**Revealing Mechanisms Implicated in Colorectal Cancer from TCGA Gene Expression Data Using Gene Set Enrichment Analysis**

Ruixi Li rl3328

**Abstract**

Colorectal cancer (CRC) is the most common malignant tumor of digestive system. The aim of this study was to identify differentially expression genes (DEGs) in CRC and uncover their potential mechanisms through enrichment analysis. The transcriptome profiling of 5 CRC tumor samples and 5 independent normal samples was downloaded from TCGA database through Bioconductor. Generalized linear model likelihood ratio test was performed to identify DEGs. The gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes pathway (KEGG) enrichment analyses were performed. In total, 10802 DEGs were identified in CRC, including 9760 up-regulated genes and 922 down-regulated genes. GO and KEGG pathway analysis results showed that up-regulated DEGs were significantly enriched in cell proliferation, cell migration, cell adhesion and ion channel activity; the down-regulated DEGs were significantly enriched in oxidation reduction, lipid metabolism and ion binding. After discussion and literature review, I think ferroptosis might be a potential mechanism and a potential treatment target of CRC.

**Introduction**

CRC is one of the most frequent malignancies worldwide with close to 2 million new cases and about 1 million CRC deaths estimated in 2020, corresponding to 10.7% and 9.5% of all new cancer cases and deaths worldwide, respectively[1]. Incidence of CRC in developed countries is five times as that in developing countries and constitutes over 90% of the global new cases[1]. Like other cancers, CRC is considered as a heterogeneous disease in which gene aberrations, cellular context, and environmental influences concur to tumor initiation, progression, and metastasis[2]. Accumulating evidence also has demonstrated that multiple genes and cellular pathways participate in the occurrence and development of CRC. To date, a lack of knowledge regarding the precise molecular mechanisms underlying CRC progression limits the ability to treat advanced disease. Therefore, understanding the molecular mechanism involved in proliferation, apoptosis, and invasion of CRC is extremely important for the development of more effective diagnostic and therapeutic strategies. Instead of only focusing on the association between individual gene expression, consider genes within gene sets regarding biological functions and pathways may light the further insight of CRC development at molecular level and explored the potential candidate biomarkers for diagnosis, prognosis, and drug targets.

**Materials and Methods**

**TCGA transcriptome profiling**

Transcriptomic data of CRC was downloaded from TCGA using the query of barcodes "TCGA-W5-AA2X-01A-11R-A41I-07","TCGA-W5-AA30-01A-31R-A41I-07", "TCGA-W5-AA2R-11A-11R-A41I-07","TCGA-ZH-A8Y2-01A-11R-A41I-07", "TCGA-W5-AA2G-01A-11R-A41I-07","TCGA-W5-AA2Z-01A-11R-A41I-07", "TCGA-W5-AA31-11A-11R-A41I-07","TCGA-W5-AA2Q-11A-11R-A41I-07", "TCGA-W5-AA2X-11A-11R-A41I-07","TCGA-W5-AA33-01A-11R-A41I-07". A total of 5 tumor samples and 5 healthy samples were included in these datasets. Since the health samples were independent healthy samples, the experiment design is case control study. These datasets were subjected to quantitative normalization and log2 transformation using the TCGAbiolinks’ package.

**Data Preprocessing**

For transcriptome data, a boxplot correlation and AAIC plot were used to define outliers, which were removed if existed. Normalization was performed on data using the ‘TCGAbiolinks’ package. Following that, quantile filter was conducted to filter out genes whose expression levels fall below the 25th percentile across all samples, effectively reducing the dimensionality of your data and potentially removing noise or uninformative genes. Details were shown in the R markdown file attached together.

**Differential Expression Analysis**

Generalized linear model likelihood ratio test was performed using ‘TCGAbiolinks’ package in differential gene expression analysis between CRC tumor samples and normal controls. Genes with an FDR < 0.05(FDR was applied for multiple comparison to correct for the inflated type I error) and |log2FC (fold change)| > 2 were considered DEGs. The volcano plot of DEGs were constructed using the “ggplot2” package in R4.3.3.

**Gene ontology and pathway enrichment analysis of DEGs**

GO is a common useful method for annotating genes and gene products and for identifying characteristic biological attributes for high-throughput transcriptome data[3, 4]. KEGG is a knowledge base for systematic analysis of gene functions, linking genomic information with higher-order functional information[5]. In order to analyze the DEGs at the functional level, GO enrichment and KEGG pathway analysis were performed using ‘TCGAbiolinks’ package. P<0.05 was considered statistically significant.

**Results**

**Identification of DEGs**

A total number of samples analyzed were 5 CRC samples and 5 normal samples. Using P<0.05 and fold control (FC) >2.0 criteria, a total of 10682 genes were identified, of which 9760 were up-regulated and 922 were down-regulated. The volcano plot for DEGs is shown in Figure 1.

图表, 散点图

描述已自动生成

**Figure 1. Volcano plot of the differentially expressed genes (DEGs).** Y‑axis: logFC (fold change); X‑axis: ‑log10 (FDR) for each gene; Y-axis: -1og10FDR. the color of the data points denotes the status of DEGs (green points: up-regulated genes with FC >2 and FDR <0.05; orange points: down-regulated genes with FC <2 and FDR <0.05); grey points: no significant change in gene expression. FDR, false discovery rate; T, tumor tissues; NT, independent non‑tumor tissues.

**GO term enrichment analysis**

We used DEGs to identify overrepresented GO categories and KEGG pathways. Enrichment analyses of the upregulated and downregulated genes were performed separately, as previously recommended[6]. GO analysis results showed that up-regulated DEGs were significantly enriched in biological processes (BP), including sensory perceptions, cell adhesions and ion transportation (Figure 2); the down-regulated DEGs were significantly enriched in biological processes, including oxidation reduction, lipid metabolic process (Figure 3). For molecular function (MF), the up-regulated DEGs were enriched in calcium ion binding and channel activities (Figure 2); the down-regulated DEGs were enriched in electron carrier activity, heme binding and iron ion binding (Figure 3). In addition, GO cell component (CC) analysis also displayed that the up-regulated DEGs were significantly enriched in the plasma membrane part, iron channel complex and mitochondrial (Figure 2); the down-regulated DEGs enriched in extracellular region and vascular transport (Figure 3).

**KEGG pathway analysis**

The plot at the right-bottom corner of Figure 2 and figure 3 contain the most significantly enriched pathways of the up-regulated DEGs and down-regulated DEGs analyzed by KEGG analysis. The up-regulated DEGs were enriched in EIF2 signaling, Regulation of the Epithelial−Mesenchymal Transition Pathway, RXR Activation and Acute Phase Response Signaling (Figure 2) while the down-regulated DEGs were enriched in pathways that increase RXR inhibition and decreased RXR activation and melatonin degradation (Figure 3). Figure 2 and Figure 3 were also attached together with the final report.

图表, 条形图

描述已自动生成

Figure 2. Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of upregulated differentially expressed genes. X‑axis: ‑log10 (FDR) for each gene set(domain). N: Number of genes related to the enriched GO or KEGG pathway.

图表, 条形图

描述已自动生成

Figure 3. Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of downregulated differentially expressed genes. X‑axis: ‑log10 (FDR) for each gene set(domain). N: Number of genes related to the enriched GO or KEGG pathway.

**Discussion and Conclusion**

Understanding the molecular mechanism of CRC is of critical importance for diagnosis and treatment. In this study, I extract the data from 5 tumor sample and 5 independent healthy sample of TCGA database and identify 9760 up-regulated and 922 down-regulated DEGs between CRC and normal control using differential expression analysis. Numerous abnormally modified GO and KEGG pathways were closely associated with cancer. Function analysis showed that the up-regulated DEGs were involved in cell adhesion, Chromosomal Replication, Epithelial−Mesenchymal Transition and EIF2 signaling, indicating the increased cell proliferation and migration in tumor tissue whereas the down-regulated DEGs were involved in oxidation reduction and lipid metabolism process, indicating the excessive oxidative stress and lipid metabolic aberrance in tumor tissue, which were reported as common biological responses in CRC[7, 8]. Interestingly, the functional analysis also revealed some potential clues about the mechanisms of CRC. Elevated iron-related transportation and reduced iron binding displayed in GO-molecular function analysis suggested the accumulation of cellular iron overload, which along with lipid peroxidation can lead to ferroptosis[9]. Retinoid X Receptor (RXR) activation in tumor was shown in KEGG pathway analysis through genes enriched in multiple RXR-related pathways, including FXR/RXR Activation, which was verified in mice model and in vivo liver cells that might suppresses lipid peroxidation and ferroptosis[10]. The protein ubiquitination pathway and melatonin degradation shown in KEGG pathway analysis also suggested a pivotal role of ferroptosis in CRC because melatonin is a ferroptosis prohibitor and the ubiquitin-proteasome system (UPS) and autophagy have a complex interaction with ferroptosis and have a dual effect in ferroptosis dependent on conditions[11, 12]. In summary, ferroptosis might be a potential mechanism of CRC according to the function analysis. Some evidence in biomedical research also provide rationale for considering ferroptosis as potential mechanisms and effective therapeutic targets[13]. But further research is needed to further explore the mechanisms. We can find interaction between ferroptosis and CRC-associated genes and construct a protein-protein network (PPI) to identify specific genes that might be associated with both. Using the significant associated genes, we can further explore the expression of genes with survival risk (survival analysis) or using prediction model (e.g. logistic regression, machine learning algorithms) to assess their ability to predict diagnosis. (These are common epidemiological research questions and methods.)

**Reference**

1. Morgan E, Arnold M, Gini A, Lorenzoni V, Cabasag CJ, Laversanne M, Vignat J, Ferlay J, Murphy N, Bray F: **Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from GLOBOCAN**. *Gut* 2023, **72**(2):338-344.

2. Aran V, Victorino AP, Thuler LC, Ferreira CG: **Colorectal Cancer: Epidemiology, Disease Mechanisms and Interventions to Reduce Onset and Mortality**. *Clin Colorectal Cancer* 2016, **15**(3):195-203.

3. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.

4. Gene Ontology C: **The Gene Ontology (GO) project in 2006**. *Nucleic Acids Res* 2006, **34**(Database issue):D322-326.

5. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic Acids Res* 2000, **28**(1):27-30.

6. Hong G, Zhang W, Li H, Shen X, Guo Z: **Separate enrichment analysis of pathways for up- and downregulated genes**. *J R Soc Interface* 2014, **11**(92):20130950.

7. Basak D, Uddin MN, Hancock J: **The Role of Oxidative Stress and Its Counteractive Utility in Colorectal Cancer (CRC)**. *Cancers (Basel)* 2020, **12**(11).

8. Coleman O, Ecker M, Haller D: **Dysregulated lipid metabolism in colorectal cancer**. *Curr Opin Gastroenterol* 2022, **38**(2):162-167.

9. Li J, Cao F, Yin HL, Huang ZJ, Lin ZT, Mao N, Sun B, Wang G: **Ferroptosis: past, present and future**. *Cell Death Dis* 2020, **11**(2):88.

10. Tschuck J, Theilacker L, Rothenaigner I, Weiss SAI, Akdogan B, Lam VT, Muller C, Graf R, Brandner S, Putz C *et al*: **Farnesoid X receptor activation by bile acids suppresses lipid peroxidation and ferroptosis**. *Nat Commun* 2023, **14**(1):6908.

11. Zhang D, Jia X, Lin D, Ma J: **Melatonin and ferroptosis: Mechanisms and therapeutic implications**. *Biochem Pharmacol* 2023, **218**:115909.

12. Chen X, Yu C, Kang R, Kroemer G, Tang D: **Cellular degradation systems in ferroptosis**. *Cell Death Differ* 2021, **28**(4):1135-1148.

13. Wang Y, Zhang Z, Sun W, Zhang J, Xu Q, Zhou X, Mao L: **Ferroptosis in colorectal cancer: Potential mechanisms and effective therapeutic targets**. *Biomed Pharmacother* 2022, **153**:113524.