Analysis of microarray gene expression data from 251 patients presenting with breast cancer

Matthew Finster
Master of Data Science student
LaTrobe University

October 2024

Contents

1 Background

Breast cancer remains one of the most common types of cancer for women worldwide. In fact, the number of cases has been increasing in Australia in recent years, with over 21,000 new cases expected in 2024[1], making it the second most common type of cancer suffered by Australians.

Despite medical advances, breast cancer still proves difficult to treat effectively. Evidence suggests that different sub-types of breast cancer exist, which has led to the use of multi-modal therapy in situations where these sub-types are unknown or unidentified in patients.

Establishing a patient's breast cancer sub-type early in their treatment can reduce their treatment cost, and could potentially improve the efficacy of their treatment and their prognosis. If there are molecularly different subtypes of breast cancer, personalised medicine could potentially be developed. This analysis therefore seeks to detect various sub-types of breast cancer; understand which genes contribute to these sub-types and to estimate prognoses related to each sub-type. Insights gained from this research can contribute to a personalised approach to cancer detection and treatment in the future.

To conduct the analysis, two corresponding datasets were used. The primary dataset consists of clinical and single-channel microarray gene expression data from 251 patients presenting with breast cancer, recorded as logged intensity values. Each row corresponds to a probe, including several dozen AFFX quality-control probes. Each column corresponds to a patient.

The second dataset consists of various clinical measures for each patient. Each of these measures have been shown to be related to the survival outcomes of patients, as described in Table 1.

Variable	**Description**	**Relevance**
sampleID	Patient's ID	Identification variable.
histgrade	Histological tumour grade	Higher grades (G3) are associated with poorer survival outcomes than lower grades (G1 or G2).[2]
ERstatus	Estrogen receptor status	ER+ status has been associated with better survival compared to ER- status in some studies.[3]
PRstatus	Progesterone receptor status	PR+ status has been associated with better survival outcomes compared to PR- status, in some studies.[4][5]
age	Patient's age in years	Logically, older age is associated with worse survival outcomes.
tumor size mm	Size of the cancer tumour in mm	Smaller tumor size is strongly associated with better survival outcomes.[6]
LNstatus	Lymph node metastasis status	LN- is strongly associated with better survival outcomes.[7]
Surv time	Survival time of the patient	Outcome variable.
event	Indicates if the patient died during the study	Outcome variable.

Table 1: Clinical data variables and their relevance to survival outcomes in breast cancer patients

2 Methods

A cluster analysis was conducted to determine whether different sub-types of breast cancer were present among the gene expression data. Prior to this cluster analysis, pre-processing was performed. The gene expression dataset contained 68 AFFX quality-control probes, which were removed because they do not carry biological relevance.

```r
## -- Pre-processing: Removing the control probes in the expression data
filteredData <- expressionData[!startsWith(rownames(expressionData), "AFFX"), ]
dim(filteredData) # 22215 genes, 251 patients
```

Listing 1: Removal of AFFX quality-control probes

Since the microarray data consists of single-channel arrays (rather than dual-channel), within-array normalisation was not necessary. However, to account for possible technical variability that might have arisen during the microarray experiment or sample processing (and that could therefore obscure true biological differences), between-array normalisation was performed using the `normalizeBetweenArrays()` function from the `limma` package. Specifically, the scale method was used, which adjusts the standard deviation of each array to make them comparable while preserving the mean expression levels. This is important because the mean expression levels might be biologically relevant when finding differences between individuals and thus potential subtypes of breast cancer. Before clustering, the data was further processed by standardising the expression values of each gene across all individuals using the `scale()` function from the base R package.

```r
## -- Pre-processing: normalising the expression data
normalisedData <- normalizeBetweenArrays(filteredData, method = "scale")
## -- Pre-processing: scaling the expression data
scaled.expressionData <- t(scale(t(normalisedData), center = TRUE))
```

Listing 2: Normalisation of data

This ensured that each gene had approximately the same variance across patients. The `scale()` function standardises each gene according to the following formula:

$$x_{\text{scaled}} = \frac{x - \bar{x}}{s}$$

After ensuring that no single gene would disproportionately influence the clustering process, the Euclidean distances between individuals' scaled gene expression data were calculated using the `dist()` function, which uses the following formula:

$$d(A, B) = \sqrt{\sum_{i=1}^{k} (A_i - B_i)^2}$$

where k is the number of genes, and A_i and B_i represent the scaled expression values of the i-th gene for individuals A and B, respectively.

```r
## -- Euclidean distances between each individual's gene expression data
distances <- dist(t(scaled.expressionData))
```

Listing 3: Calculating distances between individuals' gene expression data

Using the Euclidean distances between the 251 patients, hierarchical clustering was conducted using the `hclust()` function in the `cluster` package in R. Different linkage techniques (Ward; single; complete; average) were trialled to group the individuals. Dendrogram plots were generated to view these clusters.

```r
## -- Trialling various clustering techniques
clusteringWard <- hclust(distances, method = "ward.D2")
clusteringSingle <- hclust(distances, method = "single")
clusteringComplete <- hclust(distances, method = "complete")
clusteringAverage <- hclust(distances, method = "average")
# -- Plot example for Ward linkage
plot(clusteringWard, cex = 0.5, main = "Ward linkage", sub = "", xlab = "", ylab = "Distances")
```

Listing 4: Clustering

In hierarchical clustering, each of these linkage methods treats the 251 patients as their own cluster. Each method then merges clusters slightly differently (refer to Table 2), which becomes more apparent after a few clusters have already been merged.

Linkage method	Description
Single	Merges the two clusters that have the smallest minimum distance between any two points in the clusters.
Complete	Merges the two clusters where the maximum distance between any two points is the smallest.
Average	Merges the two clusters where the average distance between all points in each cluster is the smallest.
Ward	Merges the pair of clusters that results in the smallest increase in the total within-cluster variance.

Table 2: Description of different linkage methods

Each linkage type produced hierarchical clusters, where the number of clusters ranged from 1 to 251. Neither extreme of 1 cluster (all breast cancer patients in one group) or 251 clusters (each individual in their own cluster) is useful in identifying meaningful biological sub-types of breast cancer. Therefore, to determine the optimal number of clusters, a number of steps were taken, including silhouette analyses and the production of heatmaps and Principal Component Analysis (PCA) plots.

To conduct the silhouette analyses of each linkage type, the `silhouette()` function in R was used. This function helps in identifying the optimal number of clusters as it measures how similar a point is to its own cluster compared to other clusters. The silhouette of observation A_i is defined as:

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}} \in [0, 1]$$

where a_i is the average distance between i and all other points in the same cluster and b_i is the minimum average distance between i and the points in the nearest neighboring cluster. Larger silhouette values s_i (which result from a larger b_i value and a smaller a_i value) indicate better separation, as points within a cluster are more similar to each other, while points between clusters are more dissimilar. After calculating the silhouette values for each of the 251 individuals, the median silhouette value was taken as a measure of cluster quality. These median silhouette values for 2, 3, 4 and 5 clusters were then compared, with larger median silhouette values indicating more well-defined clusters.

```
K <- 2:5
silW <- NULL
## -- Example of the Ward silhouette analysis
for (i in K) {
  silW <- c(silW, median(silhouette(cutree(clusteringWard, k = i),dist = distances)[,3],na.rm =
      T))}
plot(K, silW, type = "l", main = "Ward linkage", xlab = "Number of clusters", ylab = "Median
    silhouette value")
## -- Example of obtaining optimal clusters for Ward linkage
clW <- cutree(clusteringWard, k = K[which.max(silW)])
```

Listing 5: Example of silhouette analysis using Ward linkage (silhouette analyses using other linkage types not shown here)

After the optimal numbers of clusters were obtained according to the silhouette analysis, a Principal Component Analysis (PCA) was conducted using the `princomp()` function from the `stats` package. A PCA plot, which reduced the dimensionality of the data to the first two principal components (PC1 and PC2), was generated to visualise how well the clusters were separated for each of the four linkage types.

```
# -- PCA plot example using Ward linkage
par(bg = "white")
pcaW <- princomp(scaled.expressionData[, ])
plot(pcaW$loadings[, 1:2], col = clW)
title("Ward linkage (k = 3)")
```

Listing 6: PCA plot example using Ward linkage (PCA plots using other linkage types not shown here)

Subsequently, to better understand the biological relevance of these clusters, heatmaps were also produced. To produce heatmaps of gene expression data, the top 5% most highly variable genes (n = 1111) were extracted. The `rowVars()` function from the `genefilter` Bioconductor package was used to calculate the variance for each gene prior to extraction. Theoretically, if the expression of these 1111 genes was similar within clusters, but different between clusters, then the clusters could be considered to have biological relevance. Heatmaps of these 1111 genes were produced using the `heatmap.2()` function from the `gplots` package.

```
## -- Finding the 5% of genes with the highest variance using the unscaled gene expression data
22215/100*5
rowVariances <- rowVars(normalisedData)
highVarianceGenes <- order(-rowVariances)[1:1111]
## -- Example of heatmap generation using Ward linkage
colsW <- c("blue", "purple", "orange")
head(cbind(colnames(normalisedData), colsW))
heatmap.2(scaled.expressionData[highVarianceGenes,],labCol = clW, trace = "none", ColSideColors
    = colsW[clW], col = redgreen(100), Colv = as.dendrogram(clusteringWard))
```

Listing 7: Heatmap example using Ward linkage (heatmaps using other linkage types not shown here)

After assessing the biological relevance of the clusters via heatmaps, a differential gene expression (DE) analysis was conducted using the `limma` package. To determine which genes were differentially expressed at a statistically significant level between the clusters, a design matrix was specified that corresponded with the clusters identified from the previous steps. Next, the gene expression data was fitted to the linear model using the `lmFit()` function. The contrasts between the different clusters were then fitted to this linear model using the `contrasts.fit()`. Empirical Bayes was then applied using the `eBayes()` function to estimate a posterior distribution for the variances of each gene, improving the reliability of detecting DE genes. Following this, the `qvalue()` function from the `qvalue` package was used to identify genes that were differentially expressed with an FDR (False Discovery Rate) adjusted q-value of ≤ 0.05. The FDR correction reduces the chance of false positives when testing multiple genes at once. Volcano plots were generated using the `volcanoplot()` function to highlight genes with both large fold changes and low p-values. The number of statistically significant DE genes were counted for each contrast based on the identified q-values. The top 100 genes (sorted by q-value) were also extracted for each contrast, ready for the next step in the analysis.

```
# -- Specifying a design matrix
design <- model.matrix(~0 + as.factor(clW))
colnames(design) <- c("A", "B", "C")
# -- Constructing DE object
DE.object <- lmFit(normalisedData, design)
# -- Making contrasts between Cluster 1 and Cluster 2, then reorienting model to this contrast
AvsB.contrast <- makeContrasts(BvsA = B - A, levels = design)
AvsB.DE.object <- contrasts.fit(DE.object, AvsB.contrast)
# -- Performing Empirical Bayes on each contrast
AvsB.DE.object <- eBayes(AvsB.DE.object)
# -- Volcano plots of each contrast
volcanoplot(AvsB.DE.object, main = "Cluster 1 vs Cluster 2 contrast")
abline(h = -log10(0.01), col = "red", lty = 2)
abline(v = c(-1, 1), col = "red", lty = 2)
# -- Obtaining DE genes with qvalue <= 0.05
AvsB.qval <- qvalue(AvsB.DE.object$p.value[,1], fdr.level = 0.05)
## -- Number of DE genes
sum(AvsB.qval$qvalues < 0.05)
## -- Top 100 DE genes by q-value
top_AvsB_qvals <- AvsB.DE.object[order(AvsB.qval$qvalues)[1:100], ]
```

Listing 8: Example of finding DE genes between Cluster 1 & 2 (comparisons between other clusters not shown here)

Following the DE analysis, Gene Ontology (GO) and KEGG pathway analyses were performed to identify the biological processes and pathways associated with the DE genes in the identified sub-types (clusters) of breast cancer. Probe IDs for the previously identified top 100 DE genes (sorted by q-values) were extracted. An annotation file was used to map these probe IDs to Ensembl gene IDs. These Ensembl gene IDs were then mapped to Entrez IDs, using the `select()` function from the `org.Hs.eg.db` package, which provides access to a database of human gene information. Any duplicated Entrez IDs were then removed, as well as any NA values. With the Entrez IDs for the top DE genes between each cluster, GO enrichment analyses were undertaken using the `goana()` function from the `limma` package, focusing on biological processes (BP). Similarly, KEGG pathway enrichment analyses were conducted using the `kegga()` function from the `limma` package

to identify enriched pathways. For both analyses, the top 20 enriched GO terms and KEGG pathways were reported for each contrast using the `topGO()` and `topKEGG()` functions, in order to provide an overview of the most significant biological processes and pathways associated with the DE genes across the clusters.

```
## -- Reading the annotation data
annotationData <- readRDS("STA5MB_2024_BC_annotations.RDS")
## -- Getting entrez IDs for top 100 DE genes by q values
# Cluster 1 vs 2
top_AvsB_probeIDs <- rownames(top_AvsB_qvals)
AvsB_ensemblIDs <- annotationData[annotationData$affy_hg_u133_plus_2 %in% top_AvsB_probeIDs, c(
    "ensembl_gene_id")]
homo_sapiens <- org.Hs.eg.db
AvsB_entrezIDs <- select(
  homo_sapiens,
  keys = AvsB_ensemblIDs,
  columns = c("ENTREZID", "ENSEMBL"),
  keytype = "ENSEMBL"
)
AvsB_entrezIDs <- unique(na.omit(AvsB_entrezIDs$ENTREZID))
## -- GO for each contrast
GO_AvsB <- goana(AvsB_entrezIDs, species='Hs')
GO_top_AvsB <- topGO(GO_AvsB, n=20, ontology='BP')
## -- KEGG for each contrast
KEGG_AvsB <- kegga(AvsB_entrezIDs, species='Hs')
KEGG_top_AvsB <- topKEGG(KEGG_AvsB, n=20)
```

Listing 9: Example of GO and KEGG analyses for DE genes between Cluster 1 & 2 (comparisons between other clusters not shown here)

Lastly, a survival analysis was conducted to assess whether individuals with potentially different breast cancer subtypes (as identified by their cluster memberships) had varying survival outcomes. DE genes that were significant across all three contrasts were identified using the `intersect()` function. Gene scores were then calculated by summing the normalised expression values of these significant genes for each individual, using the `colSums()` function. A survival object was created using the `Surv()` function from the `survival` package, using the *surv_time* and *event* variables from the clinical dataset. Prior to applying a Cox proportional hazards model, convergence issues were encountered due to 9 unknown *LNstatus* values and 4 unknown *ERstatus* values. Since all 13 of these values did not have an event (death), these were merged into groups that are associated with better survival outcomes in breast cancer[3][7]. This meant that `"ER?"` values were assigned an `"ER+"` status while `"LN?"` values were assigned a `"LN-"` status.

```
# -- Identifying significant genes common to all three contrasts
significant_genes <- intersect(intersect(rownames(AvsB.DE.object)[AvsB.qval$significant],
                                         rownames(BvsC.DE.object)[BvsC.qval$significant]),
                               rownames(AvsC.DE.object)[AvsC.qval$significant])
## -- Calculating gene scores by summing across these common significant genes for each sample
gene.scores <- colSums(normalisedData[significant_genes, ])
## -- Creating a survival object (using Surv_time and event from the clinical data)
Y <- Surv(time = clinicalData$Surv_time, event = clinicalData$event)
## -- Merging unknown ER? and LN? values into the ER+ and LN- groups.
clinicalData$ERstatus[clinicalData$ERstatus == "ER?"] <- "ER+"
clinicalData$LNstatus[clinicalData$LNstatus == "LN?"] <- "LN-"
```

Listing 10: Survival analyses for Cluster 1 vs Cluster 2 vs Cluster 3 (grouped survival analysis not shown here)

A Cox proportional hazards regression model was applied to the individuals, testing various combinations of all other clinical variables (*age*, *tumor size*, *histological grade*, *PR status*, *ER status*, and *LN status*) as covariates, given their known impact on breast cancer survival (see Table 1 for further details). Since no alternative combination of these variables resulted in a statistically significant improvement in the model's performance, all covariates were included in the final model to ensure robustness. Cluster 2 was set as the reference level as it was the most dissimilar to the other two clusters. A second Cox model was also fitted after combining Cluster 1 and Cluster 3 into a single group to explore whether grouping these clusters would reveal clearer survival differences. Lastly, to understand and visualise survival outcomes for different cluster memberships, Kaplan–Meier (K-M) survival estimates were generated using the `survfit()` function from the `survival` package. K-M survival curves were visualised using the `autoplot()` and `labs()` functions from the `ggfortify` and `ggplot2` packages, respectively. All results can be found in the next section.

```r
## -- Changing the reference level to cluster 2
clinicalData$clW <- relevel(as.factor(clW), ref = "2")
## -- Fitting the Cox proportional hazards model
cox.modelRobust <- coxph(Y ~ gene.scores + as.factor(clW) + age + as.factor(histgrade) + tumor_size_mm + as.factor(PRstatus) + as.factor(ERstatus) + as.factor(LNstatus), data = clinicalData)
## -- Kaplan-Meier survival curves
km_fit <- survfit(Y ~ clW, data = clinicalData)
## -- Plotting the Kaplan-Meier curves
autoplot(km_fit)+labs(x = "\n Survival time (years) since diagnosis ", y = "Survival probabilities",
                      title = "K-M survival curves for breast cancer patients (Clusters 1 vs 2 vs 3) \n", color = "Cluster Group", fill = "Cluster Group")
```

Listing 11: Survival analyses for Cluster 1 vs Cluster 2 vs Cluster 3 (grouped survival analysis not shown here)

3 Results

Each of the four linkage methods (Ward, single, complete and average) clustered the 251 individuals in different ways, as expected. An initial inspection of dendrograms (Figure 1) showed that Ward linkage produced more evenly distributed clusters compared the other methods.

The optimal number of clusters for each linkage type, determined by the largest median silhouette values, are displayed in Figure 2 (Ward = 3; single = 2; complete = 5; average = 2). The larger median silhouette values for single and average linkage ($s_i \approx 0.24$) suggest that using these methods to separate the individuals into 2 clusters would produce the most well-defined clusters compared to Ward linkage for three clusters ($s_i \approx 0.035$) or complete linkage for 5 clusters ($s_i \approx 0.014$). However, using only 2 clusters with single or average linkage would group 250 individuals into one cluster and 1 individual (85A03) into the other. This could be a problem, or it could be important if 85A03 has a unique or rare subtype of breast cancer that the other 250 individuals do not have. Further analysis via inspection of PCA plots and heatmaps is generally required before making decisions on the relevance of such clusters.

To further understand how well the clusters were separated for each of the four linkage methods; PCA plots were generated. For single and average linkage, one individual (85A03) was positioned in the centre of the PCA plot, surrounded by the other 250 individuals. This suggests that the single outlier was not distinctly separated during the clustering process. The PCA plots of the clusters generated by Ward and complete linkage indicated that these methods showed better separation between their clusters, although a significant amount of overlap between individuals from different clusters was still present. Although it does appears that using three clusters (Ward linkage) or five clusters (complete linkage) produced the most well-defined clusters; PCA reduces the dimensionality of the data to only the first two principal components (PC1 and PC2) that explain the most variability. Therefore, it is possible that the clusters are better separated along other principal components not captured in these plots. Therefore, like the silhouette analyses results; the PCA results alone should not be used as the final determinant to decide the number of clusters present or the linkage method to group these clusters. See Figure 3 for more details.

To then further understand the biological relevance of the clustering; heatmaps for the top 5% most variable genes were generated and analysed. The Ward method appears to have divided individuals into clusters in the most biologically relevant way, as indicated by commonly under- and over-expressed genes of individuals within its clusters. Figure 4 shows the relative expression of the top 5% variable genes clustered using Ward linkage, while Figure 5 is an annotated version of this same heatmap. Heatmaps for the clusters generated by the other linkage methods can be found in the Appendix, along with additional notes.

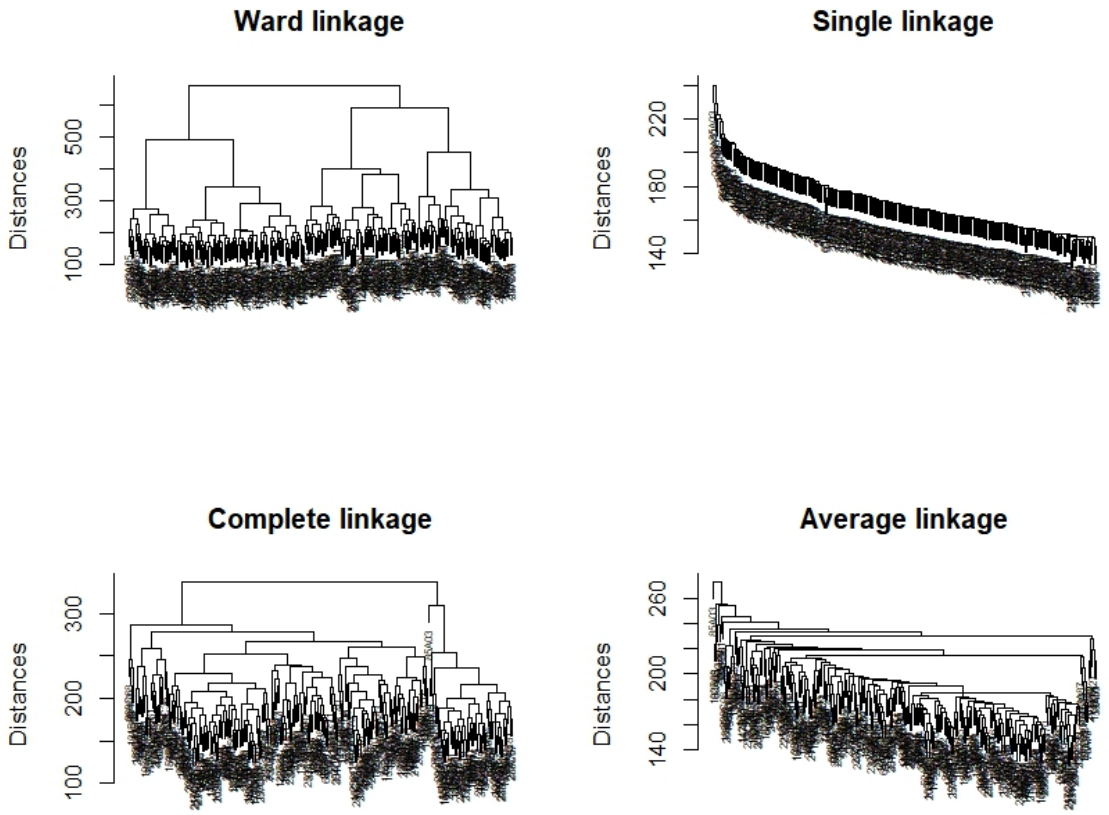


Figure 1: Comparison of different linkage techniques used to cluster the 251 individuals by their gene expression data

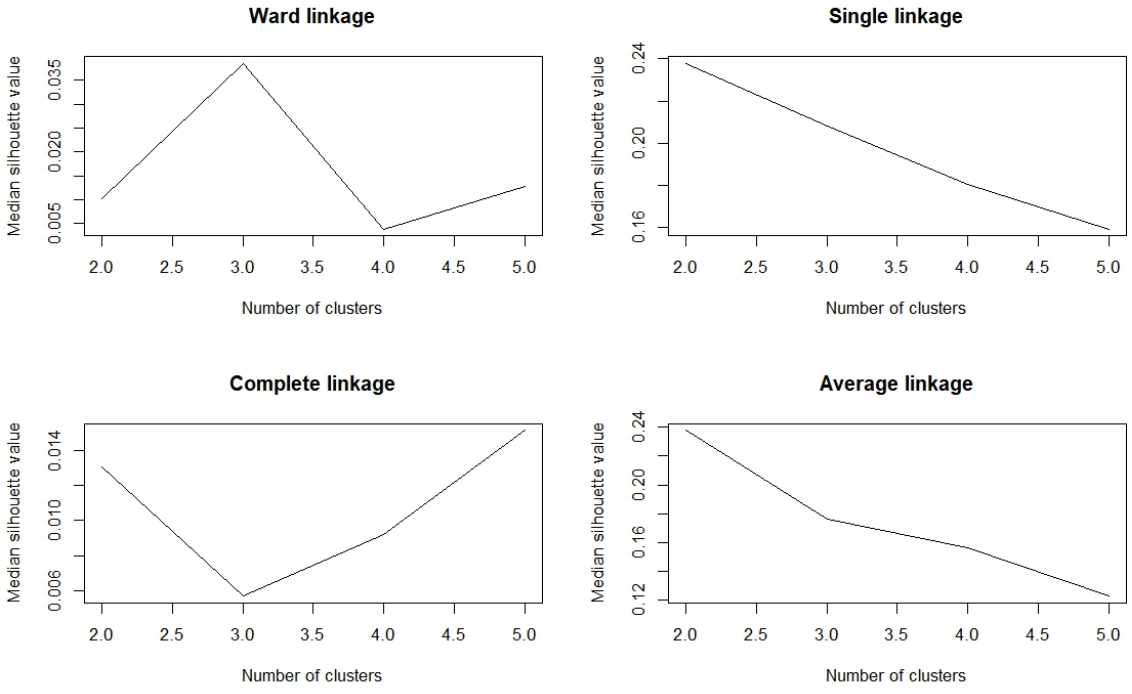


Figure 2: Median silhouette values for each linkage type and different numbers of clusters

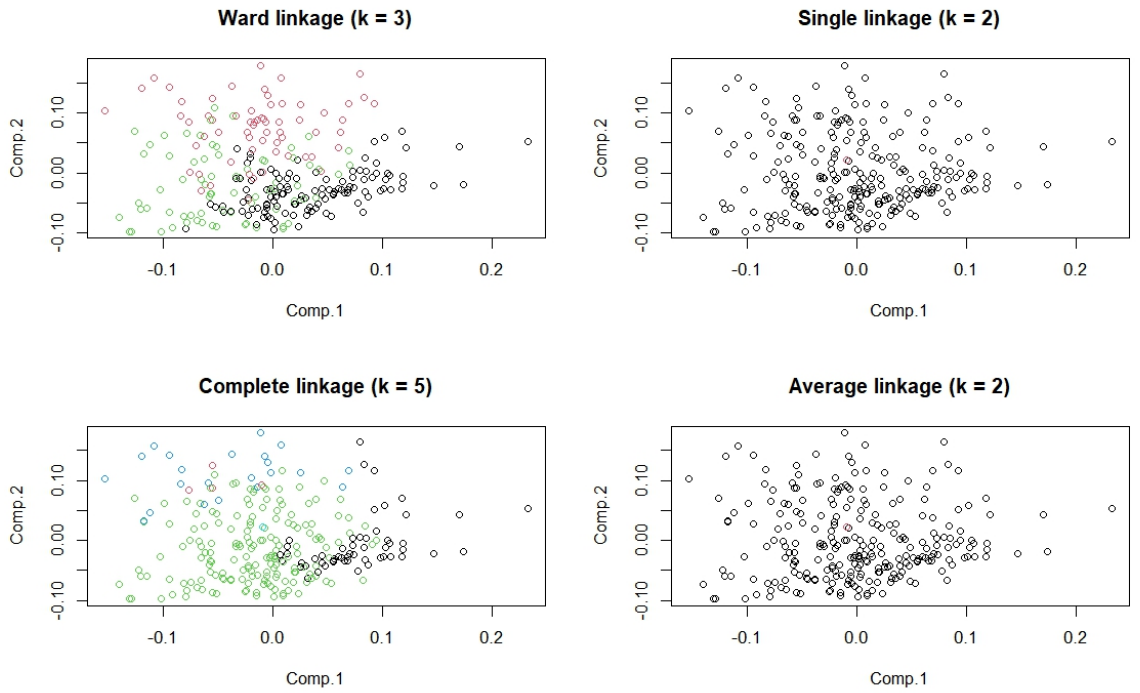


Figure 3: PCs 1 and 2 of breast cancer data, coloured by clusters

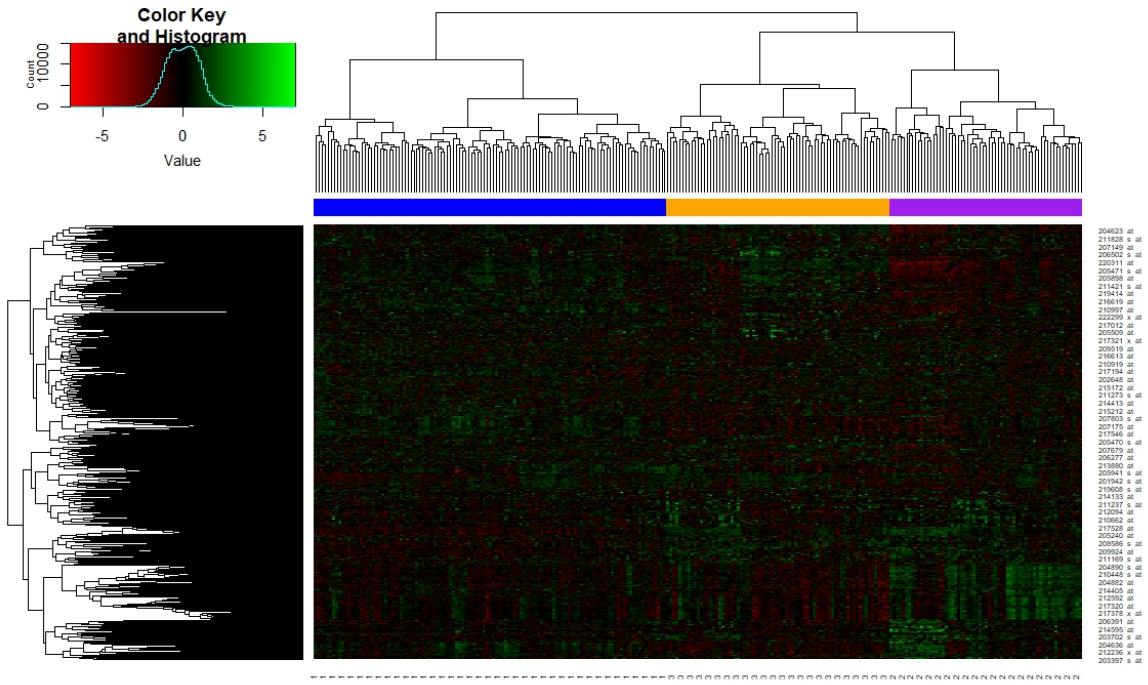


Figure 4: Heatmap for the top 5% (n = 1111) most variable genes, by cluster (k = 3), using Ward linkage

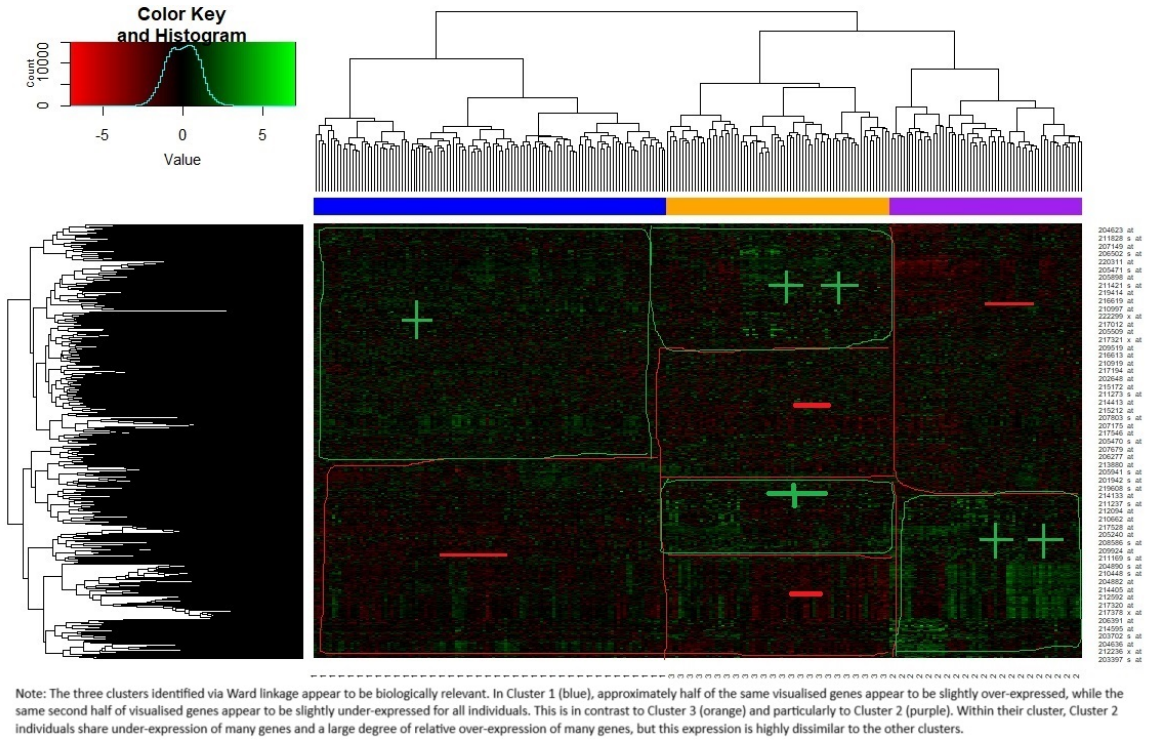


Figure 5: Annotated heatmap for the top 5% (n = 1111) most variable genes, by cluster, using Ward linkage

Due to the similarity in gene expression within clusters and the dissimilarity between clusters, as illustrated by the heatmap, the three clusters identified using the Ward linkage method were used to conduct a DE analysis. The DE analysis revealed that each cluster contrast contained a substantial number of statistically significant DE genes. The number of these statistically significant DE genes, along with the percentages relative to the entire dataset, can be found in Table 3. The top five differentially expressed genes for each contrast, ranked by their q-values, can be found in Table 4. Volcano plots, displayed in Figure 6, further highlight these genes that were statistically significant and which also contained biologically meaningful fold changes. Interestingly, as shown in both Table 4 and Figure 6, the log fold change (logFC) values among the statistically significant genes are greater when Cluster 2 is involved in the contrast, indicating that this cluster is substantially different from the other two.

Contrast	**Number of significant DE genes**	**Percentage of total genes (22,215)**
Cluster 1 vs Cluster 2	11,303	50.9%
Cluster 2 vs Cluster 3	9,671	43.5%
Cluster 1 vs Cluster 3	11,855	53.4%

Table 3: Number of significant DE genes (q-value ≤ 0.05) and their percentage of the total (22,215) for each cluster comparison.

Cluster 1 vs Cluster 2			Cluster 2 vs Cluster 3			Cluster 1 vs Cluster 3		
Gene	**logFC**	**q-value**	**Gene**	**logFC**	**q-value**	**Gene**	**logFC**	**q-value**
209408_at	0.86	7.26×10^{-28}	205529_s_at	1.58	1.48×10^{-27}	206134_at	-0.75	9.99×10^{-25}
202410_x_at	-2.09	3.59×10^{-26}	203880_at	1.46	7.59×10^{-25}	201693_at	0.71	5.25×10^{-24}
206134_at	1.54	3.59×10^{-26}	201694_at	1.73	9.99×10^{-25}	205168_at	-0.67	7.05×10^{-24}
210029_at	1.02	3.59×10^{-26}	205168_at	1.63	5.25×10^{-24}	201694_at	-0.83	1.48×10^{-24}
209603_at	-1.50	7.04×10^{-26}	215867_x_at	2.08	4.34×10^{-25}	214164_x_at	-1.70	4.69×10^{-28}

Table 4: Top five significant DE genes, sorted by lowest q-values. Negative logFC values indicate an under-expression in the second-named cluster in the contrast.

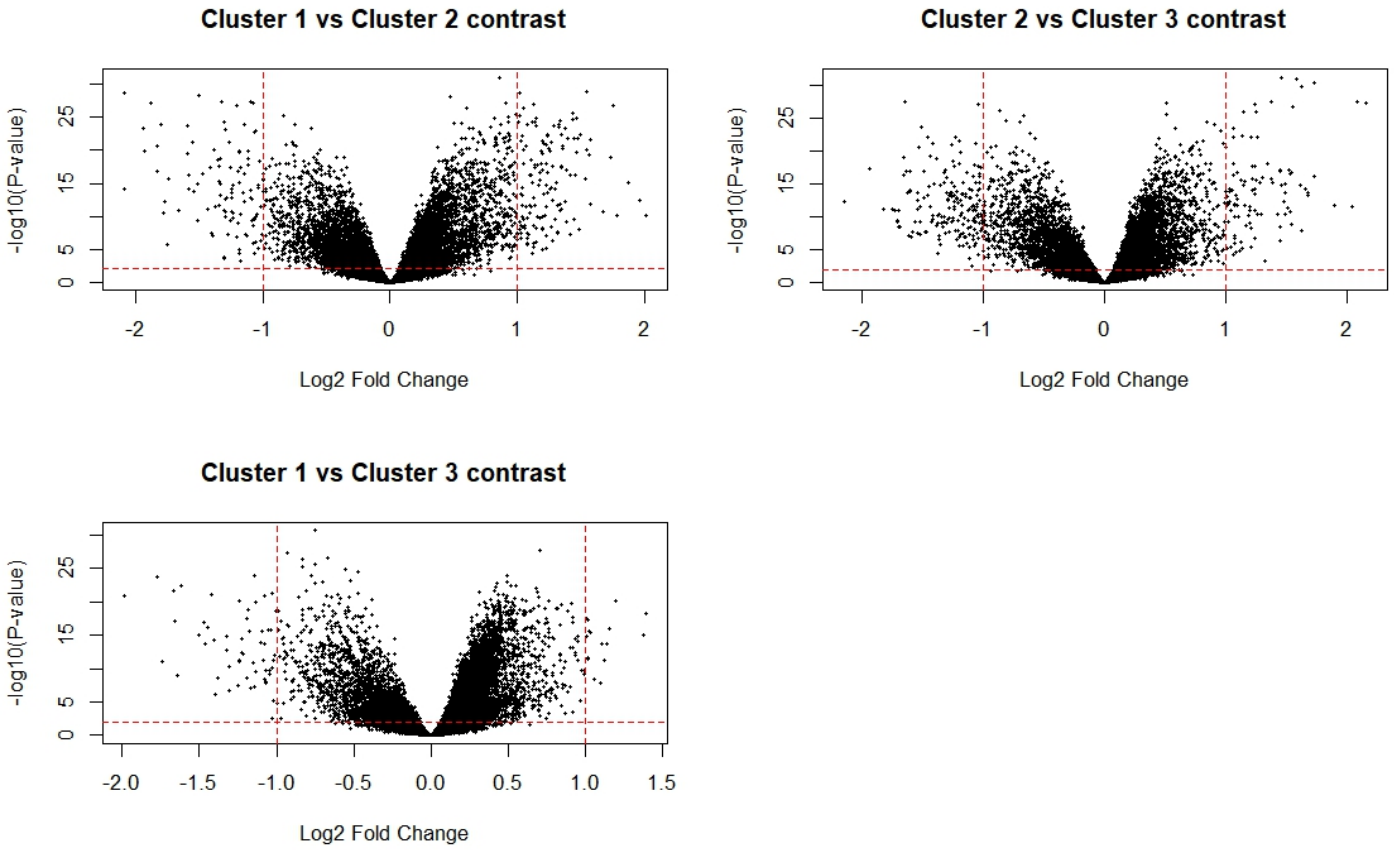


Figure 6: Volcano plots, where points above the dotted red horizontal line indicate statistical significance (p-value ≤ 0.01). Points outside of the vertical red dotted lines ($\leq$ -1.0 or ≥ 1.0) indicate at least a two-fold change in expression of that gene between clusters.

After extracting the IDs of the top 100 DE genes (ranked by smallest q-values), GO and KEGG analyses were conducted to understand what biological pathways these genes are involved in to hopefully explain the differences between the clusters. The top 20 GO terms and top 20 KEGG pathways of the top 100 DE genes between each cluster were collected and are presented in the Appendix (Tables 7-12).

While it is beyond the scope of this analysis to thoroughly explain each of these biological pathways, it is noteworthy that cell cycle processes and regulation emerged in the GO analysis of the top 100 DE genes between Cluster 1 and Cluster 2, while the cell cycle was also highlighted in the KEGG pathways of the DE genes between these clusters. These findings are particularly relevant as *"deregulation of the cell cycle is a hallmark of cancer that enables limitless cell division."*[8] It is also interesting to note the identification of lymphocyte activation in the GO analysis of the top DE genes between Cluster 2 and Cluster 3. Lymphocytes are cells that are commonly known to fight disease and healthy lymph nodes are strongly associated with better survival outcomes in breast cancer patients[7]. Thus, greater or lesser lymphocyte activation between Cluster 2 and 3 may provide an explanation if there is a difference in the prognoses and survival outcomes between these two clusters.

A survival analysis was undertaken to understand whether cluster membership resulted in better or worse survival outcomes. As Cluster 2 had the best survival outcomes, it was set as the reference level. The hazard ratio for membership in Cluster 1 was 1.4219 ($p = 0.3366$), meaning that patients in Cluster 1 had a 42.19% higher risk of death compared to those in Cluster 2. Although this suggests that Cluster 1 membership may be associated with worse survival outcomes, the result was not statistically significant ($p > 0.05$). Similarly, the hazard ratio for membership in Cluster 3 was 1.3523 ($p = 0.4648$), which indicates that patients in this cluster had a 35.23% higher risk of death compared to those in Cluster 2. Again, this result was not statistically significant ($p > 0.05$). The complete results of the Cox proportional hazards model can be found in Table 5.

Following the findings from this initial survival analysis, it was considered that combining Clusters 1 and 3 into a single group may provide a more effective classification for breast cancer patients. This decision was supported by the observation that Cluster 2 exhibited better survival outcomes compared to Clusters 1 and 3. Additionally, earlier evidence, such as the larger logFC values observed when Cluster 2 was included in the contrast (see Figure 6), along with the identification of

lymphocyte activation in the GO analysis, reinforced this consideration. A second Cox proportional hazards model was fitted using these new groupings. Patients with combined membership in either Cluster 1 or Cluster 3 had a 39.74% higher risk of death compared to those in Cluster 2. However, similar to the previous analysis, this finding was not statistically significant ($p = 0.3359$). The results can be found in Table 6.

Covariate	**Hazard Ratio**	**p-value**
gene_scores	1.0007	0.1662
as.factor(clW)1	1.4219	0.3366
as.factor(clW)3	1.3523	0.4648
age	1.0033	0.7507
as.factor(histgrade)G1	0.2015	0.1474
as.factor(histgrade)G2	0.3528	0.321
as.factor(histgrade)G3	0.4215	0.429
tumor_size_mm	1.0322	0.0111*
as.factor(PRstatus)PgR+	0.6831	0.3914
as.factor(ERstatus)ER+	1.6372	0.3687
as.factor(LNstatus)LN+	2.9528	0.0005***

Table 5: Hazard Ratios and p-values from Cox Proportional Hazards Model (using Cluster 2 as the reference cluster)

Covariate	**Hazard Ratio**	**p-value**
as.factor(clW_grouped)1&3	1.3974	0.3359

Table 6: Hazard ratio and p-value for grouped clusters (Cluster 2 vs Clusters 1 & 3 combined)

Finally, Kaplan-Meier (K-M) survival curves, showing survival outcomes over time for different cluster memberships, can be found in Figures 7 and 8. Overall, the K-M plots, as well as the the Cox proportional hazards models, indicate that belonging to Cluster 2 had a protective effect and led to longer prognoses, although this result wasn't significant ($p > 0.05$). Despite the fact that the differences in survival outcomes between clusters were not statistically significant, there is enough evidence to suggest that further data collection could lead to a statistically significant result in the future, and that these clusters are potentially valid sub-types of breast cancer.

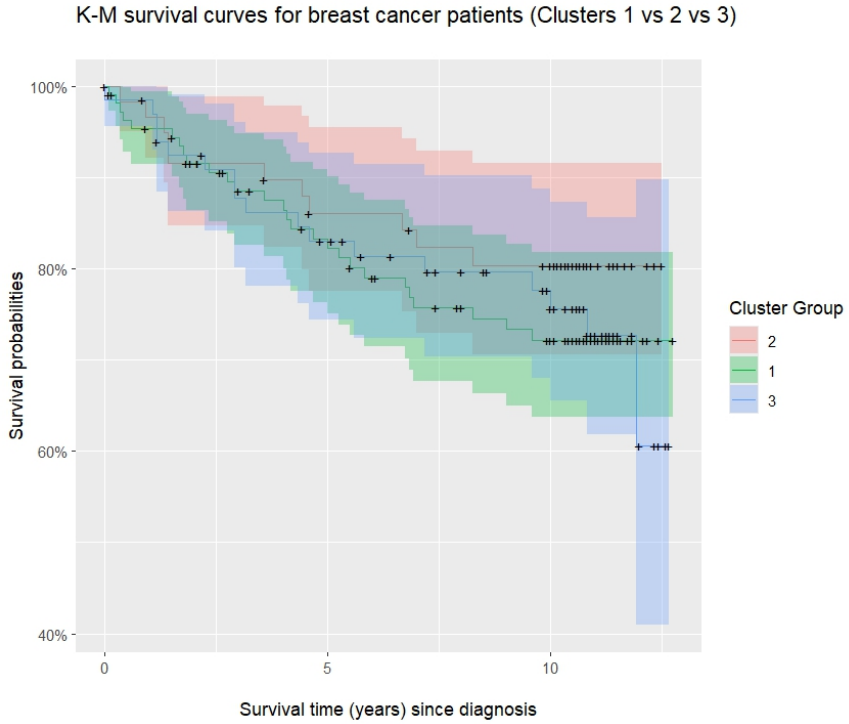


Figure 7: K-M survival curves for breast cancer patients (Clusters 1 vs 2 vs 3)

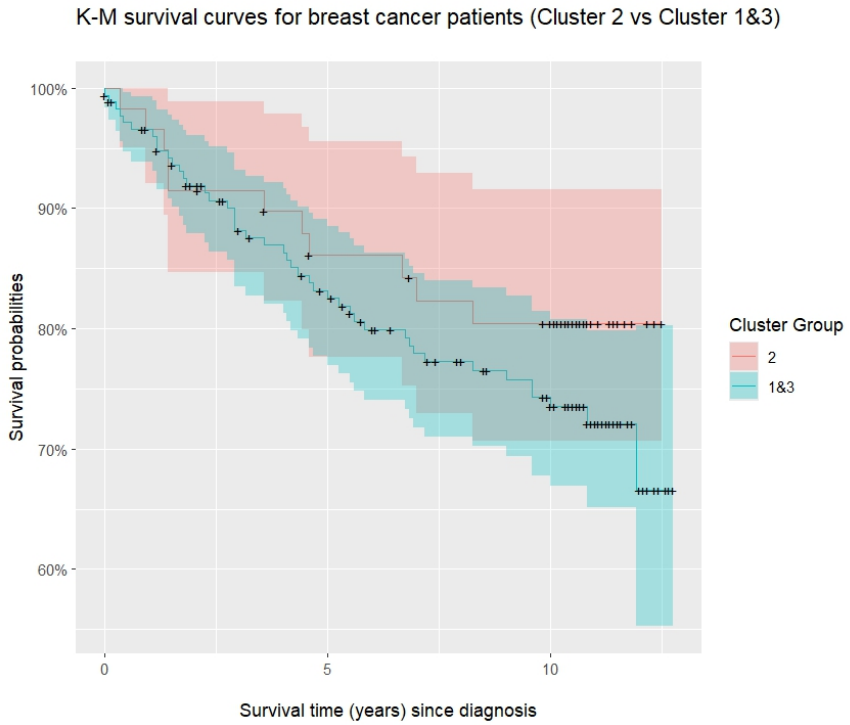


Figure 8: K-M survival curves for breast cancer patients (Cluster 2 vs Cluster 1 & 3)

4 Conclusion

This analysis aimed to identify subtypes of breast cancer using gene expression data and to assess how these subtypes relate to patient survival outcomes. The results from the clustering analysis, followed by differential gene expression (DE) analysis, revealed significant proportions of differentially expressed genes among the identified clusters. Volcano plots further highlighted a substantial number of DE genes with large log fold change (logFC) values, underscoring the biological relevance of these clusters, particularly in the contrasting gene expression profiles of Cluster 2 compared to Clusters 1 and 3.

Although the survival analysis did not yield statistically significant differences in outcomes, there is evidence suggesting that membership in Cluster 2 may be associated with improved prognoses. This finding aligns with the results of the GO and KEGG pathway analyses, which may help explain these observed differences in survival outcomes. Notably, cell cycle processes emerged prominently in the GO analysis for the comparison between Cluster 1 and Cluster 2. The association of DE genes with cell cycle processes likely correlates with the poorer prognosis for Cluster 1 compared to Cluster 2, highlighting a potential target for treatment, given that dysregulation of the cell cycle is recognized as a major factor in cancer. Additionally, the identification of lymphocyte activation among DE genes suggests that this process is more pronounced in Cluster 2 than in Cluster 3, contributing to the observed differences in survival outcomes and again highlighting a potential target for treatment.

Overall, the three sub-types of breast cancer identified via this cluster analysis do appear to be valid, as evidenced by the differential gene expression analysis. The absence of statistically significant findings in the survival analyses does, however, underscore the need for larger datasets and additional biological insights. Thus, future research should focus on further data collection, which may yield statistically significant results, as well as refining the identification of these subtypes and their implications for prognosis in breast cancer patients. This approach could potentially enable tailored treatments that support the relevant biological pathways, and ultimately, better outcomes for breast cancer patients.

5 References

[1] National Breast Cancer Foundation. (2024). Retrieved from https://nbcf.org.au/about/

[2] Rakha, E. A., Reis-Filho, J. S., Baehner, F., Dabbs, D. J., Decker, T., Eusebi, V., ... Ellis, I. O. (2010). Breast cancer prognostic classification in the molecular era: the role of histological grade. *Breast Cancer Research*, 12, 1-12.

[3] Samaan, N. A., Buzdar, A. U., Aldinger, K. A., Schultz, P. N., Yang, K.-P., Romsdahl, M. M., Martin, R. (1981). Estrogen receptor: A prognostic factor in breast cancer. *Cancer*, 47, 554-560.

[4] Mason, B. H., Holdaway, I. M., Mullins, P. R., Yee, L. H., Kay, R. G. (1983). Progesterone and estrogen receptors as prognostic variables in breast cancer. *Cancer Research*, 43(6), 2985-2990.

[5] Bardou, V. J., Arpino, G., Elledge, R. M., Osborne, C. K., Clark, G. M. (2003). Progesterone receptor status significantly improves outcome prediction over estrogen receptor status alone for adjuvant endocrine therapy in two large breast cancer databases. *Journal of Clinical Oncology*, 21(10), 1973-1979. https://doi.org/10.1200/JCO.2003.09.099.

[6] Narod, S. A. (2012). Tumour size predicts long-term survival among women with lymph node-positive breast cancer. *Current Oncology*, 19(5), 249-253. https://doi.org/10.3747/co.19.1043

[7] Carter, C. L., Allen, C., Henson, D. E. (1989). Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. *Cancer*, 63(1), 181-187.

[8] Thu, K., Soria-Bretones, I., Mak, T., Cescon, D. (2018). Targeting the cell cycle in breast cancer: towards the next phase. *Cell Cycle*, 17(15), 1871–1885. https://doi.org/10.1080/15384101.2018.1502567

6 Appendix

It should be noted that the complete linkage method identified at least 2-3 potentially interesting clusters for further analysis and these could be investigated in future research (Figures 9 & 10). However, the clusters are not as clear as those found using the Ward linkage method.

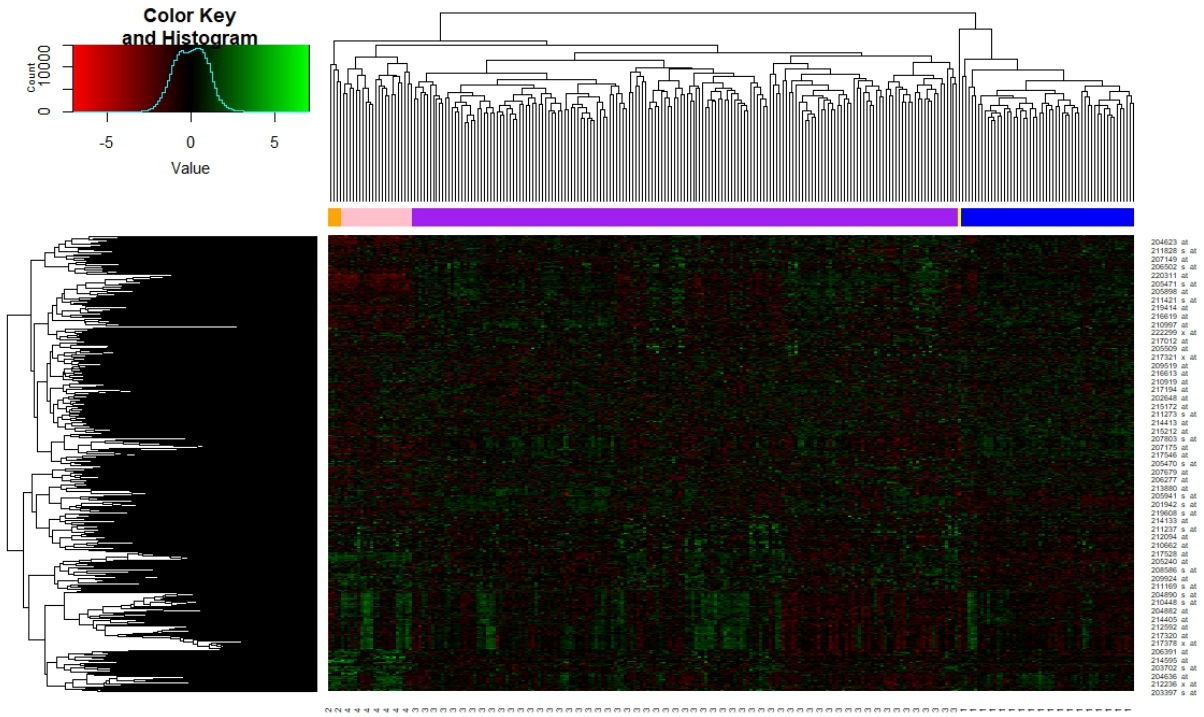
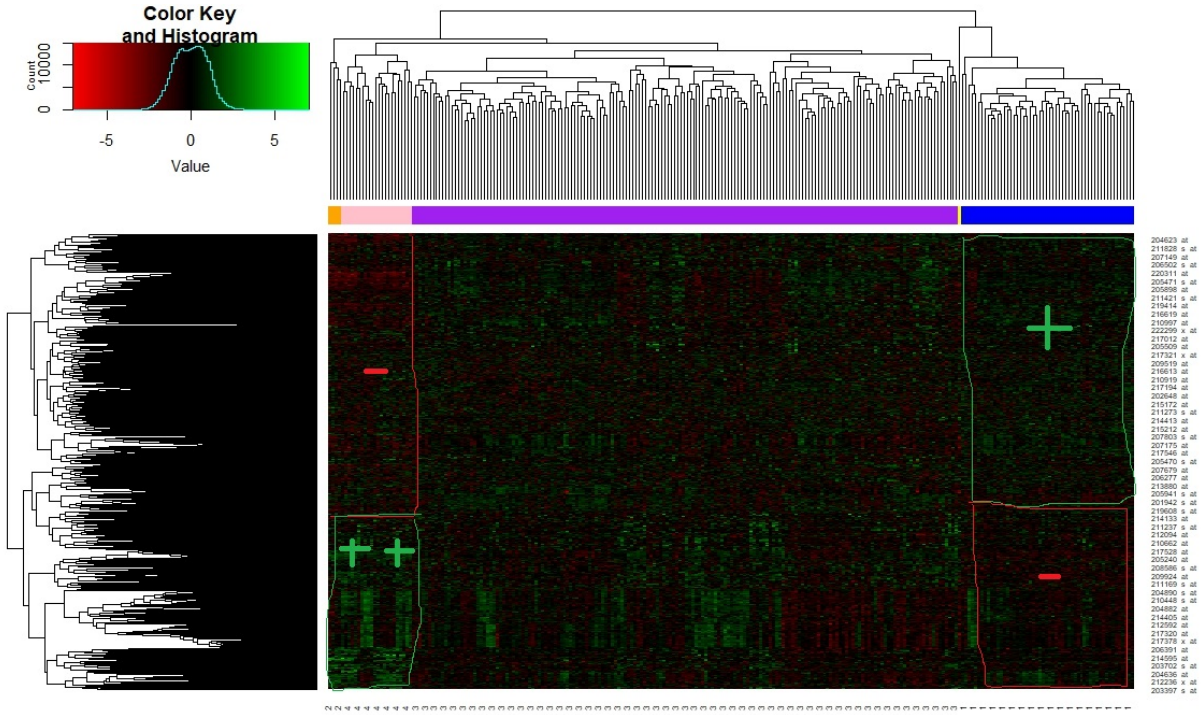


Figure 9: Heatmap for the top 5% (n = 1111) most variable genes, by cluster (k = 5), using **complete** linkage



Note: A case certainly could be made for two or three separate clusters using complete linkage as opposed to the 5 suggested by the median silhouette value. In the first proposed cluster (orange and pink) there are some genes that are commonly over-expressed by the individuals in that group. There is also clearly genes that are commonly under-expressed by this proposed cluster. This gene expression is clearly different to the other proposed clusters.

Figure 10: Annotated heatmap for the top 5% (n = 1111) most variable genes, by cluster (k = 5), using **complete** linkage

Single and average linkage methods did not appear to divide the individuals into biologically meaningful clusters as the one individual (85A03) did not appear to have pronounced differences in gene expression compared to the other 250 individuals (Figures 11 & 12).

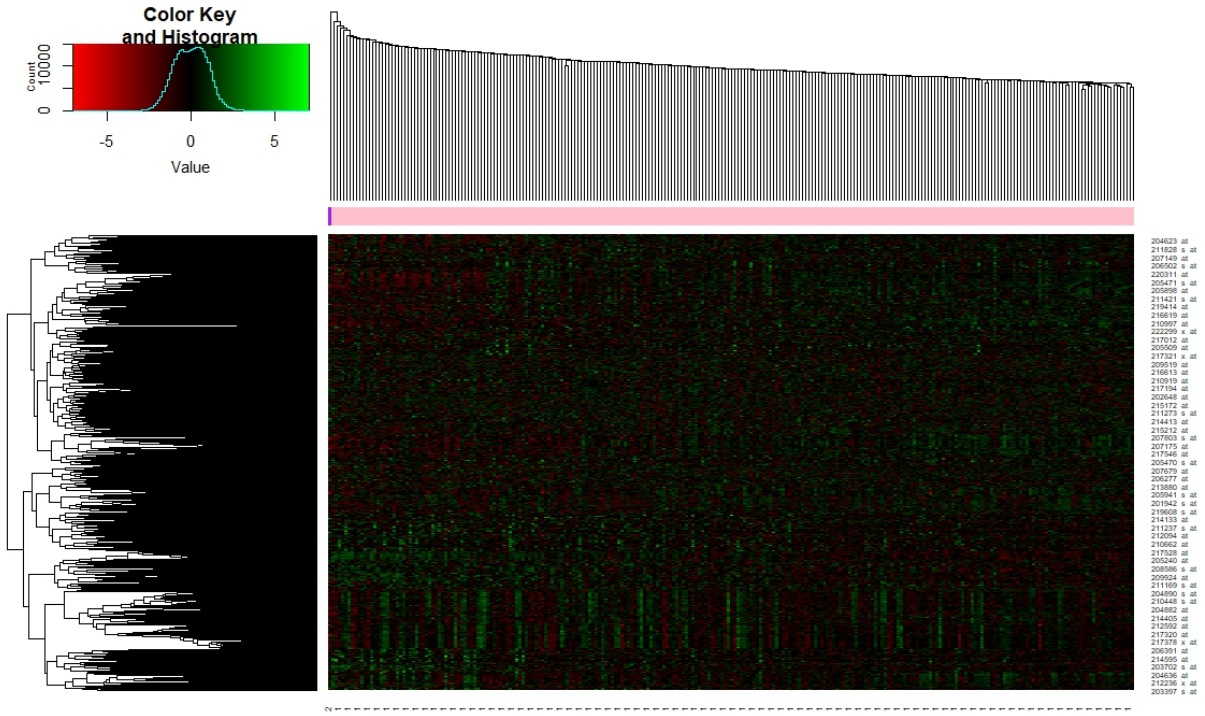


Figure 11: Heatmap for the top 5% (n = 1111) most variable genes, by cluster, using **single** linkage

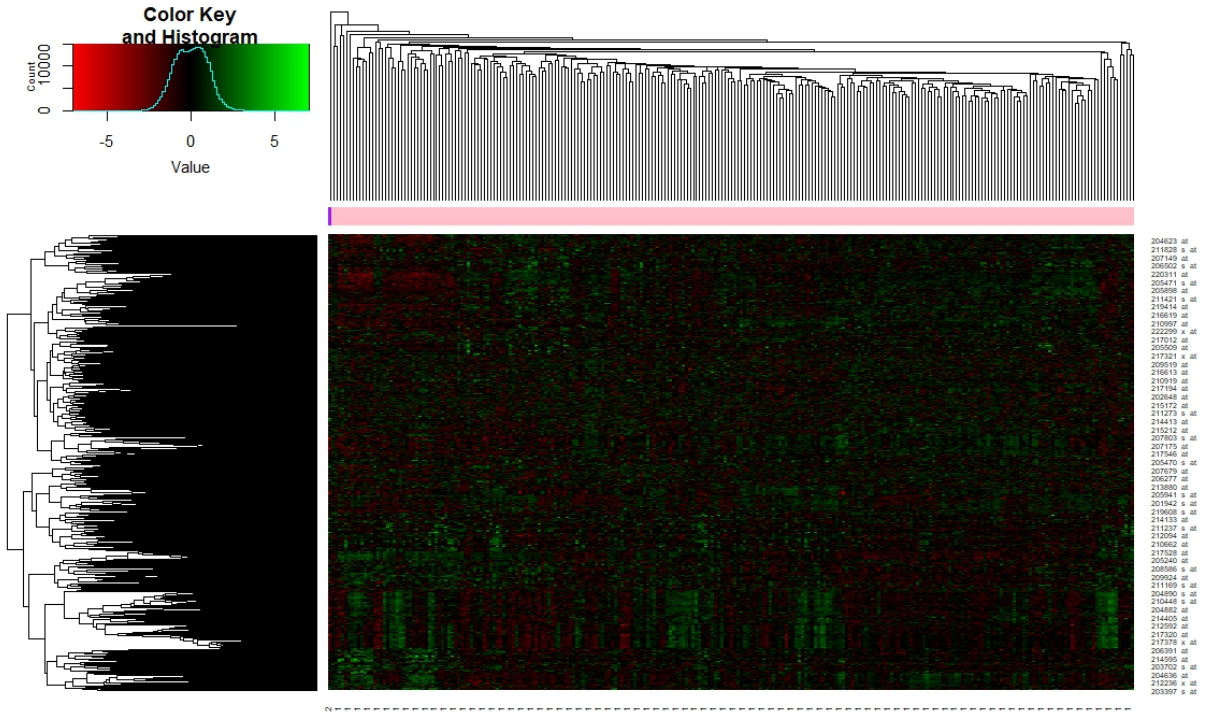


Figure 12: Heatmap for the top 5% (n = 1111) most variable genes, by cluster, using **average** linkage

GO Code	GO Term	N	DE	P.DE
GO:1903047	mitotic cell cycle process	761	43	5.76×10^{-38}
GO:0007059	chromosome segregation	423	35	4.10×10^{-36}
GO:0098813	nuclear chromosome segregation	317	32	1.12×10^{-35}
GO:0000278	mitotic cell cycle	911	43	1.15×10^{-34}
GO:0022402	cell cycle process	1294	48	2.02×10^{-34}
GO:0000280	nuclear division	446	34	8.34×10^{-34}
GO:0048285	organelle fission	493	35	8.98×10^{-34}
GO:0140014	mitotic nuclear division	277	29	1.07×10^{-32}
GO:0000070	mitotic sister chromatid segregation	188	26	1.62×10^{-32}
GO:0007049	cell cycle	1816	51	4.21×10^{-31}
GO:0051301	cell division	643	35	8.90×10^{-30}
GO:0051276	chromosome organization	635	34	1.23×10^{-28}
GO:0051726	regulation of cell cycle	1116	39	3.21×10^{-26}
GO:0051983	regulation of chromosome segregation	131	19	3.21×10^{-24}
GO:0044772	mitotic cell cycle phase transition	451	26	1.65×10^{-22}
GO:1902850	microtubule cytoskeleton organization involved in mitosis	164	19	2.75×10^{-22}
GO:0010564	regulation of cell cycle process	746	30	1.58×10^{-21}
GO:0044770	cell cycle phase transition	555	26	3.07×10^{-20}
GO:0007088	regulation of mitotic nuclear division	118	16	5.18×10^{-20}

Table 7: Top 20 GO terms of the top 100 DE genes between Cluster 1 and Cluster 2

GO Code	GO Term	N	DE	P.DE
GO:0046649	lymphocyte activation	799	21	2.33×10^{-12}
GO:0050852	T cell receptor signaling pathway	146	11	1.13×10^{-11}
GO:0050851	antigen receptor-mediated signaling pathway	208	12	2.87×10^{-11}
GO:0045321	leukocyte activation	964	21	7.83×10^{-11}
GO:0002684	positive regulation of immune system process	1098	22	1.25×10^{-10}
GO:0002429	immune response-activating cell surface receptor signaling pathway	332	13	5.01×10^{-10}
GO:0042110	T cell activation	570	16	5.47×10^{-10}
GO:0050778	positive regulation of immune response	762	18	6.26×10^{-10}
GO:0002682	regulation of immune system process	1569	25	6.33×10^{-10}
GO:0001775	cell activation	1113	21	1.07×10^{-9}
GO:0002768	immune response-regulating cell surface receptor signaling pathway	360	13	1.34×10^{-9}
GO:0050776	regulation of immune response	925	19	1.97×10^{-9}
GO:0002757	immune response-activating signaling pathway	500	14	7.69×10^{-9}
GO:0002764	immune response-regulating signaling pathway	528	14	1.53×10^{-8}
GO:0050867	positive regulation of cell activation	380	12	2.70×10^{-8}
GO:0006955	immune response	2046	26	3.03×10^{-8}
GO:0002376	immune system process	2844	31	3.10×10^{-8}
GO:0002253	activation of immune response	567	14	3.72×10^{-8}
GO:0050870	positive regulation of T cell activation	253	10	5.32×10^{-8}
GO:0051251	positive regulation of lymphocyte activation	328	11	5.87×10^{-8}

Table 8: Top 20 GO terms of the top 100 DE genes between Cluster 2 and Cluster 3

GO Code	GO Term	N	DE	P.DE
GO:0045787	Positive regulation of cell cycle	332	11	1.59×10^{-7}
GO:0090068	Positive regulation of cell cycle process	263	10	1.70×10^{-7}
GO:0016043	Cellular component organization	6581	51	3.25×10^{-7}
GO:0071840	Cellular component organization or biogenesis	6791	51	9.54×10^{-7}
GO:0033044	Regulation of chromosome organization	247	9	1.04×10^{-6}
GO:0043467	Regulation of generation of precursor metabolites and energy	140	7	2.20×10^{-6}
GO:0140014	Mitotic nuclear division	277	9	2.68×10^{-6}
GO:0006119	Oxidative phosphorylation	148	7	3.18×10^{-6}
GO:0034614	Cellular response to reactive oxygen species	151	7	3.63×10^{-6}
GO:0008283	Cell population proliferation	2090	24	4.08×10^{-6}
GO:0048522	Positive regulation of cellular process	5861	45	4.69×10^{-6}
GO:0048144	Fibroblast proliferation	110	6	7.35×10^{-6}
GO:0007088	Regulation of mitotic nuclear division	118	6	1.10×10^{-5}
GO:0030198	Extracellular matrix organization	332	9	1.16×10^{-5}
GO:0051256	Mitotic spindle midzone assembly	11	3	1.19×10^{-5}
GO:0043062	Extracellular structure organization	333	9	1.19×10^{-5}
GO:0045229	External encapsulating structure organization	334	9	1.21×10^{-5}
GO:0015980	Energy derivation by oxidation of organic compounds	338	9	1.33×10^{-5}
GO:0051276	Chromosome organization	635	12	1.49×10^{-5}
GO:0000070	Mitotic sister chromatid segregation	188	7	1.53×10^{-5}

Table 9: Top 20 GO terms of the top 100 DE genes between Cluster 1 and Cluster 3

KEGG Code	KEGG Pathway	N	DE	P.DE
hsa04110	Cell cycle	158	14	4.33×10^{-13}
hsa04114	Oocyte meiosis	139	7	2.19×10^{-5}
hsa04914	Progesterone-mediated oocyte maturation	111	6	5.98×10^{-5}
hsa05166	Human T-cell leukemia virus 1 infection	223	7	4.26×10^{-4}
hsa04814	Motor proteins	197	6	1.32×10^{-3}
hsa04218	Cellular senescence	157	5	2.81×10^{-3}
hsa04115	p53 signaling pathway	75	3	1.12×10^{-2}
hsa03030	DNA replication	36	2	2.09×10^{-2}
hsa04061	Viral protein interaction with cytokine and cytokine receptor	100	3	2.40×10^{-2}
hsa04620	Toll-like receptor signaling pathway	109	3	3.00×10^{-2}
hsa00232	Caffeine metabolism	6	1	3.67×10^{-2}
hsa05169	Epstein-Barr virus infection	203	4	3.69×10^{-2}
hsa00983	Drug metabolism - other enzymes	81	2	9.01×10^{-2}
hsa00910	Nitrogen metabolism	17	1	1.01×10^{-1}
hsa04260	Cardiac muscle contraction	87	2	1.02×10^{-1}
hsa04658	Th1 and Th2 cell differentiation	92	2	1.12×10^{-1}
hsa05200	Pathways in cancer	533	6	1.12×10^{-1}
hsa04062	Chemokine signaling pathway	193	3	1.18×10^{-1}
hsa05203	Viral carcinogenesis	205	3	1.34×10^{-1}
hsa05170	Human immunodeficiency virus 1 infection	213	3	1.46×10^{-1}

Table 10: Top 20 KEGG pathways of the top 100 DE genes between Cluster 1 and Cluster 2

KEGG Code	KEGG Pathway	N	DE	P.DE
hsa04658	Th1 and Th2 cell differentiation	92	5	1.18×10^{-4}
hsa04064	NF-kappa B signaling pathway	105	5	2.21×10^{-4}
hsa04071	Sphingolipid signaling pathway	122	5	4.43×10^{-4}
hsa04660	T cell receptor signaling pathway	122	5	4.43×10^{-4}
hsa05169	Epstein-Barr virus infection	203	6	6.63×10^{-4}
hsa04650	Natural killer cell mediated cytotoxicity	135	5	7.03×10^{-4}
hsa01521	EGFR tyrosine kinase inhibitor resistance	80	4	8.22×10^{-4}
hsa05170	Human immunodeficiency virus 1 infection	213	6	8.53×10^{-4}
hsa05340	Primary immunodeficiency	38	3	1.04×10^{-3}
hsa04012	ErbB signaling pathway	86	4	1.08×10^{-3}
hsa04933	AGE-RAGE signaling pathway in diabetic complications	101	4	1.96×10^{-3}
hsa04659	Th17 cell differentiation	108	4	2.50×10^{-3}
hsa04062	Chemokine signaling pathway	193	5	3.42×10^{-3}
hsa04919	Thyroid hormone signaling pathway	122	4	3.88×10^{-3}
hsa04370	VEGF signaling pathway	60	3	3.89×10^{-3}
hsa05205	Proteoglycans in cancer	204	5	4.33×10^{-3}
hsa04664	Fc epsilon RI signaling pathway	69	3	5.76×10^{-3}
hsa05162	Measles	139	4	6.16×10^{-3}
hsa05200	Pathways in cancer	533	8	6.43×10^{-3}
hsa04380	Osteoclast differentiation	142	4	6.64×10^{-3}

Table 11: Top 20 KEGG pathways of the top 100 DE genes between Cluster 2 and Cluster 3

KEGG Code	KEGG Pathway	N	DE	P.DE
hsa00190	Oxidative phosphorylation	138	8	2.83×10^{-6}
hsa05020	Prion disease	278	10	1.09×10^{-5}
hsa05016	Huntington disease	311	9	1.70×10^{-4}
hsa03050	Proteasome	46	4	2.21×10^{-4}
hsa05012	Parkinson disease	271	8	3.54×10^{-4}
hsa05014	Amyotrophic lateral sclerosis	371	9	6.25×10^{-4}
hsa05010	Alzheimer disease	391	9	9.10×10^{-4}
hsa05022	Pathways of neurodegeneration - multiple diseases	483	10	1.04×10^{-3}
hsa03018	RNA degradation	78	4	1.65×10^{-3}
hsa04114	Oocyte meiosis	139	5	2.09×10^{-3}
hsa05323	Rheumatoid arthritis	94	4	3.27×10^{-3}
hsa04932	Non-alcoholic fatty liver disease	157	5	3.54×10^{-3}
hsa05110	Vibrio cholerae infection	51	3	4.45×10^{-3}
hsa04914	Progesterone-mediated oocyte maturation	111	4	5.91×10^{-3}
hsa05152	Tuberculosis	180	5	6.31×10^{-3}
hsa05200	Pathways in cancer	533	9	7.36×10^{-3}
hsa05120	Epithelial cell signaling in Helicobacter pylori infection	71	3	1.11×10^{-2}
hsa04115	p53 signaling pathway	75	3	1.29×10^{-2}
hsa05017	Spinocerebellar ataxia	144	4	1.45×10^{-2}
hsa05208	Chemical carcinogenesis - reactive oxygen species	226	5	1.59×10^{-2}

Table 12: Top 20 KEGG pathways of the top 100 DE genes between Cluster 1 and Cluster 3