for CRC.

We used GO and KEGG enrichment analyses to explore the functions of the DEGs identified by overlapping the DEGs in the 3 datasets. GO analysis indicated that negative regulation of growth, bicarbonate transport, and transporter activity (38–40) were closely related to the development and growth of cancer; some KEGG pathways, such as nitrogen metabolism and retinol metabolism, were also linked to the pathogenesis of CRC. Nitrogen is an essential biomolecule in humans and regulates

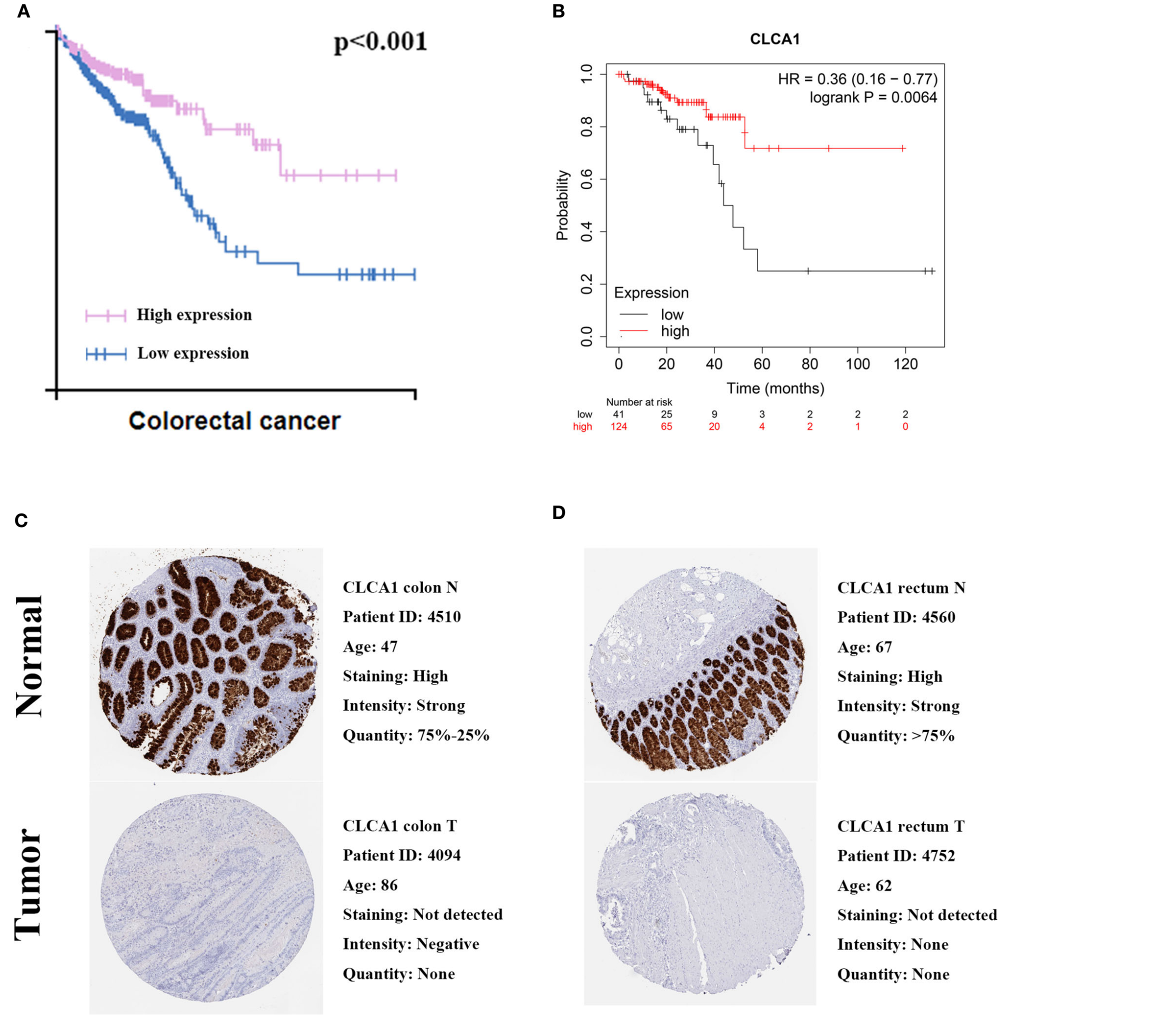


FIGURE 10 | Validation prognostic value. **(A)** Prognostic value in The Human Protein Atlas. **(B)** Prognostic value in Kaplan-Meier Plotter. **(C)** Immunohistochemical in The Human Protein Atlas of colon. **(D)** Immunohistochemical in The Human Protein Atlas of rectum.

cellular metabolism, and retinol is a form of vitamin A closely related to immune functions (41, 42). Based on the results of the GO and KEGG enrichment analyses, the DEGs were closely associated with CRC occurrence and development.

Cytohubba can extract key sub-networks, and MCC is a newer algorithm of cytohubba. To identify the key genes among 129 DEGs, we utilized MCC to determine the top 20 hub genes (CLCA1, TMIGD1, SLC30A10, MT1F, MT1M, MT1G, MT1H, MT1E, GUCA2B, GUCA2A, SLC26A3, CLCA4, MS4A12, SI, SLC9A2, GCG, PYY, SST, SLC4A4, and SLC16A9). To explore the potential functions of the hub genes, we utilized R packages, ROC curves, K-M analysis, and GO and KEGG analyses. According to the results, all the hub genes were closely related to each other and had high diagnostic value, but only CLCA1 was associated with survival time. In addition, the hub genes were closely associated with the development of CRC.

To determine the hub genes significantly associated with overall survival, we utilized Cox proportional hazards regression analysis to develop a prognostic model. We explored each gene's characteristics. Used ROC curve and random forest analysis to verify the model. The AUC values were high for 1-year, 3-year, and 5-year survival, all of which demonstrated the intermediate value of the prognostic model. Then, we calculated the risk score of each patient and divided the patients into a high-risk group and a low-risk group. K-M risk survival analysis showed that the model can predict survival time. Then we utilized the random forest method to validate the prognostic model for CRC, which showed high prognostic value for CRC. The yielded the following values: accuracy = 79.3%, sensitivity = 85% which reflecting good prognositc value for CRC. Among the hub genes, only CLCA1 was associated with a good prognosis in CRC but their dignostic value is very high. In accordance with the expected results, the expression of CLCA1 protein was down-regulated in colorectal cancer tissues. To demonstrate the prognostic value of CLCA1, we conducted the external validation in three online databases and the results of The Human Protein Atlas and KMplotter showed that CLCA1 has a high prognostic value. The results were inconsistent with GEO database, and the unsatisfactory results of validation of GEO database may be related to the possible influencing factors such as sample size, experimental environment and methods.

Although CLCA1 has a high prognostic value for CRC, the mechanism of its influence is unclear. To further explore its characteristics, we analyzed differences in CLCA1 expression levels between tumor and normal tissues, across clinical stages, and across TNM stages. There was a large difference between tumor and normal tissues, but no significant differences were found across the different stages of CRC. This pattern indicates that CLCA1 levels decrease starting from the initial development of CRC and have diagnostic value at every stage of CRC. The characteristic expression of CLCA1 may provide a new perspective for exploring CRC at the gene level and serve as a useful diagnostic biomarker for CRC.

To explore the mechanisms of the hub genes in CRC, we utilized TISIDB and the R package "estimate" to assess immune infiltration and GSEA data of biological functions for CLCA1. TISIDB and the estimated score analysis indicated that CLCA1 had a weak relationship with lymphocyte expression and was expressed mainly in CRC cells. GSEA indicated that CLCA1 was enriched in "ascorbate and aldarate metabolism", "butanoate metabolism", "fatty acid metabolism", "starch and sucrose metabolism", and "valine, leucine, and isoleucine degradation", suggesting that CLCA1 can influence CRC development and progression through different metabolic pathways. This result provides new insight into the mechanism and pathology of CRC.

In summary, we determined that CLCA1 could be used as a prognostic marker for CRC and correlated with immune infiltration. It may be a potential therapeutic target for CRC to improve the prognosis of patients. However, our work has some limitations. First, more work needs to be done on the pathogenic immune responses and gene expression in CRC cells to identify the mechanism linking the immune response with the development of CRC. Second, validation in GEO datasets is not ideal and pure bioinformatics analysis cannot well prove the prognostic significance of CLCA1 in colorectal cancer, in future research we will focus on large-scale population for further investigation. Furthermore, basic research needs to be done to verify our model and the regulatory mechanism *in vitro* and *in vivo*.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

F-ZW designed the research. Z-JW, S-WM, JL, and H-YS organized the data. J-NC, F-QZ, and ZL analyzed and visualized the data. F-ZW drafted the article. QL revised the paper. All authors contributed to the article and approved the submitted version.

FUNDING

Key Project of National Key R & D Plan "Research on Prevention and Control of Major Chronic Non-Communicable Diseases" (No. 2019YFC1315705), China Cancer Foundation Beijing Hope Marathon Special Fund (No. LC2017L07), Medical and Health Science and Technology Innovation Project of the Chinese Academy of Medical Sciences (No. 2017-12M-1-006).

ACKNOWLEDGMENTS

The study was supported and funded by the Key Project of National Key R & D Plan "Research on Prevention and Control of Major Chronic Non-Communicable Diseases" (No. 2019YFC1315705), the China Cancer Foundation Beijing Hope

Marathon Special Fund (No. LC2017L07), and the Medical and Health Science and Technology Innovation Project of the Chinese Academy of Medical Sciences (No. 2017-12M-1-006) of China. In addition, the authors wish to thank all teachers in the Department of Colorectal Surgery, National Cancer Center/Cancer Hospital.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2020.573295/full#supplementary-material

REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* (2018) 68:394–424. doi: 10.3322/caac.21492
2. Arnold M, Sierra MS, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global patterns and trends in colorectal cancer incidence and mortality. *Gut* (2017) 66:683–91. doi: 10.1136/gutjnl-2015-310912
3. Picardo F, Romanelli A, Muinelo-Romay L, Mazza T, Fusilli C, Parrella P, et al. Diagnostic and prognostic value of B4GALT1 hypermethylation and its clinical significance as a novel circulating cell-free DNA biomarker in colorectal cancer. *Cancers* (2019) 11:1598. doi: 10.3390/cancers11101598
4. Edoo MIA, Chutturghoon VK, Wusu-Ansah GK, Zhu H, Zhen TY, Xie HY, et al. Serum Biomarkers AFP, CEA and CA19-9 Combined Detection for Early Diagnosis of Hepatocellular Carcinoma. *Iran J Public Health* (2019) 48:314–22.
5. Fang T, Wang Y, Yin X, Zhai Z, Zhang Y, Yang Y, et al. Diagnostic sensitivity of NLR and PLR in early diagnosis of gastric cancer. *J Immunol Res* (2020) 2020:9146042. doi: 10.1155/2020/9146042
6. Adler A, Geiger S, Keil A, Bias H, Schatz P, deVos T, et al. Improving compliance to colorectal cancer screening using blood and stool based tests in patients refusing screening colonoscopy in Germany. *BMC Gastroenterol* (2014) 14:183. doi: 10.1186/1471-230x-14-183
7. Liu Q, Deng J, Wei X, Yuan W, Ma J. Integrated analysis of competing endogenous RNA networks revealing five prognostic biomarkers associated with colorectal cancer. *J Cell Biochem* (2019) 120:11256–64. doi: 10.1002/jcb.28403
8. Song ZY, Chao F, Zhuo Z, Ma Z, Li W, Chen G. Identification of hub genes in prostate cancer using robust rank aggregation and weighted gene co-expression network analysis. *Aging (Albany NY)* (2019) 11:4736–56. doi: 10.18632/aging.102087
9. Kolde R, Laur S, Adler P, Vilo J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinf (Oxford England)* (2012) 28:573–80. doi: 10.1093/bioinformatics/btr709
10. Hong Y, Downey T, Eu KW, Koh PK, Cheah PY. A 'metastasis-prone' signature for early-stage mismatch-repair proficient sporadic colorectal cancer patients and its implications for possible therapeutics. *Clin Exp Metastasis* (2010) 27:83–90. doi: 10.1007/s10585-010-9305-4
11. Moreno V, Alonso MH, Closa A, Valles X, Diez-Villanueva A, Valle L, et al. Colon-specific eQTL analysis to inform on functional SNPs. *Br J Cancer* (2018) 119:971–7. doi: 10.1038/s41416-018-0018-9
12. Galamb O, Wichmann B, Sipos F, Spisak S, Krenacs T, Toth K, et al. Dysplasia-carcinoma transition specific transcripts in colonic biopsy samples. *PLoS One* (2012) 7:e48547. doi: 10.1371/journal.pone.0048547
13. Skrzypczak M, Goryca K, Rubel T, Paziewska A, Mikula M, Jarosz D, et al. Modeling oncogenic signaling in colon tumors by multidirectional analyses of microarray data directed for maximization of analytical reliability. *PLoS One* (2010) 5:e13091. doi: 10.1371/journal.pone.0013091
14. Molnar B, Galamb O, Peterfia B, Wichmann B, Csabai I, Bodor A, et al. Gene promoter and exon DNA methylation changes in colon cancer development - mRNA expression and tumor mutation alterations. *BMC Cancer* (2018) 18:695. doi: 10.1186/s12885-018-4609-x
15. Ryan BM, Zanetti KA, Robles AI, Schetter AJ, Goodman J, Hayes RB, et al. Germline variation in NCF4, an innate immunity gene, is associated with an increased risk of colorectal cancer. *Int J Cancer* (2014) 134:1399–407. doi: 10.1002/ijc.28457
16. Sayagues JM, Corchete LA, Gutierrez ML, Sarasquete ME, Del Mar Abad M, Bengoechea O, et al. Genomic characterization of liver metastases from colorectal cancer patients. *Oncotarget* (2016) 7:72908–22. doi: 10.18632/oncotarget.12140
17. Okazaki S, Ishikawa T, Iida S, Ishiguro M, Kobayashi H, Higuchi T, et al. Clinical significance of UNC5B expression in colorectal cancer. *Int J Oncol* (2012) 40:209–16. doi: 10.3892/ijo.2011.1201
18. Vlachavas EI, Pilalis E, Papadodima O, Koczan D, Willis S, Klippel S, et al. Radiogenomic analysis of F-18-fluorodeoxyglucose positron emission tomography and gene expression data elucidates the epidemiological complexity of colorectal cancer landscape. *Comput Struct Biotechnol J* (2019) 17:177–85. doi: 10.1016/j.csbj.2019.01.007
19. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. *Nucleic Acids Res* (2012) 40:4288–97. doi: 10.1093/nar/gks042
20. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinf* (2008) 9:559. doi: 10.1186/1471-2105-9-559
21. Li L, Lei Q, Zhang S, Kong L, Qin B. Screening and identification of key biomarkers in hepatocellular carcinoma: evidence from bioinformatic analysis. *Oncol Rep* (2017) 38:2607–18. doi: 10.3892/or.2017.5946
22. Li J, Wang Y, Wang X, Yang Q. CDK1 and CDC20 overexpression in patients with colorectal cancer are associated with poor prognosis: evidence from integrated bioinformatics analysis. *World J Surg Oncol* (2020) 18:11–50. doi: 10.1186/s12957-020-01817-8
23. Song W, Fu T. Circular RNA-associated competing endogenous RNA network and prognostic nomogram for patients with colorectal cancer. *Front Oncol* (2019) 9:1181. doi: 10.3389/fonc.2019.01181
24. Yao X, Lan Z, Lai Q, Li A, Liu S, Wang X. LncRNA SNHG6 plays an oncogenic role in colorectal cancer and can be used as a prognostic biomarker for solid tumors. *J Cell Physiol* (2020) 235(10):7620–34. doi: 10.1002/jcp.29672
25. Zhang Z, Wang Z, Huang Y. Identification of potential prognostic long non-coding RNA for predicting survival in intrahepatic cholangiocarcinoma. *Med (Baltimore)* (2020) 99:e19606. doi: 10.1097/md.0000000000019606
26. Sperr WR, Kundi M, Alvarez-Twose I, Van Anrooij B, Oude Elberink JNG, Gorska A, et al. International prognostic scoring system for mastocytosis (IPSM): a retrospective cohort study. *Lancet Haematol* (2019) 6:e638–49. doi: 10.1016/s2352-3026(19)30166-8
27. Ru B, Wong CN, Tong Y, Zhong JY, Zhong SSW, Wu WC, et al. TISIDB: an integrated repository portal for tumor-immune system interactions. *Bioinformatics* (2019) 35:4200–2. doi: 10.1093/bioinformatics/btz210
28. Gou R, Zhu L, Zheng M, Guo Q, Hu Y, Li X, et al. Annexin A8 can serve as potential prognostic biomarker and therapeutic target for ovarian cancer: based on the comprehensive analysis of Annexins. *J Transl Med* (2019) 17:222–75. doi: 10.1186/s12967-019-2023-z
29. Liu X, Bing Z, Wu J, Zhang J, Zhou W, Ni M, et al. Integrative gene expression profiling analysis to investigate potential prognostic biomarkers for colorectal cancer. *Med Sci Monit* (2020) 26:e918906. doi: 10.12659/msm.918906
30. Arnold M, Abnet CC, Neale RE, Vignat J, Giovannucci EL, McGlynn KA, et al. Global burden of 5 major types of gastrointestinal cancer. *Gastroenterology* (2020) 159(1):335–49.e15. doi: 10.1053/j.gastro.2020.02.068
31. Lech Pedersen N, Mertz Petersen M, Ladd JJ, Lampe PD, Bresalier RS, Davis GJ, et al. Development of blood-based biomarker tests for early detection of colorectal neoplasia: influence of blood collection timing and handling procedures. *Clin Chim Acta* (2020) 507:39–53. doi: 10.1016/j.cca.2020.03.035
32. Barbieri CE, Chinnaiyan AM, Lerner SP, Swanton C, Rubin MA. The emergence of precision urologic oncology: a collaborative review on biomarker-driven therapeutics. *Eur Urol* (2017) 71:237–46. doi: 10.1016/j.eururo.2016.08.024
33. Kottorou AE, Dimitrakopoulos FD, Antonacopoulou AG, Diamantopoulou G, Tsoumas D, Koutras A, et al. Differentially methylated ultra-conserved regions Uc160 and Uc283 in adenomas and adenocarcinomas are associated with overall survival of colorectal cancer patients. *Cancers (Basel)* (2020) 12: E895. doi: 10.3390/cancers12040895

34. Muller M, Hansmannel F, Arnone D, Choukour M, Ndiaye NC, Kokten T, et al. Genomic and molecular alterations in human inflammatory bowel disease-associated colorectal cancer. *United Eur Gastroenterol J* (2020) 8 (6):675–84. doi: 10.1177/2050640620919254

35. Hohla F, Buchholz S, Schally AV, Krishan A, Rick FG, Szalontay L, et al. Targeted cytotoxic somatostatin analog AN-162 inhibits growth of human colon carcinomas and increases sensitivity of doxorubicin resistant murine leukemia cells. *Cancer Lett* (2010) 294:35–42. doi: 10.1016/j.canlet.2010.01.018

36. Ostasiewicz B, Ostasiewicz P, Duś-Szachniewicz K, Ostasiewicz K, Ziółkowski P. Quantitative analysis of gene expression in fixed colorectal carcinoma samples as a method for biomarker validation. *Mol Med Rep* (2016) 13:5084–92. doi: 10.3892/mmr.2016.5200

37. Mencia N, Selga E, Noe V, Ciudad CJ. Underexpression of miR-224 in methotrexate resistant human colon cancer cells. *Biochem Pharmacol* (2011) 82:1572–82. doi: 10.1016/j.bcp.2011.08.009

38. Cheng C, Huang Z, Zhou R, An H, Cao G, Ye J, et al. Numb negatively regulates the epithelial-to-mesenchymal transition in colorectal cancer through the Wnt signaling pathway. *Am J Physiol Gastrointest Liver Physiol* (2020) 318:G841–53. doi: 10.1152/ajpgi.00178.2019

39. Wang YR, Meng LB, Su F, Qiu Y, Shi JH, Xu X, et al. Insights regarding novel biomarkers and the pathogenesis of primary colorectal carcinoma based on bioinformatic analysis. *Comput Biol Chem* (2020) 85:107229. doi: 10.1016/j.compbiolchem.2020.107229

40. Cheng CY, Zhou Z, Stone M, Lu B, Flesken-Nikitin A, Nanus DM, et al. Membrane metalloendopeptidase suppresses prostate carcinogenesis by attenuating effects of gastrin-releasing peptide on stem/progenitor cells. *Oncogenesis* (2020) 9:38. doi: 10.1038/s41389-020-0222-3

41. Hu Y, Wang L, Li Z, Wan Z, Shao M, Wu S, et al. Potential prognostic and diagnostic values of CDC6, CDC45, ORC6 and SNHG7 in colorectal cancer. *Onco Targets Ther* (2019) 12:11609–21. doi: 10.2147/OTT.S231941

42. Jing C, Wang T, Ma R, Cao H, Wang Z, Liu S, et al. New genetic variations discovered in KRAS wild-type cetuximab resistant chinese colorectal cancer patients. *Mol Carcinogen* (2020) 59:478–91. doi: 10.1002/mc.23172

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.