Integrated Bioinformatics and Machine Learning Analysis Reveals Shared Key Candidate Biomarkers and Therapeutic Targets in Ulcerative Colitis and Colorectal Cancer

Conference Paper · October 2024

DOI: 10.1109/ICICT64387.2024.10839675

CITATIONS
0

READS
23

4 authors, including:

Sakib Sarker
Rajshahi University of Engineering and Technology
3 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Md. Faruk Hosen
Daffodil International University
17 PUBLICATIONS **47** CITATIONS
SEE PROFILE

Md. Abul Basar
Uttara University
12 PUBLICATIONS **35** CITATIONS
SEE PROFILE

2024 2nd International Conference on Information and Communication Technology (ICICT)
October 21-22, Dhaka, Bangladesh

Integrated Bioinformatics and Machine Learning Analysis Reveals Shared Key Candidate Biomarkers and Therapeutic Targets in Ulcerative Colitis and Colorectal Cancer

Sakib Sarker
Department of CSE
Uttara University
Dhaka, Bangladesh
sakib.sarker@uttarauniversity.edu.bd

Md. Faruk Hosen
Department of CIS
Daffodil International University
Dhaka, Bangladesh
faruk.cis@diu.edu.bd

Md Abul Bashar
Department of CSE
Uttara University
Dhaka, Bangladesh
a.bashar@uttarauniversity.edu.bd

Emon Ahammed
Department of CSE
Mymensingh Engineering College
Mymensingh, Bangladesh
emonahammed1999@gmail.com

***Abstract*—The interplay between ulcerative colitis (UC) and colorectal cancer (CRC) has garnered significant research interest due to their potential shared molecular mechanisms. This study aims to identify common significant biomarkers and potential therapeutic targets for UC and CRC. We utilized two microarray datasets to perform differential expression analysis, identifying DEGs for both conditions. Subsequent ML-based gene selection was conducted using SHapley Additive exPlanations (SHAP) algorithm models on the respective datasets. Common ML-based DEGs were then identified and a protein-protein interaction (PPI) network was constructed using the STRING database. The PPI network was visualized and analyzed in Cytoscape, with the top ten hub genes identified using the Degree method in the cytoHubba plugin. The hub genes identified were CDC20, ANLN, HMMR, CCNB1, CDK1, KIF20A, ECT2, KIF11, NUF2, and CCNA2. These genes were further validated through survival analysis, establishing their significance in patient outcomes. Finally, we explored the drug-gene interaction network to identify potential therapeutic drugs targeting these hub genes. This comprehensive bioinformatics approach provides insights into the shared molecular pathways in UC and CRC and highlights potential therapeutic targets for future research and drug development.**

***Index Terms*—Biomarkers, Colorectal cancer, Protein-Protein Interaction (PPI), Hub Genes, Survival Analysis**

I. Introduction

Ulcerative colitis (UC) is a chronic and recurrent inflammatory bowel disease (IBD) with a rising incidence globally. The pathogenesis of UC remains unclear, and there is a lack of specific diagnostic markers, presenting significant challenges. This underscores the urgent need to identify predictive biomarkers for the diagnosis and treatment of UC [1]. Colorectal cancer (CRC) is a prevalent gastrointestinal malignancy characterized by high morbidity and mortality. It is one of the leading causes of cancer-related deaths worldwide. Its incidence has been increasing, driven by changes in lifestyle and dietary habits. The overall survival rate for patients with advanced CRC is very low [2]. Data from population studies indicate that individuals with ulcerative colitis (UC) face a two- to three-fold increased risk of developing colorectal cancer (CRC) compared to the general population. This elevated risk underscores the importance of regular surveillance and early detection strategies in UC patients to prevent the progression to CRC [3]. This highlights that UC and CRC share common molecular pathways, suggesting that significant biomarkers could be valuable for early prognosis and diagnosis of both conditions.

Recently, gene expression analysis has become a popular method for identifying core hub genes in various carcinomas. For instance, Li et al. [2] identified four hub genes (CCNA2, EXO1, CCNB1, and AURKA) in CRC and observed elevated expression of these genes in patients with the disease. Using co-expression network analysis, Mortezapour et al. [4] identified eight hub genes for CRC. Hu et al. [5] pinpointed 13 hub genes related to immune-infiltrated cells in UC through a PPI network analysis. Yuan et al. [6] also utilized a co-expression network and suggested XBP1 and PLPP1 as effective biomarkers after evaluating their diagnostic potential. Many studies have utilized bioinformatics methods to identify and confirm hub genes, but there remain challenges and prospects for advancing research in this area. Incorporating machine learning into these bioinformatics methodologies could bolster their applicability and precision [7].

In this study, we utilized two microarray datasets to explore critical genes in both UC and CRC. Our methodology involved initially identifying differentially expressed genes (DEGs) within each dataset, categorizing them as upregulated

979-8-3315-0822-7/24/$31.00 ©2024 IEEE

or downregulated. We then applied SHAP to pinpoint significant ML-based DEGs among the identified sets. Subsequently, we identified common ML-based DEGs shared between the two datasets. We constructed a protein-protein interaction (PPI) network incorporating these shared ML-DEGs using the STRING database. Analysis of this network in Cytoscape facilitated the identification of core hub genes. To validate their significance, we conducted a survival analysis to assess their roles in the progression and development of UC and CRC. Finally, we analyzed potential therapeutic drugs targeting these hub genes.

II. Materials and Methods

A. Dataset Description

Two publicly accessible microarray datasets were obtained from the Gene Expression Omnibus (GEO) database, identified by their GEO accession numbers GSE87473 [8] and GSE21510 [9]. These datasets were employed to assist in the identification of hub genes. They were structured on the GPL13158 [HT_HG-U133_Plus_PM] Affymetrix HT HG-U133+ PM Array Plate and the GPL570 platform [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array. Additional details regarding the datasets used are summarized in Table I.

TABLE I: Summary of the configuration of the datasets

Dataset	Tumor Samples	Normal Samples	Total Samples
GSE87473	106	21	127
GSE21510	123	25	148

B. Data Preprocessing and DEG Selection

The datasets were subjected to log2 transformation and quantile normalization. Differentially expressed genes (DEGs) were identified in the GSE87473 and GSE21510 datasets using the limma package in R. Criteria for selecting DEGs included an adjusted $p-value < 0.05$ and $|logFC| > 1$. P-values were adjusted using the Benjamini-Hochberg method.

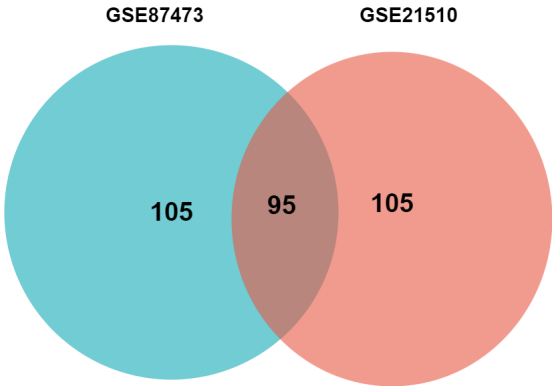


Fig. 1: Illustration of common ML-based DEGs

C. DEGs Selection using Machine Learning

ML-based DEGs were identified using SHAP from the set of DEGs in each dataset, treating each DEG as an individual feature. The SHAP approach is an additive feature attribution technique where each individual prediction is interpreted based on the contribution of the features, and then the features are ranked according to the relevance they have. SHAP assigns a significance value to each characteristic for a specific prediction. SHAP values provide a detailed breakdown of a prediction by quantifying the influence of each individual component. In this study, we identified the top 200 significant DEGs applying SHAP. We identified the common ML-based DEGs from both datasets and proceeded to further analysis.

D. Composition of PPI Network

The Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) enhances understanding of the intricate molecular mechanisms underlying various biological processes by providing a comprehensive collection of data on functional protein interactions. A protein-protein interaction (PPI) network was constructed using STRING and the common DEGs identified. The PPI network was established with a confidence score exceeding 0.70, and the upper limit of interactions was set to 0. Disconnected nodes were then meticulously removed from the PPI network to refine its structure.

E. Hub Genes Identification

The PPI networks were visualized and analyzed using Cytoscape (version 3.10.2). The top 10 hub genes were determined using the Degree method, which is one of the centrality measures available through the cytoHubba plugin.

F. Validation of the Identified Hub Genes

1) Survival Analysis: The predictive relevance of the central genes in colorectal cancer was determined through Kaplan-Meier survival analysis, using KMPlot. For this analysis, individuals were categorized into two distinct groups based on low and high expression levels of the hub genes.

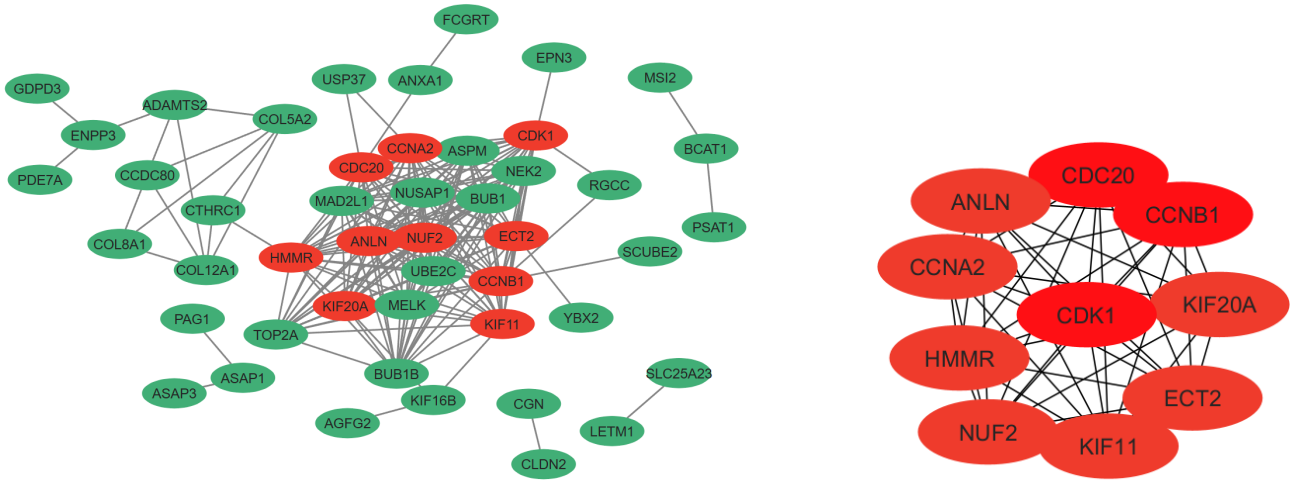
G. Drug-Gene Interaction Network Construction

We analyzed potential therapeutic drugs for these hub genes using the DGIdb database, focusing on FDA-approved drugs. Subsequently, the network was analyzed in Cytoscape.

III. Obtained Results

A. DEGs identification

Differentially expressed genes (DEGs) were selected using the limma package in R. From the GSE87473 dataset, we identified 2,149 DEGs, including 1,007 upregulated and 1,142 downregulated genes. Similarly, from the GSE21510 dataset, 5,760 DEGs were identified, comprising 3,606 upregulated and 2,154 downregulated genes.



(a) (b)

Fig. 2: Analysis of PPI Network and Hub Genes: (a) Visualization of the PPI Network in Cytoscape, (b) Top 10 hub genes identified relying on Degree

B. Identification of ML-based DEGs

We identified ML-based DEGs from the initial DEGs by applying SHapley Additive exPlanations (SHAP). From both datasets, we selected the top 200 DEGs as the most significant. Ultimately, we identified 95 common ML-based DEGs between the two datasets. The corresponding Venn diagram is shown in Fig. 1.

C. PPI Network Evaluation and Hub Genes Identification

The PPI network was constructed using the STRING database and analyzed with Cytoscape, comprising 47 nodes and 273 edges. The visualization of the PPI network is shown in Fig. 2. Using the Degree method, we identified the top 10 hub genes: CDC20, ANLN, HMMR, CCNB1, CDK1, KIF20A, ECT2, KIF11, NUF2, and CCNA2. These hub genes have potential as valuable biomarkers for studying and diagnosing UC and CRC.

D. Validation of the Identified Hub Genes

1) Survival Analysis: The Kaplan-Meier survival analysis demonstrated a correlation between alterations in ten pivotal genes and increased mortality rates, as well as reduced overall survival times for patients with CRC. Specifically, elevated expression levels of CDC20, ANLN, HMMR, CCNB1, CDK1, KIF20A, ECT2, KIF11, NUF2, and CCNA2 were associated with poorer survival outcomes compared to their lower expression levels. This relationship is graphically depicted in Fig. 3.

E. Drug-Gene Interaction Analysis

In the Drug-Gene interaction network, there are 74 nodes and 74 edges. As depicted in Fig. 4, we identified 70 potential drugs targeting four hub genes. This opens avenues for further analysis towards identifying potential therapeutic options for UC and CRC.

IV. Discussion

The primary objective of this research was to identify crucial biomarkers in UC and CRC. We conducted differential gene expression analysis on two microarray datasets to identify DEGs. Next, we applied SVM as a gene selection algorithm to identify ML-based DEGs from each dataset. Using the 95 common ML-based DEGs, we constructed a PPI network. Ten hub genes were identified from the PPI network using the Degree method in Cytoscape. These top 10 hub genes were critically evaluated for their roles in CRC through Kaplan-Meier survival analysis, suggesting their potential as prognostic biomarkers.

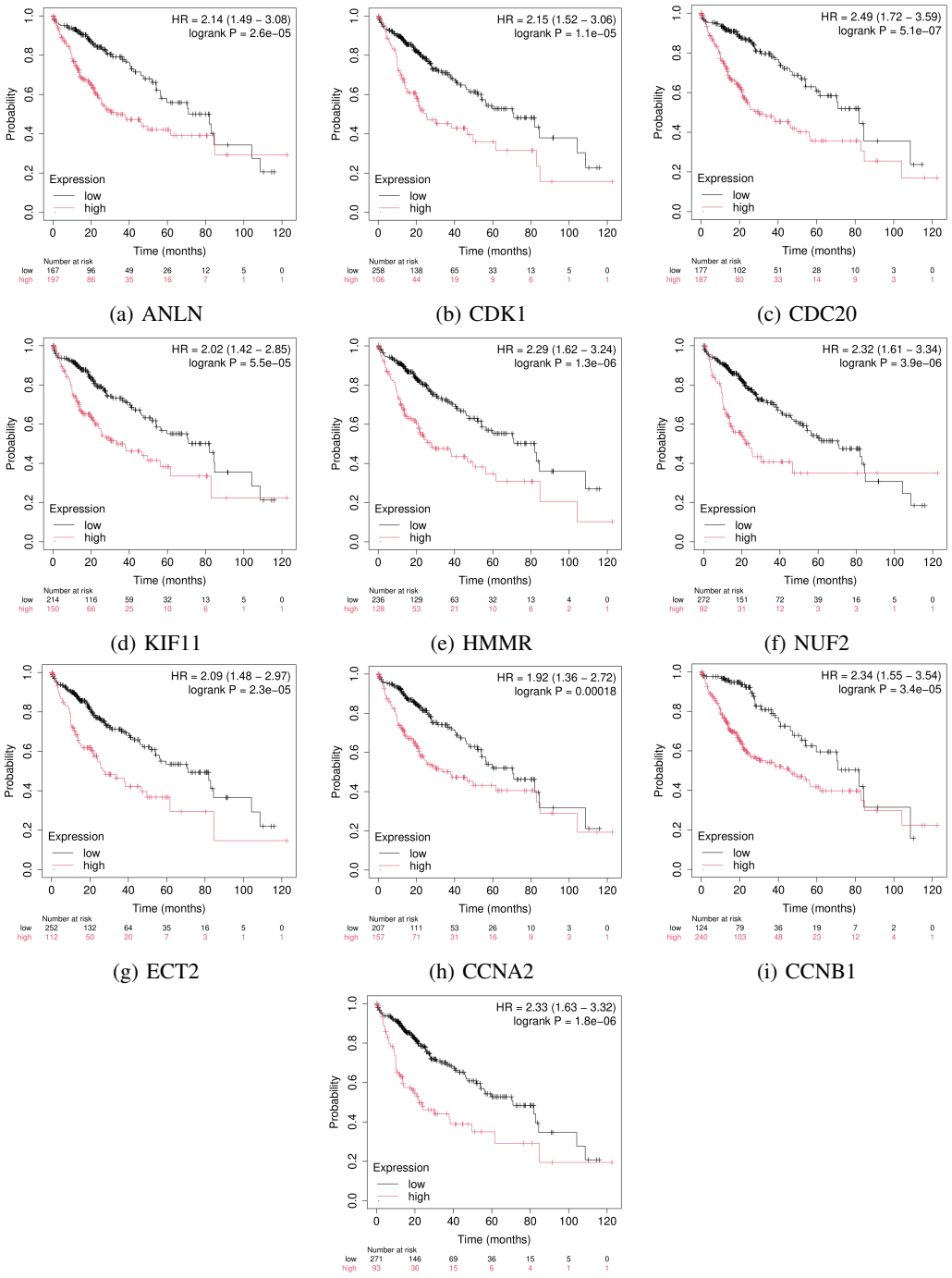
Cyclin-dependent Kinases (CDKs) are often elevated in cancerous tissues, with CDK1 showing increased levels in approximately 70.8% of reported tumors. CDK1 promotes tumor growth by interacting with proteins involved in various signaling pathways. Its overexpression has been observed in colorectal malignancies [10].

Cell Division Cycle Protein 20 (CDC20) initiates the activation of the anaphase-promoting complex and the E3 ubiquitin ligase complex at the onset of mitosis, crucial for ensuring proper progression through cell division. Its regulatory role in these complexes underscores its significance in maintaining genomic stability and cell cycle control. Its overexpression in various cancers is correlated with poor prognosis [7].

Kinesin Family Member 11 (KIF11) is significantly expressed in CRC tissues and is linked to patient prognosis, impacting the development of various cancers [11].

NUF2 exhibits elevated expression in cancerous tissues, including CRC. It also actively contributes to abnormal cell growth and death inhibition [12].

Cyclin A2 (CCNA2) acts as a tumor-promoting gene in various cancers, including UC and CRC, where its mRNA level is notably elevated.



(a) ANLN (b) CDK1 (c) CDC20

(d) KIF11 (e) HMMR (f) NUF2

(g) ECT2 (h) CCNA2 (i) CCNB1

(j) KIF20A

Fig. 3: Kaplan-Meier survival analysis of 10 hub genes. (a) ANLN, (b) CDK1, (c) CDC20, (d) KIF11, (e) HMMR, (f) NUF2, (g) ECT2, (h) CCNA2, (i) CCNB1, and (j) KIF20A

Cyclin B1 (CCNB1) overexpression has been identified as a critical gene across various cancer types. This gene ranked highly in the PPI network, strongly suggesting that CCNB1 can serve as a pivotal biomarker in multiple malignancies, including UC and CRC [13].

Anillin (ANLN) is an actin-binding protein crucial for cell division and plays a significant role in cell growth and migration. It is involved in the formation and stability of the contractile ring during cytokinesis, making it essential for successful cell division. Elevated levels of ANLN have been observed in various cancers, highlighting its potential as a biomarker for tumor progression and metastasis, including UC

and CRC [14].

Kinesin family member 20A (KIF20A) belongs to the kinesin family and is responsible for transporting chromosomes during mitosis, playing a critical role in cell division. It is essential for ensuring proper spindle formation and accurate alignment of chromosomes, which are vital processes for maintaining genomic stability and preventing aberrant cell division. Overexpression of KIF20A has been linked to several types of cancer, indicating its potential as a therapeutic target and a biomarker for cancer diagnosis and prognosis.

Hyaluronan-mediated motility receptor (HMMR) is a hyaluronic acid receptor that influences cell movement and regulates spindle formation during brain development. Elevated HMMR expression is consistently associated with a poor prognosis in several carcinomas, including breast, ovarian, and colorectal cancers. Its role in promoting cell motility and invasion makes HMMR a potential target for therapeutic interventions and a valuable biomarker for cancer prognosis.

ECT2 is an activator of RHO GTPases crucial for cytokinesis. Nuclear ECT2 promotes ribosomal DNA transcription and ribosome biogenesis in CRC.

These findings provide valuable insights into potential molecular targets for UC and CRC, serving as critical markers for early prognosis and diagnosis to improve timely and effective treatment strategies.

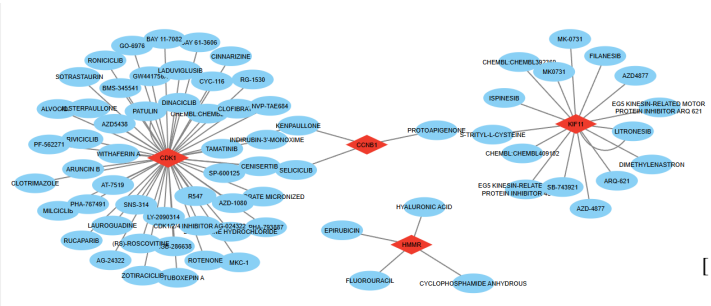


Fig. 4: Drug-Gene interaction network

V. Conclusion

The main objective of this study was to identify and validate pivotal biomarkers associated with ulcerative colitis (UC) and colorectal cancer (CRC) using bioinformatics and machine learning techniques. A protein-protein interaction (PPI) network was constructed using common ML-based DEGs from two microarray datasets, leading to the identification of 10 hub genes. While previous investigations primarily used bioinformatics techniques for hub gene identification, integrating machine learning in this study allowed for a more comprehensive identification of these genes. Kaplan-Meier survival analysis revealed a significant association between the overexpression of these hub genes and reduced survival rates in UC and CRC patients. These genes, known to be associated with UC, CRC, and other cancers, highlight their potential as biomarkers, offering avenues for advancements in early diagnosis, prognosis, and potentially therapeutic interventions for UC and CRC. Future research directions could include applying advanced machine-learning algorithms to strengthen the study's findings further. Additionally, integrating functional annotation methods could provide deeper insights into the molecular mechanisms of UC and CRC.

References

[1] L. Tan, X. Li, H. Qin, Q. Zhang, J. Wang, T. Chen, C. Zhang, X. Zhang, and Y. Tan, "Identified s100a9 as a target for diagnosis and treatment of ulcerative colitis by bioinformatics analysis," *Scientific Reports*, vol. 14, no. 1, p. 5517, 2024.

[2] S. Li, T. Li, Y.-Q. Shi, B.-J. Xu, Y.-Y. Deng, and X.-G. Sun, "Identification of hub genes with prognostic values in colorectal cancer by integrated bioinformatics analysis," *Cancer Biomarkers*, no. Preprint, pp. 1–19, 2024.

[3] W. Li, T. Zhao, D. Wu, J. Li, M. Wang, Y. Sun, and S. Hou, "Colorectal cancer in ulcerative colitis: mechanisms, surveillance and chemoprevention," *Current Oncology*, vol. 29, no. 9, pp. 6091–6114, 2022.

[4] M. Mortezapour, L. Tapak, F. Bahreini, R. Najafi, and S. Afshar, "Identification of key genes in colorectal cancer diagnosis by weighted gene co-expression network analysis," *Computers in Biology and Medicine*, vol. 157, p. 106779, 2023.

[5] W. Hu, T. Fang, M. Zhou, and X. Chen, "Identification of hub genes and immune infiltration in ulcerative colitis using bioinformatics," *Scientific Reports*, vol. 13, no. 1, p. 6039, 2023.

[6] Y. Yuan, N. Li, M. Fu, and M. Ye, "Identification of critical modules and biomarkers of ulcerative colitis by using wgcna," *Journal of inflammation research*, pp. 1611–1628, 2023.

[7] S. Sarker, M. A. M. Hasan, and A. Y. Srizon, "Identification and validation of the most prevalent biomarkers in hepatocellular carcinoma," in *2023 26th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2023, pp. 1–6.

[8] K. Li, R. Strauss, J. Ouahed, D. Chan, S. E. Telesco, D. S. Shouval, J. B. Canavan, C. Brodmerkel, S. B. Snapper, and J. R. Friedman, "Molecular comparison of adult and pediatric ulcerative colitis indicates broad similarity of molecular pathways in disease tissue," *Journal of Pediatric Gastroenterology and Nutrition*, vol. 67, no. 1, pp. 45–52, 2018.

[9] S. Tsukamoto, T. Ishikawa, S. Iida, M. Ishiguro, K. Mogushi, H. Mizushima, H. Uetake, H. Tanaka, and K. Sugihara, "Clinical significance of osteoprotegerin expression in human colorectal cancer," *Clinical cancer research*, vol. 17, no. 8, pp. 2444–2450, 2011.

[10] K. Zeng, W. Li, Y. Wang, Z. Zhang, L. Zhang, W. Zhang, Y. Xing, and C. Zhou, "Inhibition of cdk1 overcomes oxaliplatin resistance by regulating acsl4-mediated ferroptosis in colorectal cancer," *Advanced Science*, vol. 10, no. 25, p. 2301088, 2023.

[11] Y. Zhou, L. Yang, L. Xiong, K. Wang, X. Hou, Q. Li, F. Kong, X. Liu, and J. He, "Kif11 is upregulated in colorectal cancer and silencing of it impairs tumor growth and sensitizes colorectal cancer cells to oxaliplatin via p53/gsk3β signaling," *Journal of Cancer*, vol. 12, no. 12, p. 3741, 2021.

[12] R. Li, J. He, Z. Ni, J. Zhang, X. Chi, C. Kang, Z. Li, and X. Li, "Mining and exploration of rehabilitation nursing targets for colorectal cancer," *Aging (Albany NY)*, vol. 16, no. 8, p. 7022, 2024.

[13] P. Dai, L. Xiong, Y. Wei, X. Wei, X. Zhou, J. Zhao, and H. Tang, "A pancancer analysis of the oncogenic role of cyclin b1 (ccnb1) in human tumors," *Scientific Reports*, vol. 13, no. 1, p. 16226, 2023.

[14] Y. Shi, X. Ma, M. Wang, S. Lan, H. Jian, Y. Wang, Q. Wei, and F. Zhong, "Comprehensive analyses reveal the carcinogenic and immunological roles of anln in human cancers," *Cancer Cell International*, vol. 22, no. 1, p. 188, 2022.