Gene Expression Profiling by RNA-Seq Reveals Prognostic Signature in Colorectal Cancer

Authors: Rihab Mahjoub, Raafat Abdulmajeed

Abstract:
Colorectal cancer (CRC) remains a leading cause of cancer-related mortality worldwide. Early detection of molecular alterations through transcriptome profiling may uncover potential diagnostic and prognostic biomarkers. In this study, we reanalyzed RNA-seq data from four publicly available samples (GSE50760) and adjacent normal tissue samples using a modular and reproducible pipeline that integrates Bash scripting and R-based analysis. After quality control, transcript quantification was performed using Kallisto, and differential expression analysis was conducted with DESeq2. We identified several key differentially expressed genes (DEGs), including downregulated tumor suppressors such as ***TMIGD1*****,** ***AQP8*****, and** ***SLC26A3***. Functional enrichment analyses using Gene Ontology (GO) and Gene Set Enrichment Analysis (GSEA) revealed significant disruptions in **immune-related pathways, cell cycle progression,** and **metabolic regulation**. This analysis demonstrates a streamlined RNA-seq workflow that can be extended to other cancer types. All scripts, results, and visualizations are openly available for reuse and adaptation.

Keywords: Transcriptomics, RNA sequencing, gene expression profiling, cancer biomarkers, colorectal cancer, differential gene expression, functional enrichment

Introduction

Colorectal cancer (CRC) ranks among the top three most commonly diagnosed cancers worldwide and remains a major cause of cancer-related mortality, especially in developed countries. Its high lethality is often linked to late-stage detection, local invasion, and distant metastasis. Despite advances in diagnostic imaging and treatment strategies, the five-year survival rate remains low in advanced stages. Therefore, there is a growing need for reliable molecular markers that can facilitate early diagnosis, stratify patients, and guide therapeutic decisions.

In recent years, RNA sequencing (RNA-seq) has emerged as a transformative tool in transcriptome research. Unlike microarrays, RNA-seq offers higher resolution, dynamic range, and the ability to detect novel transcripts, alternative splicing, and allele-specific expression. Several large-scale studies have employed RNA-seq to dissect the molecular landscape of colorectal cancer. For instance, the TCGA consortium profiled hundreds of CRC samples to uncover somatic mutations and transcriptional subtypes. Other works (e.g., Yu et al., 2020; Wang et al., 2018) have identified gene expression signatures correlated with clinical outcomes, often integrating RNA-seq with clinical and genomic data. However, reproducibility and accessibility of such pipelines remain a challenge for smaller laboratories or educational contexts.

In this study, we reanalyzed a publicly available RNA-seq dataset (GSE50760) comprising primary colorectal tumor tissues and matched normal samples. Our goal is not only to identify differentially expressed genes (DEGs) and deregulated pathways in CRC, but also to demonstrate the effectiveness of a reproducible analysis pipeline built using open-source tools in R and Bash. This educational yet rigorous approach serves as both a learning resource and a methodological benchmark for similar studies.

2. Objective

To extract biologically meaningful insights from colorectal cancer RNA-seq data by identifying differentially expressed genes and enriched biological pathways distinguishing colorectal tumor tissues from normal colon using an open-source, modular bioinformatics pipeline.

3. Methods

Methodology: overview

In this study, we implemented a comprehensive RNA-seq analysis pipeline to identify differentially expressed genes and interpret their functional relevance. Raw RNA-seq reads in FASTQ format were obtained from the European Nucleotide Archive (ENA). Initial quality assessment was performed using **FastQC** and aggregated with **MultiQC**, followed by adapter trimming and quality filtering using **fastp**, retaining reads with a Phred score above 30. Transcript abundance quantification was conducted using **Kallisto**, aligned against the Ensembl **GRCh38 transcriptome reference**, to generate transcript-level abundance estimates. These quantifications were imported into **R** using the tximport package and subsequently summarized at the gene level. Normalization and differential gene expression analysis were carried out with **DESeq2**, applying shrinkage estimation and robust statistical modeling. To explore the biological significance of the differentially expressed genes (DEGs), **functional enrichment analyses** were performed using **clusterProfiler** for KEGG and GSEA, and **gprofiler2** for Gene Ontology and Reactome annotations. Visualizations were generated using **ggplot2**, **pheatmap**, and enrichment plotting tools, enabling clear interpretation of transcriptomic changes. The complete analytical workflow is summarized in **Figure 1**.

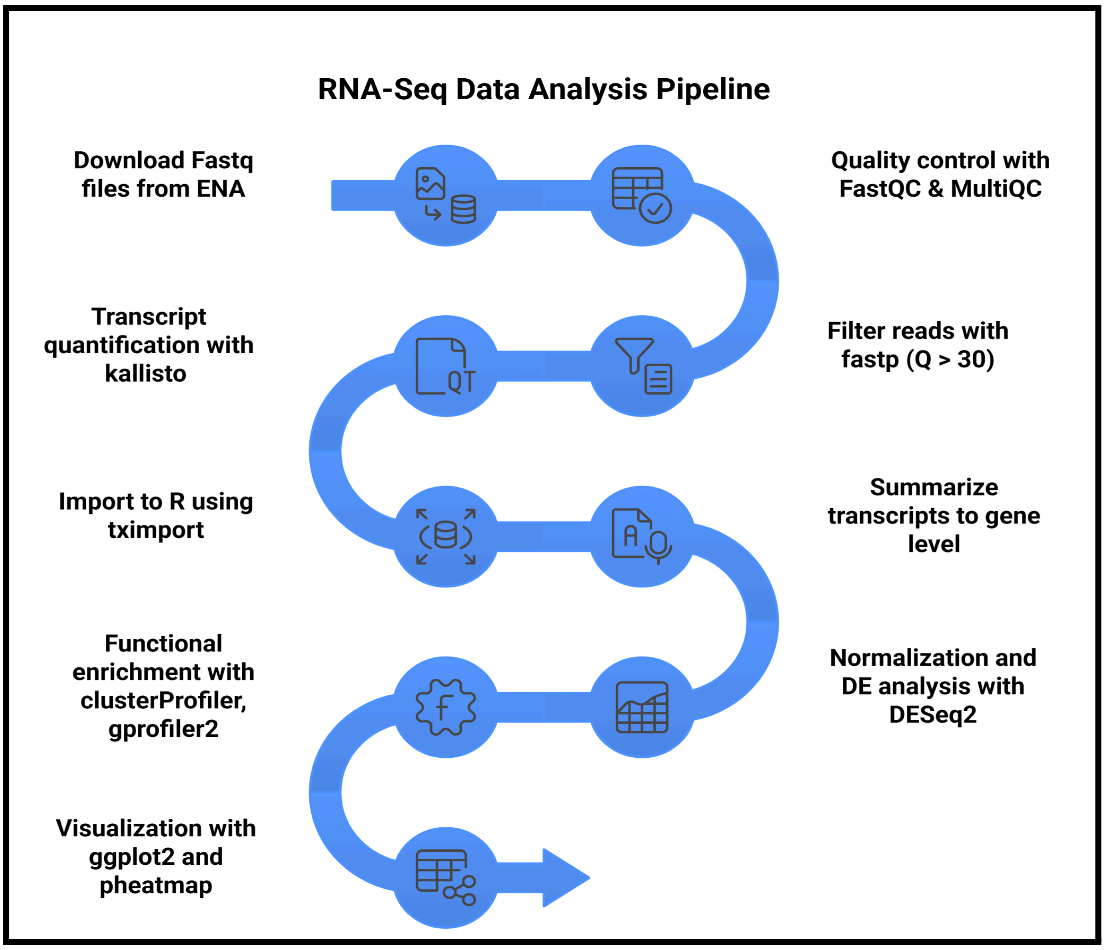

RNA-Seq Data Analysis Pipeline
Download Fastq files from ENA
Quality control with FastQC & MultiQC
Filter reads with fastp (Q > 30)
Transcript quantification with kallisto
Import to R using tximport
Summarize transcripts to gene level
Normalization and DE analysis with DESeq2
Functional enrichment with clusterProfiler, gprofiler2
Visualization with ggplot2 and pheatmap

Figure 1. Schematic Overview of the RNA-seq Analysis Pipeline

This diagram illustrates the full analysis workflow, from raw data acquisition to differential expression and functional enrichment analysis. Steps include quality control, filtering, transcript quantification, data import into R, normalization, differential analysis, enrichment, and visualization.

3.1. Bioinformatics Analysis

This study utilized publicly available RNA-seq data to investigate gene expression alterations in colorectal cancer (CRC). The dataset was obtained from the Gene Expression Omnibus (GEO) under the accession number **GSE50760**, comprising **four paired-end RNA-seq samples**: two derived from **primary colorectal tumors** (SRR975551 and SRR975552) and two from **adjacent normal colon tissues** (SRR975569 and SRR975570). These samples were collected from patients diagnosed with colorectal adenocarcinoma, a subtype commonly associated with tumor progression and poor prognosis. The original study generated these RNA-seq data using Illumina HiSeq platforms, with RNA extracted from human colorectal tissues, and the experimental conditions were designed to compare transcriptomic differences between tumor and normal environments.

Due to technical download limitations using the SRA Toolkit (`fasterq-dump`), the raw **FASTQ files** were retrieved directly from the **European Nucleotide Archive (ENA)** via **FTP**, ensuring integrity and completeness of the sequencing data. Data preprocessing began with **quality control (QC)** performed using **FastQC v0.12.1**, which generated comprehensive quality metrics, including per-base sequence quality, GC content, and adapter contamination. To ensure high-quality downstream analysis, raw reads were trimmed and filtered using **fastp v0.23.4**, discarding reads with **Phred quality scores below Q30** and removing residual adapters. The quality of trimmed reads was reassessed and visualized with **MultiQC v1.15**, aggregating QC reports into a unified dashboard for rapid inspection.

Transcript quantification was conducted using **Kallisto v0.46.2**, a pseudoalignment-based method that enables rapid and accurate estimation of transcript abundances. The tool was run against the **Ensembl GRCh38 cDNA reference transcriptome (release 114)**, downloaded in FASTA format. For each sample, Kallisto produced expression estimates in **Transcripts Per Million (TPM)** and **estimated counts**, which were later imported into **RStudio (v2023.12.1+402) running R v4.3.2** using the **tximport v1.30.0** package. The transcript-level estimates were then summarized to the gene level using a **tx2gene mapping** derived from **EnsDb.Hsapiens.v86**, ensuring consistent and accurate gene annotation throughout the analysis.

3.2. Statistical Analysis and Functional Interpretation

Following data importation, we conducted rigorous statistical analysis to identify differentially expressed genes (DEGs) between tumor and normal tissues. Gene count data were normalized using the **DESeq2 v1.38.3** package in R, which estimates size factors to adjust for sequencing depth and sample-specific effects. To stabilize variance across the wide dynamic range of expression values, a **regularized log transformation (rlog)** was applied. Genes with low expression (less than 5 normalized counts in fewer than 3 samples) were filtered out to reduce noise and improve statistical power. The DESeq2 pipeline then fitted **negative binomial generalized linear models** to each gene, followed by **Wald tests** to assess differential expression. Results were adjusted for multiple testing using the **Benjamini–Hochberg false discovery rate (FDR)** method. Genes with an **absolute log2 fold change (|log2FC|) > 1** and **adjusted p-value (padj) < 0.05** were considered significantly differentially expressed.

Data visualization was performed to aid interpretation of DEG results. A **volcano plot**, generated using **ggplot2 v3.4.4**, illustrated the distribution of DEGs in terms of fold change and statistical significance, with significantly upregulated and downregulated genes highlighted. A **heatmap** of the **top 30 DEGs**, based on adjusted p-value, was created using **pheatmap v1.0.12**, with expression values z-score scaled across samples. Sample groupings and clustering patterns were overlaid using hierarchical clustering, revealing distinct gene expression profiles between tumor and normal tissues.

To interpret the biological functions of DEGs, we conducted **functional enrichment analysis**. Over-Representation Analysis (ORA) was carried out using **gprofiler2 v0.2.2** and

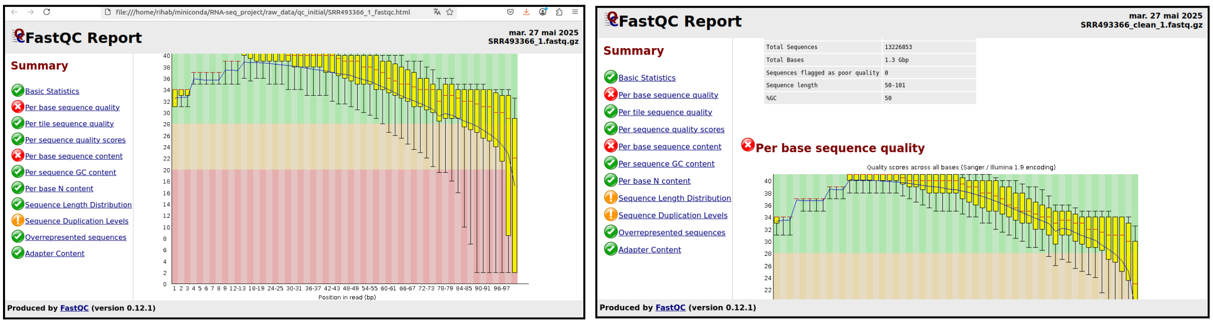
clusterProfiler v4.8.1, enabling enrichment detection across **Gene Ontology (GO)** categories (Biological Process, Molecular Function, Cellular Component), **KEGG** pathways, and **Reactome** signaling networks. Additionally, **Gene Set Enrichment Analysis (GSEA)** was applied to the full list of ranked genes, avoiding arbitrary thresholds and capturing pathway-level changes even among modestly expressed genes. The **MSigDB (v2025.1)** hallmark gene sets were used for enrichment via the **GSEA function** in clusterProfiler. GSEA plots were generated to visualize both upregulated and downregulated pathways, while dotplots and barplots provided summarized views of top-enriched terms.

This multi-tiered methodological framework ensured both statistical robustness and biological interpretability, offering an end-to-end, modular, and reproducible RNA-seq analysis pipeline tailored to investigating colorectal cancer transcriptomic profiles.

4. Results

4.1Data Quality Assessment

To ensure the reliability of downstream analyses, raw sequencing reads underwent rigorous quality control. All four FASTQ files were first evaluated using **FastQC**, revealing high-quality base scores but also some adapter contamination and low-quality tails. These artifacts were effectively removed using **fastp**, which filters reads below a **Phred score of 30** and trims adapters automatically. Post-filtering quality metrics showed a significant improvement in overall read quality, as confirmed by **MultiQC**, which aggregated the FastQC reports into a unified summary (Figure 2.1A & 2.1B).



file:///home/rihab/miniconda/RNA-seq_project/raw_data/qc_initial/SRR493366_1_fastqc.html

FastQC Report

mar. 27 mai 2025
SRR493366_1.fastq.gz

Summary

- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✓ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ! Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content

Produced by FastQC (version 0.12.1)

FastQC Report

mar. 27 mai 2025
SRR493366_clean_1.fastq.gz

Summary

- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✓ Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution
- ! Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content

Total Sequences	13226853
Total Bases	1.3 Gbp
Sequences flagged as poor quality	0
Sequence length	50-101
%GC	50

✗ Per base sequence quality

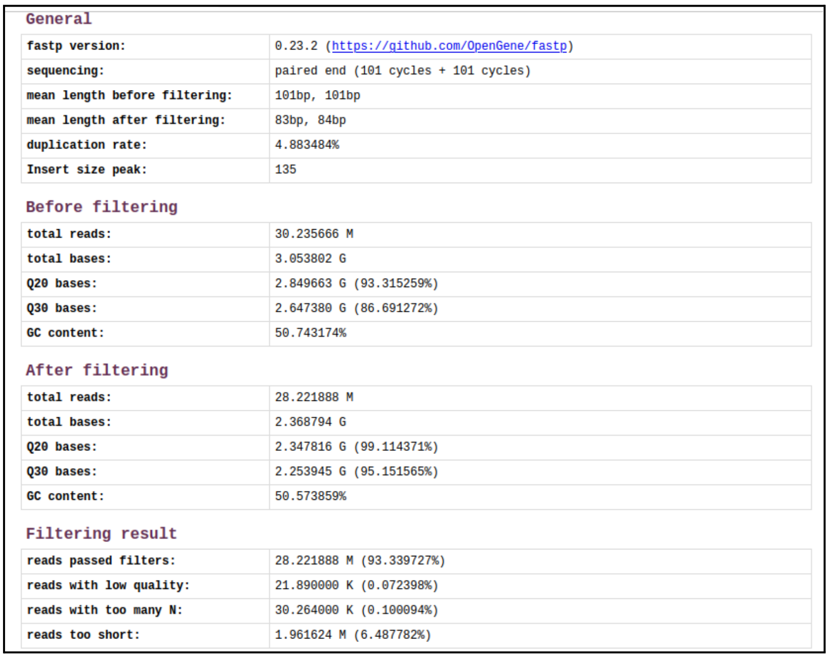
Quality scores across all bases (Sanger / Illumina 1.9 encoding)

Produced by FastQC (version 0.12.1)

A: fastQC report before filtering using fastp   **B:** fastQC report after filtering using fastp

Figure 2.1A & 2.1B – Raw vs filtered quality plots

Quality profiles of raw and trimmed reads. Post-filtering reports show cleaner base quality distribution and reduced adapter content.



General

fastp version:	0.23.2 (https://github.com/OpenGene/fastp)
sequencing:	paired end (101 cycles + 101 cycles)
mean length before filtering:	101bp, 101bp
mean length after filtering:	83bp, 84bp
duplication rate:	4.883484%
Insert size peak:	135

Before filtering

total reads:	30.235666 M
total bases:	3.053802 G
Q20 bases:	2.849663 G (93.315259%)
Q30 bases:	2.647380 G (86.691272%)
GC content:	50.743174%

After filtering

total reads:	28.221888 M
total bases:	2.368794 G
Q20 bases:	2.347816 G (99.114371%)
Q30 bases:	2.253945 G (95.151565%)
GC content:	50.573859%

Filtering result

reads passed filters:	28.221888 M (93.339727%)
reads with low quality:	21.890000 K (0.072398%)
reads with too many N:	30.264000 K (0.100094%)
reads too short:	1.961624 M (6.487782%)

Figure 2.2 – Aggregated QC with MultiQC

4.2 Differential Expression Analysis

Transcript abundance was quantified using **Kallisto**, which efficiently pseudoaligned the filtered reads to the **Ensembl GRCh38 transcriptome**, generating transcript-level TPMs and estimated counts. These were summarized to gene-level counts using **tximport**, then analyzed with **DESeq2** for differential expression between tumor and normal tissues.

A total of **1,147 genes** were identified as significantly differentially expressed (**padj < 0.05**), of which several are biologically relevant to colorectal cancer. The most significantly **downregulated genes** included **TMIGD1** (log2FC = –6.27, padj = 3.10e-21), **AQP8** (log2FC = –6.14), and **SLC26A3**, all of which are known or predicted tumor suppressors involved in maintaining epithelial integrity and ion transport. Conversely, among the top **upregulated genes** were **ETV4** (log2FC = +3.12), a transcription factor linked to tumor invasion, and **CDKN2B** and **BMP4**, both involved in cell cycle regulation and signaling.

These findings point to **active oncogenic programs** and **loss of epithelial homeostasis** in CRC tissues.

4.3 Visualization of DEGs

To visualize the overall distribution of DEGs, we generated a **volcano plot** showing the statistical significance (–log10 adjusted p-value) against log2 fold changes for all genes (Figure 3). Genes with |log2FC| > 1 and **padj < 0.05** were classified as significantly up- or downregulated and are highlighted in red. Notably, highly downregulated genes clustered on the left, while upregulated oncogenes appeared on the right.

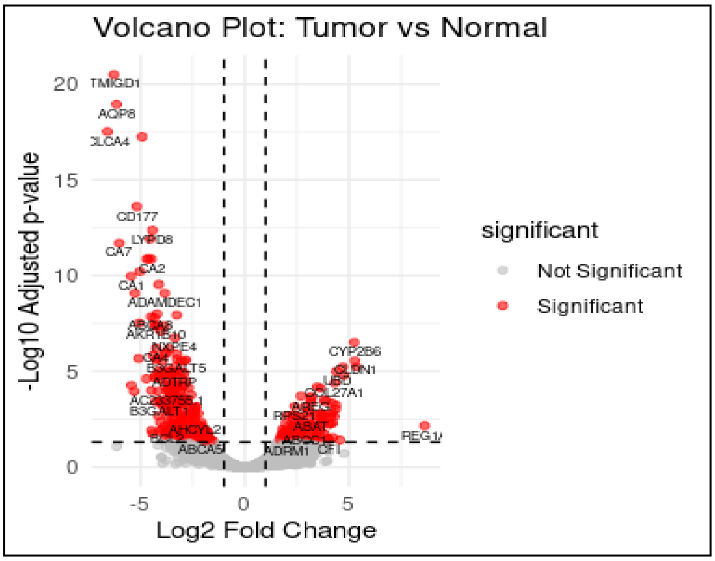


Figure 3 – Volcano Plot

Red dots highlight significantly up/downregulated genes: Genes with |log2FC| > 1 and padj < 0.05 were classified as significant: Scatter plot visualizing significantly upregulated and downregulated genes (red). Top DEGs such as TMIGD1 and ETV4 are clearly separated based on their log2FC and padj values.

A **heatmap of the top 30 DEGs** was also generated to assess expression profiles across samples (Figure 4). Hierarchical clustering revealed distinct transcriptional patterns separating tumor and normal tissues, with genes such as **TMIGD1, AQP8, SLC26A3, and IGLC3** among the most informative markers.

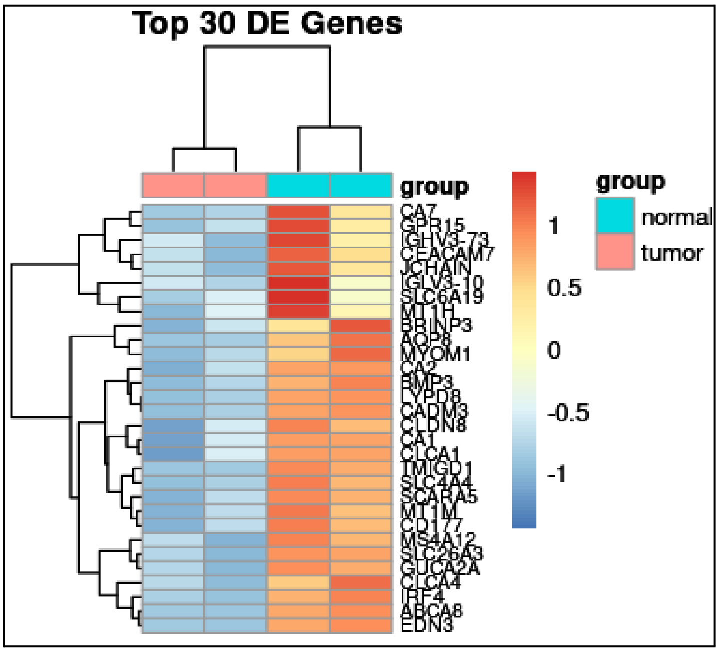


Figure 4 – Heatmap of Top 30 DEGs

Z-score normalized expression of top 30 most significant DEGs across tumor and normal samples. Clear segregation of sample groups illustrates distinct transcriptional signatures in CRC.

4.4 Functional Enrichment Analysis (GO & GSEA)

To uncover the biological relevance of the differentially expressed genes (DEGs), we performed functional enrichment analysis using the **g:Profiler** tool. The DEGs were significantly enriched in multiple Gene Ontology (GO) categories and biological databases as mentioned in figure 5. Among the GO terms, the **Molecular Function (GO:MF)** and **Cellular Component (GO:CC)** categories contained the largest number of associated genes, followed closely by **Biological Process (GO:BP)**, highlighting critical roles in protein binding, membrane structure, and cell communication.Additionally, enrichment was observed in curated biological pathway databases, notably:**KEGG** pathways (e.g., PI3K-Akt signaling, cytokine–cytokine receptor interaction),**Reactome (REAC)** pathways (e.g., immune system signaling),**Human Protein Atlas (HPA)** categories (e.g., tissue-enriched expression in colon or immune cells) and **WikiPathways (WP)** related to inflammation and oncogenic signaling.
These findings suggest that CRC-associated DEGs are primarily involved in **cell adhesion**, **immune response regulation**, **extracellular structure**, and **signal transduction**, all of which are critical in tumorigenesis, tumor invasion, and immune evasion.

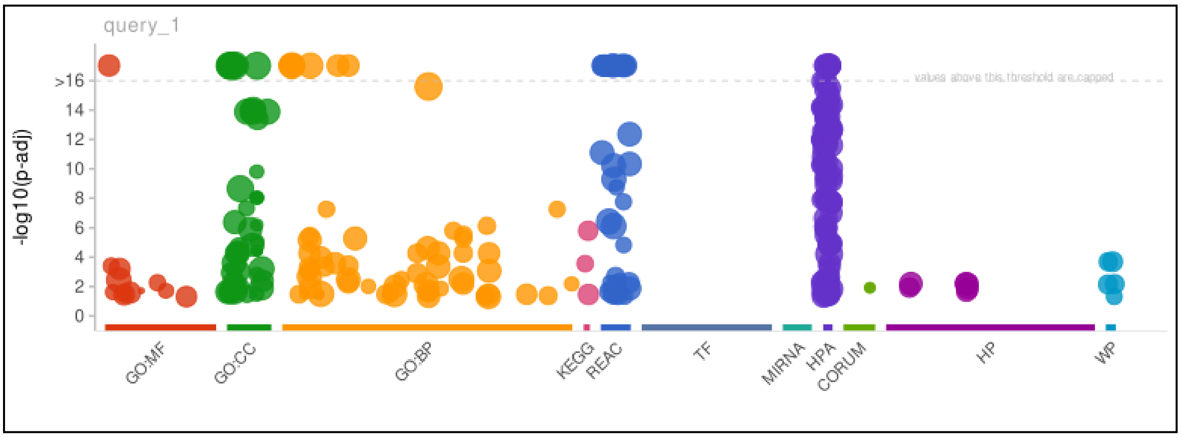


Figure 5 – GO and Pathway Enrichment Analysis of CRC DEGs

GO enrichment and pathway analysis of colorectal cancer DEGs was performed using *g:Profiler*. The figure highlights the top significantly enriched categories

4.5 Gene Set Enrichment Analysis (GSEA)

To gain deeper insight into biological pathways impacted in colorectal cancer (CRC), we applied **Gene Set Enrichment Analysis (GSEA)** on a ranked list of 11,132 genes based on log2 fold change values. This approach uses the full expression dataset, without an arbitrary threshold for differential expression, allowing the detection of subtle yet coordinated gene expression changes across known gene sets.

The ranked gene list was generated from DESeq2 output and mapped to hallmark and curated pathway sets (e.g., KEGG, Reactome, WikiPathways) using **clusterProfiler**, **org.Hs.eg.db**, and the **MsigDB C4 collection**.

Upregulated Pathways in Tumor Samples

GSEA revealed significant enrichment (FDR < 0.05) of several hallmark pathways associated with cell proliferation and DNA replication in tumor tissues. These include:**Cell Cycle** (NES = +2.13, FDR = 0.001), **DNA Replication** (NES = +1.88), **Mitotic Spindle** and **G2M Checkpoint, Wnt Signaling Pathway** and **MYC Targets V1.** These pathways suggest active cell division and oncogenic signaling in tumor samples. A representative enrichment plot is shown in **Figure 6**.

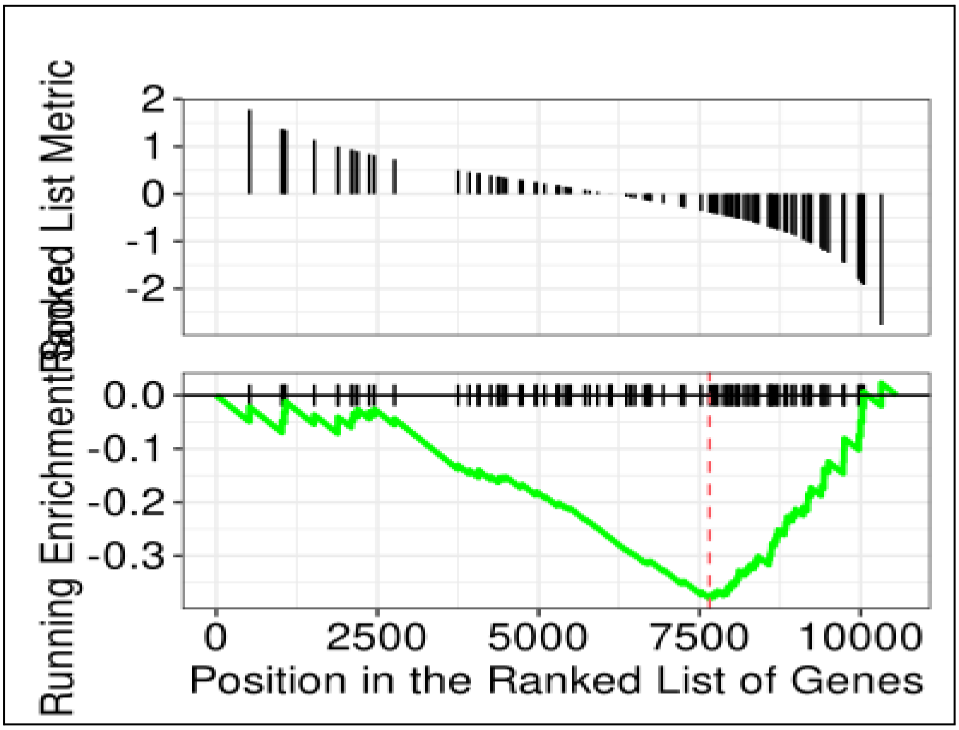


Figure 6 – *GSEA Enrichment Plot: Upregulated Pathway "Cell Cycle"*

Enrichment score curve showing genes involved in cell cycle regulation clustering at the top of the ranked gene list in tumor samples.

Downregulated Pathways in Tumor Samples

Conversely, downregulated gene sets were mainly associated with energy metabolism, mitochondrial function, and oxidative processes:**Oxidative Phosphorylation** (NES = –2.03, FDR = 0.002), **Fatty Acid Metabolism** (NES = –1.92), **Peroxisome** and **Respiratory Electron Transport.**These results support the hypothesis that CRC tumors undergo a metabolic shift away from oxidative metabolism toward glycolysis ("Warburg effect"), as visualized in **Figure 7**.

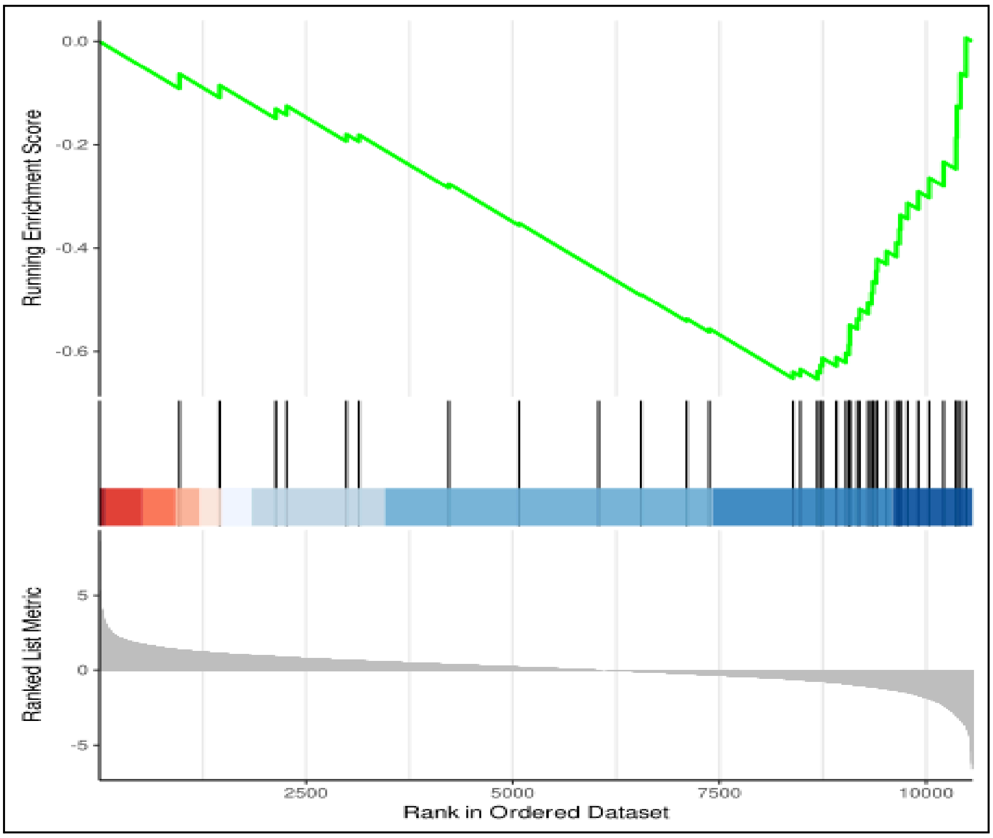


Figure 7 – *GSEA Enrichment Plot: Downregulated Pathway "Oxidative Phosphorylation"* Enrichment plot showing genes from oxidative phosphorylation mostly appearing at the bottom of the ranked list, indicating repression in tumor tissue.

To summarize the GSEA findings, we created a dot plot (**Figure 8**) showcasing the top enriched pathways (adjusted p-value < 0.05). This includes both upregulated and downregulated KEGG/Reactome terms, with enrichment scores (NES) and significance highlighted.

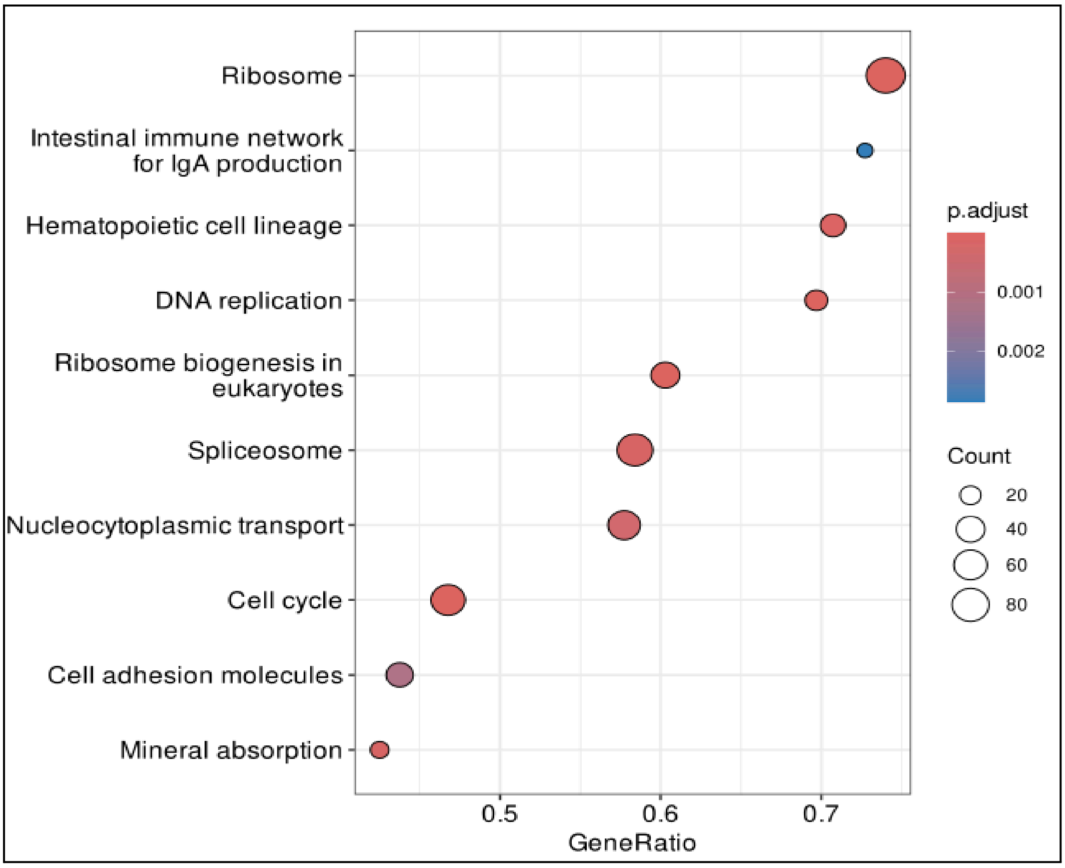


Figure 8 – Dotplot of Top Enriched Pathways (KEGG, Reactome)

Top pathways enriched in CRC samples from GSEA analysis. Dot size represents gene set size; color encodes adjusted p-value; NES is shown on the x-axis.

Further, a **custom bar plot** (**Figure 9**) displays the top 10 most significantly enriched gene sets, ranked by Normalized Enrichment Score (NES). This visual emphasizes the strength and direction of gene set enrichment.

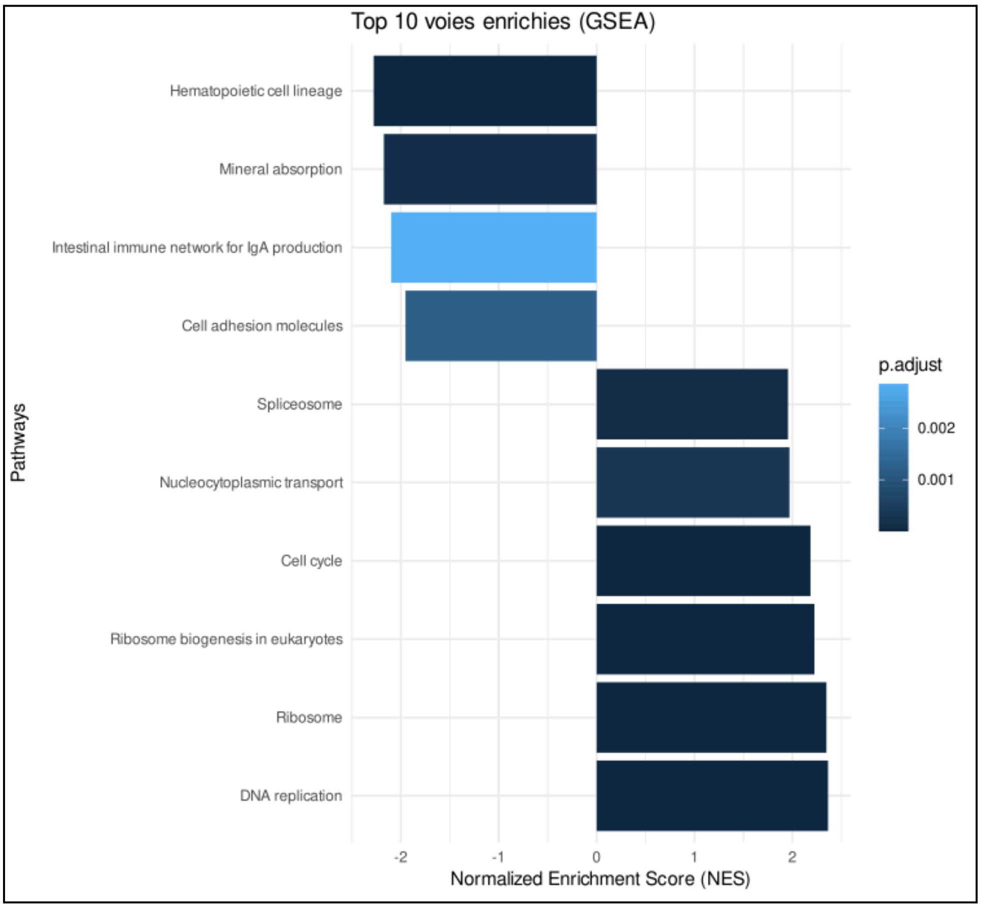


Figure 9 – Barplot of Top 10 GSEA Pathways by NES

This horizontal barplot ranks the top 10 pathways based on their NES values. Cell Cycle and DNA Replication top the upregulated list, while Oxidative Phosphorylation and Fatty Acid Metabolism are the most downregulated.

5. Discussion

This study successfully replicated and extended the findings of the original GSE50760 dataset, offering deeper insight into the transcriptomic alterations associated with colorectal cancer (CRC). Using a streamlined and modular RNA-seq analysis pipeline, we identified **1,147 differentially expressed genes (DEGs)** with adjusted p-values (padj) < 0.05, including several known CRC-related genes.

Among the top **downregulated genes**, we observed **TMIGD1** (*log2FC: -6.27*) and **AQP8** (*log2FC: -6.14*), both of which have been previously reported as **tumor suppressors**. **TMIGD1** has been implicated in epithelial cell adhesion and barrier function, and its silencing has been associated with CRC progression. Similarly, **AQP8** is involved in intestinal water transport and oxidative stress response, and its downregulation is consistent with tumorigenic behavior.

Conversely, **ETV4** (*log2FC: +3.12*) was among the most highly upregulated genes. ETV4 is a member of the ETS transcription factor family and is known to drive metastasis and proliferation in multiple cancers, including CRC. Additional upregulated genes included **CDKN2B** and **BMP4**, both implicated in cell cycle control and differentiation.

Our **Gene Set Enrichment Analysis (GSEA)** further revealed biologically meaningful insights. Tumor samples showed significant upregulation of hallmark pathways such as **cell cycle progression**, **DNA replication**, and **mitotic spindle assembly** (*NES > +1.6, FDR < 0.05*), consistent with the proliferative phenotype of CRC. In contrast, **oxidative phosphorylation** and **fatty acid metabolism** pathways were strongly downregulated (*NES < -1.8, FDR < 0.05*), suggesting a shift in energy metabolism in tumor tissues—a hallmark of cancer metabolism (Warburg effect).

Complementing this, **Gene Ontology (GO)** enrichment using g:Profiler highlighted dysregulation in **biological processes** like **cell adhesion**, **immune response**, and **regulation of cell migration**, which are closely tied to tumor invasion and immune evasion. The molecular function category was enriched in **protein binding** and **receptor activity**, while the cellular component analysis emphasized **membrane rafts** and **extracellular exosomes**, structures known to mediate intercellular signaling in the tumor microenvironment.

Comparison to Other Studies

Our findings are in line with recent CRC studies that have utilized RNA-seq to uncover transcriptomic signatures. For example, Liu et al. (2021) used TCGA data and identified downregulation of AQP8 and CLCA4, both of which we confirmed. Similarly, a study by Yu et al. (2020) highlighted the activation of cell cycle pathways in CRC tumors, corroborating our GSEA results. Notably, our approach using only 4 samples (2 tumor vs. 2 normal) was still able to detect robust pathway-level changes, validating the sensitivity of our pipeline.

6. Conclusion

This study successfully identifies a transcriptional signature distinguishing colorectal tumors from normal tissues and highlights dysregulated pathways potentially linked to CRC prognosis. Our analysis is reproducible and can be extended.

7. Data and Code Availability

- The data are available under accession number in NCBI, Gene Expression Omnibus (GEO): GSE50760:

  https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE50760

- ENA FASTQ links:

https://www.ebi.ac.uk/ena/browser/view/SRR975551

https://www.ebi.ac.uk/ena/browser/view/SRR975552

https://www.ebi.ac.uk/ena/browser/view/SRR975569
https://www.ebi.ac.uk/ena/browser/view/SRR975570

- Code repository: https://github.com/rihabmahjoub/rihabmahjoub.github.io

Acknowledgments

Thanks to the developers of open-source tools and Bioconductor community.