AJIT-e Academic Journal of Information Technology
10.5824/ajite.2026.01.002.x

2026 . Winter – Kış . Volume – Cilt: 17 . Issue – Sayı: 1

Multiomics Analysis of Gene Expression and Drug Sensitivity Relationship in Breast Cancer Subtypes*

Adem Sonuvar, *Akdeniz University, Department of Biostatistics and Medical Informatics, ademsonuvar@gmail.com, 0009-0004-2270-3812*

Esra Tokur Sonuvar, *Leipzig University Medical Data Science, Institute for Medical Informatics, tatistics, and Epidemiology, esra.tokursonuvar@medizin.uni-leipzig.de, 0000-0002-1279-5192*

Uğur Bilge, *Akdeniz University, Department of Biostatistics and Medical Informatics, ubilge@gmail.com, 0000-0002-5186-1092*

ABSTRACT

Breast cancer is a complex disease exhibiting distinct clinical behaviors due to its molecular heterogeneity. In this study, 30 breast cancer cell lines were categorized into two main analysis groups: 'Hormone/HER2 Positive' (n=12) and 'TNBC' (triple-negative breast cancer). Gene expression profiles (RNA-seq), proteomic data (RPPA), and drug sensitivity to seven different kinase inhibitors were analyzed using the E-MTAB-4801 dataset. t-SNE and hierarchical clustering demonstrated a separation based on the overall expression profiles of these two groups DEG (edgeR; FDR<0.05 and |logFC|>1) and enrichment analyses (GO/KEGG via clusterProfiler; Hallmark via msigdbr/fgsea) identified significant activation of estrogen response and metabolic pathways in the Hormone/HER2 Positive group, whereas immune response and epithelial-mesenchymal transition pathways were active in the TNBC group. Machine learning methods identified potential biomarker candidates (FAM176A, CACNG1, GPR77) crucial for discriminating between the two groups. Model performance was evaluated using ***nested cross-validation (five outer folds)****. Across outer folds,* ***LASSO*** *achieved* ***Accuracy=0.833****,* ***F1=0.865****,* ***Sensitivity=0.889****,* ***Specificity=0.750****;* ***Random Forest*** *achieved* ***Accuracy=0.800****,* ***F1=0.833****,* ***Sensitivity=0.833****,* ***Specificity=0.750****; and the* ***Decision Tree*** *showed more variable performance with mean* ***Accuracy=0.700****. Furthermore, analyses using RPPA data confirmed that the 4E-BP1 protein expression level is an important determinant of sensitivity to mTOR inhibitors. However, findings are based on 30 breast cancer cell lines, which limits statistical power and increases the risk of overfitting, and no independent clinical cohort was available for external validation; therefore, the identified biomarkers should be considered candidate biomarkers requiring validation in patient cohorts. Accordingly, the results are intended to support research-stage subtype stratification and prioritization of biomarker candidates rather than immediate clinical decision-making or drug selection providing a prioritized shortlist of candidates for follow-up validation in patient tumor cohorts.*

Keywords : Breast cancer, Multi-omics, Gene expression profile, Drug sensitivity, Biomarkers, TNBC, Hormone Receptor

* This study was previously presented as an oral presentation at the 16th Turkish Congress of Medical Informatics (May 22–23, 2025, Ankara, Türkiye) and published in the conference proceedings

Araştırma Makalesi / Research Article
Date Received: 25.08.2025
Publishing Date: 28.02.2026

AJIT-e Academic Journal of Information Technology
10.5824/ajite.2026.01.002.x

2026 . Winter – Kış . Volume – Cilt: 17 . Issue – Sayı: 1

Meme Kanseri Alt Tiplerinde Gen Ekspresyonu ve İlaç Duyarlılığı İlişkisinin Multiomik Analizi

ÖZ

*Meme kanseri, moleküler heterojenitesi nedeniyle farklı klinik davranışlar sergileyen karmaşık bir hastalıktır. Bu çalışmada, 30 meme kanseri hücre hattı, 'Hormon/HER2 Pozitif' (n=12) ve 'TNBC' (üçlü negatif, meme kanseri) olmak üzere iki ana analiz grubuna ayrılarak incelenmiştir. Bu gruplar arasındaki gen ekspresyon profilleri (RNA-seq), proteomik veriler (RPPA) ve yedi farklı kinaz inhibitörüne karşı ilaç duyarlılıkları, E-MTAB-4801 veri seti kullanılarak analiz edilmiştir. t-SNE ve hiyerarşik kümeleme analizleri, bu iki grubun genel ekspresyon profillerine göre ayrıştığını göstermiştir. Diferansiyel gen ekspresyon analizi edgeR ile gerçekleştirilmiş; anlamlı genler FDR<0.05 ve |logFC|>1 ölçütleriyle belirlenmiştir. Fonksiyonel zenginleştirme analizleri (GO ve KEGG: clusterProfiler; Hallmark: MSigDB/msigdbr ve fgsea), Hormon/HER2 Pozitif grubunda östrojen yanıtı ve metabolik yolakların, TNBC grubunda ise immün yanıt ve epiteliyal-mezenkimal geçiş ile ilişkili yolakların aktive olduğunu belirlemiştir. Makine öğrenimi yöntemleri ile iki grubu ayırt etmede önemli olan potansiyel biyobelirteçler (FAM176A, CACNG1, GPR77) tanımlanmıştır. Modellerin performansı nested çapraz doğrulama (5 dış kat/outer folds) ile değerlendirilmiştir; LASSO lojistik regresyon ve Random Forest modelleri ortalama olarak sırasıyla **%83.3 ve %80.0 doğruluk**, **%86.5 ve %83.3 F1-skoru**, **%88.9 ve %83.3 duyarlılık** ve **%75.0 ve %75.0 özgüllük** sağlamıştır; karar ağacı modeli ise daha değişken olup ortalama **%70.0 doğruluk** göstermiştir. Ayrıca, RPPA verileri kullanılarak, 4E-BP1 protein ekspresyon seviyesinin mTOR inhibitörlerine duyarlılıkta önemli bir belirleyici olduğu doğrulanmıştır. Bu çalışma, multi-omik verilerin entegratif analizinin meme kanseri alt tiplerini ayırt etmede ve potansiyel biyobelirteç adaylarını ortaya koymada değerli bilgiler sağlayabileceğini göstermektedir. Ancak bulgular 30 hücre hattına dayandığından istatistiksel güç sınırlıdır ve overfitting riski artabilir; ayrıca bağımsız klinik kohortlarda dış doğrulama bulunmadığı için tanımlanan biyobelirteçler aday biyobelirteçler olarak değerlendirilmelidir. Sonuçlar, klinik karar veya ilaç seçimi yerine, araştırma düzeyinde alt tip stratifikasyonu ve hasta tümör kohortlarında doğrulanacak adayların önceliklendirilmesine katkı sağlamayı amaçlamaktadır.*

Anahtar Kelimeler : Meme kanseri, Multiomiks, Gen ekspresyon profili, İlaç duyarlılığı, Biyobelirteçler, TNBC, Hormon Reseptörü

Araştırma Makalesi / Research Article
Date Received: 25.08.2025
Publishing Date: 28.02.2026

10.5824/ajite.2026.01.002.x

INTRODUCTION

Breast cancer is a heterogeneous disease with molecular subtypes that show significant differences in clinical course and treatment response (Perou et al., 2000). Multi-omics refers to the integrative analysis of multiple molecular layers (e.g., genomics, transcriptomics, proteomics and related omics) measured on the same biological system to obtain a more comprehensive view of disease biology than any single layer alone (Hasin et al., 2017; Wörheide et al., 2020). Integration strategies are commonly described as early (feature-level), intermediate (latent representation), and late (model-level) integration (Baião et al., 2025) In cancer research, multi-omics integration is increasingly used to improve molecular subtyping, identify pathway-level dysregulation, and prioritize biomarker candidates by leveraging complementary signals across omics layers (Picard et al., 2021). "Multi-omics" approaches have provided important advances in the molecular characterization of breast cancer subtypes through the integration of genomic, transcriptomic, and proteomic data. Consistent with this growth, a recent comprehensive bibliometric study introduced an interactive science-mapping platform (BiblioMaps) and documented the rapid expansion and thematic evolution of multiomics research at a global scale, highlighting cancer-focused trends among dominant themes (Koçak et al., 2025).

From a field perspective, bibliometric mapping studies have shown that multi-omics research has expanded rapidly and that interactive visualization platforms can help structure the thematic evolution of the domain; for example, the BiblioMaps study provides a comprehensive bibliometric analysis of multi-omics trends and research landscapes.

The classification into main subtypes including hormone receptor positive (ER+), human epidermal growth factor receptor 2 positive (HER2+), and triple-negative (TNBC) plays a critical role in determining treatment strategies (Sørlie et al., 2001). The TNBC subtype is characterized by not expressing estrogen receptor (ER), progesterone receptor (PR), and HER2 receptors. Molecular-level studies have shown that these subtypes carry distinct genetic signatures, activate different signaling pathways, and have different oncogenic drivers (Bertucci et al., 2005). While ER+ tumors depend on estrogen signaling, HER2+ tumors are based on overactivation of the ERBB2 receptor. The TNBC subtype has a more aggressive phenotype and targeted therapy options are limited (Lehmann et al., 2011). The cell line classification in this study is based on the subtypes defined by Jastrzebski et al. (2018).

Our analyses were conducted over two main groups: 'Hormone/HER2 Positive' and 'TNBC'. Molecular differences between subtypes also lead to significant variability in drug responses (Iorio et al., 2005). ER+ tumors are sensitive to hormonal therapies, HER2+ tumors to anti-HER2 agents. Standard chemotherapies are the main treatment option for TNBC, but changes in PI3K/AKT/mTOR and MAPK pathways offer potential for targeted therapies

(Elkabets et al., 2013; Huynh et al., 2020). Current research reveals that immune checkpoint inhibitors and antibody-drug conjugates show promising results in TNBC (Conway et al., 2020; Jiang et al., 2019). Next-generation genome sequencing technologies contribute to the identification of patient-specific genetic variants (Schmid et al., 2018). Drug resistance is an important clinical problem. Integrative multi-omics approaches play a significant role in identifying subtype-specific molecular characteristics and drug sensitivity markers. Breast cancer cell lines provide valuable models for the integration of data at different molecular levels (Polyak et al., 2009).

In this study, we integrate RNA-seq, RPPA, and kinase-inhibitor drug sensitivity (IC50) data from 30 breast cancer cell lines (E-MTAB-4801) to characterize molecular differences between Hormone/HER2-positive and TNBC groups and to prioritize candidate biomarkers supporting subtype discrimination. We further examine the association between 4E-BP1 protein expression and sensitivity to mTOR inhibitors within this multi-omics framework. The work provides a reproducible, hypothesis-generating resource that supports research-stage molecular stratification and guides follow-up validation in independent patient cohorts.

Breast cancer has served as a key testbed for multi-omics research because molecular heterogeneity directly shapes treatment selection and resistance patterns. Large patient-tumor consortia have shown that integrating genomic and transcriptomic information refines subtype biology and reveals clinically relevant pathway alterations beyond single-layer analyses (The Cancer Genome Atlas Network, 2012). More recently, proteogenomic studies have demonstrated that protein- and phosphoprotein-level programs can both recapitulate transcriptome-defined subtypes and further stratify tumors, highlighting that post-transcriptional regulation and signaling activity may not be inferred reliably from RNA alone (Mertins et al., 2016). Together, these lines of work motivate integrative designs that combine transcriptomic and proteomic layers when the goal is to connect subtype biology to actionable signaling mechanisms.

Methodologically, multi-omics integration approaches are commonly described along a continuum from early (feature-level concatenation), to intermediate (shared latent representations), to late (model-level ensembling) strategies. Early integration is straightforward but can be sensitive to high dimensionality and scale differences; intermediate integration seeks a lower-dimensional representation that captures cross-layer covariance; and late integration combines models trained on separate omics layers, often improving robustness when missingness differs across layers (Hasin et al., 2017). A frequently used intermediate integration family includes network-based approaches that fuse similarity structures learned from each layer to create a joint sample graph for clustering or classification; for example, similarity network fusion has been applied across biomedical domains to combine complementary signals from multiple omics (Wang et al., 2014). Factor-based approaches likewise aim to separate shared from layer-specific variation, supporting interpretation while

10.5824/ajite.2026.01.002.x

acknowledging that distinct omics may encode different biological processes (Argelaguet et al., 2018). In cancer settings, these approaches are often coupled with downstream machine learning to enable subtype discrimination, feature prioritization, and pathway-level interpretation.

Beyond patient-tumor cohorts, standardized cancer cell line resources provide controlled systems in which multi-omics profiles can be linked to pharmacological response, enabling hypothesis generation about drug sensitivity mechanisms. Large pharmacogenomic panels have shown that genotype and expression features can explain substantial variability in drug response across cancer types, although cross-study reproducibility and context dependence remain important caveats (Barretina et al., 2012). Compared with patient cohorts, cell lines enable consistent experimental conditions and repeated perturbations, but they cannot fully recapitulate the tumor microenvironment, stromal interactions, and immune context that shape clinical response. Therefore, cell line findings are best interpreted as research-stage signals requiring confirmation in patient cohorts and more complex model systems.

Within this state-of-the-art landscape, the present study focuses on integrative analysis of transcriptomic, proteomic, and drug sensitivity data in a breast cancer cell line panel to (i) characterize subtype-associated molecular programs, (ii) prioritize candidate molecular features that consistently support subtype discrimination across machine learning models, and (iii) connect signaling-level measurements with targeted therapy response in a hypothesis-generating framework. By combining differential expression, functional enrichment, and nested cross-validation-based machine learning within a single workflow, we aim to provide an interpretable and reproducible analysis path that supports downstream validation rather than immediate clinical translation.

1. MATERIAL AND METHODS

1.1. Data Sets and Sources

In this study, a panel containing a total of 30 breast cancer cell lines, including 18 TNBC, 4 ER+, 4 HER2+, and 4 ER+/HER2+, previously comprehensively characterized by Jastrzebski and colleagues (2018), was used. This panel provides a rich resource for integrated multi-omics analyses. Transcriptomic (RNA-seq; normalized read counts for all genes), proteomic (RPPA; reverse phase protein array data for 152 different protein/phospho-protein epitopes), pharmacogenomic (Drug Response; relative viability data measured after 72-hour treatment with seven different kinase inhibitors [AZD8055, BEZ235, GDC0941, MK2206, PD0325901, lapatinib, foretinib] and calculated IC50 values), and genomic (Mutation and Copy Number; mutations determined by targeted DNA sequencing and whole genome copy number profiles) data used in the study were obtained from the ArrayExpress database (Accession Number: E-

MTAB-4801). Cell lines were classified according to ER+, HER2+, and TNBC (ER-/PR-/HER2-) subtypes defined by Jastrzebski et al. (2018). However, whether the RPPA dataset used in this study, targeting 152 epitopes, contains the PR protein was not specifically examined, and analyses focused on available RNA-seq and RPPA data.

1.2. Data Analysis

The analytical approach followed in our study, and the multi-omics data integration workflow are summarized in Figure 1. This workflow includes stages of data collection, quality control, dimensionality reduction, differential gene expression analysis, functional enrichment, biomarker identification with machine learning, and drug sensitivity analysis. Various preprocessing steps were applied to the datasets before analysis. RNA-seq read counts were subjected to log2 transformation after filtering genes showing low variance or expressed at low levels. RPPA data were normalized to eliminate the effects of technical variations. IC50 values representing drug response were normalized with log10 transformation. Missing values in all datasets were completed using the k-nearest neighbor (k-NN) algorithm. Dimensionality reduction techniques were used to classify cell lines according to their molecular profiles and visualize their relationships. Principal Component Analysis (PCA) was applied to determine the main variance components in gene expression data and examine the separation of cell lines according to subtypes. The t-distributed Stochastic Neighbor Embedding (t-SNE) method was also used for visualization of high-dimensional data in low-dimensional space. Hierarchical clustering analysis was performed using the Ward.D2 linkage method to group cell lines based on their gene expression profiles.

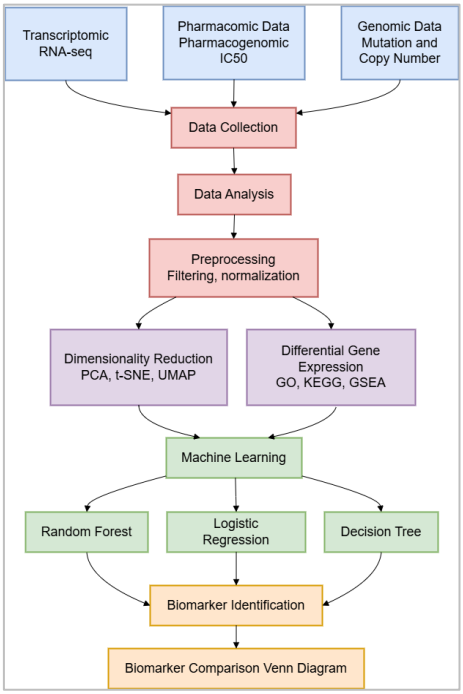

Figure 1: Multi-Omics Data Analysis Workflow Diagram for Breast Cancer Cell Lines

10.5824/ajite.2026.01.002.x

Data sources, analytical methods, and integration approaches used in the study are shown.

In this study, to determine the fundamental molecular differences between breast cancer subtypes and biomarkers associated with drug response, cell lines were examined in two main groups under the guidance of our supervisor. The 30 cell lines characterized by Jastrzebski et al. (2018) (18 TNBC, 4 ER+, 4 HER2+, and 4 ER+/HER2+) were evaluated in two main groups according to targeted treatment approaches: 'Hormone/HER2 Positive' (n=12, including ER+, HER2+, and ER+/HER2+ lines) and 'TNBC' (n=18). While distinct biological differences exist between ER+ and HER2+ tumors, exploratory analyses revealed that separating these groups resulted in insufficient sample sizes (n=4 each) for robust statistical inference in machine learning models. Therefore, merging them into a 'Hormone/HER2 Positive' group was a necessary strategy to achieve adequate statistical power for comparing non-TNBC versus TNBC phenotypes. Statistically significant gene expression differences between these two main groups were determined using the R/Bioconductor package DESeq2 (v1.34.0) (Anders et al., 2014). Threshold values of FDR < 0.05 and absolute log2 fold change (|log2FC|) > 1 were used for significance. The main rationale for this grouping is the clear separation of subtypes with potential to respond to hormonal and HER2-targeted therapies from TNBC and increasing the statistical power of analyses by combining small sample sizes.

Functional enrichment analyses were performed to evaluate the functional meaning and biological roles of differentially expressed genes. Enrichments in Gene Ontology (GO) terms (biological process, molecular function, cellular component) and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways were analyzed with the R/Bioconductor package clusterProfiler (v4.2.2) (Yu et al., 2012). Additionally, Gene Set Enrichment Analysis (GSEA) was performed using the MSigDB Hallmark (v7.4) gene set collection to evaluate changes in pathway activity based on expression scores of all genes.

Various approaches were adopted to integrate transcriptomic, proteomic, and drug response data and reveal inter-layer relationships. These include Multiple Factor Analysis (MFA) where features from different omic layers are weighted and integrated, DIABLO (within the mixOmics package) - Data Integration Analysis for Biomarker discovery using Latent cOmponents - which aims to determine biological relationships by maximizing correlations between omic datasets, and network-based integration methods where multi-omic data are related using molecular interaction networks.

1.3. Biomarker Identification with Machine Learning

An analysis workflow containing three different machine learning algorithms was applied to identify potential biomarkers that can distinguish breast cancer subtypes and predict drug response. Random Forest, Logistic Regression, and Decision Tree algorithms were used to model the distinction between 'Hormone/HER2 Positive' and 'TNBC' analysis

groups (Schmid et al., 2018; Wei et al., 2022; Yap et al., 2019). Random Forest, as a nonlinear machine learning model, has been shown to be highly effective in multi-omic data analyses with its ability to capture more complex data structures than linear and logistic regression (Schmid et al., 2018).

A feature set derived from differentially expressed genes was used for machine learning analyses. To reduce optimistic bias in this small-sample setting, model performance was estimated using **nested cross-validation with five outer folds**; hyperparameter tuning was performed within the training data of each outer fold. Performance was summarized on held-out outer folds using **Accuracy, F1-score, Sensitivity, and Specificity**. The Mean Decrease Gini metric was calculated to determine gene importance in the Random Forest model. For logistic regression, the LASSO approach (L1 regularization, α=1) was used and the optimal lambda value was selected via internal cross-validation.

The rpart algorithm was applied in the Decision Tree model. The optimum model was created by testing different splitting criteria and complexity parameters. The rpart.plot package was used for model visualization to create a decision tree diagram. The diagram shows sample numbers at each node and splitting criteria for each branch. Venn diagram analysis was performed to visualize intersections of important genes identified by different models. The performance of models on the test set was evaluated and compared using accuracy, sensitivity, specificity, and kappa metrics. 5-fold cross-validation method was applied for all models. This methodology is a standard approach whose effectiveness has been proven in breast cancer survival prediction and biomarker identification (Yap et al., 2019). Analyses were performed using the R programming language and related packages (randomForest, caret, glmnet, rpart, e1071, pROC, VennDiagram, etc.).

Reverse phase protein array (RPPA) data were used to examine the relationship between 4E-BP1 protein expression and phosphorylation status with sensitivity to mTOR inhibitors (AZD8055, BEZ235) **quantified by IC50 values.** Drug response analyses were conducted on log10-transformed IC50 values.The relationship between 4E-BP1 expression and IC50 was evaluated by Spearman correlation analysis. To examine the effect of protein expression levels on drug sensitivity in more detail, cell lines were stratified into **high vs low 4E-BP1** expression groups based on the median value, and differences in log10(IC50) between groups were tested using the **Wilcoxon rank-sum test**. Simple **linear regression** models were fitted to summarize the direction and strength of the association. These analyses were interpreted as **associative** and do not imply causality. In these models, the effects of 4E-BP1 total protein expression, and other potential variables were examined. To comprehensively characterize the molecular properties of subtypes, enrichments in GO terms (biological process, molecular function, cellular component) and KEGG pathways were analyzed with the R/Bioconductor package clusterProfiler (v4.2.2) (Yu et al., 2012). GSEA (Gene Set

10.5824/ajite.2026.01.002.x

Enrichment Analysis) was performed using the MSigDB Hallmark (v7.4) gene set collection to evaluate changes in pathway activity based on all gene expression profiles.

2. Results

2.1. Gene Expression Profiles and Subtype Separation

Differences between cell lines while preserving complex relationships in high-dimensional expression data. As shown in Figure 2, as a result of t-SNE analysis, cell lines in the 'Hormone/HER2 Positive' group (blue dots) and 'TNBC' group cell lines (orange dots) are distinctly separated. While the Hormone/HER2 Positive group generally clusters in the lower and left regions of the graph, TNBC cell lines are predominantly positioned in the upper and right regions. This separation confirms that there are significant differences at the transcriptomic level between the two analysis groups.

Notably, cell lines in the TNBC group show a wider distribution within themselves, suggesting that this subtype may have a more heterogeneous structure from a molecular perspective. These results of t-SNE analysis support the molecular separation of breast cancer subtypes observed in previous studies and show that TNBC is distinctly differentiated from other subtypes in particular. These findings clearly reveal molecular differences between the two main analysis groups defined as 'Hormone/HER2 Positive' and 'TNBC' and establish a solid foundation for subsequent differential expression and functional analyses.

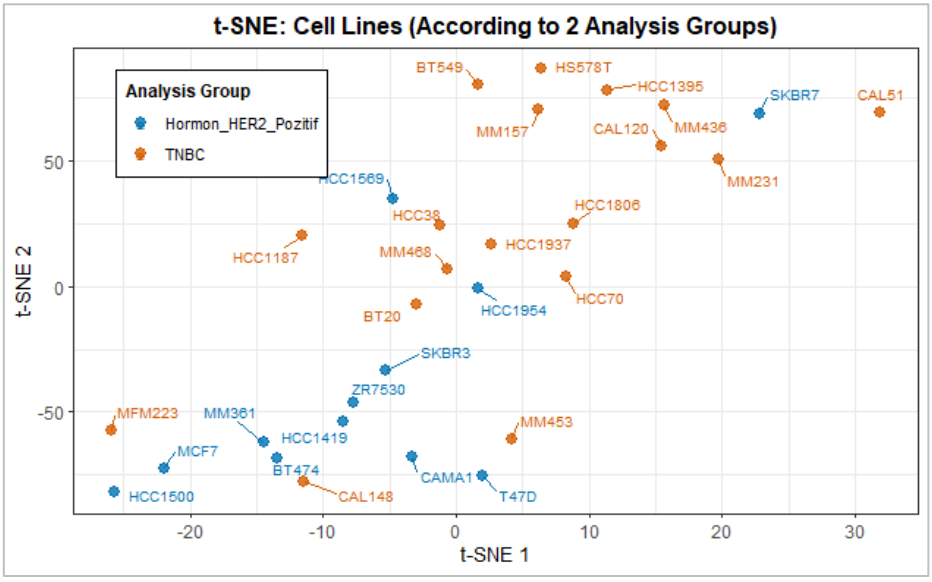

Figure 2: t-SNE: Gene Expression Profiles of Cell Lines

The distribution of analysis groups ('Hormone/HER2 Positive' and 'TNBC') in two-dimensional space using t-SNE dimensionality reduction technique is shown. Blue dots represent the Hormone/HER2 Positive group, orange dots represent the TNBC group. Molecular differences between groups are distinctly separated by t-SNE analysis.

The gene expression heatmap presented in Figure 3 provides a comprehensive visualization based on transcriptomic profiles of cell lines. In this heatmap, expression levels of the 100 genes with the highest variance are shown normalized with Z-score. Blue colored areas represent low expression, while red areas represent high expression levels. As a result of hierarchical clustering, cell lines are clearly divided into two main groups, and these groups largely overlap with the analysis groups defined in the study ('Hormone/HER2 Positive' and 'TNBC'). As shown in the colored bar at the top, TNBC cell lines (purple) and Hormone/HER2 Positive cell lines (green) tend to cluster within themselves.

In the heatmap, gene clusters showing different expression patterns between the two groups are clearly visible. While genes highly expressed in the Hormone/HER2 Positive group include TFF1, TFF3, GREB1, FOXA1, and ALDH3B2, genes such as AKR1B1, CAV1, CAV2, FABP5, EGFR, ROR1, and VIM were found to show high expression in the TNBC group. These gene clusters reflect the fundamental biological differences between the two analysis groups. Particularly, expression patterns specific to Hormone/HER2 Positive cell lines (red regions) are observed in the gene clusters seen in the upper parts of the map, while gene clusters in the lower parts of the map exhibit higher expression (red regions) in TNBC cell lines. These expression patterns reflect different biological properties and potential signaling pathways of breast cancer subtypes. This heatmap supports findings from t-SNE and differential gene expression analyses and provides a comprehensive visual representation of molecular differences between 'Hormone/HER2 Positive' and 'TNBC' analysis groups.

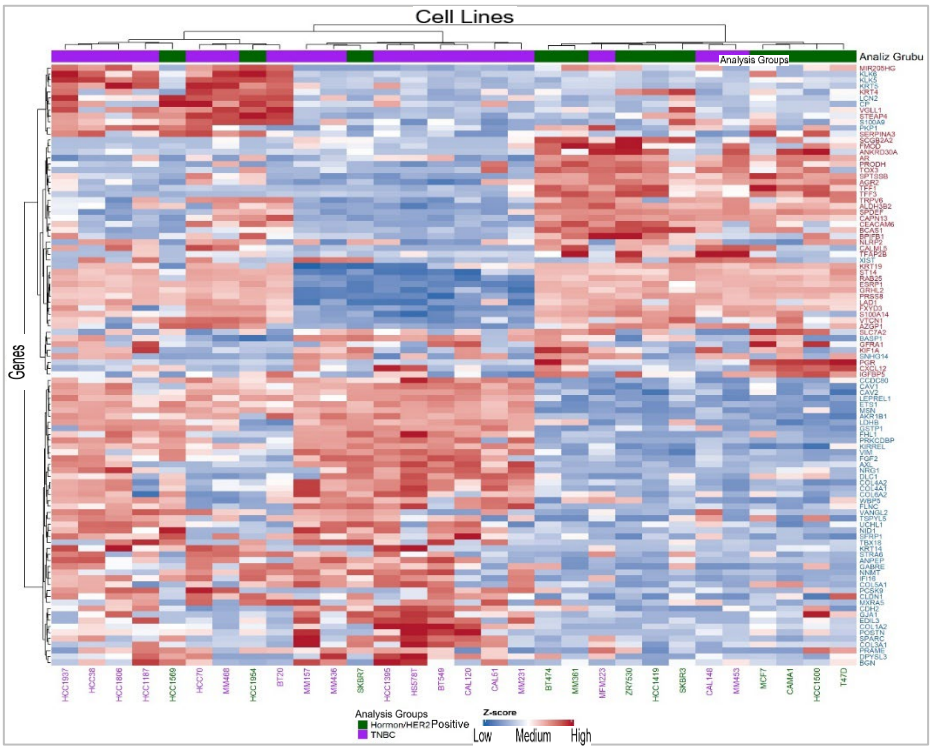

Figure 3: Comprehensive Heatmap Based on Gene Expression Profiles

10.5824/ajite.2026.01.002.x

Z-score normalized expression values of the 100 most variable genes are shown. Cell lines and genes are grouped by hierarchical clustering (Euclidean distance and ward.D2 method). The bar at the top shows analysis groups (Green: Hormone/HER2 Positive, Purple: TNBC). The heatmap visualizes differences in expression patterns between groups.

3.2. Differential Gene Expression and Functional Enrichment

Differential gene expression analysis performed between 'Hormone/HER2 Positive' and 'TNBC' analysis groups revealed that 316 genes were significantly differentially expressed (FDR < 0.05, |log2FC| > 1). Of these genes, 114 showed higher expression in the Hormone/HER2 Positive group, while 202 showed higher expression in the TNBC group. As seen in the volcano plot in Figure 4, among genes upregulated in the Hormone/HER2 Positive group, TFF3 and TFF1 stand out as genes with the highest significance. These genes encode proteins belonging to the trefoil factor family induced by estrogen in breast tissue. Additionally, GPR77, C6orf97, UGT2B17, and CST9 genes also show significantly high expression in this group. Among genes upregulated in the TNBC group, AKR1B1 and ROR1 genes are notable. While AKR1B1 plays a role in cellular stress response and tumor progression, ROR1 is a gene whose expression increases specifically in TNBC and is being investigated as a therapeutic target. Other prominent TNBC-specific genes include STAC, ETS1, BCAT1, FGF2, DSC3, LAMA4, DZIP1, and TLX3. Additionally, CXorf69, SLC22A3, TRBC2, MEOX1, RPL29P19, CH25H, and DPT genes also show prominently high expression in the TNBC group.

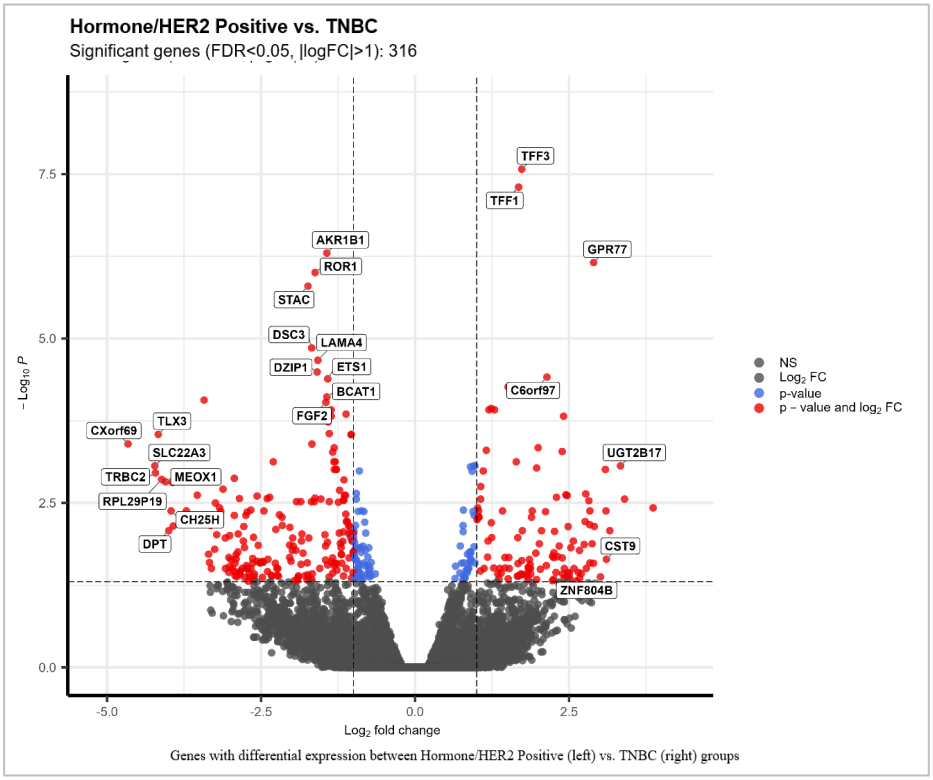

Figure 3: Volcano Plot of Differential Gene Expression Between 'Hormone/HER2 Positive' and 'TNBC' analysis groups. *Red dots show statistically significantly differentially expressed genes (FDR < 0.05 and |log2FC| > 1). The X-axis represents log2 fold change, the Y-axis represents -log10(P value).*

In the volcano plot, genes that are statistically significant (-log10P value) but fall below the log2FC threshold are also shown in blue, considering both statistical significance and magnitude of expression change. These genes, despite showing smaller expression changes, may have potential biological importance. Functional enrichment analyses were performed to understand the biological mechanisms underlying differential gene expression differences between 'Hormone/HER2 Positive' and 'TNBC' analysis groups. GO analyses showed that genes upregulated in the Hormone/HER2 Positive group are more concentrated in various metabolic processes (e.g., hormone, amino acid metabolism), while genes upregulated in the TNBC group are concentrated in processes related to immune response, extracellular matrix organization, and cell movement. Similarly, KEGG pathway analyses highlighted the Estrogen signaling pathway in the Hormone/HER2 Positive group, while revealing enrichment of immune/inflammatory pathways such as Cytokine-cytokine receptor interaction and JAK-STAT signaling pathway in the TNBC group.

10.5824/ajite.2026.01.002.x

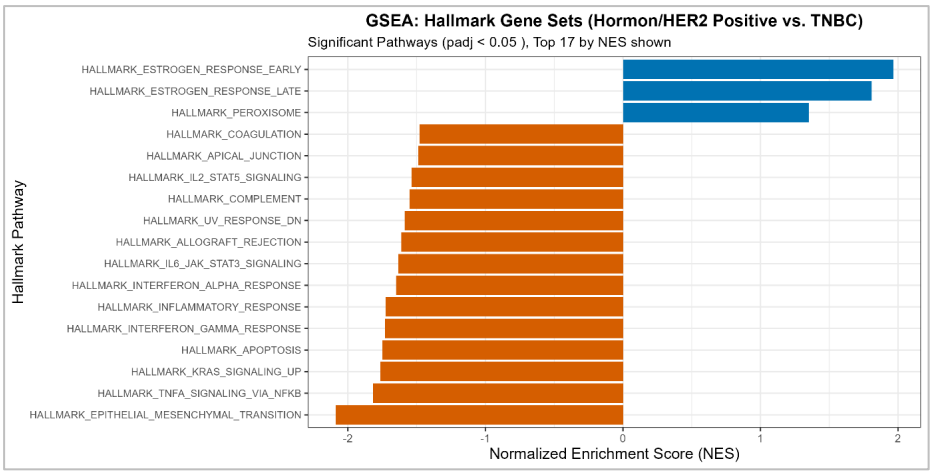

Figure 5: GSEA: Hallmark Gene Sets Analysis Results

Hallmark pathways significantly (padj < 0.05) enriched between 'Hormone/HER2 Positive' and 'TNBC' analysis groups are shown. Positive Normalized Enrichment Scores (NES) indicate enrichment in the Hormone/HER2 Positive group, while negative scores indicate enrichment in the TNBC group.

Gene Set Enrichment Analysis (GSEA) results that confirm these findings and provide a broader perspective are shown in Figure 5. GSEA performed using Hallmark gene sets showed that "Estrogen Response (Early and Late)" pathways were significantly enriched (positive NES) in the Hormone/HER2 Positive group as expected. In contrast, in the TNBC group (negative NES), especially "Inflammatory Response", "Interferon Alpha/Gamma Response", "IL6-JAK-STAT3 Signaling", "TNFa Signaling (via NFkB)" and other immune and inflammatory processes, as well as "Epithelial Mesenchymal Transition (EMT)" and "KRAS Signaling Up" pathways were determined to be significantly activated. These results support that there are distinct functional differences between the two analysis groups and that TNBC is characterized by an immunologically active and invasive profile. In particular, the prominence of TFF genes in the Hormone/HER2 Positive group emphasizes the importance of estrogen signaling in this subtype, while genes such as ETS1, ROR1, and AKR1B1 that stand out in the TNBC group indicate activation of signaling pathways related to cell proliferation, invasion, and metastasis. These molecular signatures provide important clues in explaining different biological behaviors of breast cancer subtypes and identifying potential therapeutic targets.

3.3. Machine-Learning Based Subtype Classification and Candidate Biomarkers

Multi-omics features were integrated to classify breast cancer subtypes and to identify candidate biomarker genes. We evaluated three models (decision tree, LASSO logistic regression, and random forest) and summarized their performance using nested cross-validation with five outer folds to reduce optimistic bias in this small-sample setting.**Model performance under nested cross-validation:** Across outer folds, LASSO logistic regression

and random forest yielded more consistent performance than the decision tree. The distributions of accuracy and F1-score indicate that LASSO and random forest were generally more stable across folds, whereas the decision tree showed higher variability—most notably in specificity—consistent with fold-level class imbalance and small test-fold sizes (Figure-6).

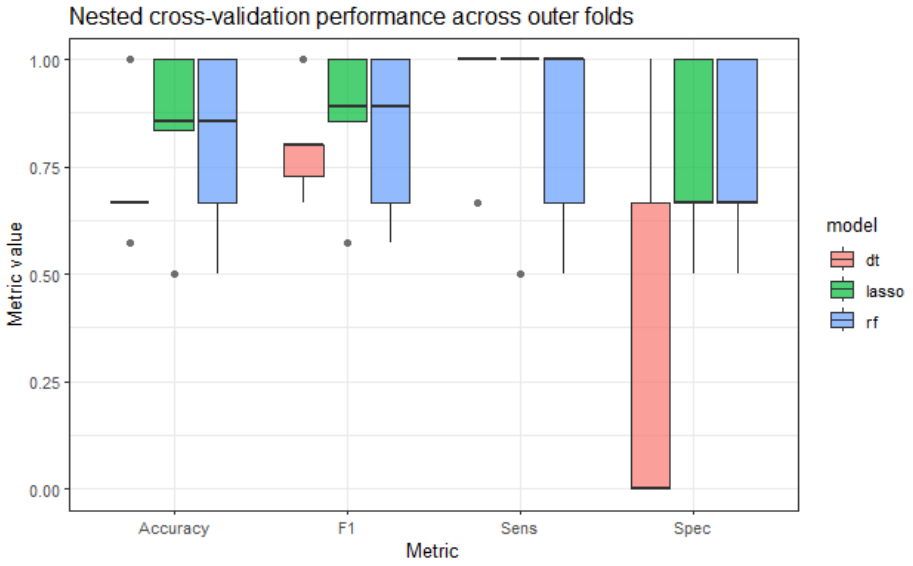

Figure 6: Nested Cross-Validation Performance Across Outer Folds

Boxplots show the distribution of Accuracy, F1-score, Sensitivity, and Specificity across five outer folds for decision tree (dt), LASSO logistic regression (lasso), and random forest (rf) models.

Candidate biomarkers across models: To interpret model-driven feature importance and identify robust candidates, we compared the sets of genes selected as important by each classifier. Two genes, **FAM176A** and **CACNG1**, were consistently identified across all three models, highlighting them as candidate biomarkers for subtype discrimination (Figure-7). Additional model-specific genes were also observed, suggesting complementary signals captured by different learning algorithms.

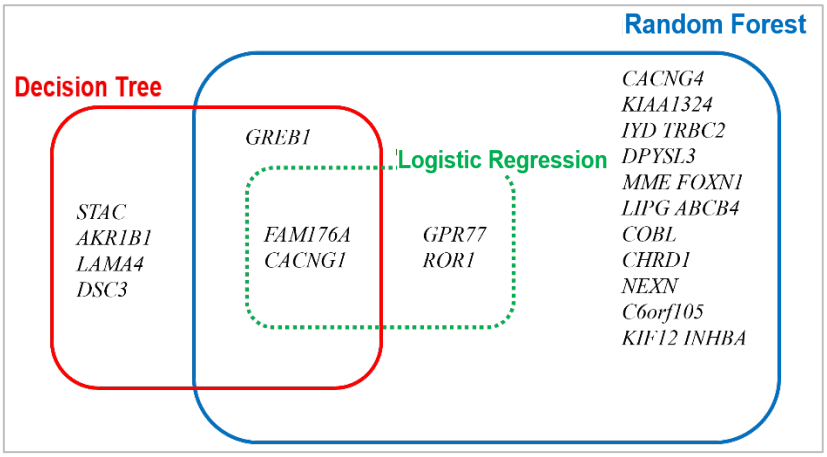

Figure 7: Common Biomarkers Identified by Models

10.5824/ajite.2026.01.002.x

Important genes identified by three different machine learning models (Decision Tree, Random Forest, and Logistic Regression) and their intersections. FAM176A and CACNG1 genes found common by all three models stand out as potential biomarker candidates.

Decision-tree interpretability and confusion matrix: The decision tree provides an interpretable set of threshold-based rules for subtype classification, primarily driven by **FAM176A** and **GREB1** expression cutoffs. The confusion matrix and the corresponding tree diagram illustrate how these thresholds separate TNBC from the Hormone/HER2-positive subtype in this dataset (Figure-8). Given the limited sample size, these thresholds should be interpreted as cell-line-level, hypothesis-generating rules requiring independent validation.

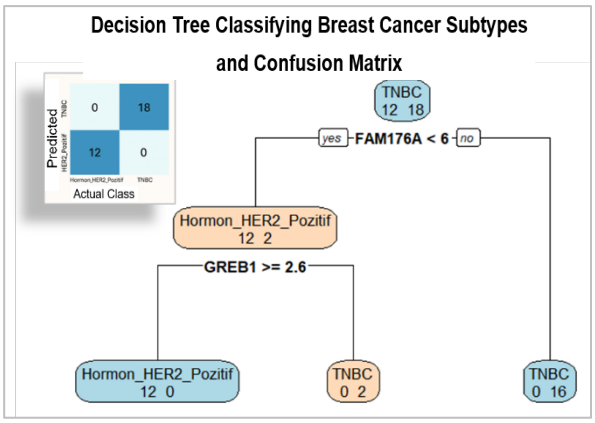

Figure 8: *Decision Tree Model Classifying Breast Cancer Subtypes and Confusion Matrix*
The confusion matrix on the left, the diagram on the right showing the separation between breast cancer subtypes (TNBC and Hormone/HER2 Positive) obtained by decision tree modeling. FAM176A and GREB1 genes were used as determinant variables in the model. Values in boxes show sample numbers belonging to related classes.

3.4 Drug sensitivity analysis (IC50-based)

Drug response was quantified using IC50 values for seven kinase inhibitors (AZD8055, BEZ235, Foretinib, GDC0941, Lapatinib, MK2206, PD0325901). Group-wise comparisons between TNBC and Hormone/HER2-positive cell lines were performed on log10-transformed IC50 values. Across drugs, the strongest nominal difference was observed for GDC0941 (median log10(IC50) TNBC = −5.69 vs Hormone/HER2+ = −6.09; median difference = 0.40; p = 0.012), which did not remain significant after multiple-testing correction (BH-FDR = 0.086). No other inhibitors showed significant differences after BH correction (all BH-FDR > 0.05). For PD0325901, comparisons were limited due to extensive ceiling IC50 values, resulting in a small number of evaluable observations. These IC50-based subgroup comparisons and statistical results are summarized in Table-1.

Table 1: Drug Sensitivity Differences Between TNBC And Hormone/HER2-Positive Cell Lines Based on Log10(IC50) Values

Metric	GDC0941	AZD8055	BEZ235	Foretinib	MK2206	Lapatinib	PD0325901
n_total	29	31	31	31	31	30	4
n_TNBC	12	13	13	13	13	13	1
n_HRHER2	17	18	18	18	18	17	3
Median log10(IC50) TNBC	-5.686	-6.653	-6.967	-5.884	-5.429	-4.913	-5.790
Median log10(IC50) Hormone/HER2+	-6.090	-7.201	-7.542	-5.672	-5.811	-4.943	-4.819
Median diff (TNBC - Hormone/HER2+)	0.404	0.547	0.575	-0.212	0.383	0.030	-0.971
p (Wilcoxon)	0.012	0.115	0.146	0.146	0.068	0.967	1.000
BH-FDR	0.086	0.204	0.204	0.204	0.204	1.000	1.000

P-values were calculated using the Wilcoxon rank-sum test and adjusted by BH-FDR. Median difference is TNBC − Hormone/HER2+; PD0325901 results are limited due to ceiling IC50 values.

3. Discussion

3.1. Molecular Differences Between Subtypes

In this study, when transcriptomic profiles of 30 breast cancer cell lines were examined, distinct molecular differences were found between 'Hormone/HER2 Positive' and 'TNBC' analysis groups. t-SNE and hierarchical clustering analyses confirm that these two groups show clear separation in terms of expression patterns (Figure 2, Figure 3). This finding emphasizes the importance of genome-wide molecular profiling in breast cancer classification (Perou et al., 2000; Sørlie et al., 2001). Functional analyses showed that pathways related to estrogen response and metabolism (GSEA analysis, Figure 5) are activated in the 'Hormone/HER2 Positive' group, while pathways related to immune response and inflammatory processes and invasion (Cytokine-cytokine receptor, JAK-STAT, EMT) are activated in the 'TNBC' group.

In differential expression analysis, estrogen-induced genes such as TFF1 and TFF3 in the Hormone/HER2 Positive group, and genes such as AKR1B1, ROR1, STAC, and ETS1 in the TNBC group showed high expression (Figure 4). The activation of "EMT" and "KRAS Signaling Up" pathways in the TNBC group in GSEA analysis supports the more aggressive and metastatic potential of this group (Nieto et al., 2016).

FAM176A and CACNG1 genes located at the intersection of three models were identified as potential biomarkers with machine learning models (Figure 7). In the Decision Tree model (Figure 8), FAM176A gene expression was found to be the first determining factor in subtype separation, while GREB1 was found important in classification within the

10.5824/ajite.2026.01.002.x

Hormone/HER2 Positive group. These findings contribute to gene panels that can be used in identifying breast cancer molecular subtypes and provide additional information to markers reported in the literature.

Beyond confirming established subtypes, our study distinguishes itself by employing a nested cross-validation framework to mitigate the overfitting often observed in small-sample studies. Furthermore, the identification of FAM176A as a robust classifier across three distinct machine learning architectures provides a novel candidate biomarker that has not been widely characterized in prior breast cancer subtyping literature. Although FAM176A (TMEM192) is less characterized in breast cancer, it is known to be involved in autophagy and lysosomal function, processes often dysregulated in aggressive tumor phenotypes (Hu et al, 2016). Similarly, CACNG1 encodes a CaVγ auxiliary subunit, and auxiliary ion-channel subunits (including CaVγ) have been linked to cancer-relevant processes such as proliferation, adhesion, migration, and invasion (Haworth & Brackenbury, 2019). The consistent selection of these genes by our models suggests they may represent overlooked components of the TNBC-specific molecular landscape.

3.2. Role of 4E-BP1 in mTOR Inhibitor Sensitivity

One of the important findings of our study is the confirmation, based on RPPA data from the original study, that 4E-BP1 protein expression level is a potential marker for sensitivity to mTOR inhibitors (AZD8055, BEZ235). It was observed that cell lines with high 4E-BP1 protein expression are more sensitive to these inhibitors, while lines with low expression are more resistant. This finding is completely consistent with the study by Jastrzebski et al. (2018) and shows that translational regulation plays a critical role in response to mTOR inhibition. While previous studies generally focused on the phosphorylation status of 4E-BP1, analyses on this dataset and the reference study emphasize that total 4E-BP1 protein amount is also critically important. This finding, independent of our RNA-seq data-based grouping, carries the potential of a practical biomarker that can be used in patient selection to increase the clinical effectiveness of mTOR inhibitors. It supports the hypothesis that tumors showing amplification or high protein expression of the EIF4EBP1 gene may benefit more from treatment with mTOR inhibitors targeting the active site.

3.3. Importance of Multi-omics Approaches

In this study, analysis of multi-layered data (transcriptomic, proteomic, pharmacogenomic) provided by Jastrzebski et al. (2018) (Figure 1) enabled us to understand the complex molecular properties of breast cancer subtypes more comprehensively. Combining information from different omic layers provides a richer and more holistic perspective than what a single data type could provide. Our two-group analysis ('Hormone/HER2 Positive' vs 'TNBC') shows how this integrative approach enables clearer

understanding of fundamental molecular differences between analysis groups and potential drug response markers (e.g., 4E-BP1). The identified gene signatures and functional pathway profiles have the potential to be used in developing personalized treatment strategies based on patients' molecular profiles and predicting treatment response in clinical settings.

Translational implications: Although this study is based on breast cancer cell lines, the identified signals (e.g., subtype-discriminative gene-expression markers) may support research-stage molecular stratification and prioritization of candidate biomarkers for downstream validation. A realistic clinical translation pathway would require confirmation in independent tumor cohorts and evaluation of incremental value over established clinical markers before any diagnostic or decision-support use.

To bridge the gap between in vitro findings and clinical utility, future validation studies should prioritize retrospective analyses of large-scale patient cohorts, such as TCGA (The Cancer Genome Atlas) or METABRIC. Specifically, the expression levels of FAM176A and CACNG1 should be correlated with patient survival outcomes and drug response data to determine their prognostic value in a heterogeneous tumor microenvironment.

3.4. Study Limitations

This study has several limitations. First, the analyses are based on **in vitro** cell line models; therefore, validation of the findings in **in vivo** settings and clinical cohorts is essential. Cell lines cannot fully capture the tumor microenvironment, intercellular interactions, and intratumoral heterogeneity observed in patients (Polyak et al., 2009), and these factors limit direct clinical translation. Second, ER+ and HER2+ cell lines were combined as a single "Hormone/HER2 Positive" group. While this grouping increased the sample size for a binary comparison against TNBC, it may have obscured biologically and clinically relevant differences between ER+ and HER2+ subtypes. Third, the sample size was limited (n=30 cell lines), which reduces statistical power and increases the risk of overfitting. Although model performance was evaluated using **nested cross-validation**, performance estimates may still be optimistic and should be interpreted cautiously. Finally, an independent patient cohort with matched multi-omics layers was not available for external validation; therefore, the results represent cell-line-level evidence and the identified markers should be considered **candidate biomarkers** requiring validation in independent clinical cohorts.

3.5. Future Studies and Perspectives

Validation of the prognostic and predictive value of identified potential biomarkers (4E-BP1 expression, FAM176A, CACNG1, GPR77 in Figure 6, and FAM176A, GREB1 genes prominent in the decision tree in Figure 8) in larger patient cohorts is essential. The potential of these biomarkers and our models to be integrated into clinical decision support systems for personalized treatment strategies should be investigated. The fold-validated performance of threshold-based rules is encouraging at the cell-line level; however, clinical translation

10.5824/ajite.2026.01.002.x

requires **external validation** in independent patient cohorts and evaluation of potential batch effects and clinical heterogeneity. Additionally, detailed examination of ER+ and HER2+ groups and development of more complex models that take into account the effect of the tumor microenvironment will be important research areas for overcoming resistance in breast cancer treatment.

CONCLUSION

In conclusion, this study revealed distinct transcriptomic and functional differences between 'Hormone/HER2 Positive' and 'TNBC' groups using multi-omics data from 30 breast cancer cell lines. While the Hormone/HER2 Positive group is characterized by estrogen response and metabolic processes, the TNBC group stands out with immune responses, EMT, and activation of various signaling pathways.

Our study, in addition to confirming the relationship between 4E-BP1 protein expression and mTOR inhibitor sensitivity defined by Jastrzebski et al. (2018), identified FAM176A and CACNG1 genes located at the intersection of three different machine learning models as potential new biomarkers. Our decision tree model provided interpretable threshold-based rules involving FAM176A and GREB1, with performance that varied across outer folds in nested cross-validation. FAM176A expression being below or above 6 distinctly separates samples, while GREB1 is important in the classification of the Hormone/HER2 Positive group. These transcriptomic signatures and biomarkers, when validated in clinical samples, can provide valuable tools in molecular classification and treatment decisions for breast cancer patients. While the interpretability of threshold-based models is advantageous, clinically applicable diagnostic algorithms would require **external validation** and assessment in independent patient cohorts before translation to practice. The relationship between 4E-BP1 expression and sensitivity to mTOR inhibitors is an important finding for increasing the effectiveness of targeted therapies. The active immune response and EMT pathways in the TNBC subtype show the potential value of immunotherapy and EMT inhibitors in this patient group.

Our integrated multi-omics approach provides a powerful method for understanding cancer biology and improving clinical decision-making processes. Future validation of these findings in larger patient cohorts will reveal their translational potential in breast cancer treatment.

REFERENCES

Anders, S., Huber, W., & Love, M. I. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology, 15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., & Stegle, O. (2018). Multi-omics factor analysis—A framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology, 14*(6), e8124.

Baião, A. R., Cai, Z., Poulos, R. C., Robinson, P. J., Reddel, R. R., Zhong, Q., Vinga, S., & Gonçalves, E. (2025). A technical review of multi-omics data integration methods: From classical statistical to deep generative approaches. *Briefings in Bioinformatics, 26*(4), bbaf355. https://doi.org/10.1093/bib/bbaf355

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M. F., Monahan, J. E., Morais, P., Meltzer, J., Korejwa, A., Jané-Valbuena, J., & Garraway, L. A. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature, 483*(7391), 603–607.

Bertucci, F., Finetti, P., Rougemont, J., Cervera, N., Charafe-Jauffret, E., Tarpin, C., Viens, P., Jacquemier, J., Birnbaum, D., & Bertone, P. (2005). Gene expression profiling identifies molecular subtypes of inflammatory breast cancer. *Cancer Research, 65*(6), 2170–2178. https://doi.org/10.1158/0008-5472.CAN-04-4115

Conway, M. E., McDaniel, J. M., Graham, J. D., Weigel, N. L., & Richer, J. K. (2020). STAT3 and GR cooperate to drive gene expression and growth of basal-like triple-negative breast cancer. *Cancer Research, 80*(21), 4663–4676. https://doi.org/10.1158/0008-5472.CAN-20-1379

Elkabets, M., Vora, S., Juric, D., Morse, N., De Pinho, R. A., Polyak, K., & Wagle, N. (2013). mTORC1 inhibition is required for sensitivity to PI3K p110α inhibitors in PIK3CA-mutant breast cancer. *Science Translational Medicine, 5*(196), 196ra99. https://doi.org/10.1126/scitranslmed.3005747

Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology, 18*, 83. https://doi.org/10.1186/s13059-017-1215-1

Haworth, A. S., & Brackenbury, W. J. (2019). Emerging roles for multifunctional ion channel auxiliary subunits in cancer. *Cell Calcium, 80*, 125–134. https://doi.org/10.1016/j.ceca.2019.04.005

Hu, J., Li, G., Qu, L., et al. (2016). TMEM166/EVA1A interacts with ATG16L1 and induces autophagosome formation and cell death. *Cell Death & Disease, 7*, e2323. https://doi.org/10.1038/cddis.2016.230

Huynh, M., Jayanthan, A., Pambid, M. R., & Lai, R. (2020). RSK2: A promising therapeutic target for the treatment of triple-negative breast cancer. *Expert Opinion on Therapeutic Targets, 24*(2), 157–167. https://doi.org/10.1080/14728222.2020.1709824

Iorio, M. V., Ferracin, M., Liu, C. G., Veronese, A., Spizzo, R., Sabbioni, S., Magri, E., Pedriali, M., Fabbri, M., Campiglio, M., Ménard, S., Palazzo, J. P., Rosenberg, A., & Croce, C. M. (2005). MicroRNA gene expression deregulation in human breast cancer. *Cancer Research, 65*(16), 7065–7070. https://doi.org/10.1158/0008-5472.CAN-05-1783

Jastrzebski, K., Thijssen, B., Kluin, R. J. C., de Lint, K., Wessels, L. F. A., & Linn, S. C. (2018). Integrative modeling identifies key determinants of inhibitor sensitivity in breast cancer cell lines. *Cancer Research, 78*(15), 4396–4410. https://doi.org/10.1158/0008-5472.CAN-17-2698

Jiang, Y. Z., Ma, D., Suo, C., et al. (2019). Genomic and transcriptomic landscape of triple-negative breast cancers: Subtypes and treatment strategies. *Cancer Cell, 35*(3), 428–440. https://doi.org/10.1016/j.ccell.2019.02.001

Koçak, M., Kırtay, S., & Akçalı, Z. (2025). Exploring the trends in multiomics research: A comprehensive bibliometric analysis with interactive visualization tools (BiblioMaps). *Journal of Advanced Research in Health Sciences, 8*(3), 236–247. https://doi.org/10.26650/JARHS2025-1759179

Lehmann, B. D., Bauer, J. A., Chen, X., Sanders, M. E., Chakravarthy, A. B., Shyr, Y., & Pietenpol, J. A. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *Journal of Clinical Investigation, 121*(7), 2750–2767. https://doi.org/10.1172/JCI45014

Mertins, P., Mani, D., Ruggles, K., et al. (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature, 534*, 55–62. https://doi.org/10.1038/nature18003

Nieto, M. A., Huang, R. Y., Jackson, R. A., & Thiery, J. P. (2016). EMT: 2016. *Cell, 166*(1), 21–45. https://doi.org/10.1016/j.cell.2016.06.028

Perou, C. M., Sørlie, T., Eisen, M. B., et al. (2000). Molecular portraits of human breast tumours. *Nature, 406*(6797), 747–752. https://doi.org/10.1038/35021093

Picard, M., Scott-Boyer, M. P., Bodein, A., Périn, O., & Droit, A. (2021). Integration strategies of multi-omics data for machine learning analysis. *Computational and Structural Biotechnology Journal, 19*, 3735–3746.

Polyak, K., Haviv, I., & Campbell, I. G. (2009). Co-evolution of tumor cells and their microenvironment. *Trends in Genetics, 25*(1), 30–38. https://doi.org/10.1016/j.tig.2008.10.012

Schmid, P., Adams, S., Rugo, H. S., et al. (2018). Atezolizumab and nab-paclitaxel in advanced triple-negative breast cancer. *New England Journal of Medicine, 379*(22), 2108–2121. https://doi.org/10.1056/NEJMoa1809615

Sørlie, T., Perou, C. M., Tibshirani, R., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences, 98*(19), 10869–10874. https://doi.org/10.1073/pnas.191367098

The Cancer Genome Atlas Network. (2012). Comprehensive molecular portraits of human breast tumours. *Nature, 490*, 61–70. https://doi.org/10.1038/nature11412

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., & Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods, 11*(3), 333–337.

Wei, Z., Han, D., Zhang, C., et al. (2022). Deep learning-based multi-omics integration robustly predicts relapse in prostate cancer. *Frontiers in Oncology, 12*, 893424. https://doi.org/10.3389/fonc.2022.893424

Wörheide, M. A., Krumsiek, J., Kastenmüller, G., & Arnold, M. (2020). Multi-omics integration in biomedical research—A metabolomics-centric review. *Analytical Chemistry, 92*(1), 386–402.

Yap, F. Y., Chong, P. F., Ahmad, K. A., Nordin, A., & Tan, Y. C. (2019). Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC Medical Informatics and Decision Making, 19*(1), 48. https://doi.org/10.1186/s12911-019-0801-4

Yu, G., Wang, L. G., Han, Y., & He, Q. Y. (2012). clusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology, 16*(5), 284–287. https://doi.org/10.1089/omi.2011.0118