Identifying Epigenetic Biomarkers in Colorectal Cancer: A Bioinformatics Analysis

Olajumoke Bisola Oladapo, Ferial Najiantabriz, Ujwala Vasireddy

Department of Computer Science, University of Oklahoma

These authors contributed equally to this work.

December 10, 2024

Abstract

Colorectal cancer (CRC) is a term that includes both colon and rectal cancers, which are treated as a single tumor type. About 72% of CRC cases are colon cancer, while 28% are rectal cancer. CRC is caused by both genetic and epigenetic changes in colon mucosal cells, affecting important genes like oncogenes, DNA repair genes, and tumor suppressor genes. Currently, only two DNA methylation-based biomarkers have received FDA approval: SEPT9 for blood tests and a combination of NDRG4 and BMP3 for stool tests. However, these biomarkers have limitations, especially for early detection and precancerous stages. To address this, we used a bioinformatics pipeline to find new DNA methylation-regulated genes (MRGs) related to CRC.

We applied Weighted Gene Co-expression Network Analysis (WGCNA) to find gene modules associated with rectal cancer conditions. We also identified differentially methylated CpG sites (DMCs) and differentially expressed genes (DEGs) using the Limma R package. By overlapping DMCs and DEGs, eight MRGs were identified: *LY6G6D*, *GNG7*, *HKDC1*, *AZGP1*, *ALG1L*, *PITX2*, *KCNQ1*, and *PDX1*. Functional analysis showed these genes are involved in pathways like Type II Diabetes Mellitus, Cholinergic Synapse, and biosynthesis of certain antibiotics. Gene Ontology analysis also revealed their role in glucose metabolism and cell differentiation. This study shows that combining WGCNA and epigenetic data can help discover new biomarkers for CRC and improve understanding of the disease.

1. Introduction

Colorectal cancer (CRC) includes both colon and rectal cancers, which are often treated as a single type of tumor. It is one of the most common cancers worldwide, ranking as the third leading type of cancer and the second highest in mortality. Approximately 72% of CRC cases are colon cancer, while 28% are rectal cancer. Early detection is crucial because patients diagnosed at early stages have a 90% survival rate over five years, compared to only 13% for those diagnosed at advanced stages with metastasis. Biomarkers, which are measurable biological indicators, play an important role in improving CRC diagnosis and treatment. However, very few biomarkers have been successfully used in clinical practice, which highlights the need for further research.

Epigenetics, which refers to changes in gene expression without altering the underlying DNA sequence, has become a key area for identifying CRC biomarkers. DNA methylation, a major type of epigenetic modification, involves adding a methyl group to CpG sites (regions where cytosine and guanine are adjacent). In many diseases, including CRC, abnormal DNA methylation patterns have been observed. These patterns can either suppress or increase gene expression, depending on whether the CpG sites are hypermethylated or hypomethylated. Such changes make DNA methylation a promising candidate for identifying biomarkers.

DNA methylation has already shown clinical potential, with two FDA-approved biomarkers for CRC. The SEPT9 gene is used in blood-based tests, while a combination of NDRG4 and BMP3 is used in stool-based tests. Despite this progress, the current biomarkers are limited in their ability to detect early-stage CRC and precancerous conditions. This limitation emphasizes the need to find more sensitive and specific biomarkers. Advances in bioinformatics methods, which combine biological data with computational tools, provide opportunities to address this challenge.

This study aims to identify novel biomarkers by integrating DNA methylation and gene expression data using a bioinformatics pipeline. Weighted Gene Co-expression Network Analysis (WGCNA) was used to group genes into modules based on their expression patterns. By correlating these

modules with rectal cancer traits, key gene clusters were identified. These modules, combined with differentially methylated CpG sites and differentially expressed genes, revealed methylation-regulated genes (MRGs) that could serve as biomarkers for CRC. This approach provides insights into the molecular mechanisms of CRC and offers a pathway for discovering new diagnostic tools.

2. Methods

2.1 Experimental Design

We aim to identify DNA methylation biomarkers through methylation-regulated genes in colorectal cancer (CRC) by following a bioinformatics pipeline adopted from Li et al. [15], who studied biomarkers in varicose vein disease. All analyses were carried out using R [16].

2.2 Data Collection

The datasets used in this study were retrieved from the GEO Database using the GEOquery package [17]. Specifically, the GSE75548 and GSE75546 datasets, which represent expression profiling by array and methylation profiling by genome tiling array, respectively, were collected from six tissue samples of patients with rectal cancer and paired normal tissues.

2.3 Identifying and Mapping Differentially Methylated CpG Sites

Differentially methylated CpG sites (DMCs) between normal tissues and rectal cancer samples were identified using the Limma package [18]. The results were considered statistically significant if $P < 0.05$ and $\log_2\text{FC} > 0.4$. Differentially methylated regions (DMRs) were identified between normal and cancer samples. A design matrix was constructed using metadata to group samples, and CpG sites were annotated using the DMRcate package [19] with an FDR threshold of 0.001. DMRs were classified as hypermethylated or hypomethylated based on mean methylation differences. Genomic coordinates were validated using the BSgenome.Hsapiens.UCSC.hg19 package [20], ensuring all regions were within standard chromosomes. A karyogram visualizing hypermethylated (red) and hypomethylated (blue) regions was generated using the karyoploteR package [21].

2.4 Identification of Differentially Expressed Genes (DEGs)

Differentially expressed genes (DEGs) were identified using the Limma package [18]. The results were considered statistically significant if $P < 0.05$ and $\log_2\text{FC} > 0.5$. Results were visualized using a volcano plot to highlight upregulated, downregulated, and non-significant genes.

2.5 Identification and Analysis of Methylation-Regulated Genes (MRGs)

Gene annotations were linked to genomic regions using the methylKit package [22]. A gene annotation BED file was utilized for transcript features to ensure accurate identification of overlapping and nearby genes. Gene symbols from the annotated DMRs were compared with significantly differentially expressed genes (DEGs). This integration identified common genes, referred to as methylation-regulated genes (MRGs), which showed both methylation alterations and differential expression patterns. The overlapping genes were visualized using a Venn diagram created with the VennDiagram package [23].

2.6 KEGG Pathway and Gene Ontology (GO) Enrichment Analysis

Entrez gene IDs for the identified MRGs were retrieved using the org.Hs.eg.db package [24]. These IDs were used for pathway and functional enrichment analysis using the clusterProfiler package [25]. KEGG pathway analysis was conducted to identify enriched biological pathways with significance determined by a P-value cutoff of 0.05, and results were visualized as dot plots. Similarly, Gene Ontology (GO) analysis focused on biological processes, identifying functional categories enriched in the MRGs, with results visualized in high-resolution dot plots.

2.7 WGCNA

We do WGCNA to find groups of genes that work together and are related to rectal cancer. It helps us organize genes into smaller groups, called modules, based on their behavior. This way, we can focus on important genes instead of looking at all genes at once.

WGCNA also helps us find key genes in each group, which might play a big role in the condition. By doing this, we can understand how the disease works and discover potential biomarkers or targets for treatment.

1. Results

Differentially Methylated CpG Sites (DMCs)

3.1 Differentially Methylated CpG Sites (DMCs)

In the figure 1 shows the differentially methylated CpG sites (DMCs) in a volcano plot. The plot highlights the hypomethylated CpG sites on the left wing in blue, the hypermethylated CpG sites on the right wing in red, and the ash-colored points represent CpG sites that do not significantly differ in methylation between cancer and normal tissues.

The boxplot in Figure illustrates the top ten DMCs in normal and cancer tissues. These data provide a focused view of the most significantly altered CpG sites. Figure presents a karyogram showing the distribution of differentially methylated regions (DMRs), with hypermethylated regions shown in red and hypomethylated regions in blue.

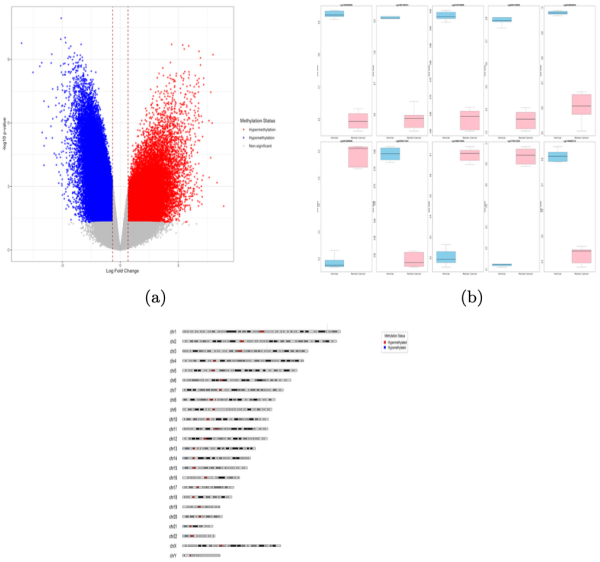

Figure 1: The DMCs from the GSE75546 dataset.(a) Volcano plot showing the DMCs. (b) Boxplot showing the top ten DMCs. (c) Karyogram showing the DMRs.

3.2 Analysis of Gene Expression Profiles

In the figure 2, panel (a) represents a heatmap depicting the clustering of samples based on their gene expression profiles. Panel (b) shows the results of Principal Component Analysis (PCA), where the scatter plot illustrates the separation of samples along the first two principal components (PC1 and PC2). Group-specific coloring of the points confirms the distinct global expression profiles of rectal cancer and normal samples.

Panel (c) is a volcano plot visualizing the differentially expressed genes (DEGs) between rectal cancer and normal tissues. Each point represents a gene, with the x-axis showing the $\log_2$ fold change ($\log_2$FC) and the y-axis representing the -$\log_{10}$ of the p-value. Genes with significant upregulation (red points) and downregulation (blue points) are highlighted. Panel (d) displays a heatmap of the top DEGs, where each row represents a gene and each column represents a sample.

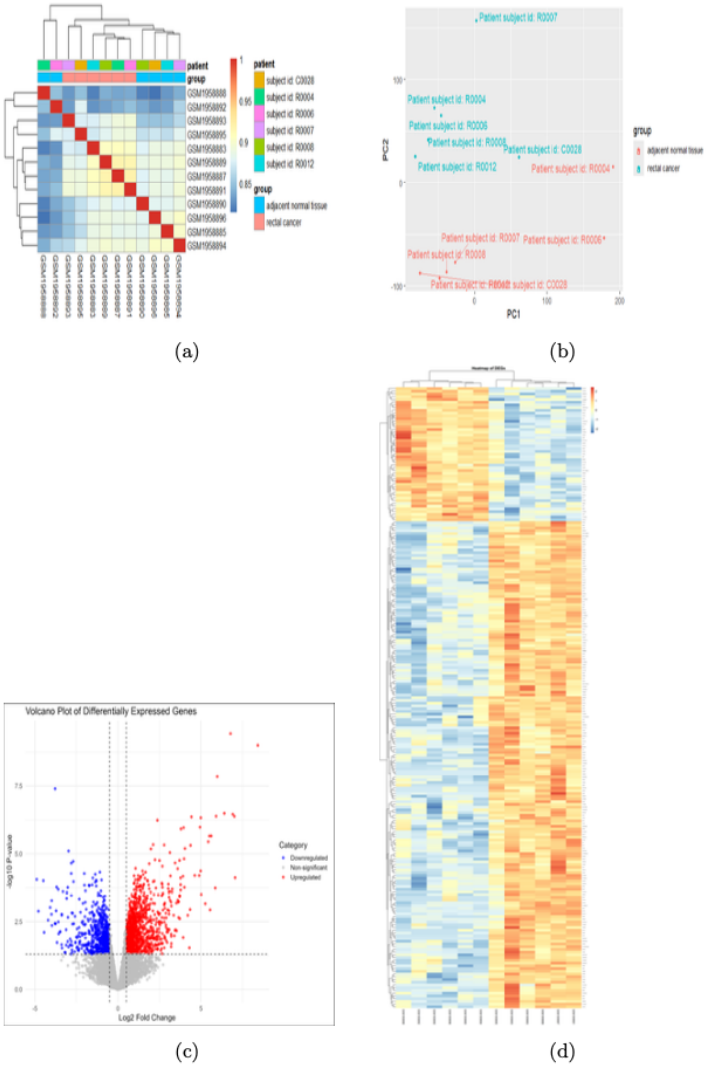

Figure 2: Heatmap showing expression profiles of samples. (a) Heatmap showing expression profiles of samples. (b) Principal Component Analysis (PCA) of Gene Expression Profiles. (c) Volcano plot of DEGs. (d) Heatmap showing DEGs among samples.

3.3 Methylation Regulated Genes (MRGs)

In the figure 3 we can see that, Out of the 776 differentially methylated genes, eight were found to overlap with the 120 differentially expressed genes . These eight overlapping genes are considered MRGs and represent candidates for further functional and pathway analysis. The identified MRGs, including their $\log_2$ fold change, average expression, and adjusted p-values, are summarized in the following table.

Figure 3: Venn Diagram Showing MRGs.

3.4 KEGG Pathway and Gene Ontology

Figure 4 and 5 illustrates the KEGG pathway enrichment analysis, showing the top enriched pathways associated with methylation-regulated genes (MRGs). These pathways represent significant biological processes and systems related to disease and metabolism, offering insights into the functional roles of the identified genes.

Figure 5 displays the Gene Ontology (GO) enrichment analysis results, highlighting the top biological processes associated with MRGs. These processes provide valuable information about the cellular mechanisms influenced by these genes, including their roles in metabolism, cellular differentiation, and signaling pathways.

3.5 Soft-Thresholding Power Selection in WGCNA

In the figure 6 we can see that The soft-thresholding plots help us choose the best power for creating a network that matches how genes work in real life.

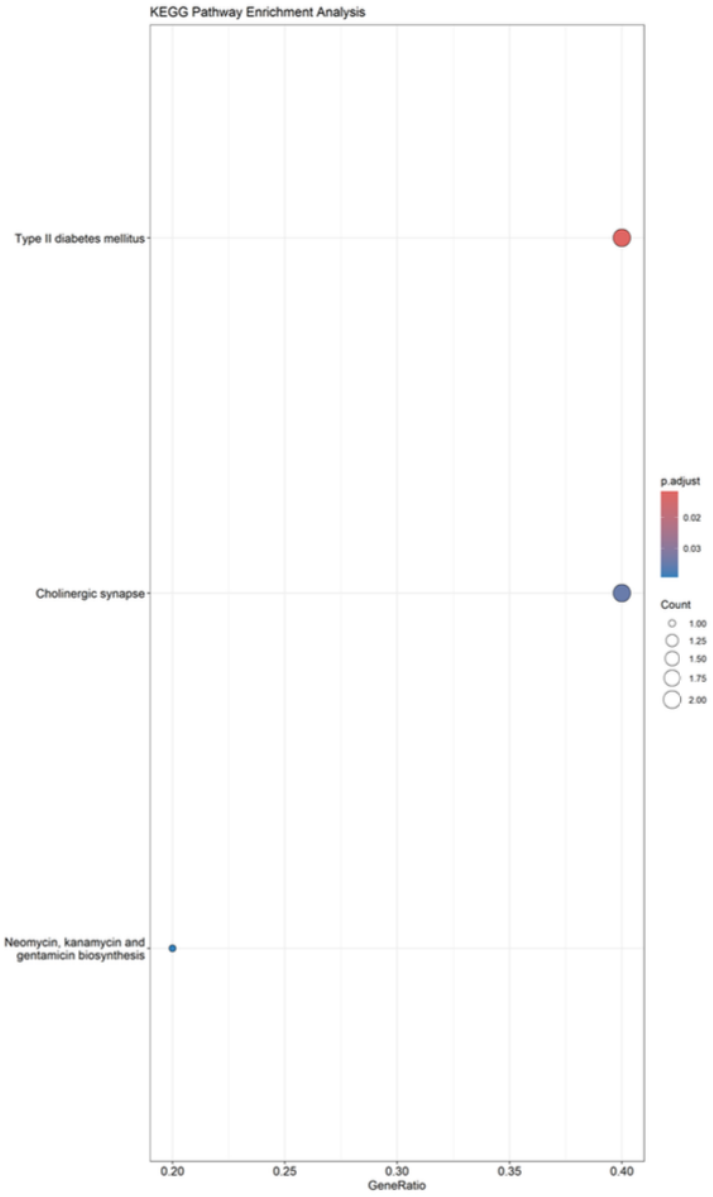

Figure 4: MRGS and associated KEGG pathway

In the *Scale Independence* plot, we check how well the network follows a scale-free structure by looking at the R^2 value. A network is scale-free if R^2 is above 0.8. In our plot, R^2 reaches 0.8 at power 6, so we picked 6 as the best power.

The *Mean Connectivity* plot shows how the connections between genes change as the power increases. When the power goes up, the average connections become fewer, which is expected. This ensures the network isn't too crowded and focuses on the strongest relationships between genes. Picking the right power helps make the network biologically meaningful and reliable for the next steps.

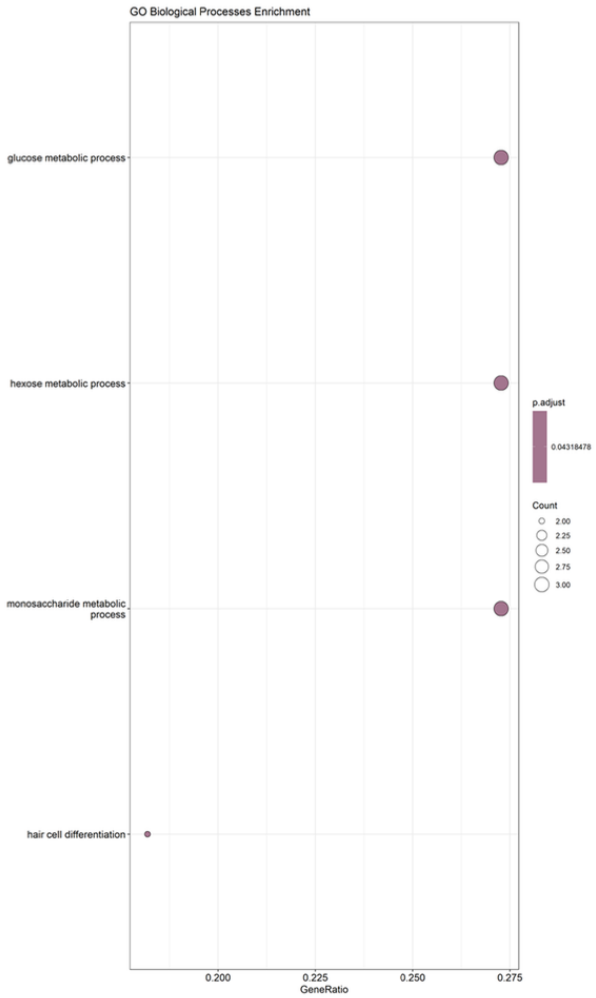

Figure 5: MRGS and associated functional biological process in Gene Ontology.

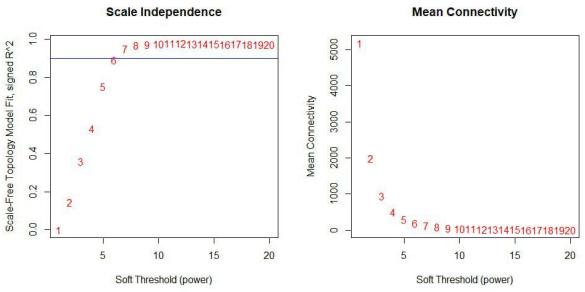

Figure 6: Scale Independence and Mean Connectivity Plots.

3.6 Hierarchical Clustering Dendrogram

The figure 7 shows a hierarchical clustering dendrogram of the samples, which helps to identify if there are any outliers. Each branch on the tree represents a sample, and the height scale on the left side indicates how similar or different the samples are from each other. Samples that cluster together at lower heights are more alike, while those that only join at a higher height are more different.

By examining this clustering, we can quickly see if there are any samples that behave very differently from the rest. These outlier samples could affect the analysis and may need to be removed or checked before continuing. In this case, we have a clear structure of samples grouped together, which is a good sign that most samples are consistent with each other.

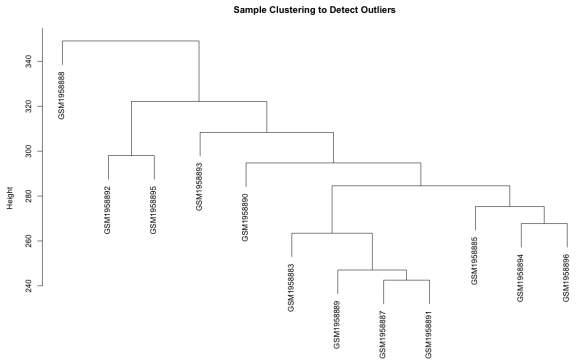

Figure 7: Sample Clustering

3.7 Identifying gene modules through Cluster dendogram Analysis

figure 8 shows a cluster dendrogram of all the genes we analyzed, where each branch corresponds to a group of genes that have similar expression patterns. By arranging genes according to their similarity, we can identify modules—groups of genes that might be working together or regulated by similar mechanisms. The colors at the bottom of the plot come from the "Dynamic Tree Cut" method, which automatically finds these modules without requiring us to set a fixed number of clusters in advance.

Performing this step is important because it helps us simplify a large, complex dataset into a manageable set of modules. Instead of looking at thousands of genes individually, we can focus on a smaller number of gene clusters. Understanding these modules can give us insights into the underlying biological processes, such as pathways involved in disease progression or in the body's response to treatment. This approach ultimately helps us identify key targets for further study.

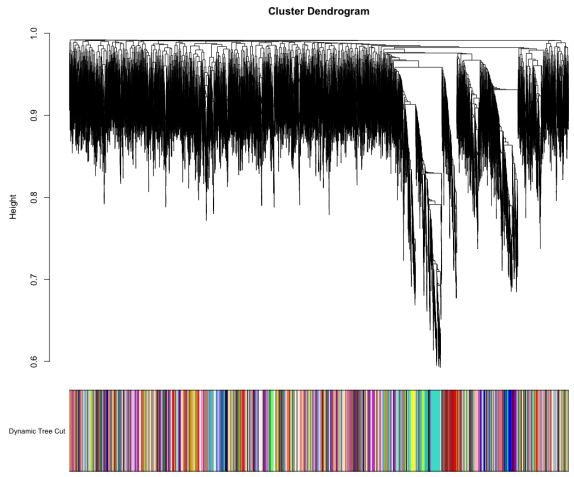

Figure 8: Identifying Gene Modules

3.8 Identifying Key Gene Modules Associated with Rectal Cancer

figure 9 heatmap shows the relationship between different groups of genes and rectal cancer. Each module is labeled ('ME19', 'ME13', etc.) and is listed along the left side of the chart. The colors and numbers in the heatmap show how strongly each group of genes is linked to rectal cancer. A positive number means the module is directly related to rectal cancer, while a negative number means the module is inversely related. For example, one module might have a strong positive relationship , while another might have a negative relationship. In the middle are p-values.

By analyzing this heatmap, researchers can identify which modules have the strongest and most significant relationships with rectal cancer. These important modules may help pinpoint genes or pathways involved in the disease. For instance, a module with a high positive correlation and a very small p-value could contain genes that play a key role in cancer development. Understanding these connections can guide further research to find biomarkers or develop new treatments for rectal cancer.

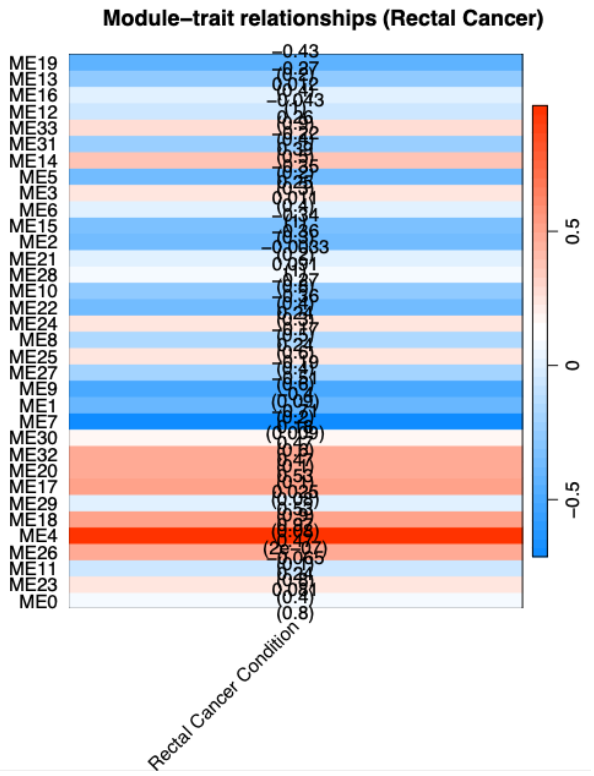

Figure 9: Heat Map

4. Conclusion

Colorectal cancer (CRC) develops through the transformation of normal colon and rectal epithelial cells into precancerous lesions and advanced carcinoma, which can metastasize to other organs. This study investigated the methylation and expression profiles of rectal cancer samples to identify potential biomarkers. From a total of 384,933 CpG sites analyzed, 98,209 were found to be differentially methylated, with 30,613 hypermethylated and 67,596 hypomethylated sites. Genomic annotation revealed that the majority of differentially methylated regions (DMRs) were located in intergenic regions, suggesting their involvement in distal regulatory functions. Smaller fractions of DMRs were found in genic regions, indicating potential roles in transcription regulation, initiation, and splicing.

Through integration of differential methylation and expression analysis, we identified eight methylation-regulated genes (MRGs): *LY6G6D*, *GNG7*, *HKDC1*, *AZGP1*, *ALG1L*, *PITX2*, *KCNQ1*, and *PDX1*. These genes have been previously associated with various cancers, and some have specific links to CRC. For example, *LY6G6D* was found to be upregulated in CRC and has been identified as a tumor-associated antigen, while *GNG7* was downregulated, consistent with its tumor-suppressive role. Similarly, genes like *AZGP1*, *HKDC1*, and *PITX2* have demonstrated roles in tumor progression and aggressiveness, highlighting their potential as diagnostic and therapeutic targets for CRC.

This study provides a comprehensive analysis of DNA methylation and gene expression patterns in CRC, identifying potential biomarkers for diagnosis and prognosis. The findings emphasize the importance of DNA methylation in CRC and its influence on gene regulation. However, limitations include

the lack of experimental validation and additional computational analyses. Future research should focus on validating these findings through laboratory experiments and exploring their clinical utility as biomarkers for CRC. This work lays a foundation for further studies aimed at improving the diagnosis and management of colorectal cancer.

Acknowledgments

General

We appreciate and acknowledge the guidance of our professor, Dr. Moussa, throughout this project and the Bioinformatics class members for their useful reviews during the project proposal stage.

Author Contributions

Oladapo, O. conceptualized the idea for this project, designed the experimental approach, worked on annotating differentially methylated regions and generating karyogram plots, and performed KEGG pathway and Gene Ontology analyses. Oladapo, O. also drafted the manuscript for this project.

NajianTabri, F. worked on data collection and identification of methylation-regulated genes and performed KEGG pathway. NajianTabri, F. also contributed to weighted gene co-expression network analysis (WGCNA) to identify modules associated with CRC.also drafted the manuscript for this project.

Vasireddy, U. worked on the identification of differentially methylated CpG sites and differentially expressed genes, as well as generating respective plots for these analyses.

Funding

The authors acknowledge that they did not receive funding for this work.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Data Availability

The datasets used in this study are publicly available at the GEO Database.

Supplementary Materials

All tables generated and codes used for this analysis are provided in the supplementary materials.

References

[1] Alzahrani SM, Al Doghaither HA, and Al-Ghafari AB. General insight into cancer: An overview of colorectal cancer. *Molecular and Clinical Oncology* 2021; 15:271.

[2] Hoang T, Kim H, and Kim J. Dietary intake in association with all-cause mortality and colorectal cancer mortality among colorectal cancer survivors: A systematic review and meta-analysis of prospective studies. *Cancers* 2020; 12:3391.

[3] Housini M, Dariya B, Ahmed N, et al. Colorectal cancer: Genetic alterations, novel biomarkers, current therapeutic strategies, and clinical trials. *Gene* 2024; 892:147857.

[4] Ogunwobi OO, Mahmood F, and Akingboye A. Biomarkers in colorectal cancer: Current research and future prospects. *International Journal of Molecular Sciences* 2020; 21:5311.

[5] Zygulska AL and Pierzchalski P. Novel diagnostic biomarkers in colorectal cancer. *International Journal of Molecular Sciences* 2022; 23:852.

[6] Davalos V and Esteller M. Cancer epigenetics in clinical practice. *CA: A Cancer Journal for Clinicians* 2023; 73:376–424.

[7] Xu W, Wang F, Yu Z, and Xin F. Epigenetics and cellular metabolism. *Genetics & Epigenetics* 2016; 8:GEG–S32160.

[8] Scala G, Federico A, Palumbo D, Cocozza S, and Greco D. DNA sequence context as a marker of CpG methylation instability in normal and cancer tissues. *Scientific Reports* 2020; 10:1721.

[9] Chen C, Wang Z, Ding Y, et al. DNA methylation: From cancer biology to clinical perspectives. *Frontiers in Bioscience-Landmark* 2022; 27:326.

[10] Wang Y, Wang C, Zhong R, Wang L, and Sun L. Research progress of DNA methylation in colorectal cancer. *Molecular Medicine Reports* 2024; 30:154.

[11] Caruso FP, D'Andrea MR, Coppola L, et al. Lymphocyte antigen 6G6D-mediated modulation through p38 MAPK and DNA methylation in colorectal cancer. *Cancer Cell International* 2022; 22:253.

[12] Naqvi M, Abbasi WA, Samma MK, et al. Decoding LY6G6D in colorectal cancer: Unraveling biomarker potential and therapeutic insights. *Cellular and Molecular Biology* 2024; 70:14–20.

[13] Gu S, Guo G, Yee H, et al. QL335, a novel T cell engager with enhanced safety profile targeting MSS colorectal cancer. *Cancer Research* 2024; 84:6727–7.

[14] Sanvicente Garcia A, Pedregal M, Paniagua-Herranz L, et al. Clinical and Immunologic Characteristics of Colorectal Cancer Tumors Expressing LY6G6D. *International Journal of Molecular Sciences* 2024; 25:5345.

[15] Wei Q, Miao T, Zhang P, Jiang B, and Yan H. Comprehensive analysis to identify GNG7 as a prognostic biomarker in lung adenocarcinoma correlating with immune infiltrates. *Frontiers in Genetics* 2022; 13:984575.

[16] Irwin DM and Tan H. Molecular evolution of the vertebrate hexokinase gene family: Identification of a conserved fifth vertebrate hexokinase gene. *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics* 2008; 3:96–107.

[17] Wang X, Shi B, Zhao Y, et al. HKDC1 promotes the tumorigenesis and glycolysis in lung adenocarcinoma via regulating AMPK/mTOR signaling pathway. *Cancer Cell International* 2020; 20:1–12.

[18] Pang Q, Huang S, Wang H, and Cao J. HKDC1 promotes autophagy and proliferation in pancreatic adenocarcinoma through interaction with PARP1 and poly (ADP-ribosyl) ation. *Cellular Signalling* 2024; 124:111474.

[19] Liu P, Luo Y, Wu H, et al. HKDC1 functions as a glucose sensor and promotes metabolic adaptation and cancer growth via interaction with PHB2. *Cell Death & Differentiation* 2024:1–16.

[20] Lian H, Wang A, Shen Y, et al. Identification of novel alternative splicing isoform biomarkers and their association with overall survival in colorectal cancer. *BMC Gastroenterology* 2020; 20:1–12.