Prognostic Risk Model of Colorectal Cancer Constructed by Bioinformatics Method and Functional Study of the Gene CLCA1 & SPINK4

Shiying Li, Haowei Liu, Yongming Zhao, Yingxing Zeng, Jun Xu*

Author affiliations
College of Pharmacy, Jinan University, Guangzhou, China

Author contribution:
Shiying Li, article writing; Haowei Liu, contributed towards data collection and analysis; Yongming Zhao, investigation; Shiying Li, reference search; Yingxing Zeng, formal analysis; Jun Xu Supervision and editing proof reading and project administration.

Address reprint requests to
Jun Xu
College of Pharmacy, Jinan University, Guangzhou 510632, PR China
Email: goldstar_8209@163.com

Abstract:
Background Colorectal cancer (CRC) is the most common malignant tumor of digestive system. The metastases is the main cause of mortality in CRC patients, of whom the initial diagnosis is about 25%. In our study, we aimed to identify potential gene biomarkers based on RNA sequencing data to predict and improve CRC patient survival.
Method In this study, by screening differentially expressed genes of colon cancer related to liver metastasis, a survival prognostic risk model was constructed by bioinformatics analysis. Here, we conducted our data mining analysis for CRC by integrating the differentially expressed genes acquired from Gene Expression Omnibus (GEO) database by primary tumor versus liver metastasis (GSE81582, GSE41258, GSE49355, GSE68468) into The Cancer Genome Atlas (TCGA) database which includes 415 primary tumor and 132 liver metastasis tissue. At the same time, we used transwell, RT-PCR and western to examine the effects of CLCA1 and SPINK4 on the migration of colorectal cancer cells at the cell level.
Results We identified intersections of 197 genes (117 up-regulated and 80 down-regulated) between GEO data and TCGA data. Differentially expressed genes in TCGA-COAD by single factor cox analysis, lasso cycle training and multifactor cox analysis composed a survival prognosis prediction model consisted of 7 genes ORM1, CLCA1, C8B, SPINK4, ALDOB, GAMT, C8G. And results of transwell experiments showed that high expression of CLCA1 and SPINK4 can inhibit the migration ability of colon cancer cells LOVO and SW620, meanwhile western blotting showed that the high expression of both genes can upregulate the expression of epithelial phenotypic marker E-cadherin, and Vimentin expression is down-regulated.
Conclusion In this study, 197 differentially expressed genes were selected and a relatively robust survival prognosis prediction model was constructed. The model consisted of seven genes: GAMT, C8G, ORM1, CLCA1, C8B, SPINK4, and ALDOB. At the same time, we found that CLCA1 and SPINK4 are closely related to survival prognosis. The predictive model nomogram will enable patients with CRC to be more accurately managed in trials testing new drugs and in clinical practice.

Key words Colorectal cancer, prognostic risk model, bioinformatics.
INTRODUCTION

Colorectal cancer accounts for approximately 10% of all annually diagnosed cancers and cancer-related deaths worldwide. It is the second most common cancer diagnosed in women and third most in men. In women, incidence and mortality are approximately 25% lower than that in men. Nearly 25% of patients diagnosed with metastatic disease. The metastases are the main cause of mortality in CRC patients, of whom the initial diagnosis is about 25%. Liver and peritoneum metastases are the most common sites of CRC metastases compared with brain and bone metastases. In the beginning of metastases, cancer cells undergo epithelial-to-mesenchymal transition to attain the mesenchymal phenotype and the ability to migrate and infiltrate in the surrounding mesenchymal tissues. EMT (Epithelial-mesenchymal transition) is involved in malignant tumor progression and metastasis. EMT triggers detachment of cancer cells from the primary cancer organ and triggers invasion into lymphatic or blood vessels through the loss of intercellular junctions.

In this genomic era, a large number of genome-sequencing technologies and data have emerged. These tools have made great contributions to tumor diagnosis and prognosis prediction. GEO is a public gene expression repository that contains more than 94,000 datasets and over 2 million samples. GEO database provides an invaluable resource of publicly available gene expression data that can be integrated and analyzed to derive new hypothesis and knowledge. TCGA project was a multi-center institutional effort, supported by the National Institute of Health, with the aim of providing a comprehensive genetic analysis of different cancers and establishing correlations with clinical outcomes.

In this study, we aimed to explore the difference in the mRNA expression profiles of colorectal cancer in situ and metastatic colorectal cancer to identify potential gene biomarkers using GEO and TCGA data. We established a seven-gene prognostic model that included GAMT, C8G, ORM1, CLCA1, C8B, SPINK4, and ALDOB based on our RNA sequencing survival analysis. Functional enrichment analysis of the predictive genes was performed by STRING, KEGG. As a whole, this prognostic model and nomogram might be helpful in guiding the prognostic status of patients with CRC.

MATERIAL AND METHODS

DATA SOURCE

GEO, NCBI’s publicly available genomics database, which collects submitted high throughput gene expression data, was thoroughly queried for all datasets involving studies of CRC. Studies were considered eligible for our following analysis according to the following criteria: (1) A comparative study of liver metastasis CRC and original CRC. (2) mRNA expression profile measurement data. (3) Human-based CRC research. (4) Research data is available and undergoes a certain data quality control process. Based on these criteria, four datasets for CRC were downloaded from the repository. Principal component analysis (PCA) was done for the datasets for dimensionality reduction and quality control. If the quality of a particular sample is not good enough, it would be excluded for subsequent analysis.

The mRNA expression profiles and the corresponding clinical information from the patients with CRC were obtained from the TCGA, which was calculated on an Illumina HiSeq RNA-seq platform, containing 413 CRC tissues as of January 1, 2020. The data from the TCGA are publicly available and open-access; therefore, the local ethics committees did not need to approve the study because the current research follows the TCGA data access policies and publication guidelines.

<table>
<thead>
<tr>
<th>GSE number</th>
<th>GPL number</th>
<th>Liver metastasis vs Primary tumor</th>
<th>Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>81582</td>
<td>15207</td>
<td>19:23</td>
<td>Spain</td>
</tr>
<tr>
<td>41258</td>
<td>96</td>
<td>47:186</td>
<td>Israel</td>
</tr>
<tr>
<td>49355</td>
<td>96</td>
<td>19:20</td>
<td>France</td>
</tr>
<tr>
<td>68468</td>
<td>96</td>
<td>47:186</td>
<td>USA</td>
</tr>
</tbody>
</table>
DIFFERENTIAL EXPRESSION ANALYSIS
Heterogeneity and potential variables are commonly recognized as major sources of bias and variability in high-throughput experiments. Since the datasets we recruited for our multi-datasets analysis were based on different platforms and samples were handled on different days, in different groups or by different people. Therefore, we first integrated all samples of four datasets to significantly improve the number of samples (415 primary tumor vs. 132 Liver metastasis tumor samples) so as to avoid generating less reliable results followed by batch normalization in the R computing environment using sva package 9. Next, we performed the differential analysis (|log FC| > 1, adjusted p-value < 0.05) by comparing tumor tissues to normal tissues in the R computing environment using limma package10.

INTEGRATION OF THE DIFFERENTIALLY EXPRESSED GENES IN TCGA DATABASE
The Cancer Genome Atlas (TCGA), a project supported by the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI), has generated comprehensive, multi-dimensional maps of the key genomic changes in various types of cancers. In order to obtain a consensus of differentially expressed genes, gene expression quantification data and clinical information of CRC patients in TCGA database were downloaded using TCGAAbiowlinks11.

MIGRATION ASSAY
SW620 or LOVO (5x10^4) cells were over-expressed with SPINK4, CLCA1 and SPINK4/CLCA1 before seeded them onto 8μm pore transwell chamber (Corning). Cell migration was examined after cultured in 5% CO2, 24 h and stained with crystal violet (Beyotime).

WESTERN BLOT ANALYSIS
Cells were lysed with cell lysis buffer for western and IP(Beyotime), supplemented with protease inhibitors PMSF(Beyotime). Vertical gel electrophoresis ran for 2h and then electricity runs for 2h (Bio-Rad). E-cadherin mouse monoclonal antibodies (mAb) (Cell signaling), Vimentin rabbit mAb (Cell signaling), GAPDH rabbit mAb (Cell signaling), CLCA1 rabbit mAb (Abcom), SPINK4 rabbit mAb (Abcom). The membrane was washed and incubated with appropriate secondary antibodies (Abcom) for 1h. Immunoreactivities were visualized using the enhanced chemiluminescence system (Tanon). GAPDH was probed to ensure equal protein loading.

REAL-TIME REVERSE TRANSCRIPTION–PCR ANALYSIS
Total RNA was isolated from cells using the Trizol (Invitrogen) and concentrations were determined. The reverse transcription was performed with the PrimeScript RT reagent kit (TaKaRa, Shiga, Japan). After the resulting complementary DNA template was mixed with primers and TaKaRa SYBR Premix Ex Taq, quantitative real-time PCR was performed on an iQ5 real-time PCR detection system (Bio-Rad). Crossing threshold values for individual genes were normalized to GAPDH. Changes in mRNA expression were expressed as fold change relative to control. Gene-specific primer pairs used in this study was presented in table 1 as supplementary data.

Table 2 Primer sequence list
<table>
<thead>
<tr>
<th>Gene</th>
<th>Forward primer 5′-3′</th>
<th>Reverse primer 5′-3′</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLCA1</td>
<td>CACCGTGGAAGGACACCTT</td>
<td>CCTGGGATTGGAGGTAGGC</td>
</tr>
<tr>
<td>SPINK4</td>
<td>GAACACATGGTAGAGTCTCCAA</td>
<td>AGAGCTGCGATTCTCGTATA</td>
</tr>
</tbody>
</table>

TRANSFECTION WITH OVER-EXPRESSION PLASMIDS
CLCA1 and SPINK4 over-expression plasmids and over-expression plasmids control were purchased from Cell Signaling Technology. SW620 or LOVO cells were transfected with 100 nM over-expression plasmids at 60% confluence in RPMI 1640 complete medium. Lipofectamine 3000 (Invitrogen) was used for transfection following the manufacturer’s protocols. After transfected for 24 h, the cells were collected for other tests.
RESULTS
DIFFERENTIALLY EXPRESSED mRNAs BETWEEN PRIMARY TUMOR AND LIVER METASTASIS TISSUE
To describe our study more clearly, a flow chart of the analysis procedure was developed (Fig. 1 A). Firstly, four GEO datasets of GSE81582, GSE41258, GSE49355, and GSE68468 were screened by limma method, and 197 differentially expressed genes were obtained, of which 117 were up-regulated and 80 were down-regulated (Fig. 1 B). The differentially expressed gene set obtained single-factor cox analysis, lasso cycle training and multi-factor cox analysis in the TCGA-COAD data set, and obtained the survival prognosis prediction composed of the seven genes ORM1, CLCA1, C8B, SPINK4, ALDOB, GAMT, and C8G (Fig. 1 C).

Figure 1 (A) Overall workflow describing the process used to develop and validate the prognostic model to predict prognostic outcomes (B) Differential gene volcano (Red is up-regulated gene, blue is down-regulated gene), (C) Forest map with multi-factor cox analysis.

Visualize the survival prognosis prediction model with a Nomogram (Fig. 2A). The expression levels of 7 genes were used to estimate the 1-year, 3-year and 5-year survival rates of colon cancer patients. The calibration curves in Fig. 2 B show that the model of 3-year and 5-year forecast and actual consistency evaluation results show that the model has ideal robustness. It can be seen from Fig. 2 C that the optimal cut-off value of the risk of the training model in this study is 0.328. The risk value less than 0.328 is the low-risk group, and the risk value greater than 0.328 is the high-risk group. GAMT and C8G were up-regulated as the risk value increased, while ORM1, CLCA1, C8B, SPINK4, and ALDOB were the opposite. Fig. 2 D-E shows that the survival analysis results using TCGA data as the test set show that the Log-rank P value is less
than 0.0001, indicating that the high-risk group and the low-risk group have statistically significant differences in survival effects. The survival analysis result of GSE17536 as an external verification set is Log-rank $P=0.0076$. The survival analysis results of the test set and the verification set both indicate that the model has good reliability in predicting the survival of colon cancer.

Figure 2. (A) Nomogram of survival prediction model, (B) Model calibration curve, (C) Test set risk factor correlation diagram (A: distribution of risk values for CRC patients; B: distribution of survival status and survival time of CRC patients; C: heat map of 7 gene expression profiles), (D) Survival analysis results of internal training set, (E) Results of survival analysis of external validation set.

FUNCTIONAL ENRICHMENT ANALYSIS

In order to study the interaction and regulation mechanism of the differentially expressed genes screened, the online tool STRING was used to analyze protein protein interaction (PPI) network. As is showed in Fig. 3 A, 193 differentially expressed genes were imported, with a total of 1821 edges, 15 nodes were isolated nodes, PPI enrichment P value was less than 1.0e-16. The results show that many genes have a direct or indirect interaction relationship and it has a strong confidence. GO enrichment analysis of differentially expressed genes, as shown in Fig. 3B, the extracellular structure organization, acute inflammation response and enzyme
inhibitor activity are the main enrichment directions. The KEGG pathway analysis of the differential genes showed that the P Value values of 9 pathways were all less than 0.005 (Fig. 3C). GSVA was used to calculate the difference of KEGG pathway in each sample. The analysis results are shown in Fig. 3D complement coagulation cascade pathway, drug cytochrome P450 metabolism pathway, steroid hormone biosynthesis pathway, retinol metabolism pathway, etc. showed an upward trend in metastatic cancer samples and a downward trend in primary cancer samples. Protein export pathways, oxidative phosphorylation pathways, DNA replication, mismatch repair, etc. are showing a downward trend in metastatic cancer samples and an upward trend in primary cancer samples.

Figure 3 (A) diagram of PPI network by differential genes, (B) Enrichment analysis of differentially expressed genes GO (BP: biological process group; CC: cell localization group; MF: molecular function group), (C) Enrichment of KEGG pathway of differentially expressed genes, (D) GSVA analysis heat map (Sample grouping: blue is the metastatic cancer group, pink is the primary cancer group).

PROGNOSIS OF MODEL GENES
Through 197 differentially expressed genes, a relatively robust survival prognosis prediction model was selected and constructed. The model consisted of seven genes: GAMT, C8G, ORM1, CLCA1, C8B, SPINK4, and ALDOB. Seven genes were subjected to PCA (Principal Component Analysis) in three data sets of TCGA-COAD-Tumor, TCGA-COAD-Normal, GTEx-Colon-Sigmoid. As shown in Figure 4A, the data of colorectal cancer, normal colon tissue and normal sigmoid colon roughly show three clusters, showing that the 7 factor has a certain degree of distinction between cancer and adjacent tissues and normal tissue. The online tool GEPIA (Gene Expression Profiling Interactive Analysis) was used to analyze the survival of each gene. C8B cannot perform survival analysis due to sample size is insufficient. Survival analysis of six genes other than C8B revealed that the genes CLCA1 and SPINK4 showed close correlation with survival prognosis (Fig. 4B). The log-rank P of CLCA1 and SPINK4 were 0.012 and 0.00018, respectively, and both genes had high expression and low risk. These two genes may be protective factors.
In the correlation analysis, CLCA1 and SPINK4 also showed a positive co-expression relationship, with a Poisson coefficient of 0.622 (Fig. 5 A). At the same time, the co-expression relationship between CLCA1 and SPINK4 was confirmed in the linkedOmics database analysis (Fig. 5 B-C). In the STRING database analysis, it can be found that CLCA1 and SPINK4 have a medium-strength protein interaction relationship, and that they have interactions with SPINK4, MUCA5AC, IL13, SERPINB2, FKBP5, POSTN, SLC13A2, SI, CLCA4, CLCA2 and other proteins (Fig. 5 D).
SINGLE GENE GSEA ANALYSIS OF KEY GENES CLCA1 AND SPINK4

The limma algorithm is used to sort the gene set in descending order by log FC, and the GSEA analysis is performed on the c2.kegg gene set. The GSEA results are shown in the Fig. 6A, showing the first five enrichment channels of CLCA1 and SPINK4, respectively, and all of them are upwardly trending. According to the GSEA results ridge map (Fig. 6B), CLCA1 and SPINK4 are in the hematopoietic lineage pathway, cytokine receptor interaction pathway, P450 cytochrome drug metabolism pathway, cytochrome P450 exogenous metabolism pathway, vitamin A metabolism pathway, steroid hormone biosynthesis pathway. Single-gene GSEA results suggest that CLCA1 and SPINK4 may affect the prognosis and disease progression of colon cancer by regulating the above pathways.

Figure 6 (A) CLCA1, SPINK4 single gene GSEA analysis results (first 5 enrichment pathways), (B) CLCA1 SPINK4 single gene GSEA analysis ridge map.
EFFECTS OF CLCA1 AND SPINK4 ON TUMOR METASTASIS

CLCA1 and SPINK4 are differential genes screened from metastases and in situ tumors, this study explored the effects of CLCA1 and SPINK4 on colon cancer cell migration at the cellular level. First, use Western Blot and RT-PCR to explore the expression of these two genes in colon cancer cells (Fig. 7 A-B). According to the differential expression and cell characteristics, SW620 and LOVO cell lines were selected to transiently transfect the two genes and follow-up. Transwell experiment showed that over-expression of CLCA1 and SPINK4 can inhibit the migration of tumor cells SW620 and LOVO. And over-expression of CLCA1 and SPINK4 can up-regulation E-cadherin and down-regulate Vimentin. Experimental results show that up-regulation of CLCA1 and SPINK4 can inhibit cell migration.

Figure 7 (A) Western blot detects the expression of CLCA1 and SPINK4 in colorectal cancer cells, (B) RT-PCR detects the expression of CLCA1 and SPINK4 in colorectal cancer cells, (C) RT-PCR detects the overexpression of CLCA1 and SPINK4 in LOVO cells, (D) RT-PCR detects the overexpression of CLCA1 and SPINK4 in SW620 cells, (E) Effects of overexpression of CLCA1 and SPINK4 on Transwell migration in LOVO and SW620, (F) Western blot was used to detect the overexpression of CLCA1 and SPINK4 in LOVO and SW620, and the effect on EMT after overexpression.
DISCUSSION
Colorectal cancer (CRC), as the third most common cancer affecting the gastrointestinal tract, is a common type of cancer worldwide and is associated with a high mortality rate⁴⁶. Metastasis is the most important cause of colorectal cancer death. In this study, we mainly screened differentially expressed genes of colon cancer related to liver metastasis, and used these genes for lasso training and cox analysis to obtain a more robust prognostic survival model for colorectal cancer.

This study improves the method of lasso training, using the loop operation mode to obtain the best simulation set, and the model obtained by the transcriptome sequencing set data in this study has also been verified by the GEO transcriptome chip data set, further ensuring the stability of the model. In addition, this study also found that the genes CLCA1 and SPINK4 are related to the prognosis of colon cancer patients from the perspective of bioinformatics and there is a co-expression relationship between the two genes. At the same time, these two genes can also inhibit the migration ability of colon cancer tumor cells and affect the expression of tumor metastasis-related proteins.

The shortcoming of this study is that due to the limitations of objective factors, the model built cannot be verified by the actual data of the local hospitals, and the model has not included other relevant clinical indicators, only at the level of transcriptome gene expression. In this study, the inhibition of tumor migration by CLCA1 and SPINK4 was only discussed at the cellular level, but not verified at the animal level. At the same time, the mechanism pathways of genes CLCA1 and SPINK4 in inhibiting colon cancer metastasis have not been discussed.

CONCLUSION
In this study, through the four datasets of GSE81582, GSE41258, GSE49355, and GSE68468 in GEO, a total of 197 differentially expressed genes related to metastasis were obtained (|log FC|>1, p<0.05), of which 117 were up-regulated and down-regulated. There are 80 of them. The differentially expressed gene set obtained single-factor cox analysis, lasso cycle training and multi-factor cox analysis in the TCGA-COAD data set, and obtained the survival prognosis prediction composed of the seven genes ORM1, CLCA1, C8B, SPINK4, ALDOB, GAMT, and C8G. Among the 7 factors in the model, the genes CLCA1 and SPINK4 were found to be closely related to the prognosis of colon cancer (CLCA1: Log-rank p=0.012, HR=0.54; SPINK4: Log-rank p=0.00018, HR=0.39). And there is a positive relationship between CLCA1 and SPINK4. There is also a positive relationship between the two genes. In the single-gene GSEA analysis of CLCA1 and SPINK4. And at the cell level, the high expression of CLCA1 and SPINK4 can inhibit colon cancer cell migration.

ACKNOWLEDGMENTS
This work was supported by University's Characteristic Innovation Natural Science Project of Guangdong Provincial (No: 2018KTSCX033) and Southern Medical University research start-up plan (No: CX2018N014). Thanks to Xu Jun senior engineer for guidance.

REFERENCES
doi:10.1136/gut.2010.211557. [PubMed]