Gene expression profiles in histologically normal epithelium from breast cancer patients and cancer-free prophylactic mastectomy patients

Annalisa Xamin, MAT. 241482

July 27, 2025

This project has been developed for the *Network-Based Data Analysis* course, held by Prof. Lauria. (MSc Quantitative and Computational Biology at University of Trento, A.Y. 2024-2025).

1 Introduction

Investigating gene expression profiles in histologically normal breast epithelium from both breast cancer patients and cancer-free individuals undergoing prophylactic mastectomy offers valuable insight into early molecular changes that may underlie cancer risk or predisposition. In this project, we analyzed the publicly available dataset **GSE20437**, which includes samples from both cohorts, to identify molecular signatures and biological pathways that differentiate them.

To achieve this, we applied a combination of unsupervised and supervised machine learning techniques to explore expression patterns and classify samples. Unsupervised methods—including Principal Component Analysis (PCA), K-means clustering, and hierarchical clustering—were used to assess intrinsic structure in the data. Concurrently, supervised approaches such as Random Forest, Linear Discriminant Analysis (LDA), Lasso regression, and SCUDO were used to predict group membership and identify discriminative genes.

Following model evaluation, the most informative features were further analyzed through functional enrichment and network-based approaches. These analyses helped contextualize the findings biologically, revealing key regulatory pathways and molecular mechanisms that may be associated with increased susceptibility to breast cancer despite histologically normal tissue appearance.

2 Methods

All data processing was performed with the R programming language (version 4.5.1)[9].

2.1 Data selection and pre-processing

The analysis presented in this project is performed on the **GSE20437** dataset [4], which explores the gene expression in histologically normal epithelium from breast cancer patients and cancer-free prophylactic mastectomy patients. Data was retrieved from the *Gene Expression Omnibus* (GEO) database, which offers array- and sequence-based data. To load the dataset into R, we used the *GEOquery* R package [2]. The dataset, generated from Affymetrix HU133A microarrays, contains 42 tissue samples. In detail, the data include 18 breast epithelium samples from reduction mammoplasty (RM), 18 histologically normal epithelial samples (HN) of breast cancer patients (9 ER+ and 9 ER-), and 6 histologically normal epithelial samples from patients with prophylactic mastectomy (NlEpi). For each sample, there are 22283 genes.

The data was already converted into read count, and so the pre-processing consisted in data normalisation, using a *log-transformation* and *median-to-zero* techniques. For later analysis, the data were also annotated, retrieving the ENSEMBL gene ID and the external gene name for each gene using *biomaRt*[3].

2.2 Unsupervised Learning Methods

In order to investigate the presence of latent structures or patterns in the data, three different unsupervised methods were considered: Principal Component Analysis (PCA), K-Means, and Hierarchical Clustering.

2.2.1 Principal Component Analysis

Principal Component Analysis (PCA) is a statistical method used to reduce the complexity of data by transforming a dataset with many variables into a smaller number of new, uncorrelated variables called principal components. These components are designed to capture as much of the original data's variability as possible. In short, PCA simplifies high-dimensional data while preserving its most important patterns. The PCA has been performed with the function *prcomp* from the library *stats*, while the Scree Plot of the variance explained by each dimension was plotted thanks to the R packages *ggplot2* [13] and *factoextra* [5].

2.2.2 K-Means

K-means clustering is an unsupervised machine learning method that divides data into k distinct groups, or clusters. Each data point is assigned to the cluster with the closest center (called the centroid), which represents the average position of all points in that group. The algorithm works iteratively, updating the centroids and reassigning points until the clusters stabilize. The number of clusters (k) must be defined in advance, often guided by prior knowledge or exploratory analysis. K-means typically uses Euclidean distance to group similar observations, aiming to minimize variation within clusters while maximizing separation between them. However, this method is sensitive to outliers, so it is used for a first exploratory analysis. In this analysis, clustering was performed using the *kmeans* function from R's *stats* package. The first grouping was performed by specifying $k = 2$, knowing that there are two groups: *breast cancer* and *control.* The second grouping was then performed, setting $k = 4$, trying to include the two subtypes of breast cancer (ER+ and ER-) and the two subtypes of control (RM and NlEpi).

2.2.3 Hierarchical clustering

Hierarchical clustering is a technique used to organize data points into groups based on their similarity or distance. It constructs a hierarchical tree-like diagram called a dendrogram, which visually represents how data points are grouped at various levels. There are two main strategies: **agglomerative clustering** begins with each data point as its cluster and progressively merges the most similar clusters until all points are grouped into one; while **divisive clustering** starts with all data points in a single cluster and recursively divides them into smaller groups based on dissimilarity.

By cutting the dendrogram at different levels, one can determine an appropriate number of clusters. This method requires a distance metric and a linkage strategy to define how clusters are joined or split. In this analysis, Euclidean distance was used along with average linkage method, implemented via the *hclust* function from R's *stats* package.

2.3 Supervised Learning Methods

Supervised learning refers to a category of machine learning approaches that build predictive models using labeled datasets. In this setting, ”supervised” means that the training data includes known outcomes or labels that link input variables to their respective outputs. These methods learn from the input-output pairs to identify patterns and make accurate predictions on new, unseen data. The main goal is to generalize insights from labeled data to produce reliable forecasts.
For the supervised methods, the data was split into training and test set and the cross-validation were performed thanks to the *caret* R package[7].

2.3.1 Random Forest

Random Forest is a popular machine learning method that aggregates the predictions of multiple decision trees to produce a single outcome. Each tree is built using a bootstrap sampling of the input features. The final prediction is obtained by averaging the outputs of all trees. In addition, Random Forest can be used to identify the most influential characteristics (e.g. genes) that contribute to the prediction or clustering of the model.
Random Forest involves tuning two key parameters: *ntree*, the number of trees in the forest, and *mtry*, the number of variables randomly selected at each split. Using the *randomForest* package in R [8], the optimal number of trees was determined to be 520. Subsequently, cross-validation was applied to identify the best *mtry* value, which was found to be 41, and this configuration was used to train the final model for performance evaluation on the test set.

2.3.2 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a supervised machine learning technique commonly used for multi-class classification tasks. It aims to reduce the dimensionality of the data while maximizing the separation between different classes. LDA achieves this by finding a linear combination of features that best distinguishes between two or more groups. By projecting high-dimensional data onto a lower-dimensional space—often a single axis—it simplifies the classification process and enhances interpretability.
To carry out LDA, feature selection was required to narrow down the number of genes analyzed. A t-test was conducted using the *genefilter* R package [10], and only features with a p-value below 0.1 were retained. The LDA model was then trained using 10-fold cross-validation.

2.3.3 Lasso regression

Lasso is a regularization technique that enhances a model’s predictive performance by limiting the size of its coefficients. This constraint helps reduce model complexity and mitigates overfitting. In Lasso regression, the penalty applied can shrink some coefficients entirely to zero, effectively removing non-informative features. The result is a simpler, more interpretable model that can also be more computationally efficient.
Lasso was trained and optimized using 10-fold cross-validation. Several values of the tuning parameter λ were evaluated, while α was kept constant at 1. The best-performing model was obtained with a λ value of 0.2021.

2.3.4 SCUDO

SCUDO (Signature-based Clustering for Diagnostic Purposes) is a ranking-based method used for classification and diagnostic applications. It detects distinct gene signatures for each sample, builds a graph reflecting the similarity among these signatures, and clusters the data into groups with similar profiles. Implemented through the R package *rScudo* [1], this method compares gene signatures between individuals by pinpointing the most and least expressed genes per sample. Due to time constraints, cross-validation for parameter tuning was omitted; instead, the analysis concentrated on the top 25 most and least expressed genes from each signature.

2.4 Model comparison and feature selection

Each trained model was assessed using two metrics: accuracy and the area under the ROC curve (AUC). For the supervised models, these metrics were calculated based on predictions made on the test set, using the model trained with the optimal parameters on the training set. The R package *pROC* [11] was employed to generate the ROC curves and calculate the AUC. Additionally, feature selection was carried out to identify the most important genes from the random forest model.

2.5 Functional Enrichment analysis

Gene Set Analysis (GSA), also known as functional enrichment analysis, is a powerful approach for analyzing high-throughput experimental data and interpreting gene expression findings. Its primary goal is to detect biological functions and pathways that are significantly enriched or over-represented in a given gene list compared to a reference background. These methods usually utilize gene-level statistics to highlight meaningful expression differences across different conditions.

Functional enrichment analysis was performed using *gprofiler2* [6] to identify the most prominent molecular functions among the top-ranked genes. For this purpose, the 30 genes with the highest importance scores from the previously trained random forest model were selected. This tool conducts statistical enrichment tests to find over-represented functions and pathways from multiple databases, employing a hypergeometric test followed by correction for multiple hypothesis testing.

2.6 Network-based analysis

Network-based analysis methods investigate networks illustrating interactions among components of a complex system, aiming to reveal information about the system's connectivity and organization. This analysis was conducted using the R package *pathfindR* [12], which specializes in pathway analysis focused on active subnetworks—groups of highly connected and significant genes. To pinpoint the most important pathways, *pathfindR* integrates p-values from the input gene list (derived using *genefilter* as previously mentioned) with protein-protein interaction network data.

3 Results

3.1 Pre-processing and PCA

The data distribution for each sample was properly scaled, as illustrated in Figure 1a. The PCA results were satisfactory, with the first principal component accounting for 17.98% of the variance (Figure 1b). Although multiple combinations of principal components were visualized, they did not reveal any additional significant patterns. However, in the PCA plot (Figure 1c), only control samples tend to cluster on the left side of the latent space. A similar grouping pattern is reflected in one of the major branches of the hierarchical clustering dendrogram (Figure 1d).

3.2 Unsupervised and Supervised models for Classification

Both LDA and LASSO demonstrated moderately accurate classification of the samples, achieving accuracy and AUC values of 0.91 and 0.97 for LDA, and 0.75 and 0.91 for LASSO, respectively. These outcomes are illustrated in Figure 2. Although the random forest model did not outperform LDA and LASSO, it was still chosen for feature selection because it provides variable importance measures that allow identification of the most influential features.

The other approaches demonstrated moderately lower performance, with K-means reaching an accuracy of 0.57 and an AUC of 0.60, while SCUDO obtained values of 0.67 for accuracy and 0.54 for AUC. Hierarchical clustering performed similarly, with an accuracy of 0.60 and an AUC of 0.53. The SCUDO analysis resulted in a single network, shown in Figure 10. Although distinct clusters were not explicitly defined, breast cancer nodes appeared to be more closely connected to one another. Two control nodes located between groups of cancer samples may represent outliers. If so, this spatial arrangement could indicate greater similarity in gene expression profiles among the breast cancer cases. A summary of these results can be found in Figure 2 and Table 1.

3.3 Feature selection

Although the random forest model did not achieve superior predictive performance compared to LDA and LASSO, it was prioritized for feature selection due to its ability to compute variable importance metrics. During model training, each gene is assigned an importance score based on its contribution to reducing classification error across the ensemble of trees. The top-ranked genes, visualized in Figure 3, demonstrate clear differential expression between the two sample classes, as illustrated by the corresponding heatmap (Figure 9).

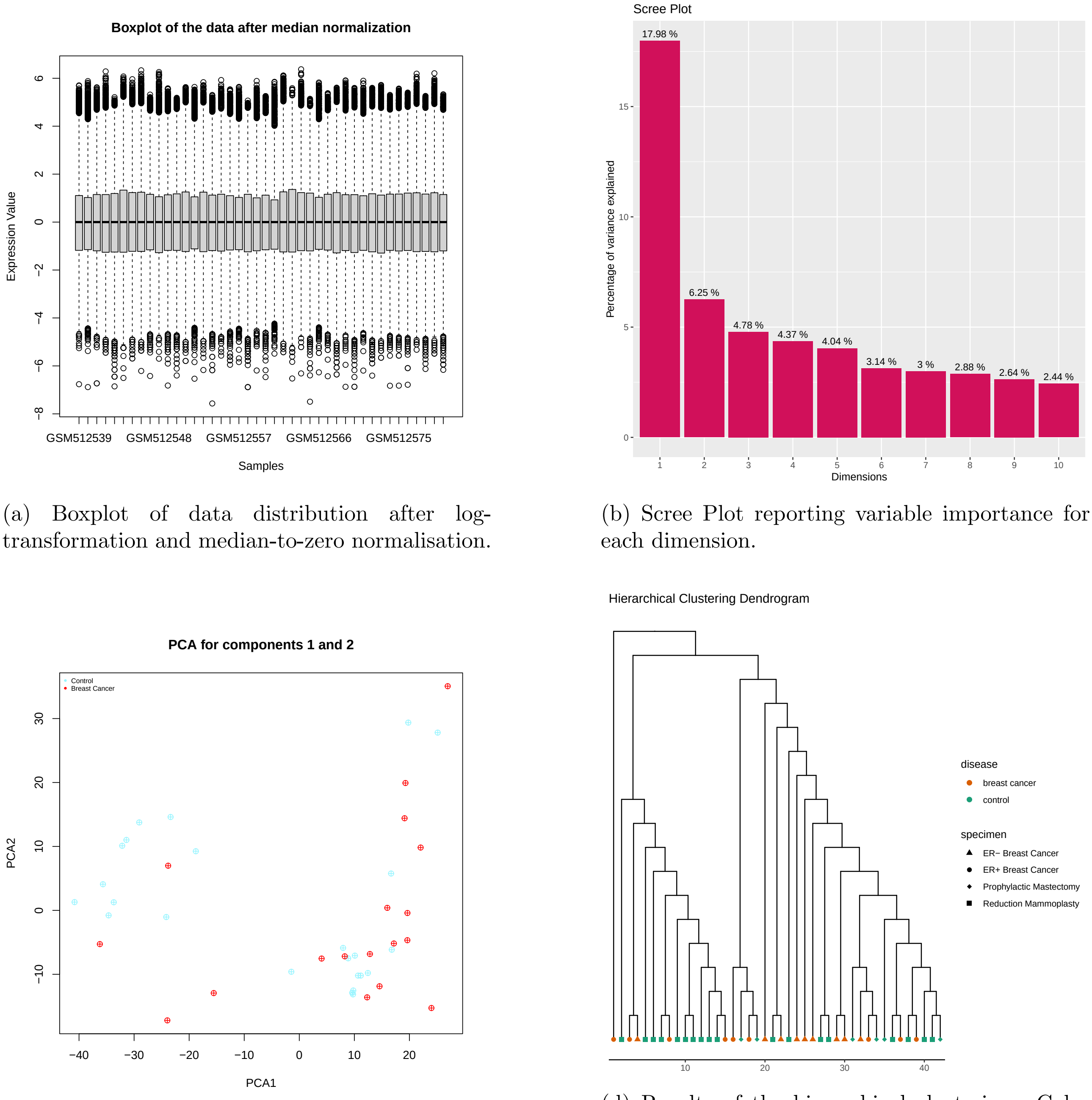
3.4 Functional Enrichment Analysis

The results of the functional enrichment analysis are shown in Figure 4a and Figure 4b. Figure 4a indicates the presence of several statistically significant p-values, primarily associated with microRNA (miRNA) databases queried through *gprofiler*. In Figure 4b, the seven most significantly enriched terms—those with the lowest adjusted p-values—are all miRNAs. These entries correspond to individual microRNAs.

3.5 Biological Network Analysis

The biological network analysis was conducted using the *pathfindR* package with enrichment based on three pathway databases: *KEGG*, *Gene Ontology (GO)*, and *Reactome*.

The **KEGG analysis** highlighted pathways associated with immune signaling and cancer-related metabolic processes, suggesting these are key themes within the differentially expressed genes (Figure 5a). Despite a small number of contributing genes per pathway, enrichment was highly significant,



(a) Boxplot of data distribution after log-transformation and median-to-zero normalisation.

(b) Scree Plot reporting variable importance for each dimension.

(c) Figure showing samples of breast cancer and control in the reduced space of the first two PCA components.

(d) Results of the hierarchical clustering. Color code identifies the origin of the sample (breast cancer or control), while the shape of the leaf identify the subtype.

Figure 1: Pre-processing and data clustering

indicating strong biological relevance.
The **GO analysis** emphasized transcriptional regulation and response to cellular stress, pointing to upstream regulatory mechanisms (Figure 6a). The high fold-enrichment values observed are likely due to the more specific nature of GO terms, which can yield strong signals even from smaller gene sets.
Reactome enrichment further supported the activation of the MAPKAP-1 signaling axis, providing additional detail through the inclusion of immune-related pathways, such as IL-17, TLR, and FCERI-mediated signaling (Figure 7a). These results reinforce the inflammatory and immune-related

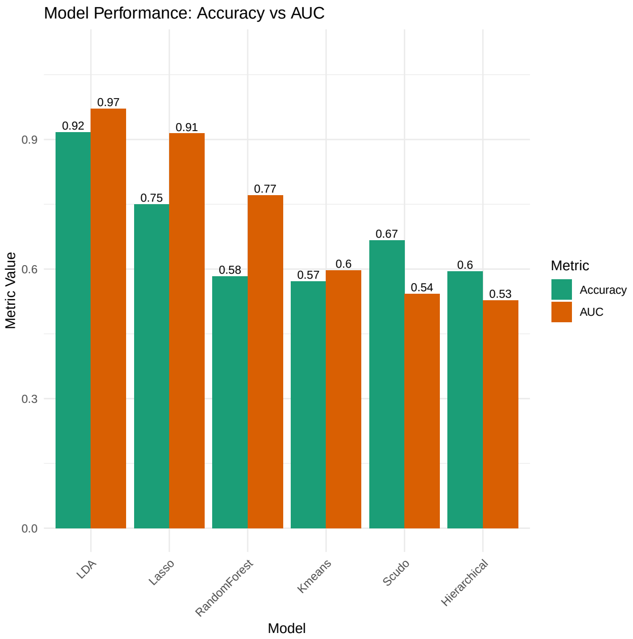


Figure 2: Comparison of the models' performances.

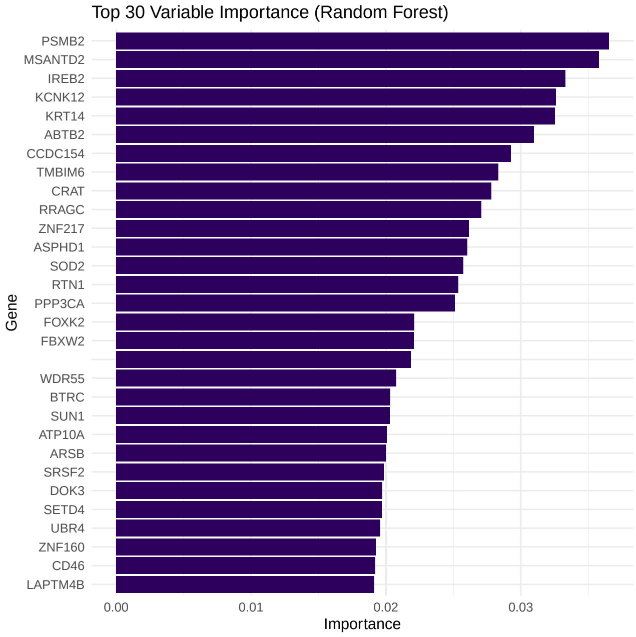
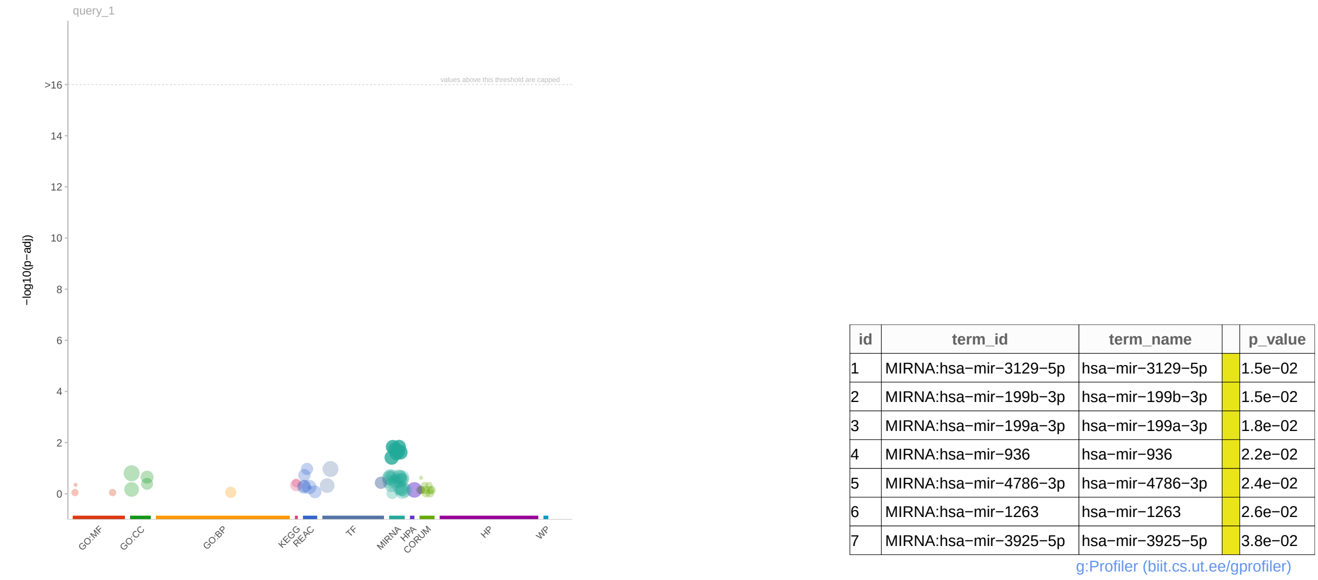


Figure 3: **Random Forest Variable importance of top** 30 **genes**. The gene absent from the y-axis, labeled as *221713_s_at*, lacks an associated external gene name in the *biomaRt* annotation database, which is why it appears as an empty entry in the visualization.

biological processes already suggested by KEGG.
Across all three databases, FOS and JUN emerged as central regulators, appearing consistently in key pathways related to signal transduction and transcriptional control (see Figure 5b, Figure 6b and Figure 7b). Together, these findings suggest a coherent biological narrative linking signaling activation to transcriptional changes and ultimately to immune and cancer-related functional outcomes, thus validating the robustness and biological consistency of the network analysis.

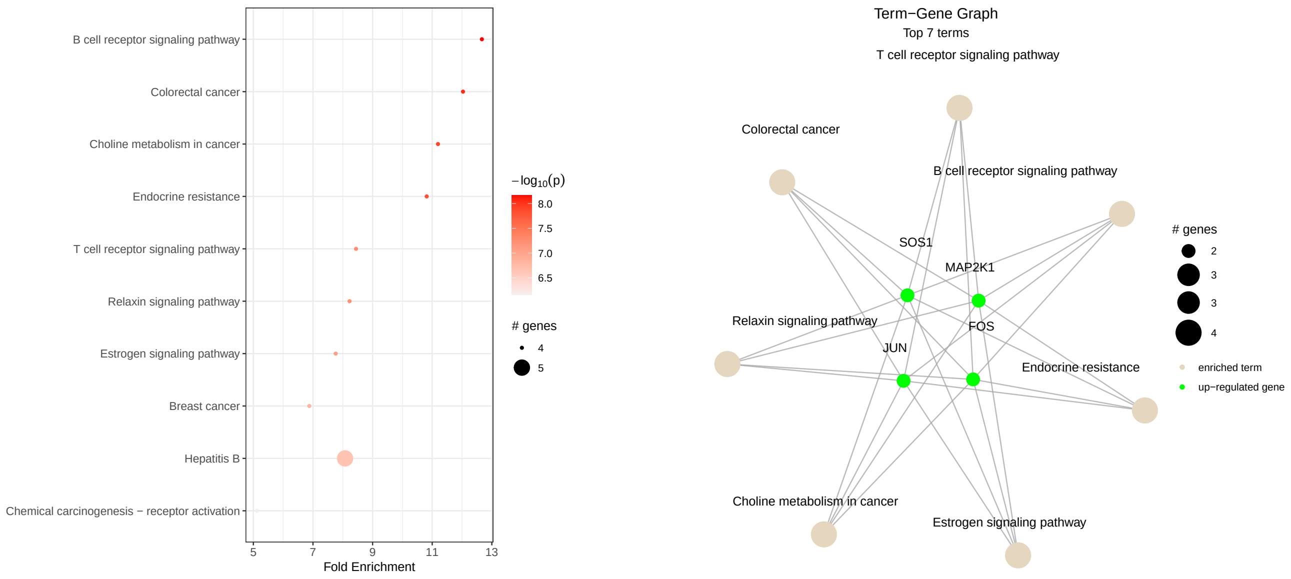


id	term_id	term_name		p_value
1	MIRNA:hsa−mir−3129−5p	hsa−mir−3129−5p	🟨	1.5e−02
2	MIRNA:hsa−mir−199b−3p	hsa−mir−199b−3p	🟨	1.5e−02
3	MIRNA:hsa−mir−199a−3p	hsa−mir−199a−3p	🟨	1.8e−02
4	MIRNA:hsa−mir−936	hsa−mir−936	🟨	2.2e−02
5	MIRNA:hsa−mir−4786−3p	hsa−mir−4786−3p	🟨	2.4e−02
6	MIRNA:hsa−mir−1263	hsa−mir−1263	🟨	2.6e−02
7	MIRNA:hsa−mir−3925−5p	hsa−mir−3925−5p	🟨	3.8e−02

(a) Visualization of enriched terms by source database with associated corrected p-values.

(b) Top 7 Significantly Enriched Functional Terms Identified by *gprofiler*.

Figure 4: **Functional Enrichment Analysis**



(a)

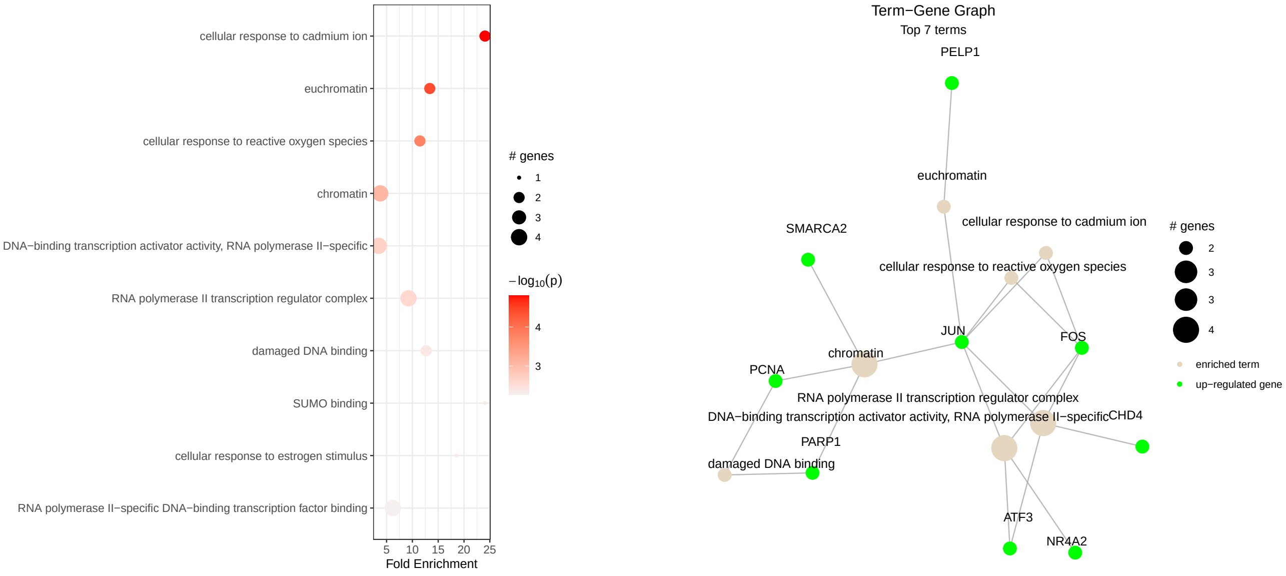
(b)

Figure 5: Biological network analysis results based on **KEGG**.

4 Discussion and conclusions

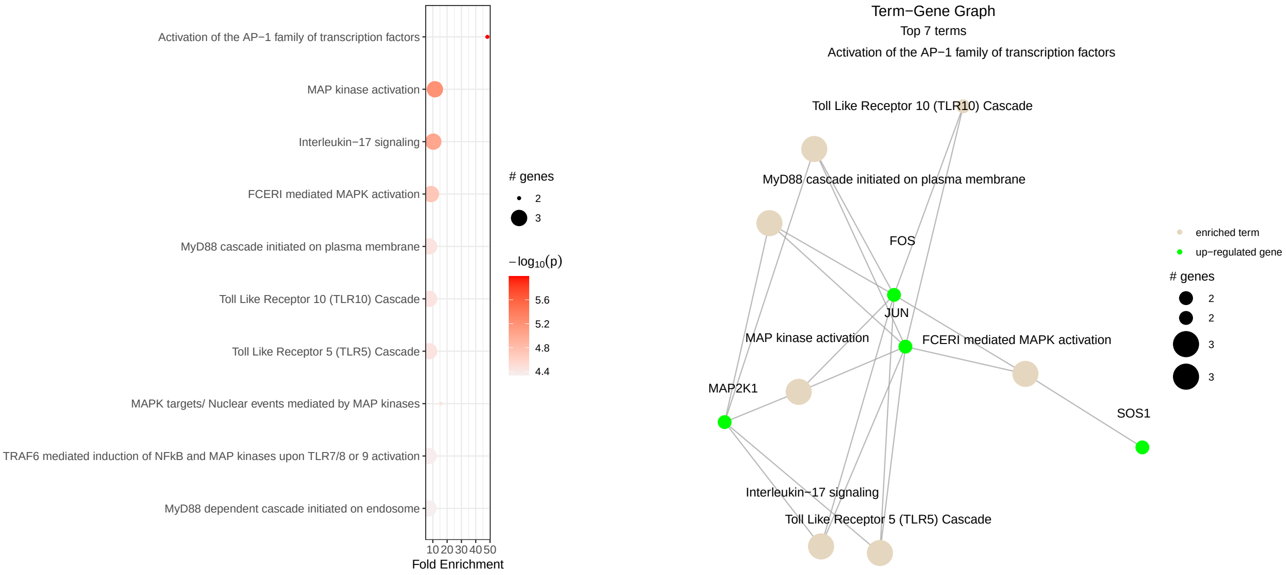
This study applied a range of unsupervised and supervised machine learning techniques to analyze gene expression profiles from histologically normal breast epithelium, with the aim of distinguishing samples from breast cancer patients and cancer-free individuals undergoing prophylactic mastectomy. Careful data preprocessing and normalization ensured high-quality inputs for subsequent analyses.

In the unsupervised analysis, PCA explained only a small portion of variance in the initial components, and clustering methods provided moderate group separation, reflecting subtle gene expression



(a) (b)

Figure 6: Biological network analysis results based on **Gene Ontology (GO)**.



(a) (b)

Figure 7: Biological network analysis results based on **Reactome**.

differences. Supervised techniques, notably LDA and Lasso regression, delivered better classification performance, reaching up to 91% accuracy and strong AUC scores, highlighting their effectiveness in distinguishing cancerous from normal samples.

Although Random Forest did not achieve the highest classification accuracy, it proved useful for selecting important features by pinpointing genes differentially expressed between groups. Functional enrichment of these genes pointed to microRNAs, indicating their possible regulatory roles in cancer biology.

Pathway and network analyses further revealed significant involvement of immune response and transcriptional regulation pathways, with central genes like FOS and JUN identified as key regulators.

These results are consistent with known mechanisms in cancer development and confirm the biological significance of the selected genes.

In conclusion, by integrating machine learning with functional interpretation, this study uncovered early molecular differences in histologically normal breast tissue that may reflect latent cancer-associated processes or predisposition. Future work should focus on validating these findings and exploring their role in breast cancer risk and initiation.

References

[1] M. Ciciani, T. Cantore, and M. Lauria. rscudo: an r package for classification of molecular profiles using rank-based signatures. *Bioinformatics*, 36(13):4095–4096, May 2020.

[2] S. Davis and P. S. Meltzer. Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor. *Bioinformatics*, 23(14):1846–1847, May 2007.

[3] S. Durinck, P. T. Spellman, E. Birney, and W. Huber. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature Protocols*, 4(8):1184–1191, July 2009.

[4] K. Graham, A. de las Morenas, A. Tripathi, C. King, M. Kavanah, J. Mendez, M. Stone, J. Slama, M. Miller, G. Antoine, H. Willers, P. Sebastiani, and C. L. Rosenberg. Gene expression in histologically normal epithelium from breast cancer patients and from cancer-free prophylactic mastectomy patients shares a similar profile. *British Journal of Cancer*, 102(8):1284–1293, Mar. 2010.

[5] A. Kassambara and F. Mundt. factoextra: Extract and visualize the results of multivariate data analyses, Apr. 2016.

[6] L. Kolberg, U. Raudvere, I. Kuzmin, J. Vilo, and H. Peterson. gprofiler2 – an r package for gene list functional enrichment analysis and namespace conversion toolset g:profiler. *F1000Research*, 9:709, Nov. 2020.

[7] M. Kuhn. Building predictive models inrusing thecaretpackage. *Journal of Statistical Software*, 28(5), 2008.

[8] A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[9] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2024.

[10] V. C. R. Gentleman. genefilter, 2017.

[11] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12(1), Mar. 2011.

[12] E. Ulgen, O. Ozisik, and O. U. Sezerman. pathfindr: An r package for comprehensive identification of enriched pathways in omics data through active subnetworks. *Frontiers in Genetics*, 10, Sept. 2019.

[13] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York, 2016.

A Data availability

All the data and scripts used in this project are available in this GitHub repository: click here.

B Supplementary figures and tables

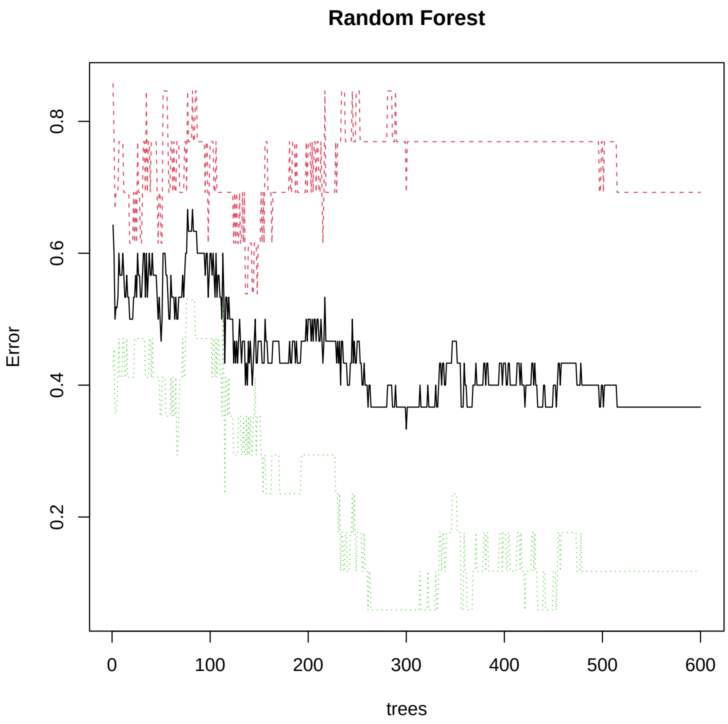


Figure 8: OOB error of the random forest model reaches a plateau with a number of trees around 520.

	Accuracy	**AUC**
Kmeans	0.5714286	0.5972222
Hierarchical	0.5952381	0.5277778
Random Forest	0.5833333	0.7714286
LDA	0.9166667	0.9714286
Lasso	0.7500000	0.9142857
Scudo	0.6666667	0.5428571

Table 1: Comparison of the models' performances.

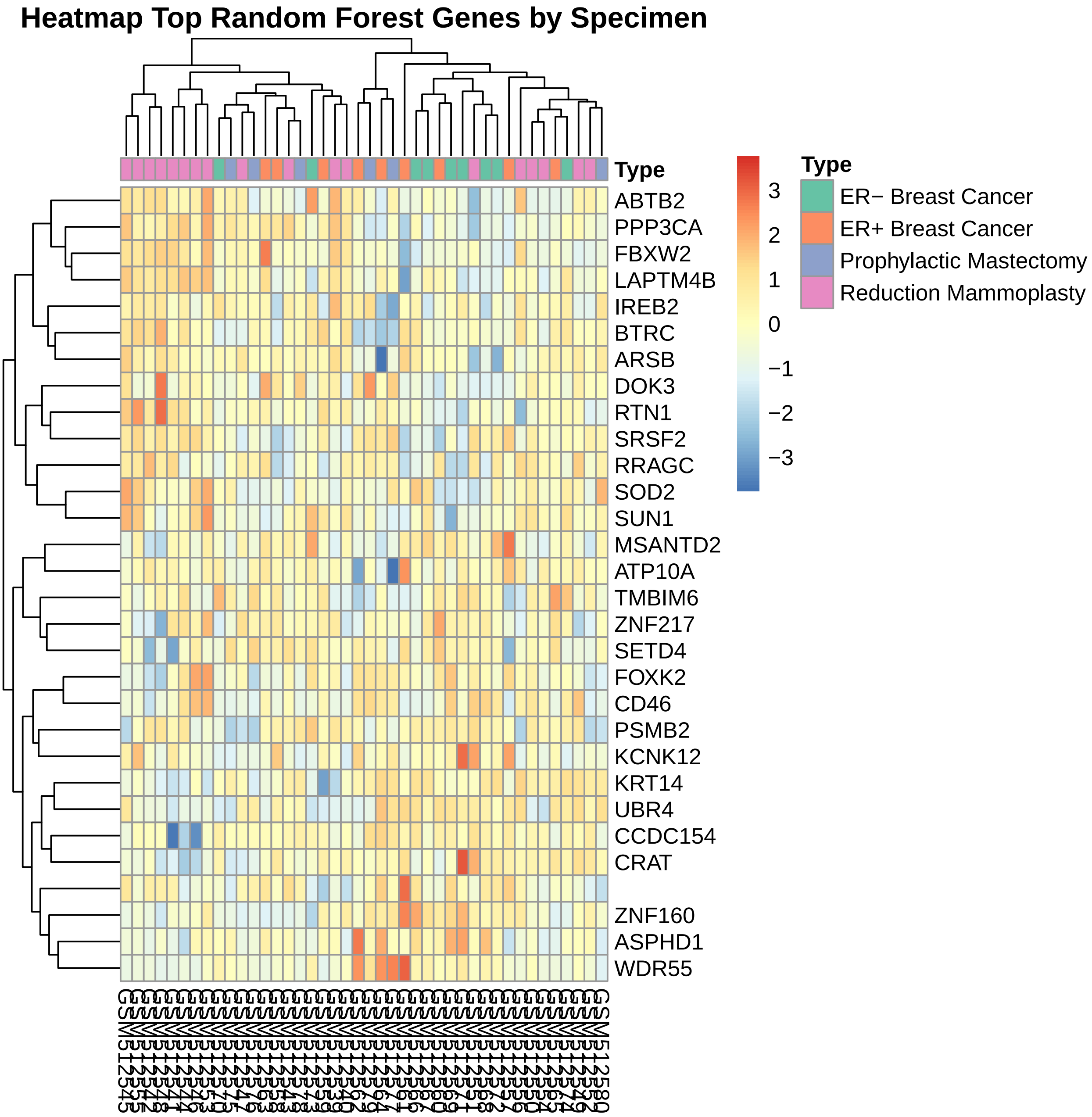


Figure 9: **Heatmap of top** 30 **ranked genes from Random Forest**. The gene absent from the y-axis, labeled as *221713_s_at*, lacks an associated external gene name in the *biomaRt* annotation database, which is why it appears as an empty entry in the visualization.

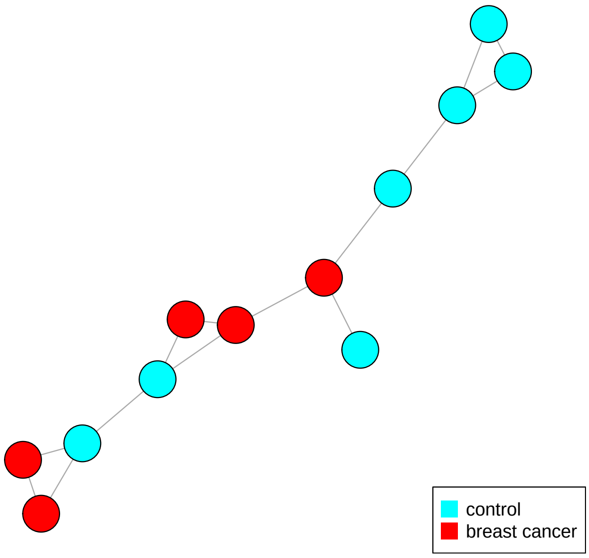


Figure 10: SCUDO results