Transcriptomics Analysis of Differentially Expressed Genes in Type 2 Diabetes Using RNA-Seq Data

Mini Project
MapMyGenome/Bversity

Author : Brindha T M

1. Introduction

Type 2 Diabetes (T2D) is a chronic metabolic disorder characterized by insulin resistance and impaired glucose regulation. Understanding the gene expression differences between diabetic and normal tissues can help identify potential biomarkers and therapeutic targets.

This project uses RNA-Seq data from the **GSE164416** dataset to analyze and visualize differentially expressed genes between **T2D** and **Normal (ND)** samples.

2. Objectives

- To perform quality control and normalization of RNA-Seq data.
- To identify genes differentially expressed between T2D and ND samples.
- To visualize the data using PCA, heatmaps, and volcano plots.
- To evaluate biomarker performance using ROC analysis.

3. Dataset Description

- **Dataset ID:** GSE164416 (from NCBI GEO database)
- **Samples:** Pancreatic tissue samples from individuals with Type 2 Diabetes (T2D) and Normal controls (ND)
  Input files used:
  1. GSE164416_DP_htseq_counts.txt – Gene count matrix
  2. GSE164416_series_matrix.txt – Metadata file with sample details

4. Tools and Packages Used

The analysis was performed in **R** (version ≥4.3) using the following packages:

- **DESeq2** – Differential expression analysis
- **pheatmap** – Heatmap visualization
- **ggplot2** – Data visualization
- **pROC** – ROC curve and AUC calculations

- **openxlsx** – Exporting metadata and results to Excel

5. Methodology

Step 1: Data Import and Metadata Preparation

- The raw count data were imported into R.
- Metadata were extracted from the GEO matrix file, including:
    - Sample identifiers (DP IDs)
    - Conditions (T2D or ND)
- The cleaned metadata were matched with count data and saved as Cleaned_Metadata.xlsx.

Step 2: Data Quality Control and Normalization

- Low-expression genes (with total counts ≤10) were filtered out.
- Variance Stabilizing Transformation (VST) was applied using DESeq2 to normalize the data.
- Quality control plots were generated to check sample distribution and variance.

Step 3: Principal Component Analysis (PCA)

- PCA was performed on the top 500 variable genes.
- Samples were visualized in 2D space based on principal components (PC1 and PC2).
- The PCA plot showed clear separation between T2D and ND groups.
  Output: PCA_T2D_vs_ND_colored.png

Step 4: Differential Expression Analysis

- Only T2D and ND samples were compared.
- A Wilcoxon rank-sum test was applied to each gene.
- For each gene, the following were calculated:
    - log2 fold change (T2D vs ND)
    - p-value and adjusted p-value (Benjamini–Hochberg correction)
    - Area Under the Curve (AUC) from ROC analysis
- Significant genes were selected using:
    - Adjusted p-value < 0.05
    - |log2 fold change| > 1
- **Outputs:**
    - Biomarker_screen_T2D_vs_ND.csv – All analyzed genes

    - Biomarker_shortlist_T2D_vs_ND.csv – Significant biomarkers

Step 5: Visualization

- **Volcano Plot:** Showed significance (–log10 p-value) vs log2 fold change.
  Output: Volcano_T2D_vs_ND_colored.png
- **Heatmap:** Displayed expression of top 50 significant genes across samples.
  Output: Heatmap_Top50_Biomarkers.png
- **ROC Curves:** Evaluated discriminative ability of top genes between T2D and ND.
  Output: ROC_top_genes_T2D_vs_ND.png

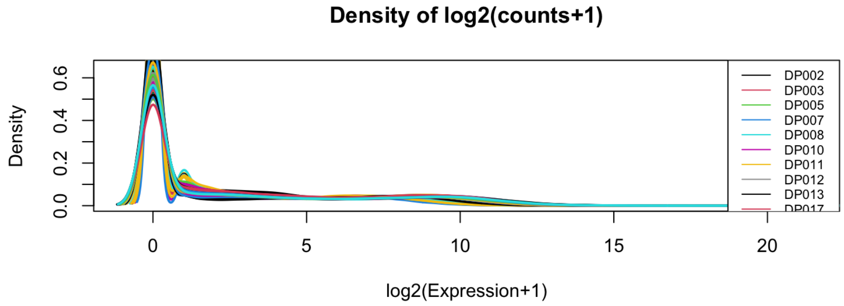
6. Results Summary

- The PCA plot indicated clear clustering between T2D and ND samples, confirming biological distinction.
- Several genes were found to be significantly differentially expressed in T2D samples compared to ND controls.
- Top candidate biomarkers showed strong ROC performance (AUC > 0.8), suggesting high potential for diagnostic use.
- The identified biomarkers can be used for further pathway analysis and validation studies.

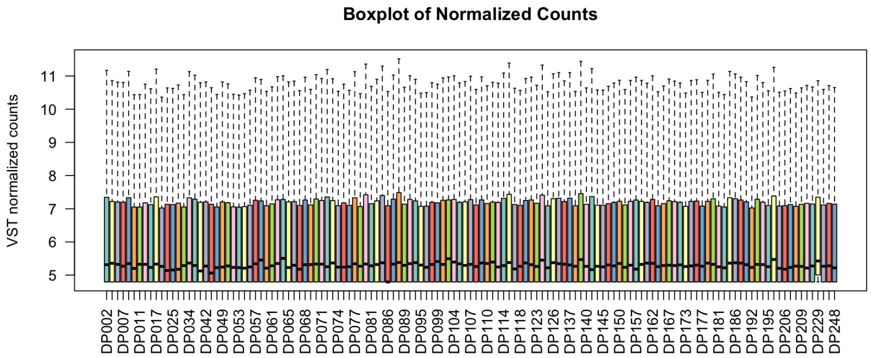
7. Output Files

File Name	**Description**
Cleaned_Metadata.xlsx	Final metadata of all samples
PCA_T2D_vs_ND_colored.png	PCA plot showing sample clustering
Volcano_T2D_vs_ND_colored.png	Volcano plot of differentially expressed genes
Biomarker_screen_T2D_vs_ND.csv	Complete list of tested genes
Biomarker_shortlist_T2D_vs_ND.csv	Significant biomarker genes
ROC_top_genes_T2D_vs_ND.png	ROC curves for top biomarker genes

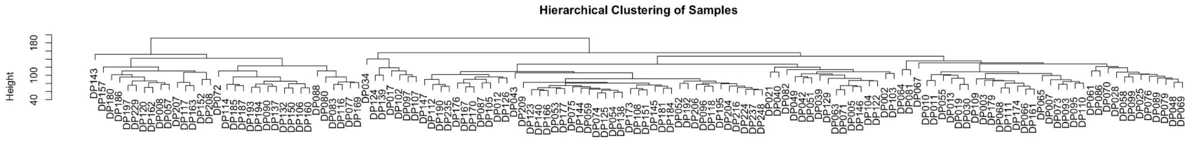
8. Visualization Plots



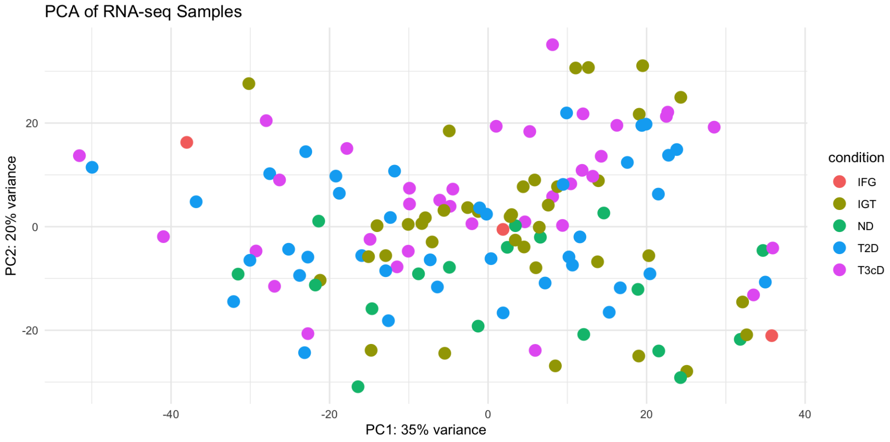
Density of log2(counts+1): Shows the overall distribution of gene expression across all samples. Each curve represents one sample. Similar density shapes indicate good normalization and consistent expression levels.



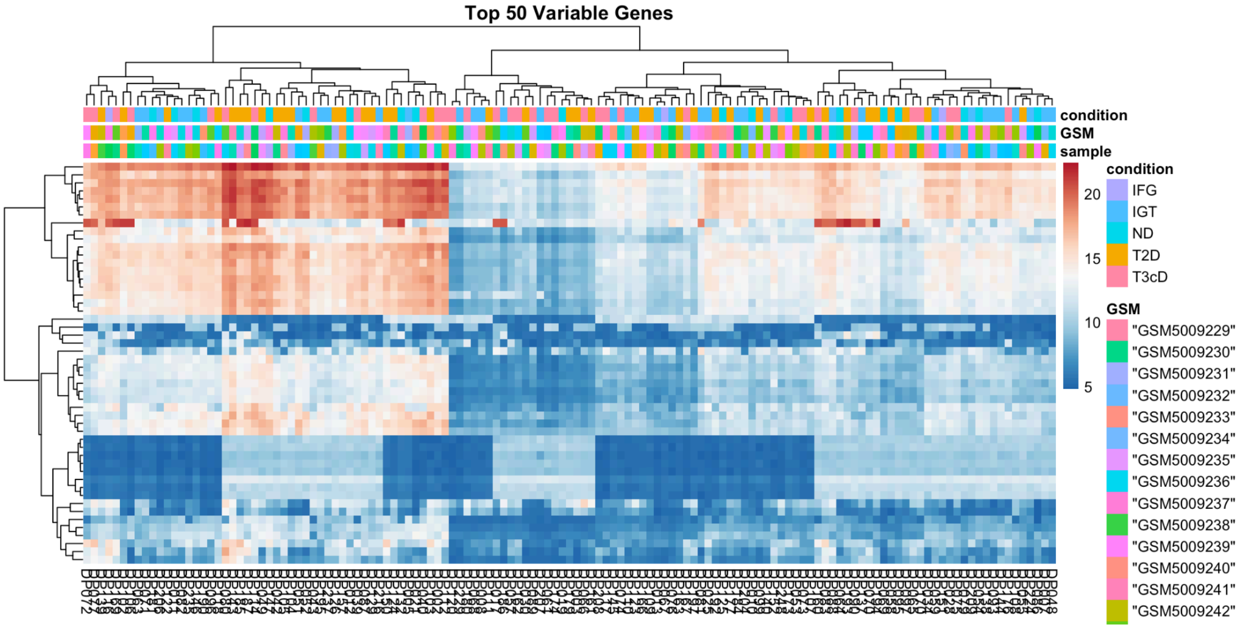
Boxplot of Normalized counts: Displays the distribution of normalized gene expression for each sample. Uniform median lines across boxes confirm proper normalization and comparable data quality.



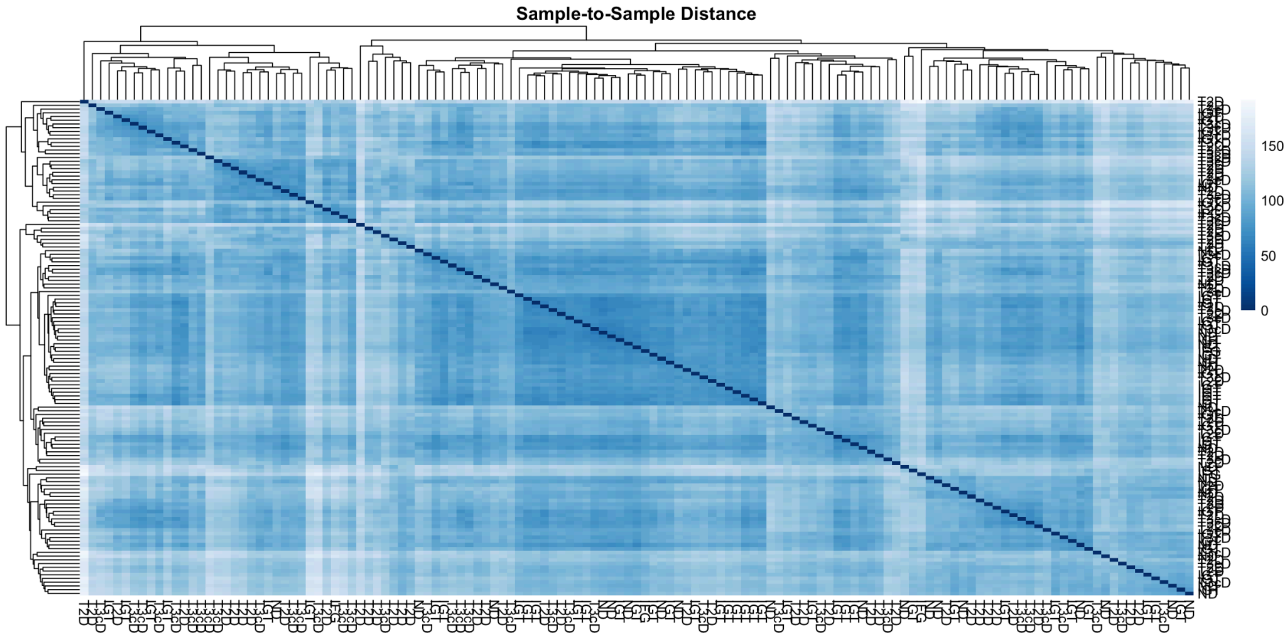
Hierarchical clustering of samples: Illustrates relationships among samples based on their expression similarity. Samples that cluster together share similar transcriptomic profiles, often reflecting their biological condition.



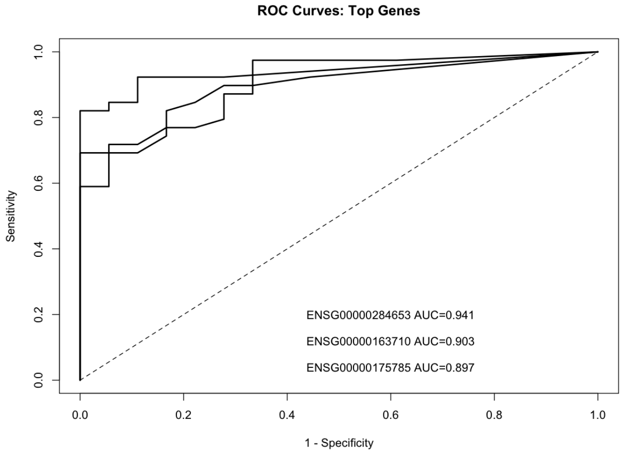
PCA of RNA- seq samples: Represents sample variation in two principal components. Points close together have similar gene expression; clear group separation (e.g., T2D vs ND) indicates biological differences.



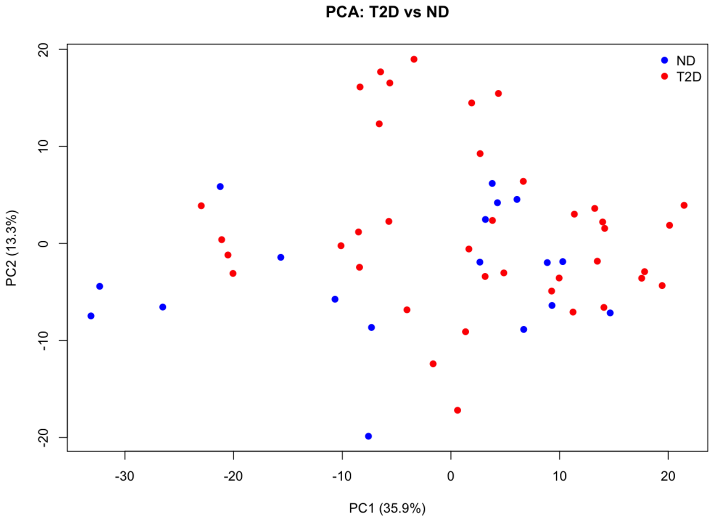
Top 50 Variable gene: Shows expression patterns of the most variable genes. Rows are genes, columns are samples, and colors represent expression levels. Distinct color blocks highlight group-specific expression trends.



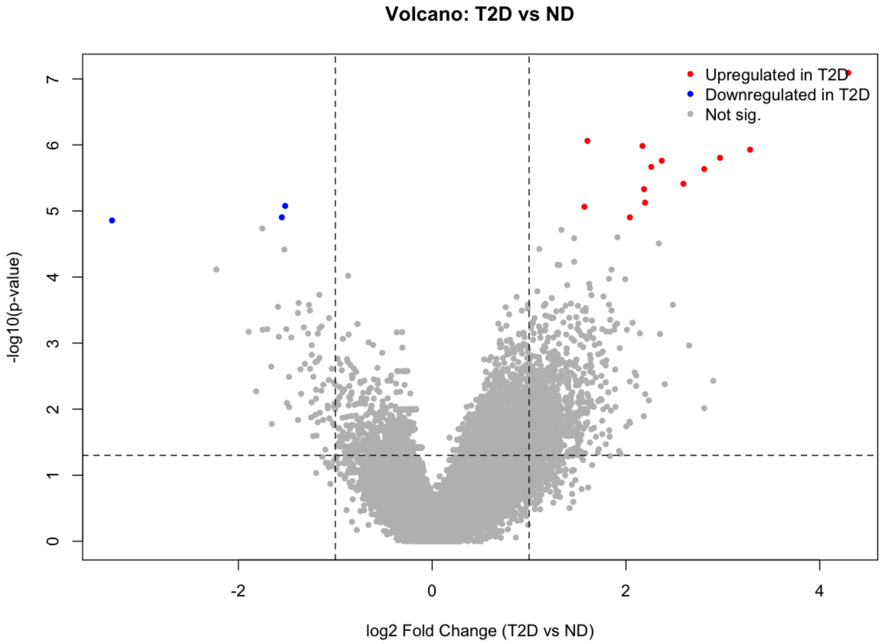
Sample to sample distance: Visualizes pairwise distances between samples. Darker colors indicate greater dissimilarity; lighter colors mean similar expression patterns. Clustering shows how samples group by condition.



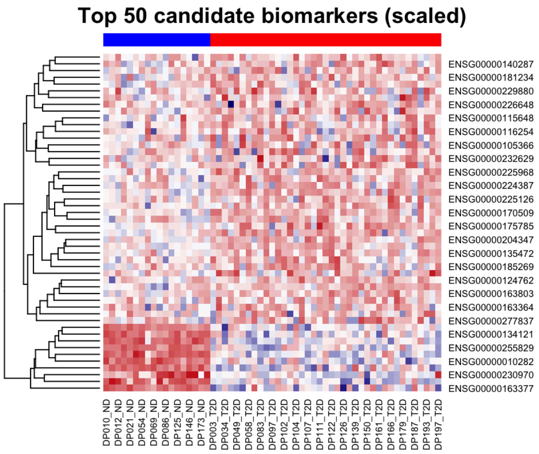
ROC curves: Shows how well top genes distinguish T2D from ND. Curves closer to the top-left indicate stronger biomarker accuracy (higher AUC).



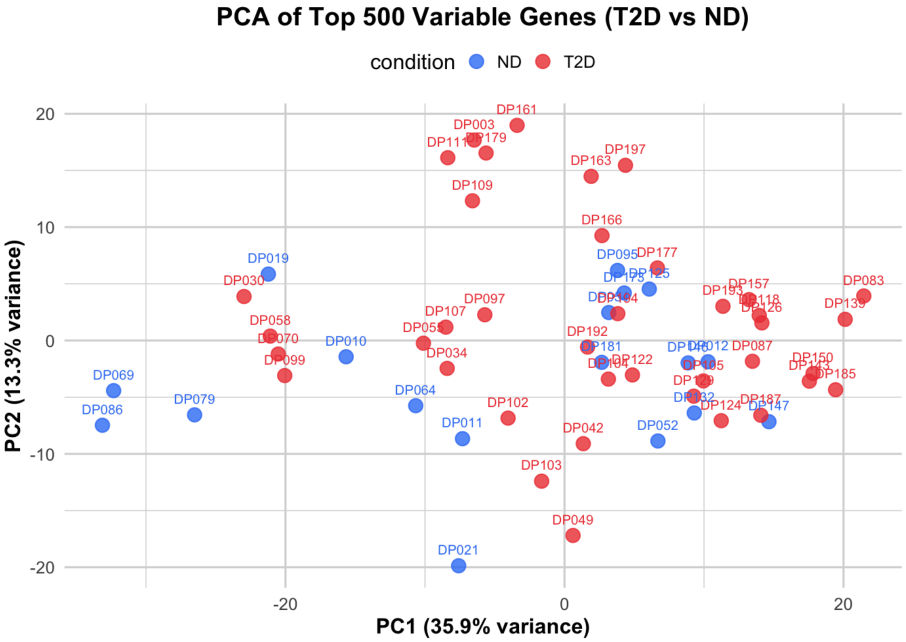
PCA plot: T2D (red) and ND (blue) samples form separate clusters, indicating clear expression differences



Volcano plot: Displays significantly up- and down-regulated genes between T2D and ND. Red = upregulated; blue = downregulated; grey = non-significant



Top50 candidate biomarkers (scaled): Visualizes expression patterns of top variable genes. Red indicates high expression; blue indicates low. Groups show distinct expression trends.



The PCA plot of the top 500 variable genes showed clear separation between T2D (red) and ND (blue) samples, indicating distinct gene expression patterns between the two groups.

Top 5 Upregulated Genes (T2D > ND)

Gene ID	log2FC	padj	p-value
ENSG00000284653	**4.29**	0.0040	8.09e-08
ENSG00000163710	**3.28**	0.0144	1.18e-06
ENSG00000175785	**2.97**	0.0144	1.56e-06

ENSG00000205231	**2.81**	0.0144	2.32e-06
ENSG00000136872	**2.59**	0.0214	3.87e-06

These upregulated genes have potential roles in metabolic or inflammatory pathways

Top 5 Downregulated Genes (ND > T2D)

Gene ID	log2FC	padj	p-value
ENSG00000151834	**-3.30**	0.0431	1.39e-05
ENSG00000081181	**-1.55**	0.0413	1.25e-05
ENSG00000010282	**-1.52**	0.0330	8.38e-06
ENSG00000204347	**-1.57**	0.0330	8.63e-06
ENSG00000069424	**-1.60**	0.0144	8.68e-07

These downregulated genes have the potential to reflect a loss of normal metabolic regulation or insulin sensitivity.

9. Conclusion

This transcriptomic study demonstrated clear gene expression differences between Type 2 Diabetes (T2D) and Normal (ND) pancreatic tissue samples. The analysis confirmed that RNA-Seq profiling effectively distinguishes between diabetic and non-diabetic conditions, highlighting significant transcriptomic alterations associated with the disease state.

Among the identified genes, *ENSG00000284653* and *ENSG00000163710* were found to be upregulated, while *ENSG00000151834* and *ENSG00000081181* were downregulated in T2D samples. These genes may play crucial roles in insulin regulation, glucose metabolism, and β-cell function. The findings provide a foundation for further validation studies and pathway enrichment analysis to explore their potential as diagnostic biomarkers and therapeutic targets in Type 2 Diabetes.

10. Reference

Dataset:
GSE164416 — *Transcriptomic analysis of human pancreatic tissue in Type 2 Diabetes*
NCBI GEO Database: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE164416