**Microarray Data Preprocessing, Normalization & DEG Analysis – Module II**

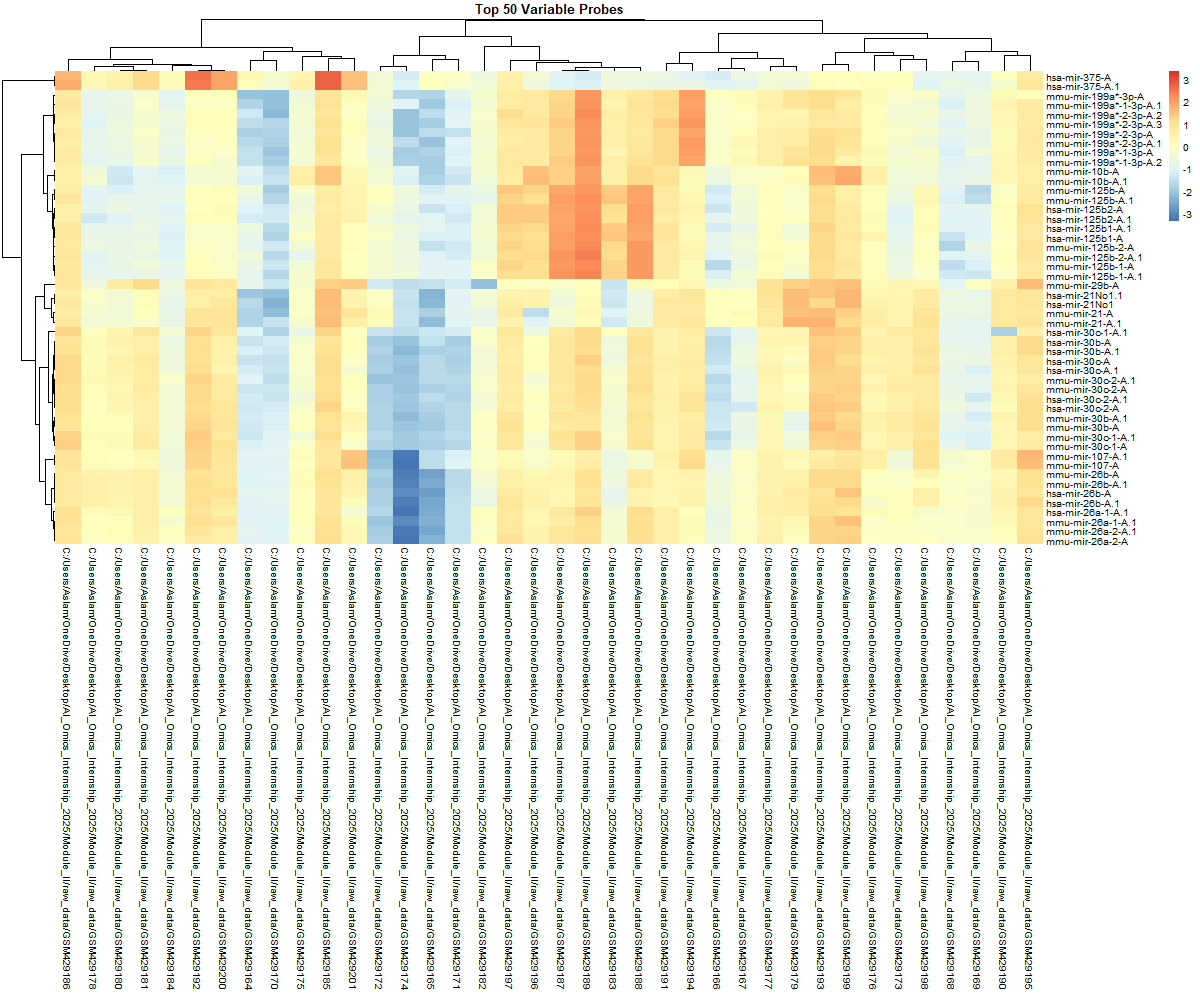
**Internship:** AI and Omics Research (2025)  
**Author:** M.K. Aysha Meharin

**1. Introduction**

This report documents the preprocessing, normalization, and differential expression analysis of microarray data from 38 GPR files. The analysis was performed using R (limma, pheatmap, ggplot2).

**2. Workflow Steps**

1. **Memory Clearance and Directory Setup**
   * Cleared R environment and set working directory for Module II data.
2. **Loading Raw Data**
   * Imported 38 GPR files using read.maimages().
   * Assigned unique probe names.
3. **Background Correction**
   * Applied backgroundCorrect() with normexp method and offset 50.
4. **Log2 Transformation & Normalization**
   * Log2 transformed expression values.
   * Performed quantile normalization using normalizeBetweenArrays().
5. **Filtering Probes**
   * Removed probes with NA, NaN, or non-finite values.
   * Remaining probes after filtering: **10,240**
6. **Top Variable Probes Heatmap**
   * Selected top 50 variable probes based on variance.
   * Plotted heatmap to assess clustering.

**Image :**

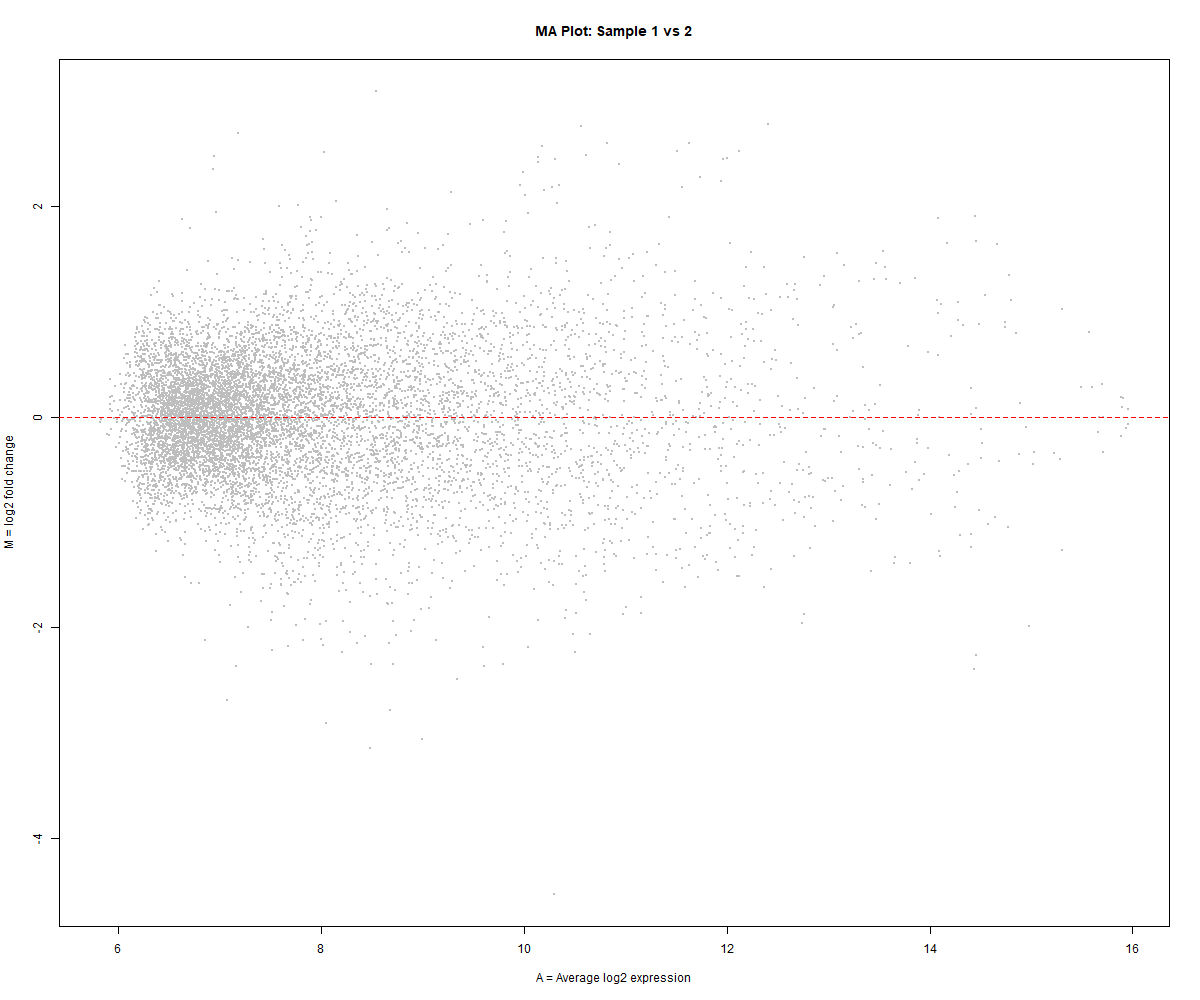
* + TopVariableProbes\_Heatmap.png

1. **Differential Expression Analysis**
   * Defined sample groups: normal (19) vs cancer (19).
   * Constructed design matrix and contrasts.
   * Applied lmFit(), contrasts.fit(), eBayes().
2. **DEG Classification**
   * Upregulated: logFC > 1 & adj.P.Val < 0.05 → **309 genes**
   * Downregulated: logFC < -1 & adj.P.Val < 0.05 → **55 genes**
   * Not significant: remaining probes

**Table 1: DEG Counts**

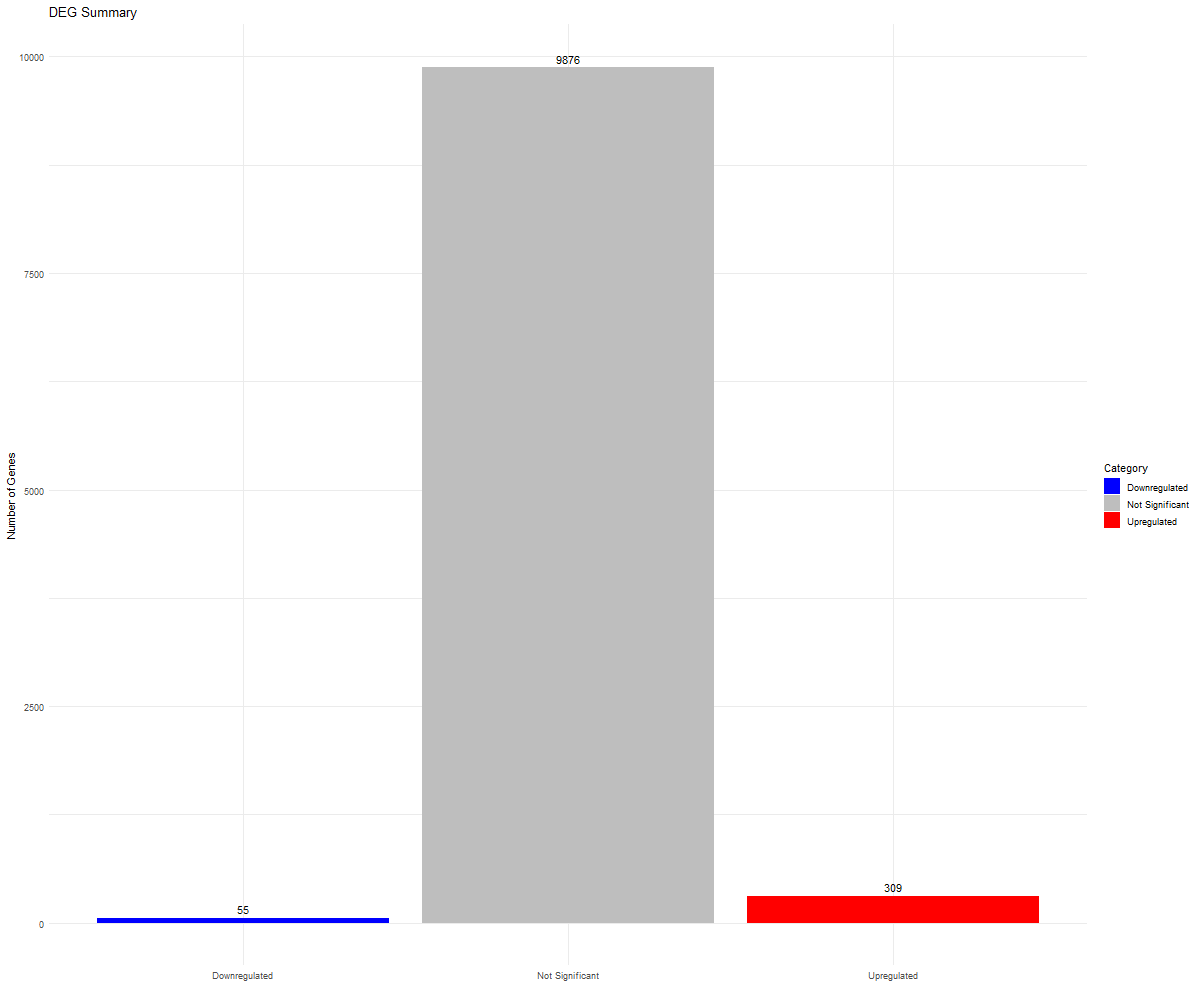
|  |  |
| --- | --- |
| **Category** | **Count** |
| Upregulated | 309 |
| Downregulated | 55 |
| Not Significant | 9876 |

1. **MA Plot**
   * Plotted average log2 expression (A) vs log2 fold change (M) for first two samples to check normalization.

**Image :**

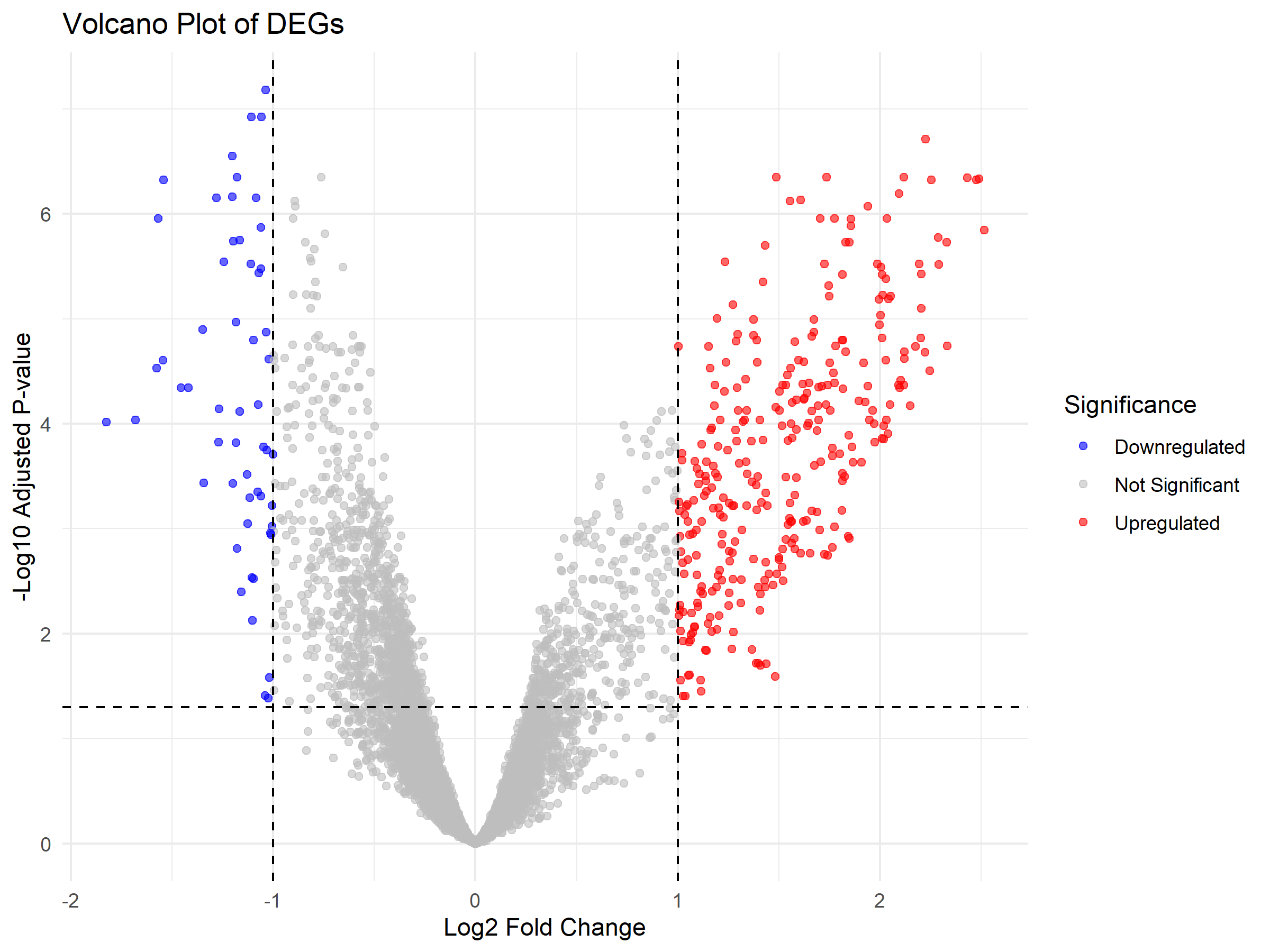
* + MA\_Plot\_Sample1\_vs\_2.jpg

1. **DEG Summary Bar Plot**
   * Visualized counts of Upregulated, Downregulated, and Not Significant genes.

**Image :**

* + DEG\_Summary\_BarPlot.jpg

1. **Volcano Plot**
   * Showed logFC vs -log10(adj.P.Val) highlighting significant DEGs.

**Image :**

* + Volcano\_Plot.png

**3. Files Generated**

1. **CSV Files (Attach with report)**
   * Processed\_Expression\_Data.csv → Normalized expression values
   * DEG\_Results.csv → Differential expression results
2. **PNG Files (Attach with report)**
   * TopVariableProbes\_Heatmap.png
   * MA\_Plot\_Sample1\_vs\_2.jpg
   * DEG\_Summary\_BarPlot.jpg
   * Volcano\_Plot.png (if generated)

**4. Conclusion**

* Normalization and filtering were successful.
* Heatmap shows proper clustering of samples.
* MA and Volcano plots indicate data quality and significant DEGs.
* DEG counts are consistent with expected patterns for this dataset.