Identification of a novel cuproptosis-related gene signature for predicting the prognosis in colon adenocarcinoma based on WGCNA and machine learning

Dongxue Wang  
The Second Affiliated Hospital of Harbin Medical University

Funing Yang  
The Second Affiliated Hospital of Harbin Medical University

Guiping Han  
The Second Affiliated Hospital of Harbin Medical University

Jifeng Zhang  
The Second Affiliated Hospital of Harbin Medical University

Hongjia Wang  
The Second Affiliated Hospital of Harbin Medical University

Zunyu Xiao  
The First Affiliated Hospital of Ningbo University

Weiyu Chen  
The Fourth Affiliated Hospital of Zhejiang University School of Medicine

Ping Li (✉ h03713@hrbmu.edu.cn)  
The Second Affiliated Hospital of Harbin Medical University

Research Article

**Keywords:** Cuproptosis, colon adenocarcinoma, gene signature, prognostic model, Dihydrolipoamide S-acetyltransferase

**Posted Date:** August 25th, 2023

**DOI:** https://doi.org/10.21203/rs.3.rs-3273593/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License. 
[Read Full License]
Abstract

Background

Colorectal cancer (CRC) is a highly aggressive malignant tumor that primarily affects the digestive system. It is frequently diagnosed at an advanced stage. Cuproptosis is a copper-dependent form cell death mechanism, distinct from all other known pathways underlying cell death, tumor progression, prognosis, and immune response. Although the role of cuproptosis in CRC has been investigated over time, there is still an urgent need to explore new methods and insights to understand its potential function.

Methods

The GEO and TCGA gene expression data were systematically explored to investigate the role of cuproptosis in colon adenocarcinoma (COAD). The weighted gene co-expression network analysis (WGCNA) was used to construct a gene co-expression network and identify the critical module and cuproptosis-related genes (CRGs) correlated with COAD prognosis. A CRGs gene prognostic signature for COAD was identified by machine learning algorithm and validated.

Results

Five hub CRGs (DLAT, CDKN2A, ATOX1, VEGFA, ULK1) were screened and a prognostic risk model for predicting overall survival (OS) was established based on these genes. The model was successfully tested in the validation cohort and the GEPIA database. COAD patients were categorized into high-risk and low-risk groups based on risk scores. The study revealed that patients with higher risk scores were more likely to have a poor prognosis.

Conclusions

These findings suggest that the newly identified five-gene signature may serve as a more reliable prognostic factor than clinical factors like age and stage of disease. These findings offer a theoretical foundation for further investigation into potential cuproptosis-related biomarkers for predicting COAD prognosis in the future.

Background

CRC ranks the third most common cancer worldwide among males and females, accounting for approximately 12% of new cancer cases and 9% of cancer-related deaths in 2020 [1]. The incidence has increased continuously and rapidly in recent decades [2]. The most frequent histological type of
colorectal cancer is colon adenocarcinoma (COAD), which accounts for approximately 95% of all cases [3]. Due to the absence of efficient diagnostic methods, most patients are diagnosed in advanced stages of the disease. In addition, it is necessary to note that nearly 50% of patients with early-stage disease will develop metastasis [2]. While radical surgery has traditionally been an effective therapy, its effectiveness has decreased significantly over time [4]. In addition to surgical intervention, drug treatments such as gene-targeted therapy and immunotherapy have been found to be effective. However, despite these options, most patients with COAD still experience tumor recurrence due to drug resistance and lack of responsiveness to current immune checkpoint therapy [5, 6]. As a significant global public health challenge, it is imperative to comprehend the molecular mechanism behind the recurrence and metastasis of COAD development. Moreover, identifying novel potential biomarkers for early detection and personalized treatment strategies is crucial.

Cuproptosis is a newly discovered type of regulated cell death (RCD) linked to copper metabolism. Unlike other known forms of cell death, such as apoptosis, ferroptosis, and necroptosis, cuproptosis has distinct molecular characteristics [7]. Copper is an essential trace element that plays a crucial role in various biological processes, though the redox properties make it both beneficial and toxic to the cell [8, 9]. Studies have shown that the levels of copper in cancer patients are significantly higher in both serum and tumor tissues compared to healthy counterparts, such as breast cancer [10], lung cancer [11], gastrointestinal [12], and thyroid cancer [13]. Tumor cells also have a higher level of copper, which indicates the potential of cuproptosis inducers in new antitumor therapy. In their study, Peter et al. demonstrated that copper could bind directly to lipid-acylated components of the tricarboxylic acid (TCA) cycle [7]. This binding increases lipoylated protein aggregation and loss of iron-sulfur cluster proteins, leading to proteotoxic stress and ultimately resulting in cell death [7]. Despite numerous studies on cuproptosis in various diseases, its potential function in COAD remains poorly understood as it is a newly discovered form of regulated cell death.

Here, the gene datasets with sample size and prognostic information for COAD from The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) datasets were adopted. We utilized the weighted gene co-expression network analysis (WGCNA) and least absolute shrinkage and selection operator (LASSO) regression analysis to identify key modules and hub cuproptosis-related genes (CRGs) correlated with COAD prognosis. A new five-gene signature to highlight the regulatory functions of CRGs in COAD progression was created. Our findings provide insights into novel strategies for predicting prognosis and suggest potential therapeutic applications for individualized treatment of patients with COAD.

**Results**

Identification of key modules in colon cancer

A detailed flowchart of this study is displayed in Fig. 1.

Fig. 1 The flowchart of the overall study
A total 952 samples and 23,412 genes in the gene expression obtained from the training sets were downloaded from GEO database. After normalization, the average RNA expression in every sample was basically the same. The WGCNA algorithm was used to build a network, including a group of genes frequently possessing high co-expression level and high topological overlap similarity, as shown in Fig. 2. A soft threshold power of $\beta = 4$ was selected to establish that the network was scale-free, and the corresponding was $R^2 = 0.9$ and with the high average connectivity (Fig. S1A, B). A gene hierarchy clustering dendrogram was constructed using the dynamic tree cut package, and genes in the new data expression profiles were allocated to eleven biologically similar modules (Fig. 2A). Meanwhile, we analyzed the correlation between modules and phenotypes, and the turquoise module containing 1935 genes had the most significant clinically correlation in COAD, based on the correlation of module feature values (Fig. 2B). The scatter plot (Fig. 2C) shows a strong correlation between the gene significance (GS) and module membership analyzes in the turquoise ($\text{Cor}=0.6$, $p = 2.3\text{e-59}$). The scatter plots of other modules are established in (Fig. S2).

Fig. 2 Identification of weighted gene coexpression network analysis (WGCNA) key module. A The cluster dendrogram of eleven colored modules based on a dissimilarity measure (1-TOM) in colon cancer patients. B The heatmap of correlation between the biologically significant modules and clinical traits (age, stage, vital status and survival time). C Scatter plot between gene salience and module membership of turquoise module ($\text{Cor}=0.6$, $p=2.3\text{e-59}$)

Functional enrichment analysis

The genes in the turquoise module were studied though GO and KEGG enrichment analysis, to explore the biological functions and potential pathways. The GO analysis results displayed that biological processes were mainly concentrated in extracellular matrix organization, extracellular structure organization, epithelial cell proliferation and ossification (Fig. 3A). Cellular components were also analyzed, including collagen-containing extracellular matrix, apical part of cell, endoplasmic reticulum lumen and many other parts (Fig. 3B). Additionally, the genes of the turquoise module enriched in the part of molecular: extracellular matrix structure constituent, integrin binding, growth factor binding (Fig. 3C). The results of KEGG signaling pathway analysis indicated that the genes of turquoise module were most enriched in Protein digestion absorption, ECM-receptor interaction, and Focal adhesion pathway (Fig. 3D).

Fig. 3 Functional enrichment analysis of GO and KEGG pathway with the turquoise. A Biological processes of GO enrichment. B Cell components process of GO enrichment. C Molecular function of GO enrichment. D KEGG pathways enrichment

Identification of hub CRGs and construction of a five-gene signature for predicting OS
To further explore the regulatory role of cuproptosis in the pathogenesis of COAD, we intersected the CRGs with turquoise module genes to obtain a total of five signature genes serving as candidates: DLAT, CDKN2A, ATOX1, VEGFA, ULK1 (Fig. 4A). Subsequently, LASSO regression module was established based on the five genes to streamline the important characteristic variables (Fig. 4B, C). The relative regression coefficients of the five genes were calculated by LASSO regression analysis in training group and the detailed information were in Table 1.

### Table 1 The detailed information of the five modeling genes and their correlation R values in the turquoise module

<table>
<thead>
<tr>
<th>Ensembl ID</th>
<th>Gene Symbol</th>
<th>Gene Description</th>
<th>Regression Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000150768.15</td>
<td>DLAT</td>
<td>Dihydrolipoamide S-acetyltransferase</td>
<td>0.06423223</td>
</tr>
<tr>
<td>ENSG00000147889.1</td>
<td>CDKN2A</td>
<td>cyclin-dependent kinase inhibitor 2A</td>
<td>0.30363910</td>
</tr>
<tr>
<td>ENSG00000177556.11</td>
<td>ATOX1</td>
<td>antioxidant 1 copper chaperone</td>
<td>-0.198567004</td>
</tr>
<tr>
<td>ENSG00000112715.20</td>
<td>VEGFA</td>
<td>vascular endothelial growth factor A</td>
<td>0.22006697</td>
</tr>
<tr>
<td>ENSG00000177169.9</td>
<td>ULK1</td>
<td>unc-51 like autophagy activating kinase 1</td>
<td>0.25098770</td>
</tr>
</tbody>
</table>

We used the relative expression level and relative regression coefficients of the five genes to calculate the risk score for overall survival (RSO) of every sample. The cut-off values of RSO divided the patients with COAD into a high-risk group and a low-risk group and the regression equation was calculated as follow:

$$\text{RSO} = (0.06423223) \times \text{expression level of DLAT} + (0.30363910) \times \text{expression level of CDKN2A} + (-0.198567004) \times \text{expression level of ATOX1} + (0.22006697) \times \text{expression level of VEGFA} + (0.25098770) \times \text{expression level of ULK1}$$

The distribution plot of RSO and survival status were analyzed by ranking and risk scores. The results indicated that individuals with higher risk scores were more likely to have dead. Moreover, the different expression profiles of the cuproptosis-related five genes were listed by the heatmap between high-risk and low-risk group (Fig. 4D). The principal component analysis (PCA) was also analyzed to prove that the five-gene signature based on RSO could distinguish two separate subgroup of COAD patients (Fig. 4E).

**Fig. 4 Prognostic module constructed by LASSO regression analysis.** A Venn diagram of the intersection of cuproptosis-related genes and turquoise module, obtaining five genes: DLAT, CDKN2A, ATOX1, VEGFA and ULK1. B, C The relative regression coefficients of the five genes. D Distribution plot of the RSO scores and survival status. E PCA plot

Prognostic value of the CRGs five-gene signature
The survival curves for the patients with high-risk and low-risk were performed by the Kaplan-Meier analysis in training and validation groups. The results of both two cohorts suggested that the patients with a higher RSO had a statistically lower probability of survival. Because of the different platforms and normalization methods for each group, the $p$ value was 0.002 in the training group (Fig. 5A) and 0.032 in the testing group (Fig. 5B). Furthermore, the ROC curve was utilized to assess the sensitivity and specificity of the five-gene signature for prognosis. The results were evaluated based on the area under the ROC curve (AUC), which showed the five-gene signature could as a possible predictive factor for OS with the AUC values for the 1-, 3-, and 5-year OS were 0.617, 0.586, and 0.57 in training group, 0.567, 0.602 and 0.573 in validation group (Fig. 5C, D), respectively. To further prove the prognostic value of the five genes and integrated signature in COAD, the GEPIA database was chosen to analyze. The high and low expression was set 70% and 30% median cutoff values. The DLAT high expression group also showed a better prognosis with the $p = 0.049$, HR = 0.52 (Fig. 5E). The CDKN2A low expression group had a better prognosis with the $p = 0.029$ and HR = 2 (Fig. 5F). Despite the ATOX1, VEGFA and ULK1 had similar results to the gene of CDKN2A, there was no statistically significant differences ($p = 0.54$, HR = 1.2; $p = 0.21$, HR = 1.5; $p = 0.098$, HR = 1.7, Fig. S3A). Meanwhile, the five-gene signature with a low expression group showed a better prognosis ($p = 0.019$, HR = 2.1, Fig. 5G). The expression levels of five genes in COAD were compared between cancer and normal tissues using this database. A $|\text{Log2FC}|$ Cutoff value of 0.5 and a $p$-value cutoff of 0.05 were set. The results indicated that the expression levels of DLAT ($p < 0.05$, Fig. 5E), CDKN2A ($p < 0.05$, Fig. 5F), ATOX1 ($p < 0.05$, Fig. S3B), and VEGFA ($p < 0.05$, Fig. S3B) in cancer tissues were significantly higher compared to those in normal tissues, the gene of ULK1 had the contrary tendency ($p < 0.05$, Fig. S3B). Moreover, the expression of the five genes in pan cancers also analyzed, the heatmap suggested the expression of each gene has a slight variation in different cancer patients (Fig. 5H).

The relationship of cuproptosis-related five-gene signature with immune infiltrates was investigated. The results showed that B cells, CD8+ T cells, dendritic cells (DCs), Mast cells, Neutrophils, NK cells, Th2 cells, TIL and Treg cells had obvious difference between high-risk and low-risk group (all $p<0.001$). Besides, T helper cells, T follicular helper (Tfh) cells (both $p<0.01$) and immune dendritic cells (iDCs) ($p<0.05$) were distinct in two groups too (Fig. 6A). In addition, immune-related functions analysis also found that nearly
all had a different expression in the two groups, such as APC co-inhibition, Cytolytic activity, T-cell co-inhibition (Fig. 6B). Immune infiltration through ssGSEA algorithm showed that the immune status in each group was totally different, which indicated that five-gene signature may be further elucidated to develop tumor immunotherapy in COAD. The TIMER database was also used to investigate the immune infiltration associated with five genes in colon cancer. The results revealed that the expression level of the five CRGs had correlation with different tumor immune cell, and \( p<0.05 \) was statistically significant (Fig. 6C).

Fig. 6 The immune status difference in the five-gene signature and each gene expression levels. (A) Comparison of immune cell abundance between the high-risk and low-risk groups of COAD patients. (B) Comparison of immune-related functions between the high-risk and low-risk groups of COAD patients. (C) Expression of the target five genes relating to immune infiltration cells in COAD. The asterisks represented the statistical \( p \) value (*\( p<0.05 \); **\( p<0.01 \); ***\( p<0.001 \))

The CRGs five-gene signature as a better independent prognostic factor for COAD

The univariate cox regression analysis was performed to investigate prognosis-related factors of colon cancer samples in the training set. The RSO calculated from the five genes and two common essential clinical characteristic factors (age and stage) were included in the cox regression analysis. The result demonstrated that RSO (\( p < 0.0001 \)), age (\( p < 0.0001 \)), and stage (\( p < 0.0001 \)) might be as independent risk factors for COAD influencing OS (Fig. 7A). Moreover, the RSO had a higher HR value of 2.745 (95%CI: 1.653–4.558) compared with other clinical factors, which indicated that the risk of death in the high RSO group was 2.745 times that in the low RSO group. Additionally, a nomogram was constructed to establish a method for quantitatively capable of predicting the survival probabilities of 1-, 3-, and 5-year OS in colon cancer patients based on the high- and low-risk score signature (Fig. 7B). The result showed that RSO had a wider range of points compared with age or stage, as a stronger factor for predicting the survival rates. Thus, it can be concluded that the predictive value of RSO is higher than that of the two important clinical factors (age and stages, respectively) in the training set.

Fig. 7 The five-gene signature as a better prognostic factor than clinical factors. A Univariate analysis of the five-gene signature and clinic pathological factors (age and stage) for overall survival in COAD patients. B Nomogram with the RSO and clinicopathological features for predicting the 1-, 3-, 5-year survival

Up-regulation of DLAT accelerates cell death of colorectal cancer cells in vitro

To ascertain the biological significance and the correlation of DLAT with the malignancy of colorectal cancer, we performed functional studies using overexpression plasmid to up-regulation DLAT in HT29 cell lines in vitro. The map of plasmid showed in Fig. S4. We transfected DLAT overexpression plasmids and empty vectors into HT29 cell, and the efficiency of interference was confirmed by mRNA levels and protein levels. Meanwhile, we found that the transfection group's protein expression was obviously elevated (Fig. 8A, B). Subsequently, through CCK-8 assays, the cell growth decreased remarkably upon
DLAT up-regulation, as shown in Fig. 8C. In addition, the colony formation assay also demonstrated that overexpression DLAT inhibited cell proliferation (Fig. 8D, E). Because the DLAT distinctly regulated HT29 cell proliferation, then the impact on the cell cycle was assessed by flow cytometry analysis and the results indicated an increased tendency in the proportion of G0/G1 phases in the DLAT overexpression group compared with the vector group (Fig. 8F, G). DLAT, as a catalytic subunit of the human the pyruvate dehydrogenase complex, considered to facilitate TCA cycle metabolism [14]. To investigate whether DLAT affected the redox reactions of COAD, ROS analysis was conducted. And the results showed that ROS accumulation was higher in DLAT overexpression cell line than the negative control group (Fig. 8H, I).

Fig. 8 Up-regulation of DLAT inhibits cell function of colorectal cancer cells in vitro. A Relative expression level of DLAT gene mRNA after transfecting DLAT overexpression plasmids. B Protein expression level of HA-DLAT. C Cell viability analysis by CCK8 showing the effect of DLAT on the proliferation of HT29 cell line at 24, 48, 72 and 96h. D, E Colony formation analysis after DLAT up-regulation. F, G Flow cytometry showing the DLAT influencing the cell cycle of HT29 cell. H, I Flow cytometry showing the accumulation of ROS after DLAT up-regulation. The asterisks represented the statistical P value (*p<0.05; **p<0.01; ***p<0.001)

Discussion

Herein, we explored the role of CRGs in colon cancer and utilized various bioinformatics tools and statistical approaches to identify a new cuproptosis-related five-gene signature biomarker for the prediction of prognostic. In addition, we further analyzed the functions of the five-gene signature involved in the model construction and other biological functions.

Colon adenocarcinoma is a heterogenous disease and dominates a major driver of morbidity and mortality in recent years. There are many molecular alterations that contribute to the dysregulation of different signaling pathways, ultimately influencing tumor onset, progression, and invasiveness [15]. Despite significant advances in clinical treatment, the exact pathogenesis of COAD remains a pressing problem. Over the past two decades, research on the genetic and epigenetic regulations has shown promising potential in improving this disease's diagnosis, therapy, and survival prediction [16]. Furthermore, numerous studies have shown that tumorigenesis and progression involve various molecular pathways and biomarkers, such as phosphatidylinositol3 kinase (PI3K)/AKT signaling pathway [17], mitogen-activated protein kinase (MAPK) pathway [18], KRAS mutation [19]. Active research on novel genes is essential for improving the prognosis of COAD.

In 2022, Tsvetkov et al. proposed cuproptosis as a non-apoptotic mode of cell death characterized by an abundance of intracellular copper [7]. This induces cell death through various subroutines, regulates mitochondrial respiration, and leads to glycolysis, insulin resistance, and lipid metabolism changes. Recently, a growing body of research indicates that cuproptosis is a significant factor in the development of various diseases [11, 20–23]. That means there may be potential therapeutic options for targeting
tumors with a specific metabolic profile using copper ions, and reliable biomarkers for detecting these tumors in humans.

This study presents the first implementation of WGCNA and LASSO algorithm to investigate the evidence the value of CRGs in COAD. The study provides a full view of the differential expression of obvious hallmark gene signatures. WGCNA is commonly used to screen key module information from chip data and conduct significant association analyses with phenotype. In the present study, we analyzed four microarray datasets from GEO as the training group and identified a key module that is significantly correlated with clinical prognosis. Functional enrichment analysis (GO and KEGG) based on the target module showed that it was enriched in multiple crucial pathways consistent with previous studies, such as epithelial cell proliferation [18] and ECM-receptor interaction [24]. These findings suggest that these functions may play important roles in the pathogenesis and tumor suppression of COAD. And we identified five pivotal genes (DLAT, CDKN2A, ATOX1, VEGFA and ULK1) that play a key regulatory role in the development of a prognosis signature.

Cyclin-dependent kinase inhibitor 2A (CDKN2A) is usually a tumor suppressor gene located on chromosome 9p21.3. It encodes two proteins, p14 and p16, which primarily function as cell cycle regulators [25]. Various studies have indicated that CDKN2A deregulation expression is linked to a negative prognosis in HPV-negative head neck squamous cell carcinomas [26]. Additionally, it has been associated with a poor prognosis in soft tissue sarcoma [27]. On the contrary, we find that CDKN2A high expression is associated with a poor prognosis, which is downregulated in COAD patients compared with the healthy controls, due to the heterogeneity between samples possibly.

ATOX1 serves as both a copper chaperone and an antioxidant, playing a crucial role in the copper-trafficking pathway that helps maintain intracellular copper homeostasis. It binds and transports cytosolic copper to ATPase proteins in the trans-Golgi network, which is subsequently incorporated into ceruloplasmin [28]. Furthermore, many activities characterize important cytology processes, such as cell proliferation, migration, autophagy, DNA damage repair, and programmed death [29, 30]. Recently, studies have found that ATOX1 plays a role in tumorigenesis in various types of cancer [9]. In human breast cancer, ATOX1 was up-regulated and localized at the lamellipodia edges of aggressive breast cancer cells [31]. Our study also found higher expression of ATOX1 in colon cancer tumor tissue. High expression of ATOX1 has been linked to poor patient outcomes [32], and we observed a consistent trend in COAD via GEPIA analysis, although statistical significance was not reached.

ULK1 is a fundamental human autophagy-related gene that encodes a serine-threonine kinase and is the mammalian counterpart of the yeast ATG1 gene. It is situated on chromosome 12q24.3 [33]. Autophagy is a recycling system in organisms regulated by autophagy-related proteins and their partners. Heavy metals, such as copper, can be cytotoxic due to the ROS generated and ROS-induced activation can trigger the autophagy pathway of cell death [34]. Furthermore, recent studies have highlighted that copper induces autophagy in normal and cancer cells, which can serve as a cellular defense against copper-mediated toxicity [35–37]. Thus, ULK1 gene was also act the CRGs involved in our study. Recent studies
have indicated that ULK1 has a distinct and specific role in mitophagy, particularly in hypoxic conditions [38]. And emerging studies demonstrate that the ULK1 expression protein is down-regulated in breast cancer patients [39], which was same as our results of low expression in colon cancer. In addition, we also found a correlation between the high expression and worsened prognosis in patients through public online analysis, but the $p > 0.5$. There need more studies to further prove the role of ULK1 in COAD.

VEGFA is a protein that belongs to the VEGF family and is closely associated with angiogenesis and development. It promotes the formation of new blood vessels and is often linked to various human diseases [40]. VEGFA binds with VEGFR-2, also known as KDR or Flk-1, the main receptor expressed on endothelial cells found in the tumor vasculature. This binding leads to receptor dimerization and trans-autophosphorylation of multiple tyrosine residues. It is important to note that the increased expression of VEGFA in tumors and the tumor microenvironment (TME) leads to increased tumor microvessel density, invasiveness, metastasis, and worsened patient prognosis [41]. VEGFA plays a crucial role in the formation of structurally abnormal and leaky tumor vasculature, while also promoting the growth of endothelial cells. This process leads to an improvement in antitumor immunity. Moreover, anti-angiogenetic drugs targeting the VEGFA pathway are used in various cancer therapy, especially in non-small cell lung cancer (NSCLC) [42]. Our study has little attention on the function of VEGFA in colon cancer, more experimental studies could be performed to verify the value in the future.

Dihydrolipoamide S-acetyltransferase (DLAT), a mitochondrial protein involved in the multi-enzyme pyruvate dehydrogenase complex and as the key molecules of cuprotosis, may play a crucial role in tumorigenesis [7]. Assuming the Warburg effect, it is intuitive to expect that DLAT would be downregulated in cancer [43]. One study suggested that gastric cancer patients had a higher expression of DLAT than usual. And knockdown in human gastric cancer cells substantially inhibited cell proliferation significantly, revealing a role for this gene in cancer development [14]. Moreover, we also found that the expression of DLAT was up-regulated in COAD. Another study demonstrated that DLAT was a novel oncogene and higher DLAT expression was related to poorer outcome in NSCLC patients [44]. But the DLAT possessed a role of cancer inhibition in COAD by GEPIA analysis and HT29 cell experiments. We constructed the over-expression plasmid of DLAT, and performed the cell cycle analysis, ROS analysis and cell proliferation analysis to investigate the role in colon cancer. Apparently, DLAT is a tumor suppressor gene that can induce cell death in the colon cancer cell line. Nevertheless, the detailed tumorigenesis mechanism research of DLAT in COAD is unknown, thus it is urgent to conduct more experimental studies.

In the present study, we used the above five CRGs characterized by copper metabolism participation to construct a signature associated with prognosis. Recently, several studies have applied signature analysis to predict characteristics or prognosis for colon cancer. For instance, Seven Ferroptosis-related lncRNAs (LINC01503, AC004687.1, AC010973.2, AP001189.3, ARRDC1-AS1, OIP5-AS1, and NCK1-DT) signature was developed as a biomarker to predict clinical outcomes and therapeutic responses in colon cancer patients [45]. Although the prognostic model was validated, this study did not present cell experiments or hospital patient data. Moreover, the gene number in this study was more than our model.
Another eight-gene signature was reported to dichotomize patients with different OS significantly, serving as independent predictor of our study [46]. Our study has shown that the prognostic influence of our five-gene signature is superior to that of clinical factors such as age and stages. Additionally, our signature was constructed using fewer genes than similar reports. Therefore, it could be potentially applied in subsequent clinical translational research or developed as a detection kit to promote clinical applications.

Although the prognostic value of the RSO related to cuproptosis has been validated, some limitations need to be acknowledged. Firstly, our analysis was conducted using retrospective public databases as the basis. While the DLAT gene was confirmed through cell function tests, it is crucial to conduct additional in vitro and in vivo experiments to further substantiate our findings. Furthermore, validation in a larger cohort of colon cancer patients in real-world settings is necessary. Secondly, our analysis was restricted to only the GEO and TCGA databases though the diversity of oncology databases. Therefore, including data from other sources in future studies would be beneficial.

**Conclusions**

In conclusion, we utilized bioinformatics analysis to identify five hub genes closely associated with cuproptosis in COAD. We systematically generated and evaluated a risk score signature based on these genes as a potential prognostic factor for colon cancer patients. These findings could enhance our comprehension of invasion, provide a theoretical foundation for exploring potential regulatory biomarkers for prognosis prediction, and even contribute to developing more precise target therapy strategies.

**Materials and Methods**

Data collection and processing

The dataset GSE17536, GSE39582, GSE29621(platform file GPL570) and GSE28722(platform file GPL13425) were downloaded from the GEO public database (https://www.ncbi.nlm.nih.gov/geo/, accessed on 26 August 2022). Then we merged all the files after gene re-annotation and normalization, removing the batch effects from different datasets using the “sva” package [47]. Above four files as the training cohort enrolled for our subsequent analysis, the expression data of 177,585, 65 and 125 samples were obtained from GSE17536, GSE39582, GSE29621 and GSE28722, respectively. All of 952 COAD cases were used for the following WGCNA and prognosis model construction. Furthermore, the RNA sequencing (Fragments per kilobase million, FMPK) and clinical data of colon cancer patients as the validation cohort were downloaded from The Cancer Genome Atlas (TCGA) Genomic Data Commons (GDC) portal (https://cancer genome.nih.gov/, accessed on 18 August 2022) database, including 451 COAD cases. To reduce the number of errors caused by confounding factors, we excluded samples without survival information (n = 3). All the data were analyzed with the R program (version 4.2.2). Forty-six CRGs were acquired from previous literature for subsequent analyses [7, 8, 48, 49] with a false discovery rate (FDR)< 0.05 as a filter condition listed in Table 2.
Weighted gene co-expression network analysis (WGCNA) was constructed with an R package named “WGCNA”, according to the method published in the literature [50]. We used the goodSampleGenes method to dislodge non-expressed genes and leveraged a standard deviation of >1 to identify expressed genes in colon cancer cases for cluster analysis. Firstly, the Person's correlation coefficient was calculated to construct a gene matrix for all pair-wise genes, investigating the relationships between gene networks and clinical traits. The power function was utilized to construct a scale-free co-expression network. After selecting appropriate $R^2$ and soft-threshold value $\beta$, the adjacency matrix was transformed into a topological overlap matrix (TOM). The TOM facilitated the measurement of a specific gene's network connectivity. Next, the gene modules were identified by hierarchical clustering, with the minimum size number of 50 genes per module for the gene dendrogram and the minimum size of 2 for deepSpilt, and the eigengene was also calculated. Furthermore, we merged all close modules into new modules with the height of 0.25 via cluster analysis. The correlations between the modules’ characteristic genes and the disease phenotypes of the training group were calculated using the eigenvectors of modules and gene expression of samples, hub genes were selected for further prognosis model building with $R > 0.7$.

Gene functional and pathway enrichment analysis

Enrichment analysis of the key module consisted of Gene Ontology (GO) analysis and Kyoto Encyclopedia of Genes (KEGG) pathway analysis, performed by “clusterProfiler” R package [51]. GO was performed three main parts, including biological process (BP), cellular component (CC), and molecular function (MF). Via ggplot2 package in R software to plot the top 10 GeneRatio of biological process, cell components or molecular function of GO analysis and the top 10 GeneRatio of KEGG pathways. The “ggplot2” package was used to plot the figures results based on the enrichment analysis [52].

Table 2. The identification of cuproptosis-related genes

<table>
<thead>
<tr>
<th>Gene symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIAS, LIPT1, LIPT2, CDKN2A, DLAT, DLST, DLD, FDX1, COA6, ATP7A, ATP7B, SLC31A1, SLC31A2, SLC25A3, PDHA1, PDHB, MAP2K1, MTF1, GLS, UBE2D1, H3C1, CD274, COX11, UBE2D2, PDK1, SCO1, COX17, UBE2D4, UBE2D3, DBT, SOD1, ULK2, CP, AOC3, DBH, LOXL2, ULK1, VEGFA, GCSH, CCS, MAP2K2, ATOX1, MTCO2P12, TYR, NF2L2, NLRP3</td>
</tr>
</tbody>
</table>
Construction of the LASSO-cox regression model and validation

Before establishing the model, taking the intersection of CRGs with candidate genes of the turquoise module screened from WGCNA, then the selected genes were used to build the eventual model. After all datasets was integrated the survival time, survival status, gene expression data and common clinical characteristics(age, sex, clinical stages), the files from GEO database, containing gene expression and overall survival(OS) related clinical prognosis information, were as the training group to construct the Least absolute shrinkage and selection operator(LASSO)-Cox regression model using the “glmnet” package [53], a machine learning algorithm applying to select the cuproptosis-related signature through the risk score. Meanwhile, we used the 10-fold cross-validation method and took the minimized $\lambda$ as lambda to obtain an optimized model. The risk score for OS (RSO) was calculated via multiplying the expression level based on the coefficients, which affected the prognosis of COAD. The calculation of RSO formula for each patient as follows: RSO = $\text{coef}_1 \times \text{Exp}_1 + \text{coef}_2 \times \text{Exp}_2 + \ldots + \text{coef}_i \times \text{Exp}_i$

($\text{coef}_i$ represents the corresponding coefficient of each cuproptosis-related signature, \text{Exp}_i identifies the expression of hub gene.)

The median RSO of training set was set as the cutoff value, all samples were categorized into high-risk and low-risk groups according to the cutoff. Next, the R package “ggplot2” was used to generate the gene expression heatmaps of RSO. In addition, the univariate survival analysis was used to construct a forestplot, containing the clinical information of COAD patients, Hazard ratio (HR) and 95%CI. Moreover, a nomogram capable of predicting the prognosis was created using the “rms” package to illustrate the independent factors. The “Points” indicate the effect of each factor on the prognosis. Meanwhile, the principal component analysis (PCA) was performed to observe the distinction between the two risk clusters. We also compared the OS difference between the two clusters by K-M survival analysis. Finally, the time dependent (1-year, 3-year, 5-year) receiver operating characteristic (ROC) curve was constructed to assess the prognostic prediction performance of this model. For further validation, TCGA cohort as the validation set, were divided into high and low risk group based on the cutoff RSO value, K-M and ROC curves were constructed respectively. We also analyzed the five target genes Survival Analysis in the Gene Expression Profiling Interactive Analysis (http://gepia.cancer-pku.cn/, GEPIA, accessed on 26 December 2022) database individually [54]. In addition, the expression of the five genes in normal tissues and tumor tissues of colon cancer was also analyzed through this dataset, the survival heatmap in pan cancers of these genes was also evaluated.

The immune cell infiltration estimation

To investigate the immune cell infiltration, the ssGSEA enrichment analysis performed by the R GSVA package was used to analyze the tumor-infiltrating immune cell subgroups and immune function between the high-risk and low-risk groups. The final visualization was presented via box plots. The open website of TIMER (https://cistrome.shinyapps.io/timer/) was used to study the correlation between the five hub CRGs and six tumor infiltrating immune cells.
Cell culture and plasmid transfection

Human CRC cell line HT29 was cell cultured in RPMI 1640 supplemented with 10% fetal bovine serum (FBS) and penicillin (100U/mL)/streptomycin (0.1mg/mL) at 37 °C in a humidified atmosphere containing 5% CO2. DLAT overexpression plasmid and the empty control were constructed using pcMV-HA-C-DLAT and vector, respectively. HT29 cell line was transfected with plasmid by Lipofecter reagents (Hanbio, Shanghai, China) according to the manufacturer’s protocol.

RNA isolation and quantitative RT-PCR

After being treated with DLAT overexpression plasmid for 48 hours, total RNA was extracted from the HT29 cell line using the RNA Quick Purification kit (Vazyme, Jiangsu, China) and 500 ng of RNA was reverse transcribed into cDNA using the PrimeScript RT Reagent Kit (Vazyme, Jiangsu, China) according to the manufacturer’s instructions. Gene expression levels were measured by reverse transcription-quantitative polymerase chain reaction (RT-qPCR) using TB Green Premix Ex Taq (TaKaRa, Japan) on a CFX96 Real-Time System (Bio-Rad, UK). The PCR primers were synthesized via Hzykang (Zhejiang, China) and the sequence were as follows: 5’ – CCGCCGCTATTACAGTCTTCC (sense) and 5’ – CTCTGCAATTAGGTCACCTTCAT (anti-sense).

Western blot

HT29 cells in 6-well plate were lysed in RIPA lysis buffer (Beyotime, Jiangsu, China) to extract total protein after transfecting DLAT overexpression plasmid for 48 hours. Protein concentrations were detected via the BCA protein assay kit (Beyotime, Jiangsu, China) and separated on an 10% SDS-PAGE gel and then transferred to polyvinylidene fluoride (PVDF) membranes. 5% skim milk was used to block the membrane at room temperature for 2 h and then incubated with the target antibodies at 4 °C overnight. Primary antibodies included: Beta-actin (66009-1-lg, Proteintech, Wuhan, China), HA (13426-1-AP, Proteintech, Wuhan, China). The next day, after washing three times, the samples were incubated with conjugated secondary antibodies (1:5000) at room temperature for 120 min. The protein bands were scanned using ECL detection system and visualized by the Chemidoc touch imaging densitometer (Bio-Rad, USA), and analyzed by Image Studio software.

Cell viability assay

Cell proliferation rate was detected with Cell Counting Kit-8 kit (Beyotime, Jiangsu, China) according to the manufacturer’s instruction. Briefly, HT29 cells were plated in 96-well plats (2500 cells/well) and transfected with plasmid. 10μl CCK-8 reagent was added to each well with 100μl medium at the indicated time points (0, 1, 2, 3,4 days after transfection) and then incubated for another 1h. 450nm of the absorbance was measured by a microplate reader (Bio Tek, China).

Colony formation
After transfection with DLAT overexpression plasmid, the cells were seeded in 6-well plates with 500 cells per plate and cultured for 7 days. Next, the cells were fixed with 4% paraformaldehyde for 30 min. Subsequently, the cells were stained using 0.5% Crystal Violet Staining Solution (Beyotime, Jiangsu, China) for 15 min. The colonies were photographed, and the Image J software was used to count the number.

Cell cycle analysis

Cell cycle analysis was performed after transfection with plasmid for 24 hours. In detail, HT29 cells were collected by trypsinization, fixed in ice-cold 70% ethanol at 4°C overnight, stained for Propidium (PI) and performed by CytoFLEX S flow cytometry (Beckman, USA) based on the protocol provided by the Cell Cycle and Apoptosis Analysis Kit (Beyotime, Jiangsu, China).

Reactive oxygen species (ROS) assay

Intracellular Reactive oxygen species (ROS) levels were detected by the ROS kit (Beyotime, Shanghai, China). After the transfecting plasmid, cells in 6-well plates were washed with PBS and incubated with 10 μM DCFH-DA for 30 min. Then washing the cells with serum-free solution and collecting, the mean fluorescence intensity of DCFH-DA was measured by flow cytometry (Beckman, USA), which was representative of ROS level. Subsequently, FlowJo software V10 was used to analyze the results.

Statistic

The student's t-test was used to compare differences between two groups. GraphPad Prism 9.0 software was used to conduct the statistical analyses. $p < 0.05$ was considered statistically significant.

Declarations

Acknowledgments: We express our gratitude to the TCGA and GEO databases for providing platforms and collaborating with us in uploading crucial datasets. Additionally, we would like to acknowledge the valuable contributions of our laboratory members, who have engaged in helpful discussions and provided technical assistance.

Author Contributions: All authors in this study have made substantial contributions to the study conception and design. GH and PL contributed to the study design. FY and JZ contributed to data collection. WC and ZX performed statistical analysis and interpretation. HW and DW drafted the manuscript. All authors read and approved the final manuscript.

Funding: This research was funded by the major subject of science and technology of Jinhua, China (grant number 2022-3-039).

Availability of data and materials The datasets collected and analyzed in the study are available in the TCGA repository, (https://gdc.cancer.gov/) and GEO portal, (https://www.ncbi.nlm.nih.gov/geo/). The
original contributions made in this study are included in the article/Supplementary Material. For any further inquiries, please contact the corresponding author.

**Ethics approval and consent to participate** The studies involving human participants were reviewed and approved by the Ethics Committee of the Second Affiliated Hospital of Harbin Medical University (Ethics approval number: YJSDW2022-027).

**Consent for publication** All authors listed in this research have taken part in the work completely and approved the final manuscript for publication.

**Competing interests** The authors declare that this study was conducted without any commercial or financial relationships that could be perceived as potential conflicts of interest.

**References**


35. Luo Q, Song Y, Kang J, et al. mtROS-mediated Akt/AMPK/mTOR pathway was involved in Copper-induced autophagy and it attenuates Copper-induced apoptosis in RAW264.7 mouse monocytes. Redox Biol. 2021;41:101912.


**Figures**
Figure 1

The flowchart of the overall study
Figure 2

Identification of weighted gene coexpression network analysis (WGCNA) key module. A The cluster dendrogram of eleven colored modules based on a dissimilarity measure (1-TOM) in colon cancer patients. B The heatmap of correlation between the biologically significant modules and clinical traits (age, stage, vital status and survival time). C Scatter plot between gene salience and module membership of turquoise module (cor=0.6, p=2.3e-59)
Figure 3

Functional enrichment analysis of GO and KEGG pathway with the turquoise. A Biological processes of GO enrichment. B Cell components process of GO enrichment. C Molecular function of GO enrichment. D KEGG pathways enrichment
Figure 4

Prognostic module constructed by LASSO regression analysis. A Venn diagram of the intersection of cuproptosis-related genes and turquoise module, obtaining five genes: DLAT, CDKN2A, ATOX1, VEGFA and ULK1. B, C The relative regression coefficients of the five genes. D Distribution plot of the RSO scores and survival status. E PCA plot
Figure 5

Prognostic value of the CRGs ve-gene signature. A Univariate survival analysis of the high-risk group and low-risk group in train cohort. B Univariate survival analysis of the high-risk group and low-risk group in validation cohort. C, D ROC curve analysis for the prediction of 1-year, 3-year, and 5-year OS in train and validation cohort. E Through the GEPIA database to analyze the two DLAT groups’ K-M survival curve and the expression difference between cancer and normal tissues. F Through the GEPIA database to analyze the two CDKN2A groups’ K-M survival curve and the expression difference between cancer and normal tissues. G K-M survival curve of the low and high ve-gene signature groups in the GEPIA database. H The difference expression of the five genes in pan cancers. The asterisks represented the statistical $p$ value (*$p<0.05$; **$p<0.01$; ***$p<0.001$)
Figure 6

The immune status difference in the five-gene signature and each gene expression levels. (A) Comparison of immune cell abundance between the high-risk and low-risk groups of COAD patients. (B) Comparison of immune-related functions between the high-risk and low-risk groups of COAD patients. (C) Expression of the target five genes relating to immune infiltration cells in COAD. The asterisks represented the statistical p value (*p<0.05; **p<0.01; ***p<0.001)
Figure 7

The five-gene signature as a better prognostic factor than clinical factors. A Univariate analysis of the five-gene signature and clinic pathological factors (age and stage) for overall survival in COAD patients. B Nomogram with the RSO and and clinicopathological features for predicting the 1-, 3-, 5-year survival.
Figure 8

Up-regulation of DLAT inhibits cell function of colorectal cancer cells in vitro. A Relative expression level of DLAT gene mRNA after transfecting DLAT overexpression plasmids. B Protein expression level of HA-DLAT. C Cell viability analysis by CCK8 showing the effect of DLAT on the proliferation of HT29 cell line at 24, 48, 72 and 96h. D, E Colony formation analysis after DLAT up-regulation. F, G Flow cytometry showing the DLAT influencing the cell cycle of HT29 cell. H, I Flow cytometry showing the accumulation
of ROS after DLAT up-regulation. The asterisks represented the statistical P value (*p<0.05; **p<0.01; ***p<0.001)

Supplementary Files

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

- FigureS1.tif
- FigureS2.tif
- FigureS3.tif
- FigureS4.tif