B. Tech. Project Report: *Phase* I

Analysis of a gene microarray dataset to analyze differential gene expression in breast cancer cells and normal cells

Submitted in partial fulfillment of requirements
for the award of the degree of Bachelor of Technology from IIT Guwahati

Under the supervision of
Prof. Anil M. Limaye

Submitted by
Sanya
210106060

November, 2024

Department of Biosciences and Bioengineering
Indian Institute of Technology Guwahati
Guwahati 781039, Assam, INDIA

Certificate

This is to certify that the work presented in the report entitled **"**Analysis of a gene microarray dataset to analyze differential gene expression in breast cancer cells and normal cells**"** by **Sanya (210106060)**, represents an original work under the guidance of Prof. Anil M. Limaye. This study has not been submitted elsewhere for a degree.

Signature of student:

_____

Date: Sanya

Place:

Signature of supervisor:

_____

Date: Prof. Anil M. Limaye

Place:

Signature of HOD:

_____

Date:

Place: Head

Department of Biosciences and Bioengineering

Indian Institute of Technology Guwahati

Guwahati, India

Table of contents

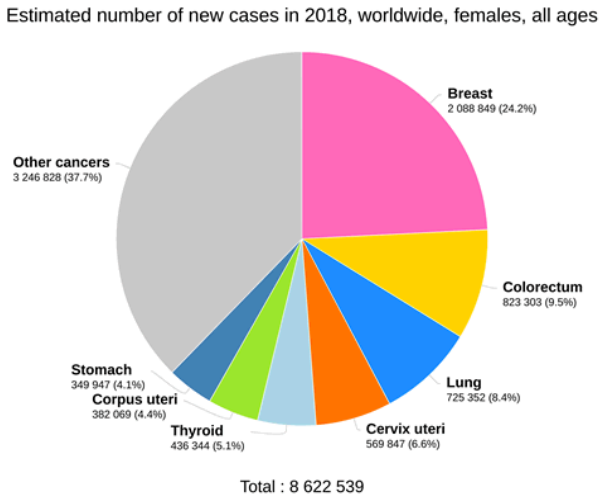
Abstract

Breast cancer is the most prevalent cancer among women worldwide, with a notable increase in incidence across Asia, particularly in younger populations. This study examines differentially expressed genes (DEGs) between cancerous and normal breast tissues using the GSE15852 dataset from a multi-ethnic Malaysian cohort. Employing the limma package for statistical analysis, we identified significant DEGs, providing insights into the molecular distinctions between tumor and normal tissues. Key findings, visualized through various bioinformatics tools, reveal a distinct separation between cancerous and normal samples, strongly supporting the presence of differential expression. These results contribute to understanding breast cancer's molecular basis, potentially aiding in early diagnosis and targeted treatment approaches.

Introduction

Breast cancer is the most common cancer among women globally, accounting for nearly 25% of all cancer cases and with an estimated 2.1 million new cases diagnosed in 2018. Over recent decades, the incidence of breast cancer has risen significantly across Asia, especially in Southeast Asia, including India. In urban regions of India, such as Delhi, Mumbai, Bangalore, and Thiruvananthapuram, breast cancer has surpassed cervical cancer as the leading cancer among women, with age-adjusted incidence rates ranging from 33 to 41 per 100,000 women. Although the age-adjusted incidence rate in India (25.8 per 100,000) remains lower than that of Western countries like the United States (93 per 100,000), India sees a disproportionately higher percentage (46.7%) of cases among women under 50, compared to 19% in the U.S. This trend aligns with broader patterns in Asian populations, where breast cancer incidence in younger women is elevated, likely due to a combination of demographic, genetic, and environmental factors.

Gene expression profiling using microarray technology is an essential method in cancer research, providing insights into the molecular basis of cancer by identifying genes that are differentially expressed between tumor and normal tissues. This study utilized a multi-ethnic Malaysian breast cancer dataset to validate the differential expression of genes in breast tumor tissues compared to normal tissues, laying the groundwork for identifying biomarkers that can improve early diagnosis and personalized treatment options.



Literature Review

The study of differential gene expression in breast cancer is grounded in several key areas:

1. **Gene Expression Profiling in Cancer**

   Gene expression profiling, particularly through microarrays, enables simultaneous quantification of thousands of genes in a single experiment. This allows researchers to identify DEGs that may contribute to tumor initiation, progression, and metastasis.

2. **Breast Cancer Subtypes and Biomarkers**

   Breast cancer is classified into molecular subtypes based on the expression of hormone receptors (ER, PR) and HER2. Identifying DEGs that are specific to each subtype enables the discovery of biomarkers that can inform treatment strategies, particularly for aggressive or treatment-resistant cancers.

3. **Differential Expression Analysis in Multi-ethnic Populations**

   Given the heterogeneity of breast cancer across populations, understanding gene expression in multi-ethnic groups is crucial. This study, conducted on a Malaysian dataset, provides insights into gene expression differences that may reveal ethnicity-specific cancer susceptibilities, offering a step towards personalized medicine.

Objectives

The main objective of this study was to identify differentially expressed genes (DEGs) between cancerous and non-cancerous tissue samples to gain insights into molecular mechanisms driving tumorigenesis. To achieve this, we aimed to:

1. **Select an appropriate dataset** from the Gene Expression Omnibus (GEO) that aligns with our research question, ensuring that it includes labeled cancerous and non-cancerous groups.
2. **Validate dataset quality** by assessing sample size, replicates, metadata richness, and batch effect information to ensure robust statistical analysis.
3. **Perform differential gene expression analysis** using GEO2R, an online tool that streamlines the process for datasets on GEO, leveraging the limma package for statistical comparisons.
4. **Identify and visualize DEGs** that could serve as potential biomarkers or therapeutic targets in cancer.

Materials and methods

Dataset Selection

- Dataset: **GSE15852** from GEO.
- Sample Groups: Cancerous (test) and non-cancerous (control) tissues, selected for direct comparison.
- Metadata: Verified for sample quality, group labels, and absence of significant batch effects.

The GSE15852 dataset was selected due to several reasons that make it suitable for differential gene expression analysis:

- It contains a balanced number of cancerous and non-cancerous tissue samples, enabling a well-powered statistical comparison.
- The dataset includes a large number of genes allowing coverage of potential biomarkers. The high-quality metadata provides essential details on age, race, etc hence subgroups can be statistically examined to detect any population-specific patterns, increasing the generalizability of findings.
- The dataset seemed compatible with well-recognized statistical methods for differential expression analysis.

Data Processing and Analysis

- **Preprocessing:** Log2 transformation and normalization were performed to reduce variance.

- **Group Assignment:** Samples were assigned into test (cancerous) and control (non-cancerous) groups.
- **Statistical Analysis:**
    - Linear modeling and hypothesis testing using the limma package.
    - False Discovery Rate (FDR) adjustment applied for multiple testing corrections.

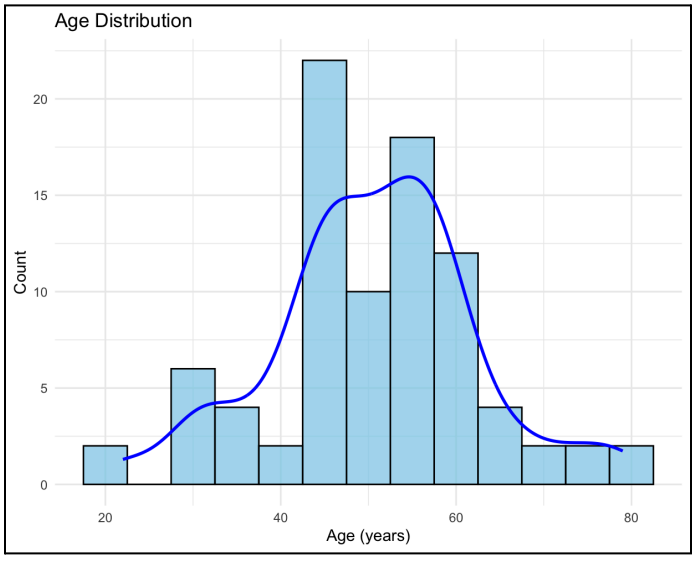
Key Metrics

- Adjusted p-value < 0.05 to determine significance.
- Log fold change threshold |log2FC| > 1 for biologically meaningful DEGs.

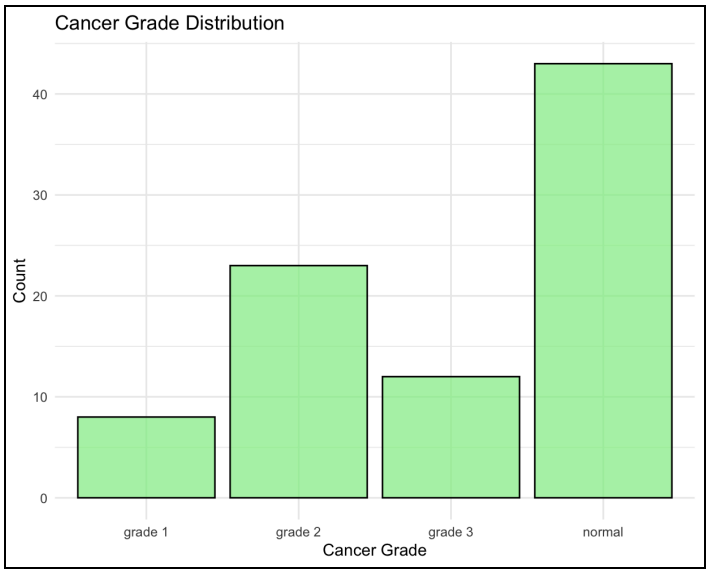
Results and Discussion

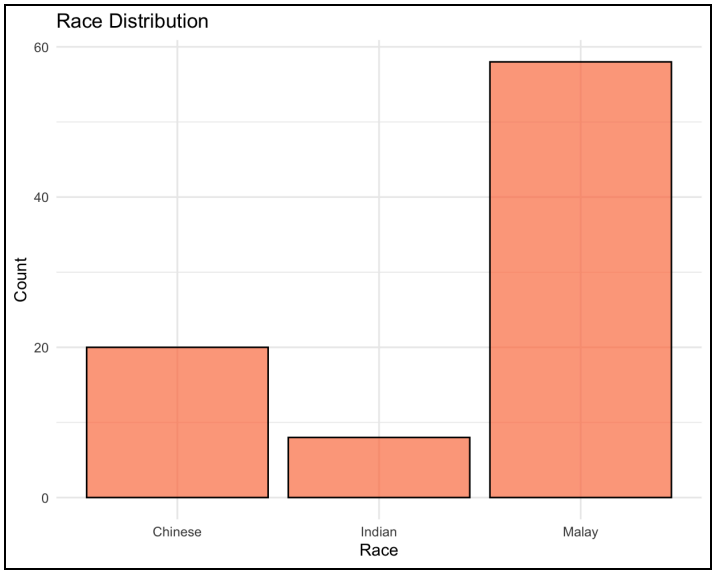
Our hypothesis was that specific genes would be differentially expressed between breast cancer tissues and normal tissues, and this differential expression could serve as potential biomarkers for breast cancer detection and treatment. The results of our analysis, visualized through various bioinformatics tools, strongly support this hypothesis.

Features of the dataset

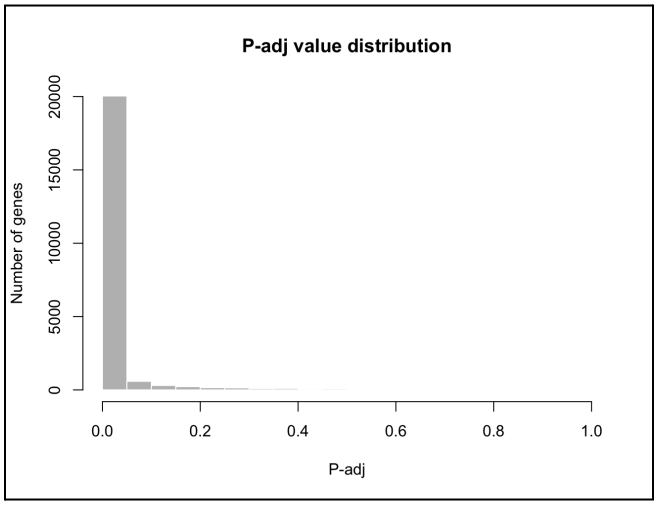


The age distribution is centered around 40-60 years, with fewer individuals at younger and older extremes. This middle-aged focus may highlight age-related trends in the dataset.



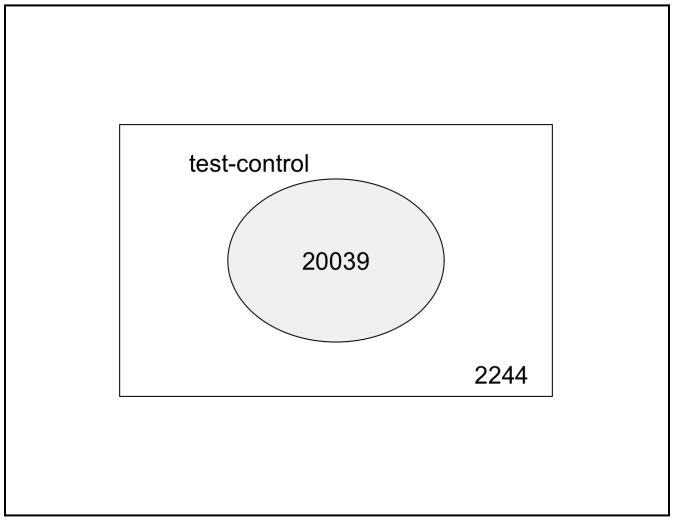


Analysing the dataset

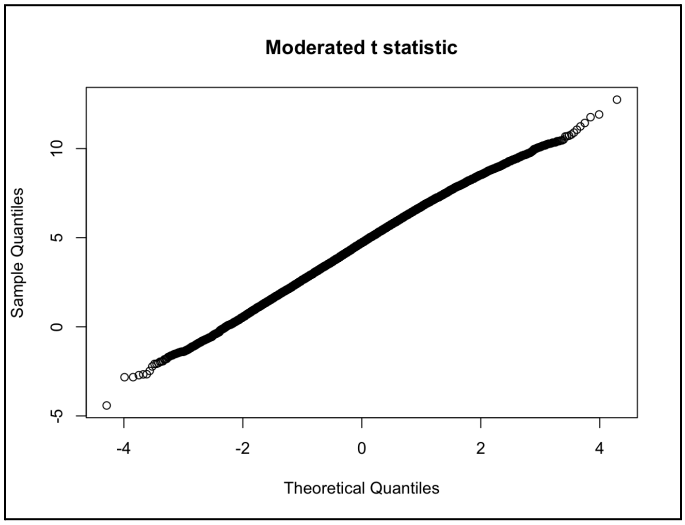


This histogram of p-adjusted values shows a high frequency of genes with low p-adjusted values. The concentration of genes with low p-adjusted values (<0.05) implies that there is a substantial number of differentially expressed genes between the conditions after adjusting for multiple testing. This indicates statistically significant differences in expression.

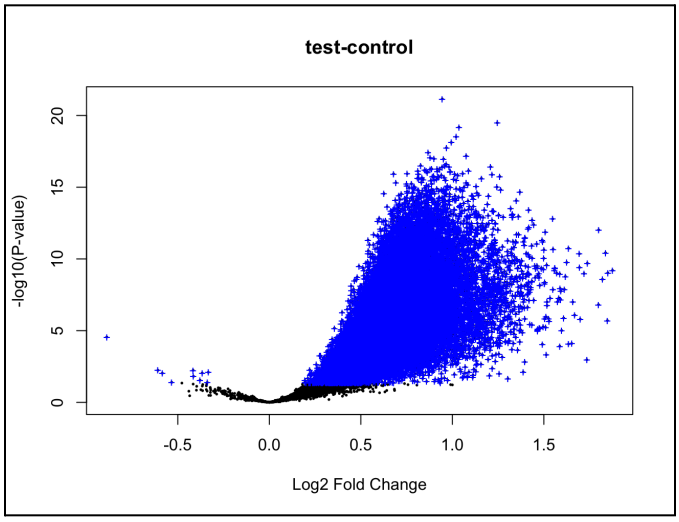
.



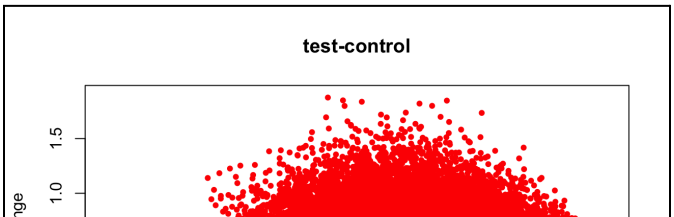
The venn diagram displays the number of differentially expressed genes unique to test and control conditions and those common to both. A large overlap (20,039 genes) indicates that these genes are expressed in both conditions, while the unique counts indicate condition-specific gene expression.



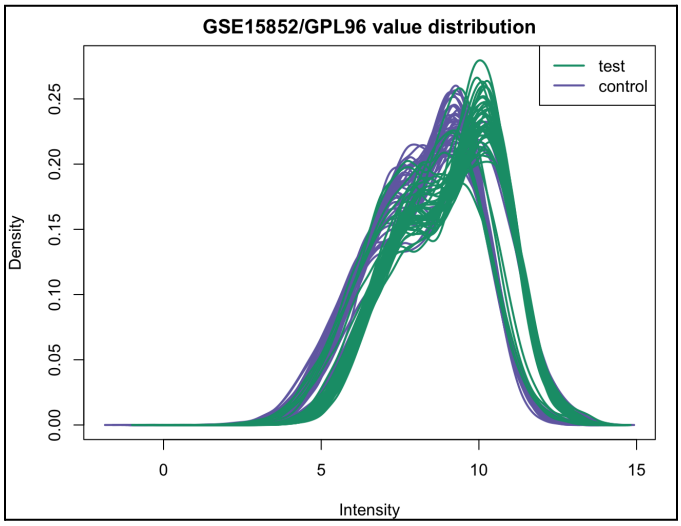
Q-Q plot compares the observed distribution of t-statistics to a theoretical normal distribution. Points along the diagonal line suggest a normal distribution of test statistics, while deviations at the extremes suggest that some genes exhibit strong differential expression (either upregulated or downregulated), validating the significance of these outliers.



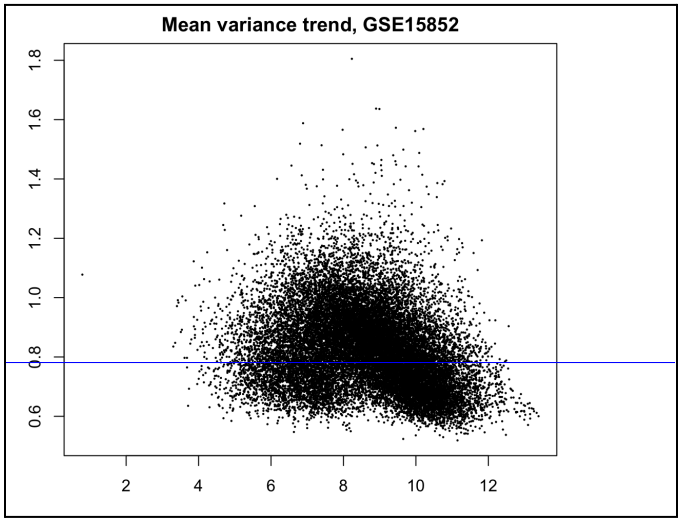
The volcano plot shows the -log10(p-value) against the log2 fold change for each gene, highlighting genes with significant changes. Genes with a high fold change and significant p-values are located at the extremes, making them potential candidates for further investigation due to their substantial differential expression.



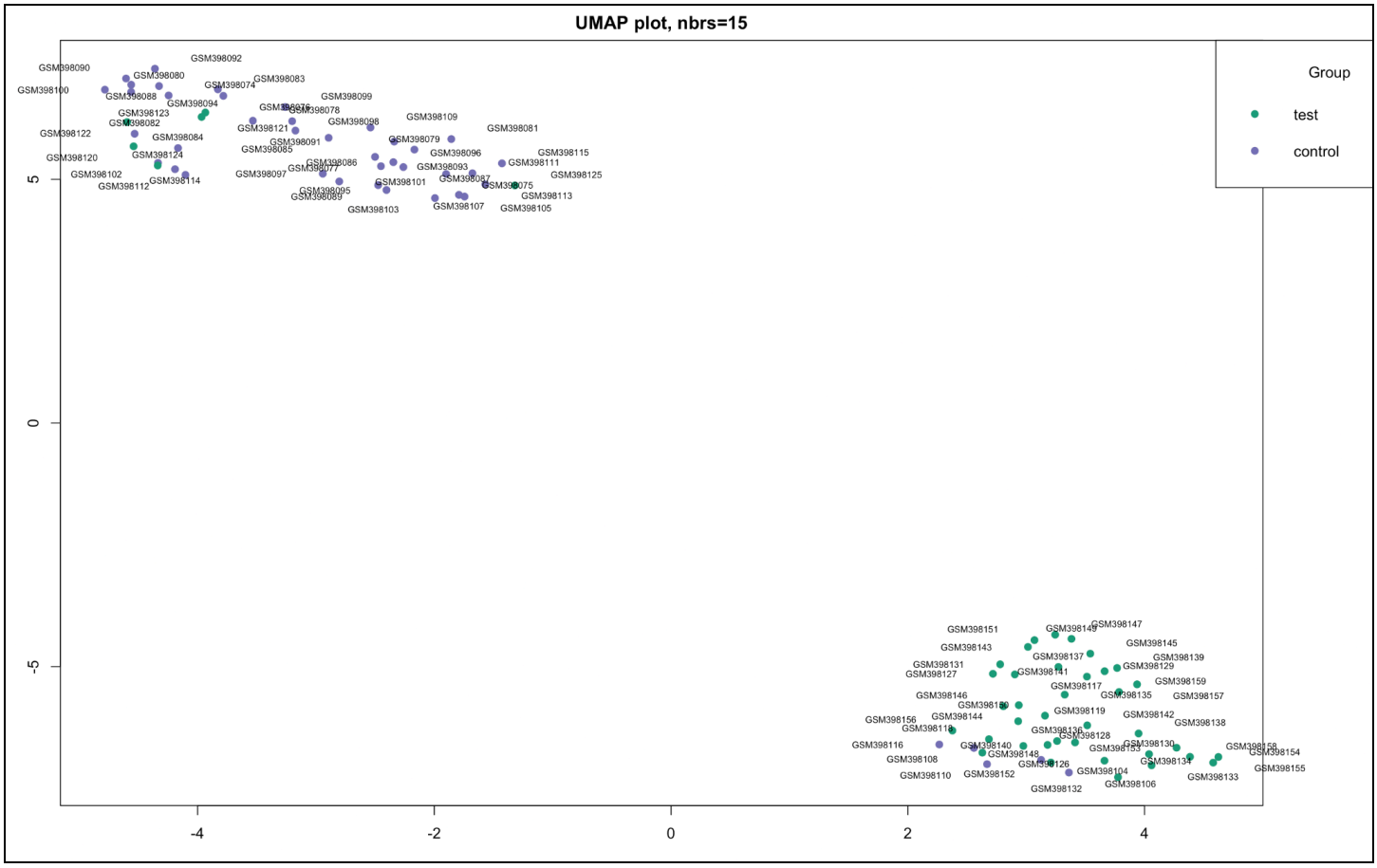
The MD plot shows the log-fold changes versus the average log-expression for each gene. The red points indicate genes that are significantly upregulated, while the blue points represent those significantly downregulated. This plot helps in visualizing which genes are differentially expressed between the test and control groups. A large number of genes are upregulated (above 0 on the y-axis), suggesting an overexpression of certain genes in the test group.



This plot shows the density distribution of expression intensities for both the test and control groups.
Both test and control samples have similar distributions, suggesting that they are comparable in terms of general expression levels. Minor deviations may indicate differential expression for specific genes rather than overall expression shifts.



Mean variance trend plot shows the relationship between mean expression and variance for each gene.
The trend line (blue line) indicates the expected mean-variance relationship. A deviation from this trend may suggest technical variations or biological variability in the data. Here, the variance tends to stabilize as the mean expression increases, suggesting a consistent expression pattern across genes.



The UMAP plot provides a visual clustering of the test and control groups based on their gene expression profiles.

- The plot shows two distinct clusters, with green points representing the "test" group and purple points representing the "control" group.
- Samples within each cluster are grouped closely together, suggesting that samples in each condition (test or control) have similar gene expression profiles.
- The separation between clusters indicates that there is a clear difference in the gene expression patterns between the test and control groups, supporting the presence of differentially expressed genes between the conditions.
- This separation could signify underlying biological differences, such as distinct responses to treatment or differences in disease status, between the groups.
- The tight clustering within each group also implies consistent data quality, as samples within each group show minimal variance in the UMAP dimensions.

Conclusion & Future Work

In conclusion, our analysis supports the initial hypothesis that specific genes are differentially expressed between breast cancer tissues and normal tissues, suggesting potential biomarkers for breast cancer detection and treatment. The clear separation of test and control groups in the UMAP visualization and the statistically significant differences in gene expression identified by differential expression analysis indicate distinct molecular profiles. These findings provide a promising foundation for further research to validate the identified biomarkers and explore their functional roles in cancer progression.

For future work, we can identify specific genes among those differentially expressed that could serve as biomarkers for early breast cancer detection and targeted treatment. Additionally, we plan to validate these biomarkers through clinical studies, potentially leading to personalized treatment approaches. Machine learning models could be incorporated to enhance the predictive accuracy of these biomarkers by learning patterns within complex gene expression data. Furthermore, integrating other data types, such as proteomics and metabolomics, may provide a more comprehensive view of breast cancer biology and improve the robustness of our findings.

References

1. Irizarry, R. A., et al. "Summaries of Affymetrix GeneChip probe level data." *Nucleic Acids Research*, 31(4): e15.
2. Smyth, G. K. "Limma: Linear Models for Microarray Data." *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*.
3. Robinson, M. D., et al. "EdgeR: A Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics*, 26(1): 139-140.