# Introduction

Breast cancer is one of the most common cancers affecting women worldwide and is a major cause of cancer-related mortality. The development of breast cancer is closely associated with abnormal gene regulation that drives uncontrolled cell growth and tumor progression. Gene expression profiling allows researchers to study molecular changes between normal and cancerous tissues by measuring mRNA abundance across thousands of genes simultaneously. Identifying these molecular differences can help discover potential biomarkers and therapeutic targets for cancer diagnosis and treatment.

In this study, gene expression data from GSE15852 were analyzed. This dataset contains microarray-based expression profiles from breast tumor tissues and matched normal breast tissues collected from Malaysian patients of different ethnic groups. The dataset consists of 43 tumor samples and 43 normal samples analyzed using the Affymetrix GeneChip U133A platform. The main objective of this analysis is to identify differentially expressed genes (DEGs) between tumor and normal tissues using statistical modeling approaches. The limma (Linear Models for Microarray Data) method was applied because it provides robust statistical testing for microarray gene expression data by combining linear modeling with empirical Bayes variance estimation.

This analysis aims to improve understanding of gene expression changes associated with breast cancer development and provide insight into molecular mechanisms underlying tumor formation.

# Methods

1. Dataset Description

The gene expression dataset used in this study was obtained from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database under accession number GSE15852. The dataset contains microarray-based gene expression profiles from human breast cancer tissues and matched normal breast tissues. The data were generated using the Affymetrix GeneChip U133A microarray platform, which measures genome-wide mRNA expression levels using probe hybridization technology. The dataset includes a total of 86 samples, consisting of 43 tumor tissue samples and 43 paired normal tissue samples. RNA was extracted from tissue samples, processed, and hybridized to microarray chips for 16 hours before signal detection and data acquisition.

## 2. Software and Package Preparation

All analyses were performed using the R programming environment. Several bioinformatics packages from Bioconductor and CRAN repositories were installed and loaded for data retrieval, preprocessing, statistical analysis, and visualization. The GEOquery package was used to download and import gene expression data from GEO, while the limma package was used for differential gene expression analysis. The hgu133a.db annotation package was used to map probe identifiers to gene symbols and gene names. Visualization of data distributions and expression patterns was performed using ggplot2 and pheatmap, while the umap package was used for dimensionality reduction analysis.

The following R packages were used:

```
library(GEOquery)
library(limma)
library(pheatmap)
library(ggplot2)
library(dplyr)
library(hgu133a.db)
library(AnnotationDbi)
library(umap)
```

## 3. Data Download and Import

The dataset was downloaded directly from the GEO database using the GEOquery function. The data were imported as an ExpressionSet object, which contains three main components: the gene expression matrix, sample metadata, and probe annotation information. This structure allows efficient downstream statistical analysis.

```
gset <- getGEO("GSE15852", GSEMatrix = TRUE, AnnotGPL = TRUE)[[1]]
```

## 4. Data Preprocessing

Data preprocessing was conducted to improve data quality and reduce technical bias. Gene expression values were first extracted from the ExpressionSet object using the exprs() function. Because microarray data can contain skewed signal intensity values, log2 transformation was evaluated and applied if necessary. Log transformation helps stabilize variance across expression levels and improves statistical modeling accuracy.

```
ex <- exprs(gset)

qx <- as.numeric(quantile(ex,
c(0,0.25,0.5,0.75,0.99,1),
na.rm = TRUE))

LogTransform <- (qx[5] > 100) ||
(qx[6] - qx[1] > 50 && qx[2] > 0)

if (LogTransform) {
ex[ex <= 0] <- NA
ex <- log2(ex)
}
```

## 5. Sample Group Classification

Sample grouping information was obtained from the dataset metadata. Samples were categorized into tumor and normal tissue groups using the source_name_ch1 metadata column. The group variable was converted into a factor variable because statistical models in R treat categorical variables differently from numerical variables.

```
group_info <- pData(gset)[["source_name_ch1"]]
groups <- make.names(group_info)
gset$group <- factor(groups)
```

## 6. Experimental Design and Contrast Matrix Construction

A design matrix was constructed to formally represent the experimental structure of the study. The design matrix is necessary for linear modeling because it specifies which samples belong to tumor and normal groups. A contrast matrix was then created to define the specific comparison between tumor and normal tissues.

```
design <- model.matrix(~0 + gset$group)
colnames(design) <- levels(gset$group)

contrast_formula <- "breast.tumor.tissue - normal.breast.tissue"

contrast_matrix <- makeContrasts(
contrasts = contrast_formula,
levels = design
)
```

## 7. Differential Expression Analysis

Differentially expressed genes were identified using the limma (Linear Models for Microarray Data) statistical framework. Limma is widely used for microarray analysis because it applies linear modeling combined with empirical Bayes variance moderation, which improves statistical reliability when analyzing high-dimensional genomic data. Linear models were fitted to each gene, and statistical significance was evaluated using moderated t-statistics. Genes were considered significantly differentially expressed if the adjusted p-value was less than 0.01.

```
fit <- lmFit(ex, design)
fit2 <- contrasts.fit(fit, contrast_matrix)
fit2 <- eBayes(fit2)

topTableResults <- topTable(
fit2,
adjust = "fdr",
sort.by = "B",
number = Inf,
p.value = 0.01
)
```

## 8. Gene Annotation

Significant probes were annotated to corresponding gene symbols and gene names using the hgu133a.db database. Gene annotation allows biological interpretation of statistical results by linking probe-level data to known gene functions.

```
probe_ids <- rownames(topTableResults)

gene_annotation <- AnnotationDbi::select(
hgu133a.db,
keys = probe_ids,
columns = c("SYMBOL", "GENENAME"),
keytype = "PROBEID"
)

topTableResults$PROBEID <- rownames(topTableResults)

topTableResults <- merge(
topTableResults,
gene_annotation,
by = "PROBEID",
all.x = TRUE
)
```

## 9. Data Visualization

Several visualization techniques were used to evaluate data quality and biological patterns. Boxplots and density plots were used to examine expression distribution across samples.

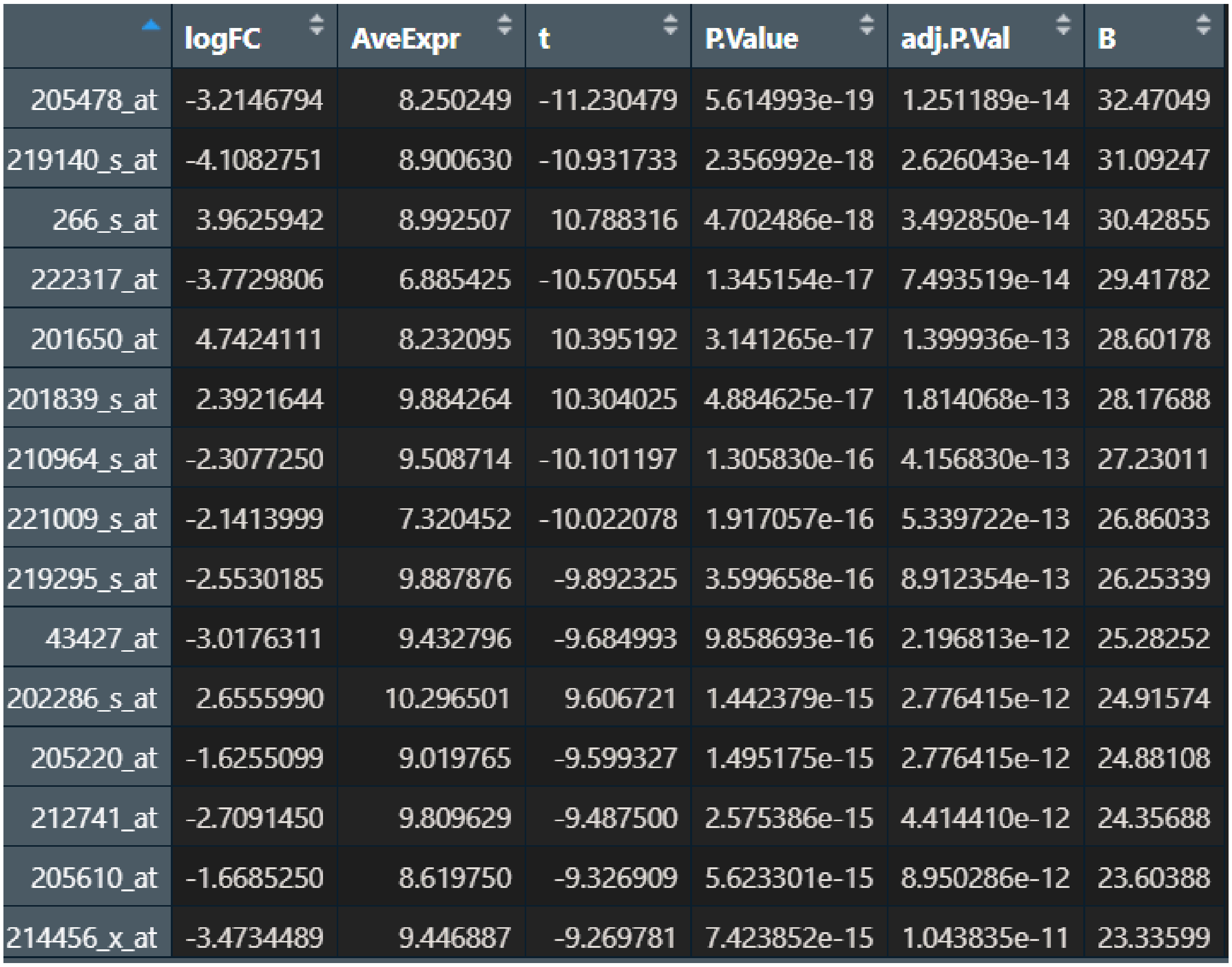
Uniform Manifold Approximation and Projection (UMAP) was applied to visualize clustering patterns between tumor and normal samples. Additionally, volcano plots and heatmaps were generated to visualize differentially expressed genes and expression patterns.

## 10. Data Export

The final list of differentially expressed genes was saved as a CSV file for downstream analysis and reporting.

```
write.csv(topTableResults, "GSE15852_DEG_results.csv")
message("Analysis completed. DEG results file has been saved.")
```

## Results and Discussion



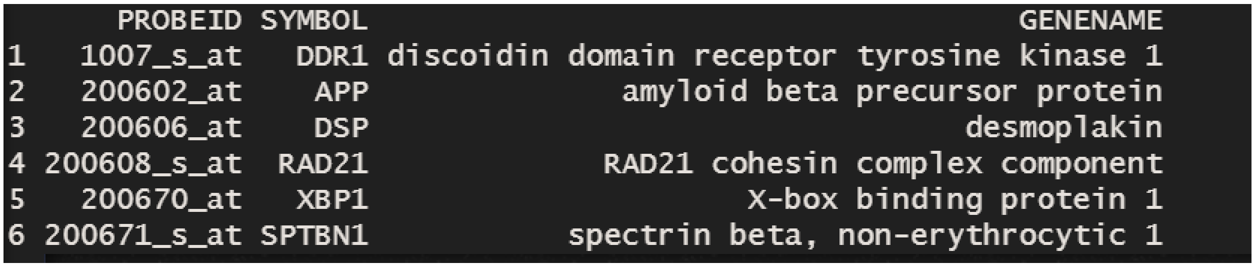
| | logFC | AveExpr | t | P.Value | adj.P.Val | B |
|---|---|---|---|---|---|---|
| 205478_at | -3.2146794 | 8.250249 | -11.230479 | 5.614993e-19 | 1.251189e-14 | 32.47049 |
| 219140_s_at | -4.1082751 | 8.900630 | -10.931733 | 2.356992e-18 | 2.626043e-14 | 31.09247 |
| 266_s_at | 3.9625942 | 8.992507 | 10.788316 | 4.702486e-18 | 3.492850e-14 | 30.42855 |
| 222317_at | -3.7729806 | 6.885425 | -10.570554 | 1.345154e-17 | 7.493519e-14 | 29.41782 |
| 201650_at | 4.7424111 | 8.232095 | 10.395192 | 3.141265e-17 | 1.399936e-13 | 28.60178 |
| 201839_s_at | 2.3921644 | 9.884264 | 10.304025 | 4.884625e-17 | 1.814068e-13 | 28.17688 |
| 210964_s_at | -2.3077250 | 9.508714 | -10.101197 | 1.305830e-16 | 4.156830e-13 | 27.23011 |
| 221009_s_at | -2.1413999 | 7.320452 | -10.022078 | 1.917057e-16 | 5.339722e-13 | 26.86033 |
| 219295_s_at | -2.5530185 | 9.887876 | -9.892325 | 3.599658e-16 | 8.912354e-13 | 26.25339 |
| 43427_at | -3.0176311 | 9.432796 | -9.684993 | 9.858693e-16 | 2.196813e-12 | 25.28252 |
| 202286_s_at | 2.6555990 | 10.296501 | 9.606721 | 1.442379e-15 | 2.776415e-12 | 24.91574 |
| 205220_at | -1.6255099 | 9.019765 | -9.599327 | 1.495175e-15 | 2.776415e-12 | 24.88108 |
| 212741_at | -2.7091450 | 9.809629 | -9.487500 | 2.575386e-15 | 4.414410e-12 | 24.35688 |
| 205610_at | -1.6685250 | 8.619750 | -9.326909 | 5.623301e-15 | 8.950286e-12 | 23.60388 |
| 214456_x_at | -3.4734489 | 9.446887 | -9.269781 | 7.423852e-15 | 1.043835e-11 | 23.33599 |

Figure 1. Differential Expression Results (Limma Output Table)

The differential expression analysis identified 33 significant differentially expressed genes (DEGs) using the limma statistical framework. Genes were considered significant when the adjusted p-value was less than 0.01, reducing the likelihood of false positive results caused by multiple testing.

The logFC (log fold change) value indicates the magnitude and direction of gene expression differences. Negative logFC values indicate genes that are downregulated in tumor tissues, while positive logFC values indicate upregulated genes. For example, a logFC value of −3.21 indicates strong downregulation in tumor tissues, corresponding to approximately an eight-fold decrease in expression.

The t-value measures statistical strength, where larger absolute values indicate stronger evidence of differential expression. The P-value represents probability by chance, while the adjusted P-value corrects for multiple testing errors. The B-value represents the log odds that a gene is truly differentially expressed, with higher values indicating greater biological confidence.



```
      PROBEID SYMBOL                                     GENENAME
1   1007_s_at   DDR1 discoidin domain receptor tyrosine kinase 1
2   200602_at    APP             amyloid beta precursor protein
3   200606_at    DSP                                 desmoplakin
4 200608_s_at  RAD21            RAD21 cohesin complex component
5   200670_at   XBP1                      X-box binding protein 1
6 200671_s_at SPTBN1          spectrin beta, non-erythrocytic 1
```

Figure 2. Gene Annotation Results

Gene annotation was performed to convert probe IDs into gene symbols and gene names for biological interpretation. Several important cancer-related genes were identified.

Genes such as CDH1, EPCAM, and KRT19 were observed to have higher expression in tumor tissues, suggesting roles in epithelial cancer cell proliferation. In contrast, ADIPOQ and LPL showed higher expression in normal tissues, indicating possible involvement in normal metabolic regulation and tumor suppression mechanisms.

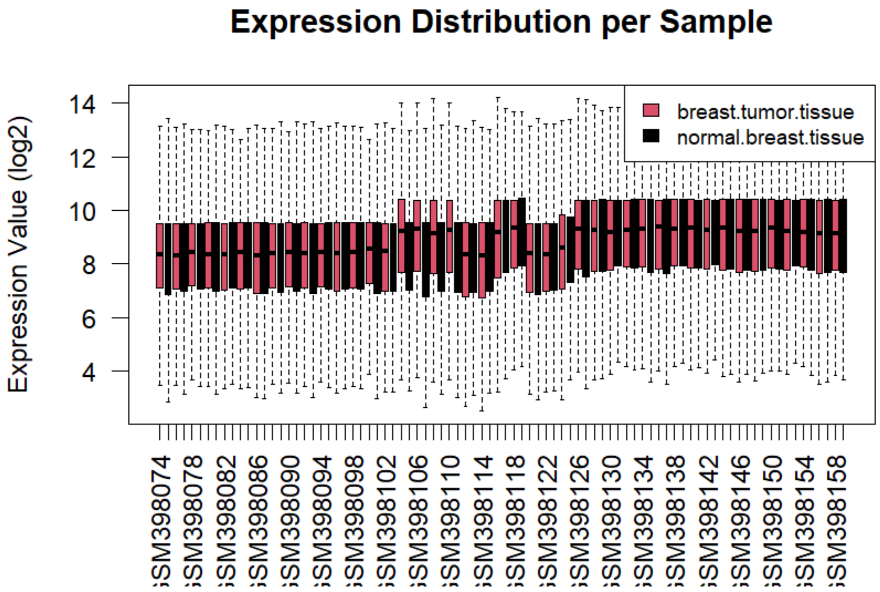


Figure 3. Expression Distribution Boxplot

The boxplot shows gene expression distribution across samples after log2 transformation and normalization. Each box represents the distribution of gene expression values within one sample. The horizontal line inside each box represents the median expression value, while the box itself represents the middle 50% of gene expression values.

The similar median lines across samples indicate that normalization was successful. Normalization is important because it removes technical bias and ensures observed differences are due to biological variation rather than experimental noise.

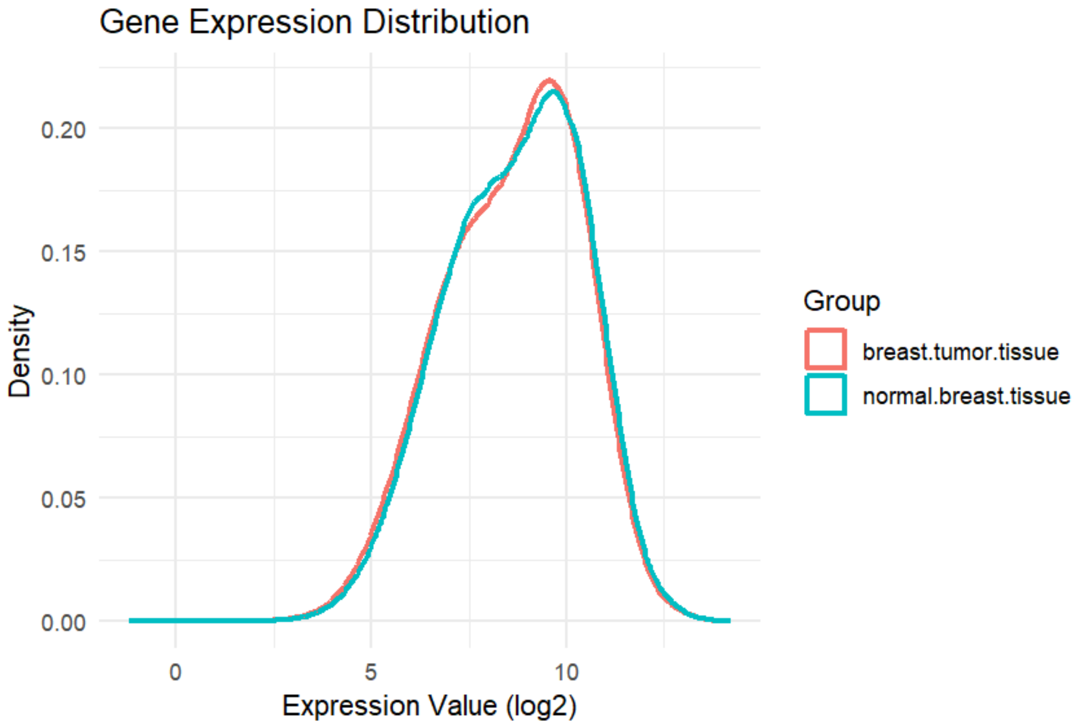


Figure 4. Gene Expression Density Plot

The density plot shows the overall distribution of gene expression values across samples. Most expression values were concentrated between log2 expression values of 7 to 10, indicating moderate to high gene expression signal intensity.

Log2 transformation was applied to stabilize variance and improve statistical modeling performance by reducing extreme outlier values and making the distribution more symmetrical.

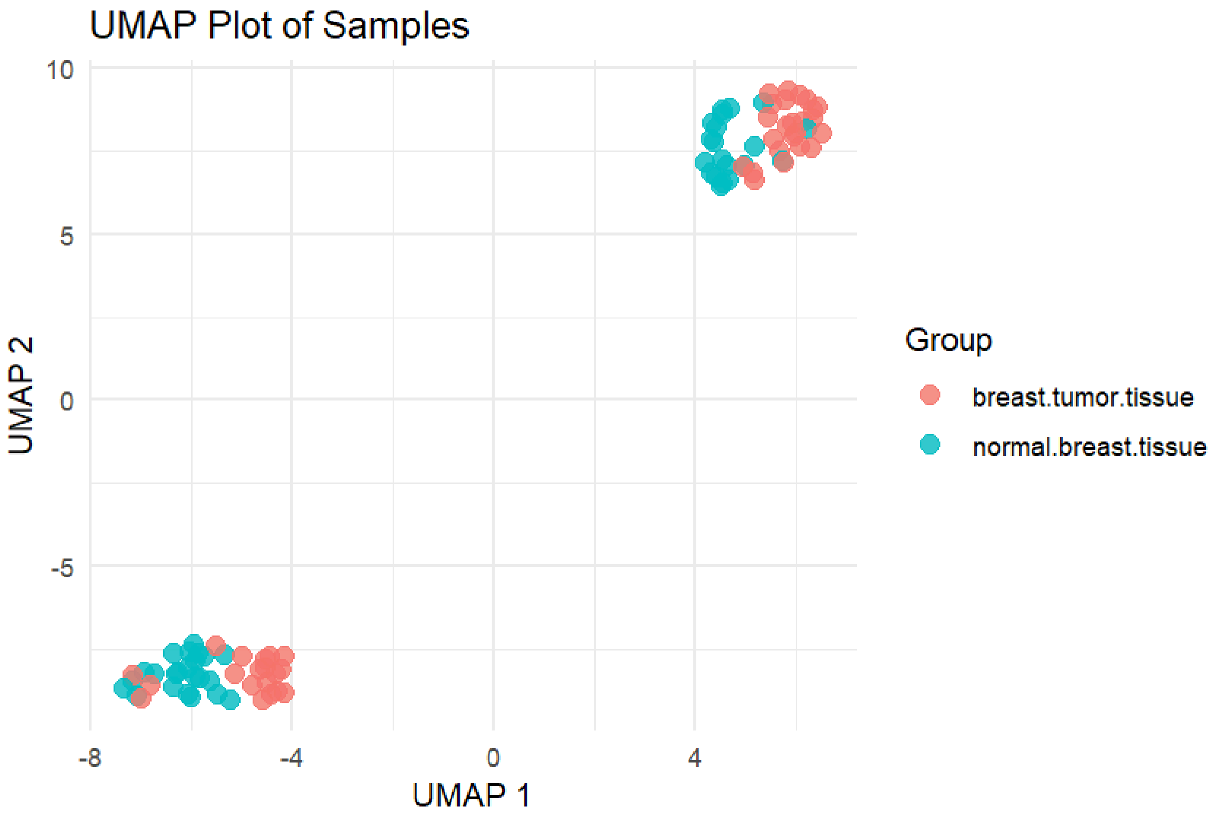


Figure 5. UMAP Sample Clustering Plot

The UMAP plot shows dimensionality reduction of gene expression data to visualize sample clustering patterns. Two major clusters were observed around UMAP1 values of approximately -6 and +5, indicating that the dataset contains meaningful biological variation.

However, tumor and normal samples showed partial overlap. This suggests that although there is biological separation between groups, tumor samples may exhibit heterogeneity. Some tumor samples may have gene expression patterns closer to normal tissue, which could be related to tumor subtype differences or disease stage variability.

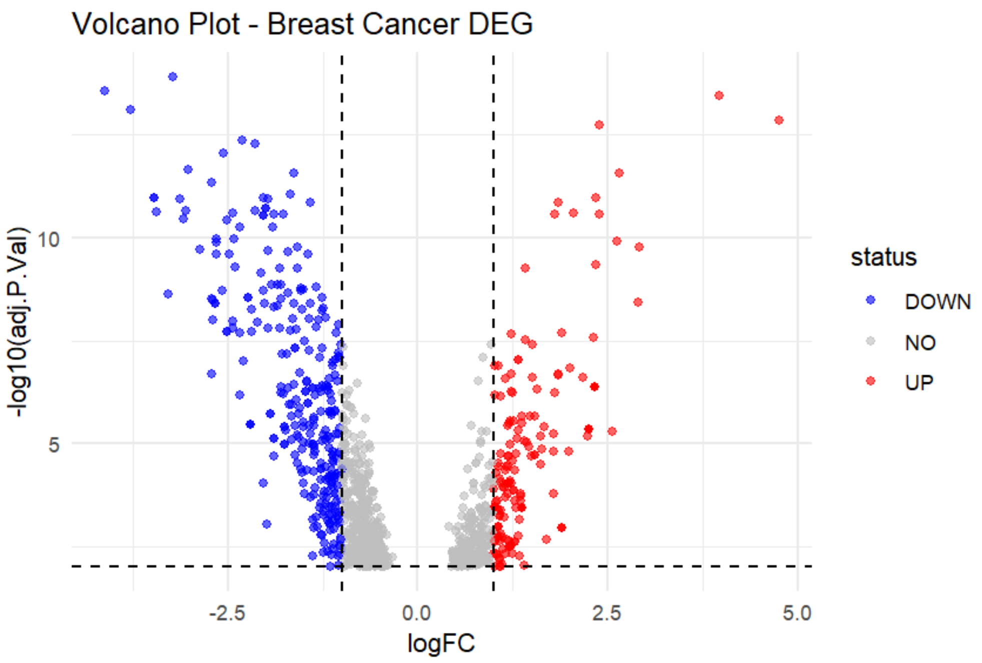


Figure 5. Volcano Plot of Differentially Expressed Genes

The volcano plot visualizes statistical significance versus magnitude of gene expression change. The x-axis represents log2 fold change, while the y-axis represents −log10 adjusted p-value.

Genes located farther from zero on the x-axis show stronger biological expression differences. Genes located higher on the y-axis have stronger statistical significance.

In this plot:

- Red dots represent upregulated genes in tumor tissue
- Blue dots represent downregulated genes in tumor tissue
- Grey dots represent genes without significant changes

The presence of many significant genes on both sides of the plot indicates strong molecular differences between tumor and normal tissues.

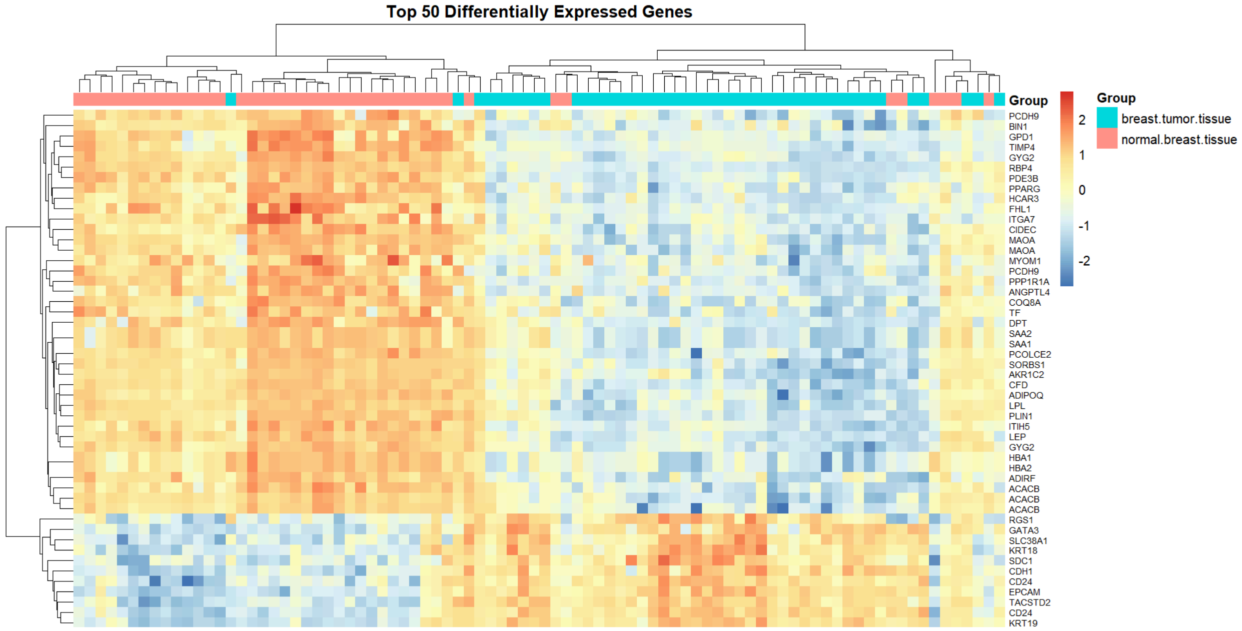


Figure 6. Heatmap of Top 50 Differentially Expressed Genes

The heatmap shows expression patterns of the top 50 DEGs across all samples. Each row represents a gene, and each column represents a sample. Red color indicates high expression, while blue color indicates low expression.

The clustering dendrograms show similarity relationships:

- Tumor samples tend to cluster together, indicating similar gene expression signatures.
- Normal samples also cluster together, suggesting consistent biological expression patterns.

Specific genes showed distinct expression patterns:

- CDH1, EPCAM, KRT19 were highly expressed in tumor samples.
- ADIPOQ and LPL were highly expressed in normal tissues.

Nevertheless, in this image, the colors are somewhat mixed between breast tumor tissue shown in blue and normal shown in red. This means that some tumor samples are clustering closer to certain normal samples than to other tumor samples. In other words, based on the expression of these top 50 genes, some tumor samples (around two-three) are more similar across groups than expected. There are several possible reasons for this, which is mainly biological variability. Not all tumor samples are identical. Some tumors may be molecularly closer to normal tissue, especially if they are less aggressive or represent specific subtypes (one of them is Luminal A is a subtype that is usually less aggressive and hormone receptor positive). On the other hand, there are many normal samples overlapping with the tumor samples, the overlap may suggest possible early molecular changes in tissue, but further clinical and experimental validation is required to confirm early cancer development.

## Conclusion

In conclusion, this study successfully analyzed gene expression differences between breast tumor tissues and normal breast tissues using the GSE15852 microarray dataset. The analysis identified 33 significant differentially expressed genes (DEGs), showing that breast cancer development is strongly associated with changes in gene regulation. The limma statistical method was effective in detecting these differences by using linear modeling and empirical Bayes variance estimation to improve statistical reliability.

The results showed that several genes were strongly associated with breast cancer. Some genes such as CDH1, EPCAM, and KRT19 were highly expressed in tumor tissues, suggesting roles in cancer cell growth and epithelial cell activity. In contrast, ADIPOQ and LPL were more highly expressed in normal tissues, suggesting they may be involved in normal metabolic functions and may have protective roles against tumor development.

Data visualization results supported the statistical findings. The boxplot and density plot showed that gene expression data were properly normalized and suitable for analysis. The UMAP plot showed two main sample clusters, but some overlap between tumor and normal samples was observed, indicating biological variability between tumor samples. The volcano plot showed many genes with strong statistical significance and large expression changes. The heatmap showed clear clustering patterns, where tumor samples generally grouped together and normal samples grouped together, indicating distinct gene expression signatures.

Overall, this study demonstrates that microarray gene expression analysis is useful for identifying molecular changes in breast cancer. The identified genes may serve as potential biomarkers for future cancer research. Further experimental validation and functional studies are recommended to confirm the biological roles of these genes in breast cancer development.

Github:
https://github.com/Angelo9453/Bioinformatics-Research-Starter-Program/tree/main/BRSP%20Work/Week%204-Final%20Project