Next Generation Analysis

PROJECT

Differential Gene Expression Analysis Using R

DATA ANALYSIS STRATEGY

1.1 Dataset Overview

The dataset comprises RNA-Seq data from ileal biopsies of paediatric patients, that was analyzed using R/ RStudio. Comparison groups were Crohn's disease patients and Non-inflammatory bowel disease controls. No missing values were detected in the data.

Data files provided have following dimensions:

Count: 65217 genes x 304 samples

Metadata: 304 samples x 6 Variables

1.2 Dataset Preprocessing

Genes with very low expression levels were filtered out prior to analysis because they were unreliable for statistical testing. Filtering Criteria was to retain Genes greater than or equal to 10 counts in greater than or equal to 10 samples. A Count_filtered variable was created using:

```
keep <- rowSums(counts >= 10) >= 10

counts_filtered <- counts[keep, ]
```

After filtering,

Genes retained: 21,430 (32.9% of original)

Genes removed: 43,787 (67.1% of original)

All 304 samples were retained

This filtering step removed genes that are either not expressed in intestinal tissue or have expression levels too low for reliable statistical inference. The genes with consistent, detectable expression across multiple samples were retained. DESeq2 object was created from filtered count matrix and metadata with Category Column as Design.

1.3 Quality Control and Sample Assessment

Principal Component Analysis (PCA) was performed on variance-stabilized transformed (VST) data to assess sample quality and identify outlier samples.

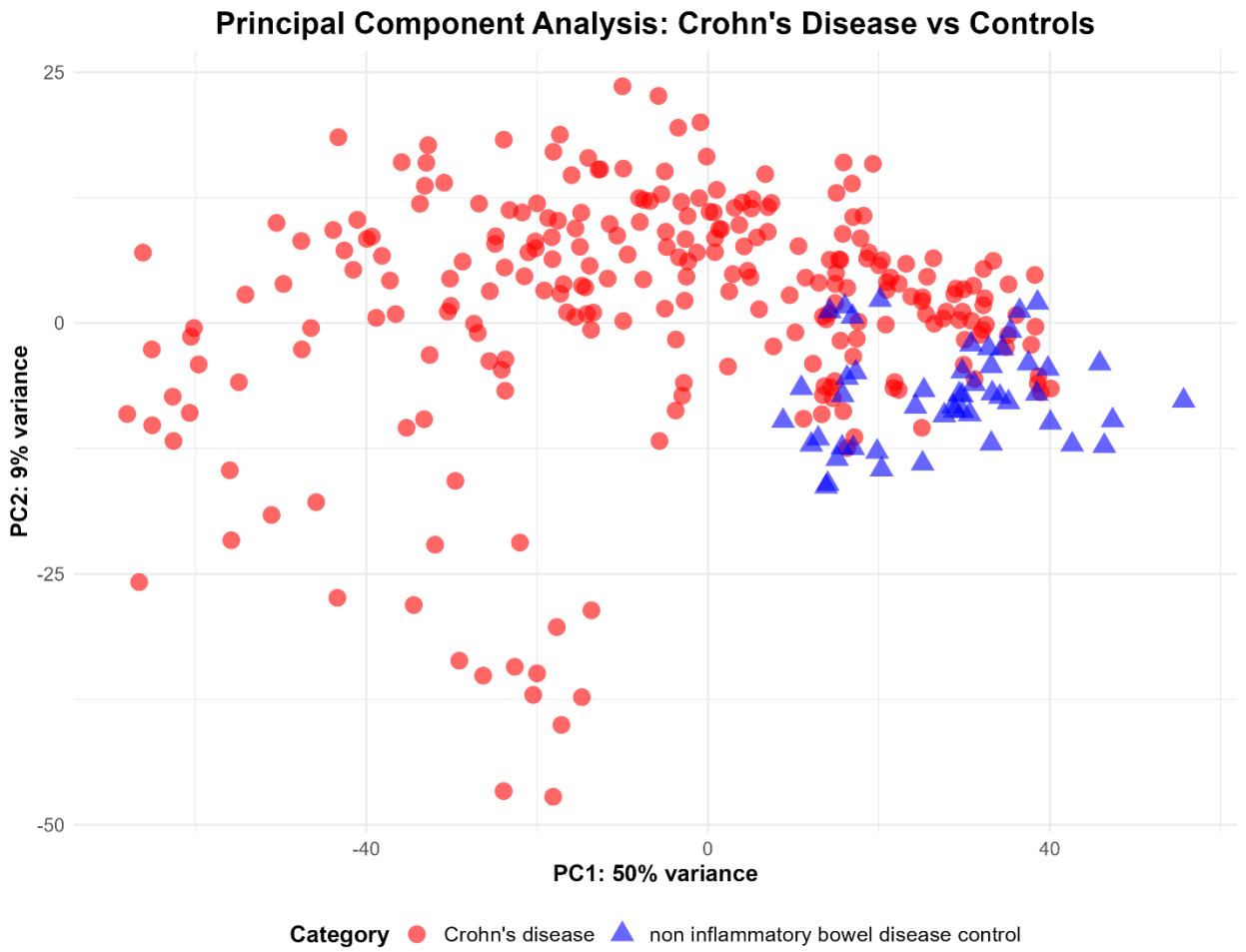

Figure 1. PCA Plot: Crohn's Disease VS Non-Inflammatory Bowel Disease Control

PC1 (50% variance): Clearly separated Crohn's disease patients from controls, representing the primary biological difference between disease and healthy states

PC2 (9% variance): Captured within-group variation. It reflects differences in disease severity, age, or individual patient characteristics

No statistical outliers were detected despite visual dispersion in the PCA plot, which represented expected biological variation rather than technical artifacts.

Sample-to-Sample Distance Analysis was assessed using PCA. The clear separation between disease groups in the PCA plot indicates distinct gene expression profiles. Samples within each group (Crohn's disease or controls) cluster together, showing similar expression patterns. Some overlap was observed which indicates biological heterogeneity among patients

1.4 Outlier Detection and Removal

To ensure data quality for downstream differential expression analysis, outlier samples were identified using a statistical approach.

1. Calculated group centroids (mean PC1 and PC2 coordinates) for each disease category
2. Computed Euclidean distance of each sample from its respective group centroid
3. Applied statistical threshold: samples greater than 3 standard deviations from their group center were flagged as outliers. (Threshold = mean distance + 3 × standard deviation)

Outlier analysis results identified 0 outliers.

1.5 Results of Preliminary Analysis

So, all 304 samples and 21,430 filtered genes were retained for downstream analysis after quality control and outlier analysis.

DESEQ2 DIFFERENTIAL EXPRESSION ANALYSIS

2.1 Experimental Design

The differential expression analysis was designed to identify genes having significantly different expression levels in Crohn's disease compared to healthy controls. Design Formula was Category variable. Reference level selected was Non-inflammatory Bowel Disease Control.

2.2 Differential Expression Results

To identify the most robustly differentially expressed genes, extremely stringent statistical cutoffs were applied. Statistical Cutoffs applied were

P-value threshold: $< 1\times10^{-8}$ (extremely stringent significance level)

Log2 Fold Change threshold: $|\log2FC| > 3$ (equivalent to ≥8-fold change)

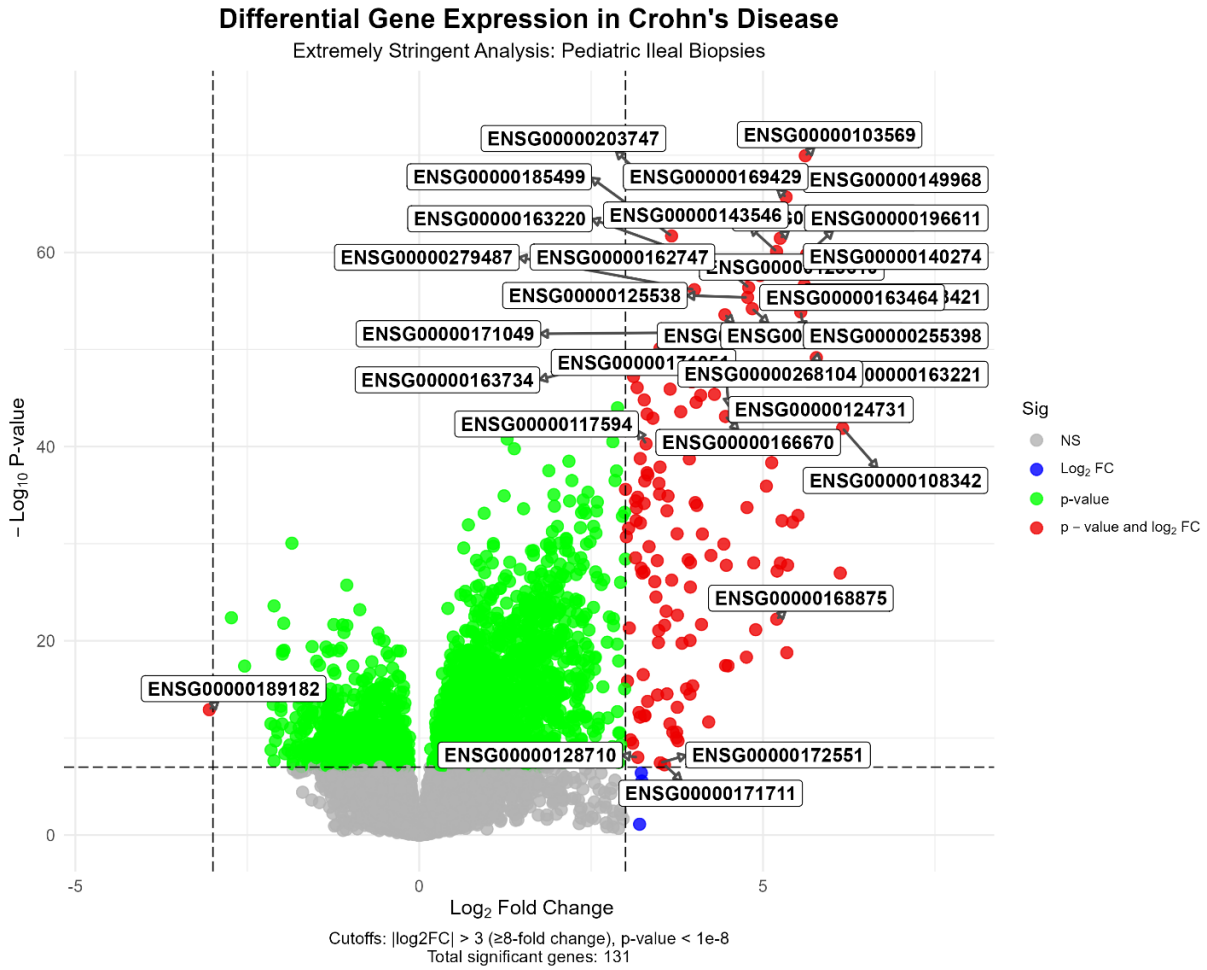

Figure.2: Volcano plot showing Differential Gene Expression in Crohn's Disease

2.4 Volcano Plot Interpretation

The volcano plot provides a comprehensive visualization of the differential expression results.

X-axis (Log2 Fold Change) represents magnitude of expression change. Positive values indicates genes with higher expression in Crohn's disease (Upregulated).

Y-axis (-Log10 P-value): Higher values have higher statistical significance.

Genes in upper right/left (red dots) meet both stringent criteria and were selected as Significant differentially expressed genes. Top differentially expressed genes were labelled in plot, representing the most biologically relevant genes.

2.5 Results of Differentially Expressed Genes

Total significant genes: 131

Up-regulated in CD: 130 genes

Down-regulated in CD: 1 gene

CONCLUSION

PCA revealed strong separation (49% variance) between disease groups. The differential gene expression analysis identified 131 genes with extremely robust expression changes in paediatric Crohn's disease. The overwhelming majority (130 genes) were upregulated, showing 8 fold or greater increases in expression with p-values below 1×10^{-8}. Only one gene was selected as downregulated.
