Identifying Epigenetic Biomarkers in Colorectal Cancer: A Bioinformatics Analysis

Olajumoke Bisola Oladapo1, Ferial NajianTabri2, and Ujwala Vasireddy2
1Department of Biomedical Engineering, University of Oklahoma.
2Department of Computer Science, University of Oklahoma.
December 6th, 2024

Abstract

Colorectal cancer (CRC) is a term that refers to the combination of colon and rectal cancer as they are being treated as a single tumor. In CRC, 72\% of tumors are colon cancer while the other 28\% represent rectal cancer. CRC is a multifactorial disease caused by both genetic and epigenetic changes in the colon mucosal cells, affecting the oncogenes, DNA repair genes, and tumor suppressor genes. Currently, two DNA methylation based biomarkers for CRC have received FDA approval: SEPT9, used in blood-based screening tests, and a combination of NDRG4 and BMP3 for stool-based tests. Although DNA methylation biomarkers have been explored in colorectal cancer (CRC), the identification of robust and clinically valuable biomarkers remains a challenge, particularly for early-stage detection and precancerous lesions. Patients often receive diagnoses at the locally advanced stage, which limits the potential utility of current biomarkers in clinical settings. This study aims to address this gap by employing a bioinformatics pipeline to identify novel DNA methylation-regulated genes associated with CRC. Datasets used in this study were retrieved from the Gene Expression Omnibus (GEO) database. The Limma R package was used to identify differentially methylated CPG sites (DMCs) and identify differentially expressed genes (DEGs). From the overlap of DMCs with DEGs in rectal cancer, we identified seven MRGs (GNG7, ZCCHC14, HKDC1, AZGP1, ALG1L, PITX2, and PDX1) as biomarkers for CRC. The KEGG pathway enrichment analysis revealed the involvement of these methylation-regulated genes in two distinct pathways, namely Type II Diabetes Mellitus, and the biosynthesis of Neomycin, Kanamycin, and Gentamicin, which are related to human disease, and metabolism, respectively. Also, through gene ontology, we identified that the MRGs were involved in in endocrine system development, intracellular glucose homeostasis, glucose metabolic process, metabolic process, and some other biological

Methods

Experimental Design

We aim to identify DNA methylation biomarkers through methylation-regulated genes in colorectal cancer (CRC) following a bioinformatics pipeline adopted from Li et al [1] study on identification of biomarkers in varicose vein disease. All analysis were carried out on R [16].

Data Collection

The datasets used in this study was retrieved from GEO Database using the GEOquery package[2]. Specifically, the GSE75548 and GSE75546 which represent Expression profiling by array and Methylation profiling by genome tiling array datasets respectively collected from six tissue samples of patients with rectal cancer with paired normal tissues.

Identifying and Mapping Differentially Methylated CpG Sites

Differentially Methylated CpG sites (DMCs) between the normal tissues and rectal cancer tissue sample were identified using the Limma package[3]. The results were considered statistically significant if $P < 0.05$. and $\log2\ FC > 0.4$. Differentially methylated regions (DMRs) were identified between normal and rectal cancer samples. A design matrix was constructed using metadata to group samples, and CpG sites were annotated using the DMRcate package [4] with a false discovery rate (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[5], ensuring all regions were within standard chromosomes. A karyogram visualizing hypermethylated (red) and hypomethylated (blue) regions was generated using karyoploteR [6].

Identification of Differentially Expressed Genes (DEGs)

The differentially expressed genes (DEGs) were identified using the Limma package[3]. The results were considered statistically significant if $P < 0.05$. and $\log2\ FC > 0.5$. Results were visualized using a volcano plot to highlight upregulated, downregulated, and non-significant genes.

Identification and Analysis of Methylation-Regulated Genes (MRGs)

Gene annotations were linked to genomic regions using the methylKit package[7]. A gene annotation BED file was utilized for transcript features, ensuring accurate identification of overlapping100and nearby genes. Gene symbols from the annotated DMRs were compared with significantly differentially expressed genes (DEGs). This integration identified common genes: methylation-regulated genes (MRGs) that showed both methylation alterations and differential expression patterns. The overlapping genes were visualized using a Venn diagram created with the VennDiagram package[8].

KEGG Pathway and Gene Ontology (GO) Enrichment Analysis

Entrez gene IDs for the identified MRGs were retrieved using the org.Hs.eg.db package [9]. These IDs were subjected to pathway and functional enrichment analysis using the clusterProfiler package[10]. KEGG pathway analysis for enrichment of biological pathways was performed with significance determined by a p-value cutoff of 0.05, results were visualized as dot plots. Similarly, Gene Ontology analysis focused on biological processes, identifying functional categories enriched in the MRGs, results were visualized in high-resolution dot plots.

Results- Differentially Methylated CpG sites

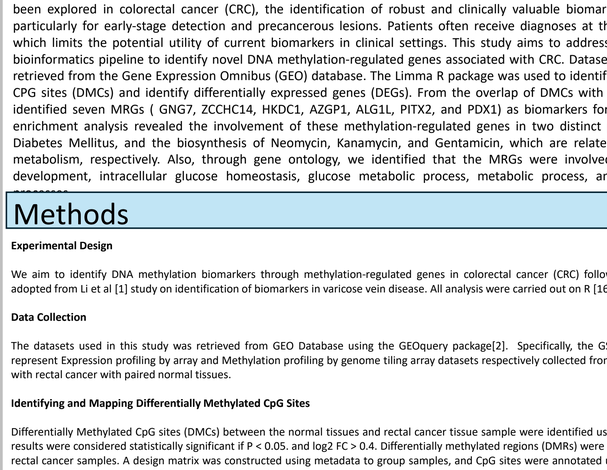


Figure 1: The DMCs from the GSE75546 dataset.(a) Volcano plot showing the DMCs. (b)Boxplot showing the top ten DMCs. (c) Karyogram showing the DMRs

Results- Methylation Regulated Genes

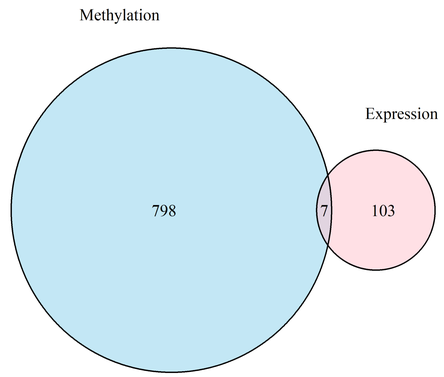


Figure 2: Venn Diagram showing MRGs

Table 1: Methylation Regulated Genes

Gene_ID	logFC	AveExpr	t	P.Value	adj.P.Val	B	Category	Symbol
ILMN_1728107	-2.02633	4.841524	-5.951859	5.200e-05	0.044556	1.721789	Downregulated	GNG7
ILMN_1743456	0.965532	6.998949	6.197013	6.469e-05	0.038755	2.035505	Upregulated	ZCCHC14
ILMN_1752502	3.600531	5.908766	6.044561	8.045e-05	0.041503	1.841604	Upregulated	HKDC1
ILMN_1797154	3.602733	2.103766	10.73252	3.337e-07	0.002552	6.294998	Upregulated	AZGP1
ILMN_2131293	3.058635	1.815741	7.235633	1.589e-05	0.020285	3.256656	Upregulated	ALG1L
ILMN_2391400	3.707355	2.942344	6.502739	4.217e-05	0.033764	2.412818	Upregulated	PITX2
ILMN_3249216	4.373601	3.777354	7.369687	1.338e-05	0.018719	3.402094	Upregulated	PDX1

Results- Differentially Expressed Genes (DEGs)

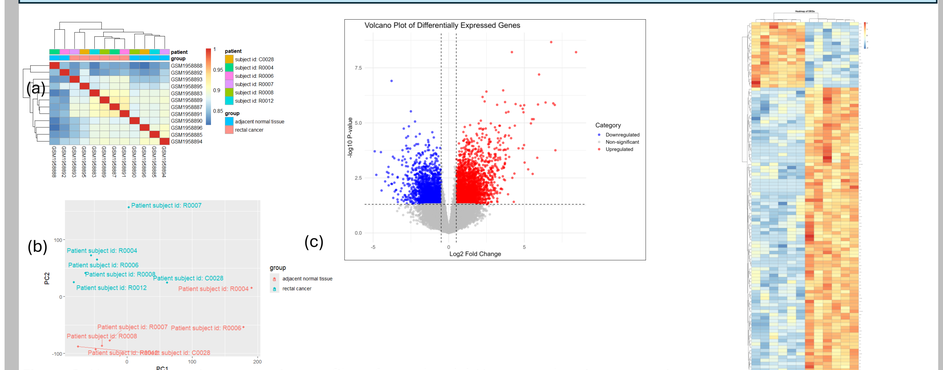
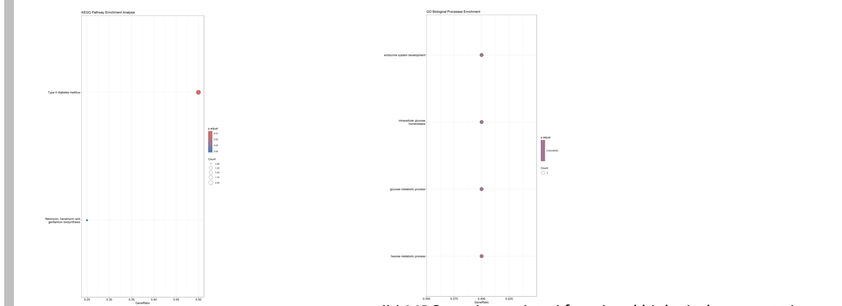


Figure 3: Heatmap showing expression profiles of samples. (a) Heatmap showing expression profilesof samples. (b) Principal Component Analysis (PCA) of Gene Expression Profiles. (c) Volcano plotof DEGs. (d) Heatmap showing DEGs among samples.

Results- KEGG Pathway and Gene Ontology



(a) MRGs and associated KEGG pathways

(b) MRGs and associated functional biological processes in Gene Ontology

Figure 4: KEGG pathways and GO biological processes associated with MRGs.

Discussion

Figure 1 shows the differentially methylated CpG sites analysis plots. A total of 384,933 CpG sites were screened with 98,209 differentially methylated CpG sites, of which 30,613 were hypermethylated and 67,596 were hypomethylated. DMCs were annotated to assess their genomic distribution relative to genic regions. Among the 874 differentially methylated regions (DMRs), the majority (95.77%) were located in intergenic regions, with smaller proportions overlapping with introns (3.89%), exons (0.80%), and promoters (0.57%). When prioritizing annotations (promoter > exon > intron), the overlap percentages were consistent, with 95.77% in intergenic regions and slight redistribution among genic parts. There were 0.40% of promoter boundaries, 0.07% of exon boundaries, and 0.31% of intron boundaries overlapped with DMRs. The distances to the nearest transcription start site (TSS) showed a median distance of 27,990 bp and a mean distance of 68,348 bp, with values ranging from 0 to 565,929 bp. These results indicate that the majority of DMRs are in intergenic regions, suggesting their potential involvement in distal regulatory functions. The smaller fraction of DMRs overlapping with promoters, exons, and introns points to possible roles in gene regulation, transcription initiation, and splicing.

Figure 2 shows the differentially expressed genes analysis plots. A total of 4,414 differentially expressed genes (DEGs) were identified, of which 2,518 were upregulated and 1,896 were downregulated. We compared the gene symbols from the annotated DMRs with the significantly downregulated DEGs which were filtered to 110 based on an adjusted P-value of 0.05. This integration identified common genes regarded as methylation-regulated. We identified genes (MRGs) that exhibited both methylation alterations and differential expression patterns. Our analysis identifies these genes, GNG7, ZCCHC14, HKDC1, AZGP1, ALG1L, PITX2, and PDX1, as methylation-regulated genes in Figure 2 above.

Our KEGG pathway enrichment analysis uncovered the involvement of these methylation-regulated genes in two distinct pathways: Type II Diabetes Mellitus, and the biosynthesis of Neomycin, Kanamycin, and Gentamicin. These pathways are associated with human disease, and metabolism, respectively From Gene Ontology, we identified that the MRGs were involved in endocrine system development, intracellular glucose homeostasis, glucose metabolic proce, metabolic process, hexose metabolic process, stem cell differentiation, monosaccharide metabolic process, glucose homeostasis, carbohydrate homeostasis, cardiac neural crest cell migration involved in outflow tract morphogenesis, carbohydrate mediated signaling, type B pancreatic cell apoptotic process, receptor guanylyl cyclase signaling pathway, cardiac neural crest cell, development involved in outflow tract morphogenesis, regulation of type B, pancreatic cell proliferation, left/right axis specification, response to leucine, cardiac neural crest cell differentiation involved in heart development, cardiac neural crest cell development involved in heart development, adenohypophysis development, negative regulation of endoplasmic, reticulum stress-induced intrinsic apoptotic signaling pathway, gland development, and cell migration involved in heart development.

Conclusion

We identified seven genes as methylation-regulated genes through a comprehensive bioinformatics analysis, suggesting that methylation affects their expression levels. These genes have been associated with a variety of tumors in literature studies, with some specifically being linked to colorectal cancer (CRC). We suggest that these genes could serve as biomarkers for CRC, and further wet lab procedures are needed to validate their functions as biomarkers for CRC. Limitations of this study include the inability to carry out further computational validation procedures and wet lab procedures, and we look forward to continuing that as a future direction for this project.

References

1. Li S, Liu Y, Liu M, Wang L, and Li X. Comprehensive bioinformatics analysis reveals biomarkers of DNA methylation-related genes in varicose veins. Frontiers in Genetics 2022;13:1013803.
2. Davis S and Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO)and BioConductor. Bioinformatics 2007;23:1846–7.
3. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research 2015;43:e47–e47.
4. Peters TJ, Buckley MJ, Statham AL, et al. De novo identification of differentially methylated293regions in the human genome. Epigenetics & chromatin 2015;8:1–16.
5. Team T. BSgenome. hsapiens. UCSC. hg19: Full genome sequences for homo sapiens (UCSC version hg19, based on GRCh37. p 13) 2020.
6. Gel B and Serra E. karyoploteR: an R/Bioconductor package to plot customizable genomes297displaying arbitrary data. Bioinformatics 2017;33:3088–90.
7. Akalin A, Kormaksson M, Li S, et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome biology 2012;13:1–9.
8. Chen H and Boutros PC. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC bioinformatics 2011;12:1–7.
9. Carlson M, Falcon S, Pages H, Li N, et al. org. Hs. eg. db: Genome wide annotation for Human. R package version 2019;3:3.
10. Yu G, Wang LG, Han Y, and He QY. clusterProfiler: an R package for comparing biological305themes among gene clusters. Omics: a journal of integrative biology 2012;16:284–7.