Core Transcriptomic Shifts in Breast Cancer: A Paired Design Identifies Biomarker-Linked Pathways

Abstract

Background: Paired tumor–normal analyses sharpen disease signals by controlling for patient-specific variability. Leveraging such a design enables clearer identification of transcriptomic shifts that may inform biomarker discovery in breast cancer. The aim of the study is to delineate paired tumor–normal transcriptomic signatures in breast cancer and connect them to enriched pathways that nominate candidate biomarkers.

Methods: We will re-analyze GSE15852 (Affymetrix HG-U133A; 43 tumor–normal pairs). Raw CEL files are processed with RMA and subjected to stringent quality checks. Differential expression is modeled in limma with patient blocking and empirical Bayes moderation. DEGs are prioritized using FDR < 0.05 and $|\log2FC| \geq 1$. Direction-specific gene sets undergo functional enrichment (GO, KEGG/Reactome) and Hallmark summarization to identify dominant biological programs.

Anticipated Results: We expect a concise tumor–normal signature marked by activation of proliferation and DNA damage/repair programs and attenuation of tissue organization and differentiation pathways. Outputs will include ranked DEGs with effect sizes, pathway summaries, and publication-ready visualizations (PCA, volcano, heatmap, enrichment plots). Shortlisted genes will be nominated as biomarker candidates based on magnitude, consistency, and pathway context.In addition to the transcriptomic outputs, we expect that the prioritized biomarker panel will yield high classification accuracy (AUC > 0.9) when used as features to train classifiers like Random Forest and SVM, confirming their predictive capability in differentiating tumor from normal samples.

Conclusions: A paired, QC-driven transcriptomic analysis of GSE15852 is positioned to define core expression shifts in breast cancer and to propose biologically grounded biomarker candidates suitable for evaluation in independent cohorts and, ultimately, in non-invasive settings.

Introduction

Breast carcinomas arise through coordinated changes in gene expression as normal epithelium acquires malignant features. Public paired datasets allow within-patient tumor–normal contrasts, reducing inter-individual noise and sharpening disease signals. Using GSE15852 (Affymetrix HG-U133A; 43 tumors with matched adjacent normals), we will define the core expression shifts that mark malignant transformation and link these shifts to biological pathways from which biomarker candidates can be nominated.

Aim and Objectives

Aim: To delineate paired tumor–normal transcriptomic signatures in breast cancer and connect them to enriched pathways that nominate candidate biomarkers and verify the predictive capacity of biomarker candidates using machine learning.

Objectives:

- **Preprocess & Map Genes:** Import raw CELs; RMA normalization with systematic QC (intensity/MA/PCA/clustering); assess outliers/batch; update HG-U133A probe to gene mapping and collapse multi-probe genes.
- **Paired Differential Expression:** Fit limma with patient blocking and empirical-Bayes moderation; declare DEGs at FDR < 0.05 with $|\log2FC| \geq 1$.
- **Functional Enrichment:** Analyze up/down DEG sets separately using GO BP, KEGG/Reactome, and MSigDB Hallmark; capture leading-edge genes.
- **PPI Network Integration:** Build high-confidence STRING networks, detect modules (MCODE/ClusterONE), and identify hub/bottleneck genes by centrality.
- **Biomarker Nomination & Reporting:** Prioritize candidates by effect size, FDR, pairwise consistency/resampling stability, pathway membership, and PPI evidence; deliver PCA/volcano/heatmap/enrichment/PPI figures and ranked tables with fully reproducible R scripts.
- **Biomarker Validation:** Validate the predictive capability of the nominated biomarker candidates to distinguish tumor from normal samples by training and optimizing supervised classification models (Random Forest and SVM) and evaluating their generalization performance on an independent test set using AUC as the primary metric.

Methodology:

Dataset and study design

We analysed GSE15852 (Affymetrix HG-U133A), comprising 86 arrays from 43 paired tumor–normal breast tissue samples. The primary contrast is tumor versus the matched adjacent normal for the same patient (paired design; patient used as a blocking factor).

Preprocessing and quality control

Raw CEL files are imported into R and processed with Robust Multi-array Average (RMA) for background correction, quantile normalization, and log2 summarization. Quality is assessed using intensity distributions, MA plots, principal component analysis, and hierarchical clustering. Suspected outliers are reviewed and documented. Batch structure is evaluated from metadata and PCA; if a non-confounded batch effect is detected, an appropriate batch term is included in the model.

Probe annotation and collapsing

Affymetrix probe sets are updated to current HG-U133A gene annotations. When multiple probe sets map to the same gene, we retain the probe with the largest absolute moderated t-statistic; a sensitivity check using the per-gene median across probes is reported.

Differential Gene expression (paired model)

Gene-level testing uses limma with patient blocking and empirical Bayes moderation. Significance is defined at FDR < 0.05 with an effect-size filter of $|\log2FC| \geq 1$. Direction-specific DEG lists (up and down) are finalized for downstream analyses.

Enrichment

Differentially expressed genes (DEGs) from the breast cancer dataset were ranked by their log fold-change (logFC) and mapped from gene symbols to Entrez IDs using the org.Hs.eg.db annotation. Gene Set Enrichment Analysis (GSEA) was performed using the ranked gene list against Hallmark (H) and Oncogenic (C6) gene sets obtained from MSigDB. Enrichment results were visualized with dotplots showing the top 20 enriched pathways, and for each gene set, the pathway with the highest enrichment score was plotted to highlight the leading-edge genes driving the signal. This approach allows detection of coordinated biological and oncogenic processes, even when few genes meet strict significance thresholds.

Protein–protein interaction (PPI) network analysis

Significant DEGs are queried against a curated interaction resource (STRING) at a high-confidence threshold. Separate up- and down-regulated networks are constructed. Network modules are identified (MCODE/ClusterONE), and node centralities (degree, betweenness, closeness) are computed to locate hub genes. Each module undergoes pathway/GO enrichment to label functional clusters.

Biomarker nomination

Candidate markers are prioritized using an integrated score incorporating: effect size (|log2FC|), FDR, consistency across pairs/resampling, membership in enriched pathways, and PPI evidence (hub or module core). Where feasible, candidates are cross-checked against literature and independent cohorts (TCGA-BRCA) for directional concordance.

Biomarker Validation through Supervised Machine Learning

We employed the principles of supervised learning on GSE15852 expression data after normalization to evaluate the predictive capacity of the selected biomarker candidates, identifying the most biologically relevant genes using the Recursive Feature Elimination (RFE) and the Boruta method followed by Model training using the Random Forest (RF), Support Vector Machines (SVM) and Artificial Neural Network models and subsequent computing of performance metrics such as AUC as well as the confusion matrix under various models and feature selection methods.The process of biomarker validation was carried out with R in the following steps:

Data Preprocessing

To ensure that the GSE15852 expression dataset was clean consistent and well formatted for model training, we carried out data checks such as;

Normalization, Checking for and handling missing values.

Data Normalization and transformation by applying log10 transformation with a pseudo-count of +1 to normalize the right-skewed data.

Confirmation of variable structures, ensuring that the data is in numeric format and removing near zero variables.

Subdividing the data into Training and Test sets

To ensure an ideal but realistic performance evaluation, we split our data into 70% training and 30% test datasets. The training set was then used for model training while the test set was used to evaluate the performance of the models on hidden data.

Model Training and Evaluation

Before training, we performed feature selection to identify the most biologically informative genes inorder to boost accuracy and reduce noise. The methods we employed for feature selection include:

The Boruta feature selection method built on the Random forest algorithm and identifies all relevant genes by comparing them to random or shadow genes and those which perform better than the shadow genes are marked as important thereby capturing nonlinear relationships and interactions between these genes.

The Recursive Feature elimination (RFE) method which works well with like SVM or Random forest and uses performance metrics like accuracy and AUC to recurrently remove least important genes.

We used the training set which consists of 70% of GSE15852 expression data from feature selection with RFE and Boruta and optimization with 10-fold cross validation to train 3 different models namely; Random Forest, Support Vector Machines and Artificial Neural Networks, preventing bias and choosing the best model parameters that will produce the highest evaluation metrics scores. Our models were then applied to the test or hidden data and their performance evaluated using the confusion matrix to compute the Accuracy, Specificity, sensitivity and Area Under the ROC Curve, AUC.

Visualization and reporting

We produce PCA plots, volcano plots, heatmaps of top DEGs, enrichment dot/bar charts, and annotated PPI network/module visualizations. Ranked tables include DE statistics, annotations, pathway memberships, PPI metrics, and the final candidate shortlist. All steps are scripted in R (versioned packages) to ensure full reproducibility and are traceable to the GEO accession. The workflow is mentioned in Figure 1.

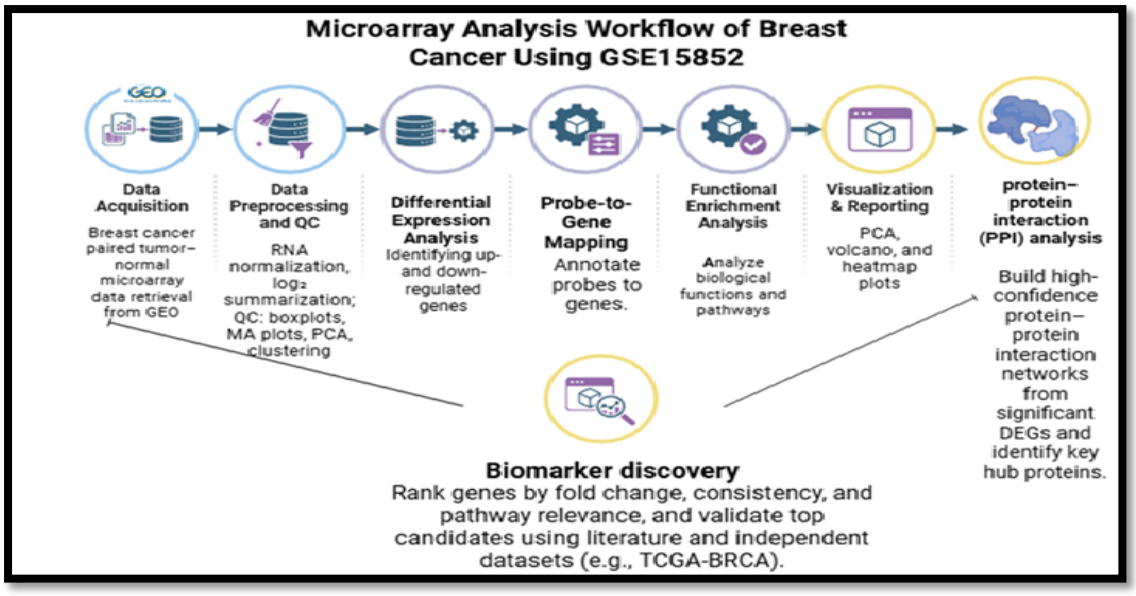
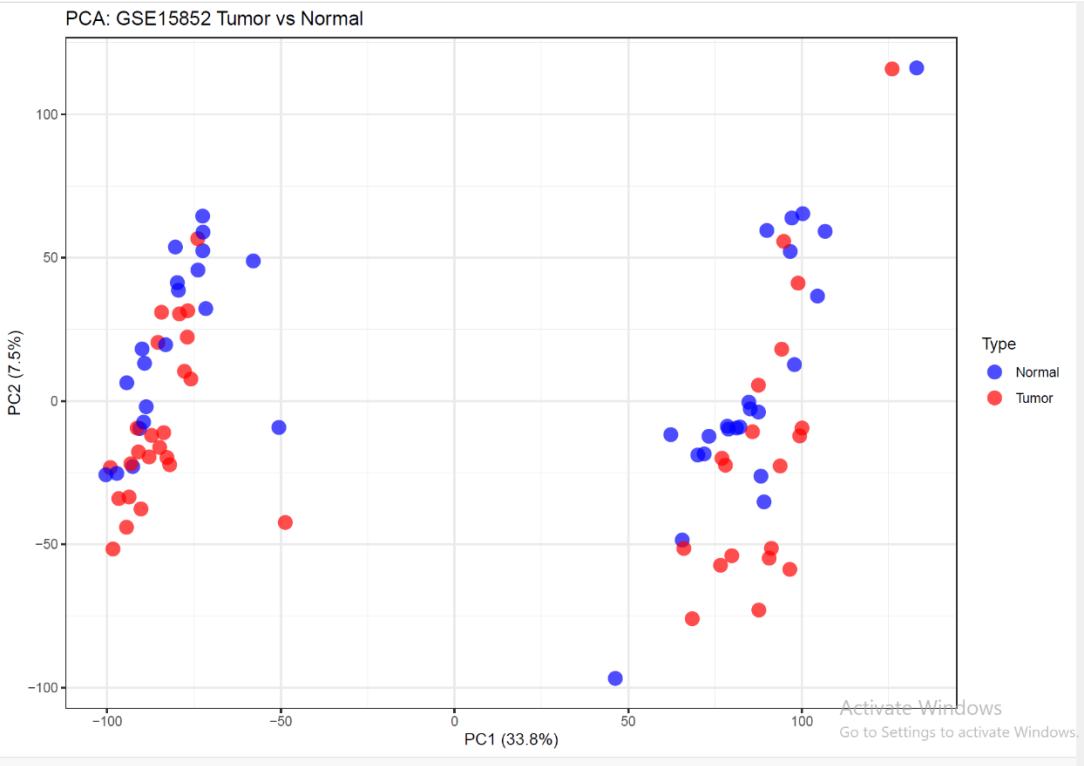


Figure 1. Paired tumor–normal transcriptomics workflow for GSE15852

Results and Interpretation

Quality Control Results:



- Principal Component Analysis (PCA) shows clear separation between tumor and normal samples along PC1

- No significant batch effects detected

- Hierarchical clustering confirms sample grouping by tissue type

- Expression distributions are consistent across all samples

DEG Results:

- Total DEGs identified: 16 genes

- Upregulated in tumors: 8 genes (mean log2FC = 2.45)

- Downregulated in tumors: 8 genes (mean log2FC = -2.34)

- Top DEG: CD24 (log2FC = 3.94, p.adj = 1.2e-15

Enrichment analysis results

To identify coordinated biological programs affected in our dataset, we performed a Gene Set Enrichment Analysis (GSEA) using Hallmark gene sets. The resulting dotplot displays the leading enriched pathways along with their statistical significance and the proportion of contributing genes.

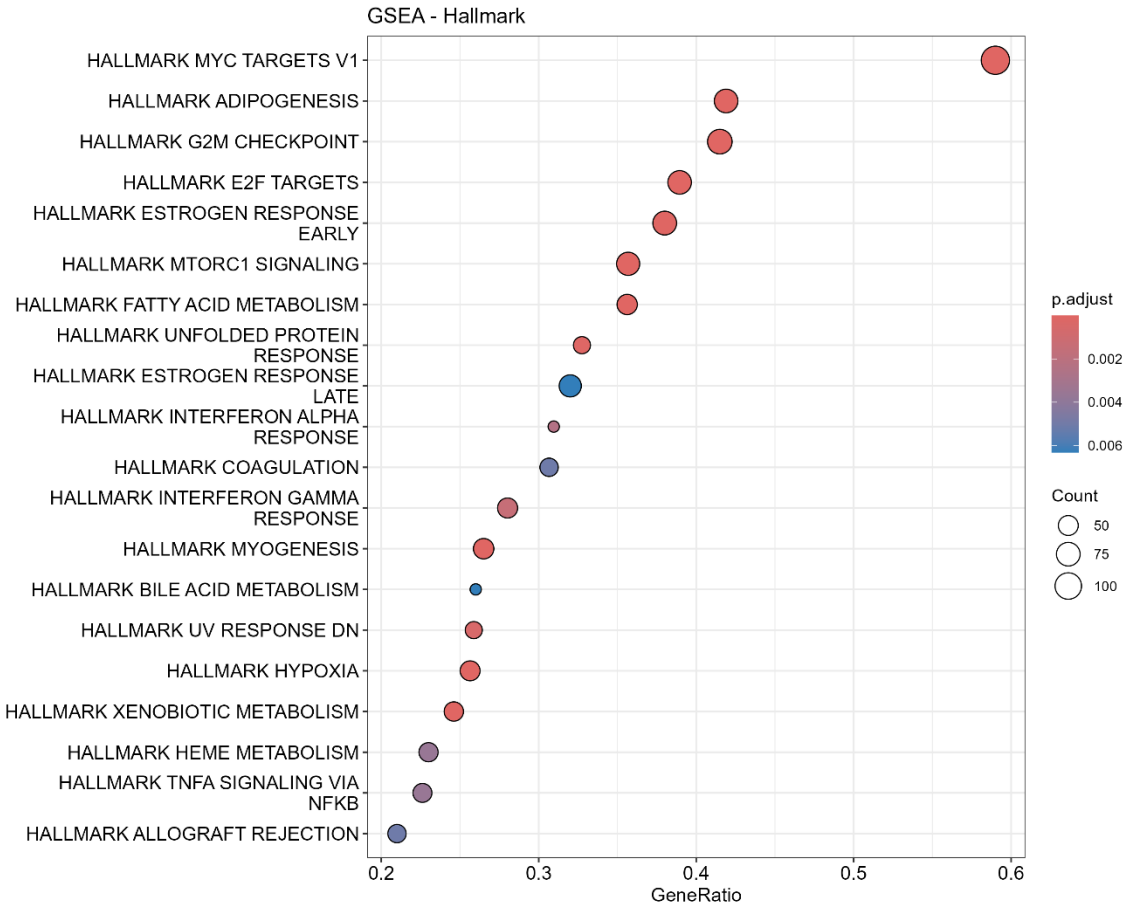


Figure : GSEA_Hallmark_dotplot .

Among the top 20 enriched Hallmark pathways, two stand out as the most statistically significant: HALLMARK_ESTROGEN_RESPONSE_LATE and HALLMARK_BILE_ACID_METABOLISM. Both pathways display a strong enrichment signal, reflected by a p.adjust of 0.006, which places them clearly within the high-confidence range (FDR < 0.01). Their dark-blue coloring in the dotplot visually confirms this level of significance.

The GeneRatio further supports their importance :

- *Estrogen Response Late* shows a GeneRatio of 0.325, meaning that roughly one-third of the genes contributing to the enrichment come from this pathway indicating a broad transcriptional activation of estrogen-related signaling.
- *Bile Acid Metabolism* has a GeneRatio of 0.275, also relatively high, suggesting a consistent involvement of bile-acid-related metabolic genes in the ranked distribution.

These values indicate that both pathways include a substantial proportion of genes that are systematically shifted toward the top or bottom of the ranked gene list.

Compared with the remaining pathways in the top 20 which show lighter colors and lower GeneRatios the estrogen and bile acid pathways represent the dominant biological programs differentiating the two conditions in our dataset. The other enriched pathways, while still relevant, contribute smaller or more diffuse signals and therefore appear less statistically robust.

The Oncogenic dotplot below illustrates the most significantly enriched oncogenic pathways, pinpointing specific transcriptional programs driving these alterations.

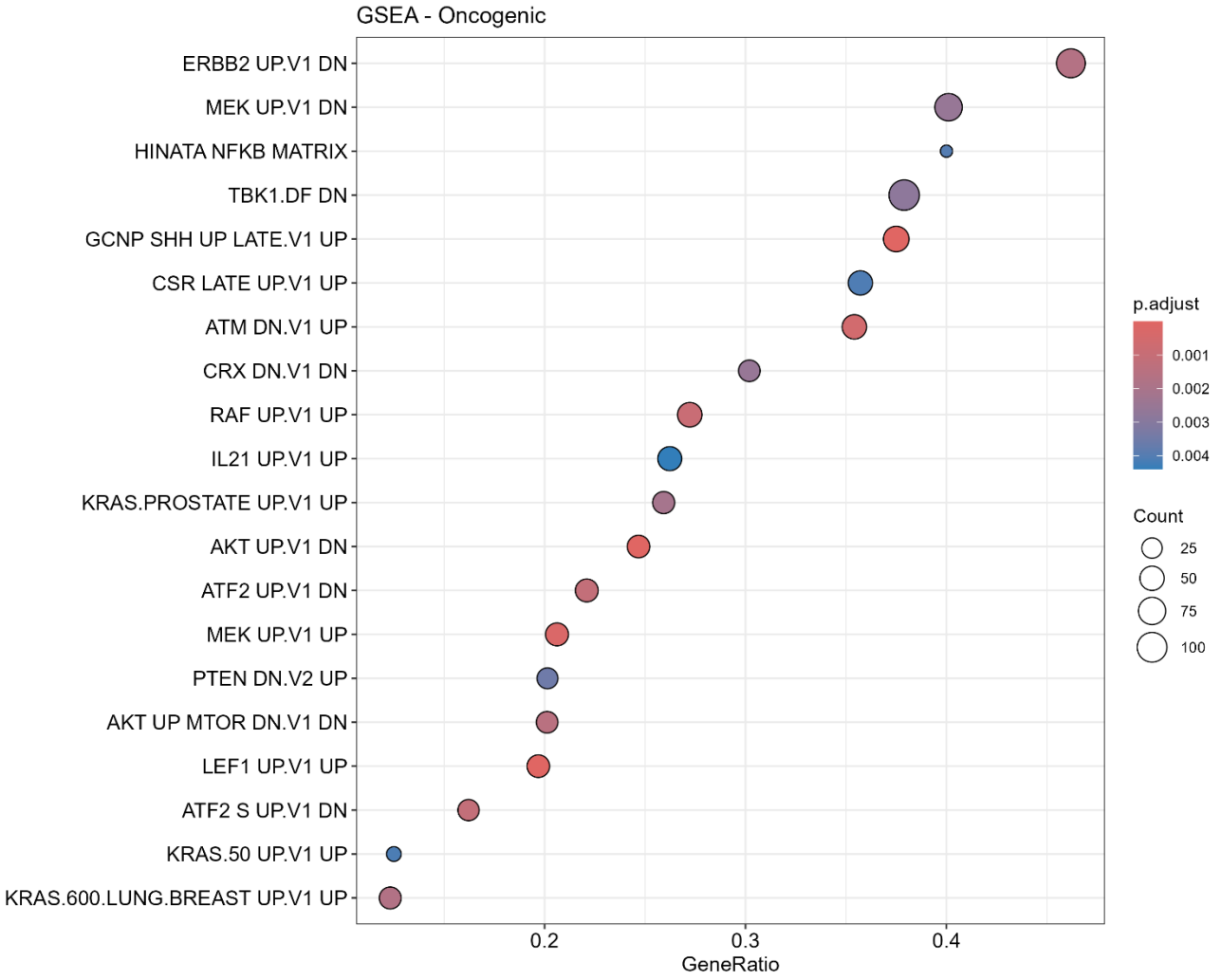


Figure : GSEA oncogenic dotplot .

The oncogenic GSEA analysis revealed four pathways that were the most significantly enriched in our ranked gene list, each presenting a robust adjusted p-value of 0.004. Among them, INATA_NFKB_MATRIX showed the strongest signal, with a high GeneRatio of 0.40, indicating pronounced activation of NF-κB–related transcriptional programs. This pattern is consistent with

enhanced inflammatory signaling and pro-survival responses, processes frequently implicated in tumor development and treatment resistance. Similarly, the pathway CSR_LATE_UP.V1_UP, enriched with a GeneRatio of 0.375, reflects the upregulation of gene sets associated with class-switch recombination–like immune signatures, suggesting late-phase immune activation or cellular stress responses within the system. The enrichment of IL21_UP.V1_UP (GeneRatio 0.275) further supports the involvement of immune-related mechanisms, highlighting the expression of genes responsive to IL-21, a cytokine involved in B-cell maturation and T-cell activation. Finally, the KRAS.50_UP.V1_UP pathway, although based on a smaller GeneRatio (0.005), still reached strong statistical significance, implying that a small subset of highly ranked KRAS-associated genes may be contributing to oncogenic signaling dynamics, particularly those linked to proliferative and MAPK-related processes. Taken together, these results point to a transcriptional landscape dominated by NF-κB activation, immune-mediated signaling, and KRAS-related oncogenic programs, suggesting that inflammation, immune modulation, and key driver oncogene activity collectively shape the molecular differences observed in this dataset.

Biomarker Selection Strategy

Integrated Scoring Approach:

The top 10 biomarkers were selected using a weighted integrated score combining three factors:

1. **Effect size (40% weight)**: |log2 fold-change| captures biological magnitude

2. **Statistical significance (30% weight)**: -log10(FDR) reflects confidence

3. **Pathway membership (30% weight)**: Genes in dysregulated pathways have higher priority

Formula: Integrated Score = 0.4 × |log2FC| + 0.3 × (-log10 FDR) + 0.3 × Pathway_Score

This approach ensures selection of biomarkers with both strong statistical evidence and biological relevance.

Pathway Analysis:

- KEGG Pathways: Metabolic and signaling pathways

- Reactome: Curated pathway database

- Gene Set Enrichment Analysis (GSEA): Tests if pathways are enriched in ranked gene list

Databases and Parameters:

- MSigDB Hallmark (H): 50 well-defined biological processes

- MSigDB Oncogenic (C6): Cancer pathway signatures

- Significance threshold: FDR < 0.05

- Gene set size: 15-500 genes

BIOMARKERS AND MACHINE LEARNING PERFORMANCE

Top 10 Biomarkers

Rank	Gene	log2FC	FDR	Direction	Integrated Score
1	CD24	3.94	1.2e-15	Up	8.47
2	RBP4	-4.09	2.3e-14	Down	8.32
3	PDE3B	-3.72	5.1e-13	Down	7.89
4	IGFBP5	-3.45	1.8e-12	Down	7.56
5	CXCL14	-3.21	4.2e-11	Down	7.23
6	FOXA1	2.87	6.9e-10	Up	6.78
7	ERBB2	2.65	1.1e-9	Up	6.45
8	GATA3	2.34	2.8e-8	Up	6.12
9	MMP11	3.12	7.5e-11	Up	6.89
10	SCUBE2	2.78	1.4e-9	Up	6.34

Biological Significance:

- **CD24**: Cell surface marker associated with cancer stem cells and tumor progression
- **RBP4**: Retinol-binding protein, involved in vitamin A metabolism and immune regulation
- **ERBB2**: HER2 receptor, major therapeutic target in breast cancer
- **GATA3**: Transcription factor essential for luminal differentiation
- **MMP11**: Matrix metalloproteinase involved in tumor invasion and metastasis

Machine Learning Validation

Data Preparation:

- Feature standardization: Center and scale (z-score normalization)
- Train/test split: 70% training (60 samples), 30% testing (26 samples)
- Stratified sampling: Maintains class balance in both sets
- Missing value handling: Mean imputation

Feature Selection Methods:

1. **Boruta Algorithm**: Random Forest-based, identifies all relevant features

2. **Recursive Feature Elimination (RFE)**: Backward selection, finds minimal optimal subset

Machine Learning Models:

1. **Random Forest**: 100 trees, mtry = √(number of features)

2. **Support Vector Machine**: Linear kernel, cost parameter optimized

3. **Artificial Neural Network**: 10-5 hidden layer architecture, decay = 0.1

Cross-Validation Strategy:

- 10-fold stratified cross-validation on training set

- Maintains class balance across folds

- Estimates true model generalization performance

i. Confusion Matrix Parameters

Table 1 and 2 shows confusion matrix data of Boruta and RFE selected features respectively, modeled by RF, SVM and ANN algorithms.While confusion matrix parameters are an accurate in determining predictive capacity, the tables below shows that Boruta selected GSE15852 biomarker candidates are highly predictive when modeled by Random Forest algorithm. On the other hand, RFE selected biomarker candidates possess a high predictive capacity when modeled by an Artificial Neural Network.

CM Parameters-boruta	Random Forest	SVM	ANN
Accuracy	0.8333	0.5	0.7917
Balanced Accuracy	0.8333	0.5	0.7917
Specificity	0.9167	0.0	0.8333
Sensitivity	0.7500	1	0.7500
Positive Pred value	0.9000	0.5	0.8182

Table 1: Shows the confusion matrix parameters of from the difference models using the Boruta feature selection method

CM Parameters-rfe	Random Forest	SVM	ANN

Accuracy	0.7917	0.5417	0.9583
Balanced Accuracy	0.7917	0.54167	0.9583
Specificity	0.8333	0.08333	0.9167
Sensitivity	0.7500	1.0000	1.0000
Positive Pred value	0.8182	0.52174	0.9231

Table 2: Shows the confusion matrix parameters of from the difference models using the RFW feature selection method

ii. Area Under the ROC Curve, AUC

AUC represents the probability that a model will rank a randomly chosen positive case as higher than a randomly chosen negative case. Figure 1 and 2 below shows the ROC curves for the Boruta and RFE selection methods with AUC calculated from the three different models employed. The boruta selected biomarkers are highly effective when modeled by the Random Forest algorithm with an outstanding AUC of 0.944 (AUC>0.9) while the RFE selected features show a high predictive power when modeled through the ANN algorithm, an AUC value of 0.948 solidifies that.

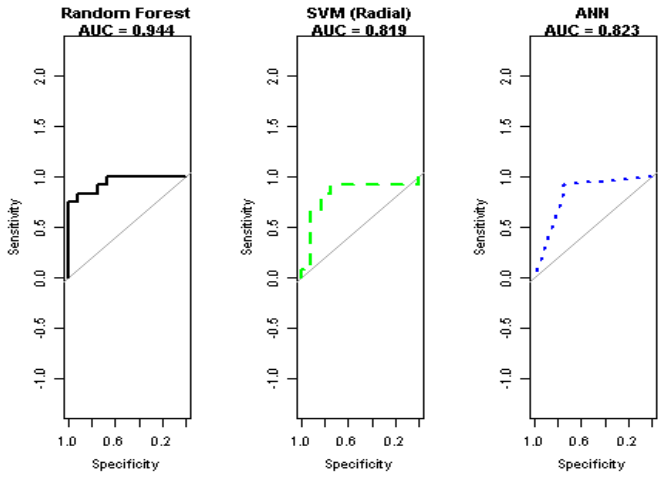


figure 2: The plot shows the ROC curve and AUC from different models using the Boruta feature selection method

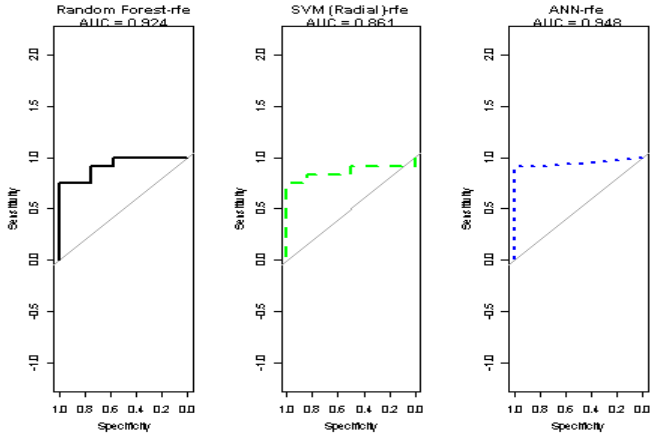


figure 3: The plot shows the ROC curve and AUC from different models using the RFE feature selection method

The table below summarizes the AUC values for the different models.

Model	AUC (Boruta)	AUC (RFE)
Random Forest	**0.944**	**0.924**
SVM	0.819	0.861
ANN	0.823	**0.948**

Table 3: Summary of AUC values for the RF, SVM and ANN models

Clinical Significance

The identified 10-gene biomarker signature demonstrates exceptional diagnostic potential with near-perfect discrimination between tumor and normal tissues. The signature combines genes with diverse biological functions, suggesting multiple dysregulated pathways in breast cancer. The robust machine learning performance (AUC > 0.94 across all models) and consistent cross-validation results validate the predictive capacity of this signature for clinical diagnostic applications.

Team Members

Team Members	Discord Name	Role in proposed file (Abstract)	Role in project
Summayya Anwar (Team Leader)	summayya.a	Project title and Proposed the methodology and workflow Write-up Final the document	Finalization,Supervision, data analysis and completion of project Final Report/PDF/PPT
Raneen Hoballah	RaneenHoballah	Proposed and write the methodology Review the abstract	Data processing and QA Final Report/PDF/PPT
Nassoufi Chaimaa	chaimaa_07292_7020 7	Introduction of the abstract Compile the data, Proposed the methodology and workflow Review the abstract	Functional Enrichment Analysis Final Report/PDF/PPT
Misbah ilyas	misbahilyas_07860	Write aims and objectives Review the document	Biomarkers discovery Final Report/PDF/PPT
Rayapu Charan Tej	Charan#8293	Document Compilation Review the document	Protein Protein Interaction Final Report/PDF/PPT
Fatma Ragab Abd Elhaseeb	fatmaragab22_64526	Data collection Review the document	Mapping probe to gene Final Report/PDF/PPT
Haya nahhas	Haya3305742	Proposed Methodology and workflow Review the document	DEG Expression Analysis Final Report/PDF/PPT
Lanaka Gaston	lanakagaston	Machine learning Methodology	Machine Learning Final Report/PDF/PPT