



Màster Universitari

**Anàlisi de Dades Òmiques /  
Omics Data Analysis**

FACULTAT DE CIÈNCIES I TECNOLOGIA

**UVIC** | UVIC-UCC

# Master of Science in Omics Data Analysis

Master Thesis

## **Machine Learning-Based Gene Expression Signature for Classification of Endocrine Therapy Sensitivity in ER+ Breast Cancer Patients**

by

**Gilles Flamen**

Supervisor: Lara Nonell, Head of Bioinformatic Unit, VHIO

Academic tutor: Malu Calle Rosingana, Prof. Biostatistics and  
Bioinformatics, UVIC

Biosciences Department

University of Vic – Central University of Catalonia

10-09-2023

---

*Gene expression*

# Machine Learning-Based Gene Expression Signature for Classification of Endocrine Therapy Sensitivity in ER+ Breast Cancer Patients

Gilles Flamen<sup>1,2\*</sup>, Laia Monserrat<sup>3</sup>, Violeta Serra<sup>3</sup>, Lara Nonell<sup>1,2</sup><sup>1</sup>Bioinformatics Unit, VHIO, Barcelona, 08035, Spain<sup>2</sup>Universitat de Vic – Universitat Central de Catalunya (UVic), Vic, 08500, Spain<sup>3</sup>Experimental Therapeutics Group, VHIO, Barcelona, 08035, Spain

\*To whom correspondence should be addressed.

**Abstract:** Endocrine therapy (ET) combined with cyclin-dependent kinase 4/6 inhibitors (CDK4/6i) is the standard treatment for metastatic estrogen receptor-positive (ER+) breast cancer. However, not all patients require this combination therapy upfront, and identifying those who would respond well to ET alone could save healthcare costs, reserving the CDK4/6i combination for ET-resistant patients, and delay the eventual need for chemotherapy. In this study, we integrated two independent bulk RNA sequencing datasets, one inhouse generated and one publicly available, and used differential expression analysis (DEA) followed by LASSO selection to identify potential biomarkers associated with estrogen response. In this study, different machine learning techniques were employed to create predictive gene signatures, and comparisons were made based on performance. Through external validation with diverse datasets, we established a neural network-based 27-gene signature capable of classifying ET-sensitive patients with F-scores of up to 0.75. While validation presented challenges, our model offers promise for personalized clinical decision-making, provided more suitable validation data can be obtained.

**Code Availability:** [https://github.com/gillesflamen/Code\\_Gene\\_Signature.git](https://github.com/gillesflamen/Code_Gene_Signature.git)**Contact:** flamen.gilles@gmail.com**Supplementary information:** [Supplementary Materials.pdf](#)

---

## 1 Introduction

Breast cancer stands as the most prevalent tumor among women worldwide, and its incidence continues to rise (Duan, Y., et al., 2023). Approximately 70% of these tumors initially rely on the hormone estrogen for their growth and proliferation. The estrogen receptor (ER) plays a critical role in this process, triggering the transcription of pro-survival genes and cellular signaling, upon binding of estrogen. While estrogen-driven mitogenic signaling naturally stimulates mammary tissue development, its dysregulation can lead to hyperplasia and tumorigenesis, contributing to breast cancer formation (Hanker, A. B., et al., 2020).

Since these tumors heavily rely on estrogen signaling, the development of endocrine therapies (ETs) targeting this pathway has become standard practice, encompassing selective modulation of the ER (tamoxifen),

degradation of the receptor (fulvestrant), or inhibition of the aromatase enzyme crucial for converting androgen to estrogen (anastrozole or letrozole). All three approaches have been used for the last 20 to 50 years for the treatment of ER+ breast cancer patients in adjuvant therapy (Hanker, A. B., et al., 2020). Nonetheless, acquired resistance to ET is typical and often associated with somatic mutations in the *ESR1* gene encoding the estrogen receptor alpha (ER $\alpha$ ). Among the extensively studied mutations, the ligand binding domain (LBD) point mutations lead to ER $\alpha$  proteins with ligand-independent activity. Notable examples include mutations such as Y537S and D538G (Ma, CX. et al., 2015; Dustin, D. et al., 2019) that typically arise after patients undergo long-term endocrine treatment and may be present in up to 40% of patients with ER $\alpha$ -positive metastatic breast cancer (MBC; Fribbens, C. et al., 2016; Spoerke, JM. et al., 2016).

First-line and second-line treatment for ER+ metastatic patients typically include ET in combination with cyclin-dependent kinase 4/6 inhibitors (CDK4/6i; Turner, N. J. et al., 2018) and most patients will receive chemotherapy after progression of the disease following ET plus CDK4/6i treatment. However, not all patients need the addition of CDK4/6i in the first-line, as some are already responsive to ET alone (Sonke, G. S., 2023). Identifying these responsive patients presents a clinical challenge. Achieving this identification offers several advantages, including avoiding the toxicity associated with CDK4/6i treatment, saving costs to the healthcare system due to the high cost of these drugs, and delaying chemotherapy as long as possible. In the proposed treatment strategy, sensitive patients would receive endocrine therapy in monotherapy in the first-line, reserving the combination with CDK4/6i for those ET-resistant patients, and then chemotherapy as a third-line treatment only when needed. This treatment strategy is designed to be more specific to the individual and optimize time, resources, and patient quality of life (Al-qasem, A. J., et al., 2021; Sammons, S., et al., 2020).

Gene expression signatures have become a valuable tool in the breast cancer field for clinical use. In essence, it is a concise collection of genes that effectively forecast the broader transcriptome alteration in a cell or tissue triggered by external stimuli, functioning as a genetic fingerprint of the cell's biological state (Sithara, S. et al., 2017). For instance, the 21-gene Recurrence Score (Oncotype DX) has transformed clinical decision-making for ER+ breast cancer patients, enabling personalized treatment strategies (Qian, Y., et al., 2021). Moreover, it is worth mentioning the MOTERA-score, an acronym for Mutant or Translocated Estrogen Receptor Alpha. This 24-gene signature identifies cases with *ESR1* gene fusions and active *ESR1* LBD point mutations, that contribute to endocrine treatment failure in metastatic breast cancer (MBC; Gou, X. et al., 2021).

Constructing genetic signatures involves various methods. A straightforward approach is regression analysis, where specific genes (predictors) are selected, and coefficients are assigned based on their importance. The patient's gene expression profile is then entered into the model, and a cut-off is applied to predict the outcome (Theilhaber, J. et al., 2020; Cantini, L. et al., 2018). For example, in the MOTERA study, genes were ranked by percentile within each sample, and scores were computed as the mean percentile of the signature gene sets, followed by ROC analysis to determine a cutoff. In cases where complex relationships need to be considered, such as in our study, non-linear regression models or machine learning algorithms like Random Forest (RF), Neural Networks (NN), and Support Vector Machines (SVM) are more appropriate (Chicco, D et al., 2022; Yu, C et al., 2021; Xu, Q et al., 2016). In this study the focus will be on machine learning techniques. These complex algorithms are trained, in a supervised manner, on the outcome-labeled data in order to classify new patients data (Kalafi, E Y et al., 2019). The specific methods used during this signature generation are explained in detail in the Methods section.

Given the importance of selecting ET-sensitive patients, we attempted to identify a gene expression signature capable of classifying responders from non-responders. To achieve this, patient-derived xenograft (PDX) models were established by implanting fresh biopsies into NOD scid gamma (NSG) mice, which are immunodeficient laboratory mice. The response to ET was measured by monitoring tumor growth in mice upon ovaries removal to simulate hormone deprivation (Montserrat, L., 2022). Eventually, according to the RECIST 1.1 guidelines samples were classified as resistant or sensitive (Eisenhauer, E. A. 2009). Concurrently, RNA sequencing (RNAseq) was performed for transcriptomic profiling, generating a data pool that will be combined with external data from ER+ breast cancer patients in whom estrogen response was also measured (Gou, X. et al., 2021). Out of this data pool biomarkers were identified through

differential expression analysis (DEA), filtered them for relevance and used to construct predictive models using machine learning techniques.

## 2 Methods

### Data description

A total of four datasets were used. Two datasets were combined for biomarker discovery, training, and testing of the machine-learning models. The other two were used for the external validation of the final model.

The first dataset was generated internally and consisted of RNAseq RSEM quantified counts obtained through the nf-core/mseq pipeline (v3.4; Harshil, P. et al., 2021). The samples were from ER+ PDX models made with fresh biopsies from breast cancer patients (Montserrat, L., 2022). In total, they comprised 6 sensitive and 29 resistant samples. For simplicity, we refer to this as the Internal data.

The second dataset was from the study by Gou, X. et al. (Gou, X. et al., 2021). The raw data was available on request, but preprocessed data can be found on GEO (GSE191158). It consisted of RNAseq data from ER+ PDX models, in the form of RSEM expected counts. In total, the set consisted of 6 sensitive and 15 resistant samples. This set will be referred to later as the Gou data.

The third dataset consisted of single-nucleus RNAseq (snRNAseq) data from 41 ER+ breast cancer patients receiving neoadjuvant ET (letrozole; Griffiths, J. I., 2021). Nuclei were isolated and snRNAseq was performed using 10X Genomics technology. After filtering for relevant cells (nucleated cells) from patients receiving ET alone, around 21k tumoral cells remained from 6 sensitive and 5 resistant samples. The data is available on GEO (GSE158724).

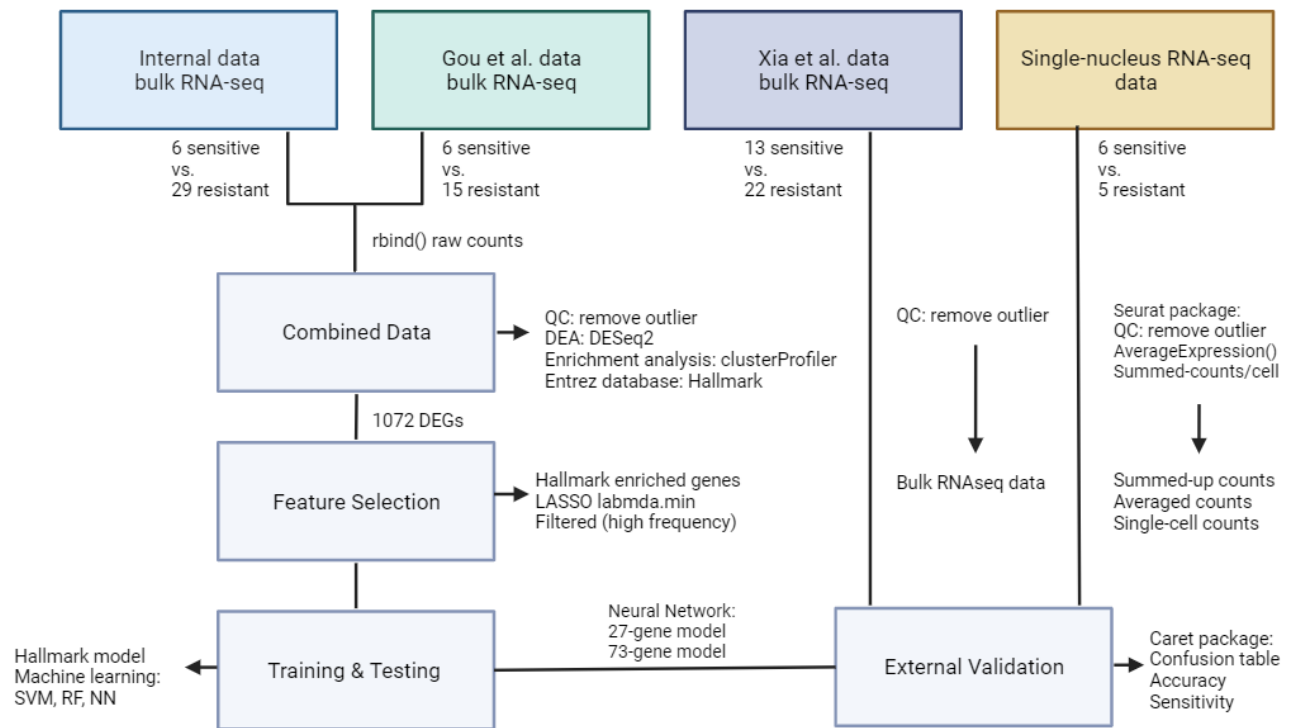
The fourth and final dataset originated from a study conducted by Youli, X. et al. (Youli, X et al., 2022), encompassing RNAseq data obtained from ER+ breast cancer patients prior to undergoing neoadjuvant therapy. Among these patients, three displayed intrinsic resistance, 19 developed acquired resistance, and 13 retained sensitivity post-treatment. Patients 18 and 21 were excluded from analysis since no expression data was present before receiving any therapy. This dataset is accessible via the EMBL-EBI ArrayExpress database under accession number E-MTAB-9917. For the sake of convenience, we will refer to this dataset as the Xia data.

### Data preprocessing

Prior to combining the two bulk RNAseq datasets into the training dataset, we performed a series of preprocessing steps to ensure data quality. These steps included checking gene expression distributions across samples, hierarchical clustering of samples using the ward.D2 method, and conducting principal component analysis (PCA).

For the Internal dataset, we used DESeq2's (v1.38.2; Love, MI. et al., 2014) regularized logarithm method (rlog) for normalization and filtered genes to include those with at least 15 reads in at least 6 samples, matching the group size of the smallest response category (i.e., the sensitive samples). In the case of the Gou dataset, the data was already  $\log(x+1)$ -transformed, so we reversed it with an anti-transformation, followed by TMM normalization.

After these initial preprocessing steps, we combined the raw counts of the Internal and Gou bulk RNAseq datasets using the `rbind()` function in



**Figure 1:** Overview of the datasets, methodology, and bioinformatic steps followed to reach the final classifying models: First, the fusion of raw counts from the Internal and Gou et al. bulk RNAseq data using the `rbind()` function was done, resulting in the creation of the "Combined data." QC of the combined data allowed us to identify and remove an outlier. DEA was then performed using DESeq2, resulting in the identification of 1072 DEGs. To gain biological insights, the set of DEGs underwent Entrez hallmark enrichment analysis through the use of the clusterProfiler package. From this initial pool of 1072 DEGs, three different sets of features were selected: genes enriched in hallmark pathways, LASSO-selected genes with lambda.min, and LASSO-selected genes filtered for high frequency. Machine learning models were constructed using these feature sets and were trained and tested on the combined data. Only the best-performing models were selected for external validation. Additionally, single-nucleus RNAseq data underwent preprocessing using the Seurat package, and pseudobulks were generated using the `AverageExpression()` function and by summing raw counts across cells. For validation purposes, Xia et al. bulk RNAseq data was also preprocessed, with outliers removed. The validation process was carried out using the Caret package.

R. This combined data, utilized for training and testing the models, underwent TMM normalization. Subsequently, we conducted another round of quality control (QC) checks, which included hierarchical clustering based on ward.D2 method and PCA, to identify and exclude any potential outlying samples.

The Xia et al. bulk RNAseq data, used for external validation of the models, were TMM-normalized and subjected to outlier detection through analyses of library sizes, hierarchical clustering using ward.D2 method, and PCA. Following the removal of outlying samples and feature selection, gene expression data were scaled before being input into the models.

The snRNAseq data was preprocessed and analyzed with the Seurat package (v4.3; Hao, Y. et al., 2021). Cells with less than 1500 and more than 7500 features and more than 5% mitochondrial DNA were filtered. The data were normalized using the default `NormalizeData()` function, which normalizes feature expression by the total expression in each cell and multiplies it by a factor of 10000, upon log transformation. The data was then scaled by a linear transformation that shifts the expression of each gene so that the mean expression and variability are 0 and 1, respectively, across cells.

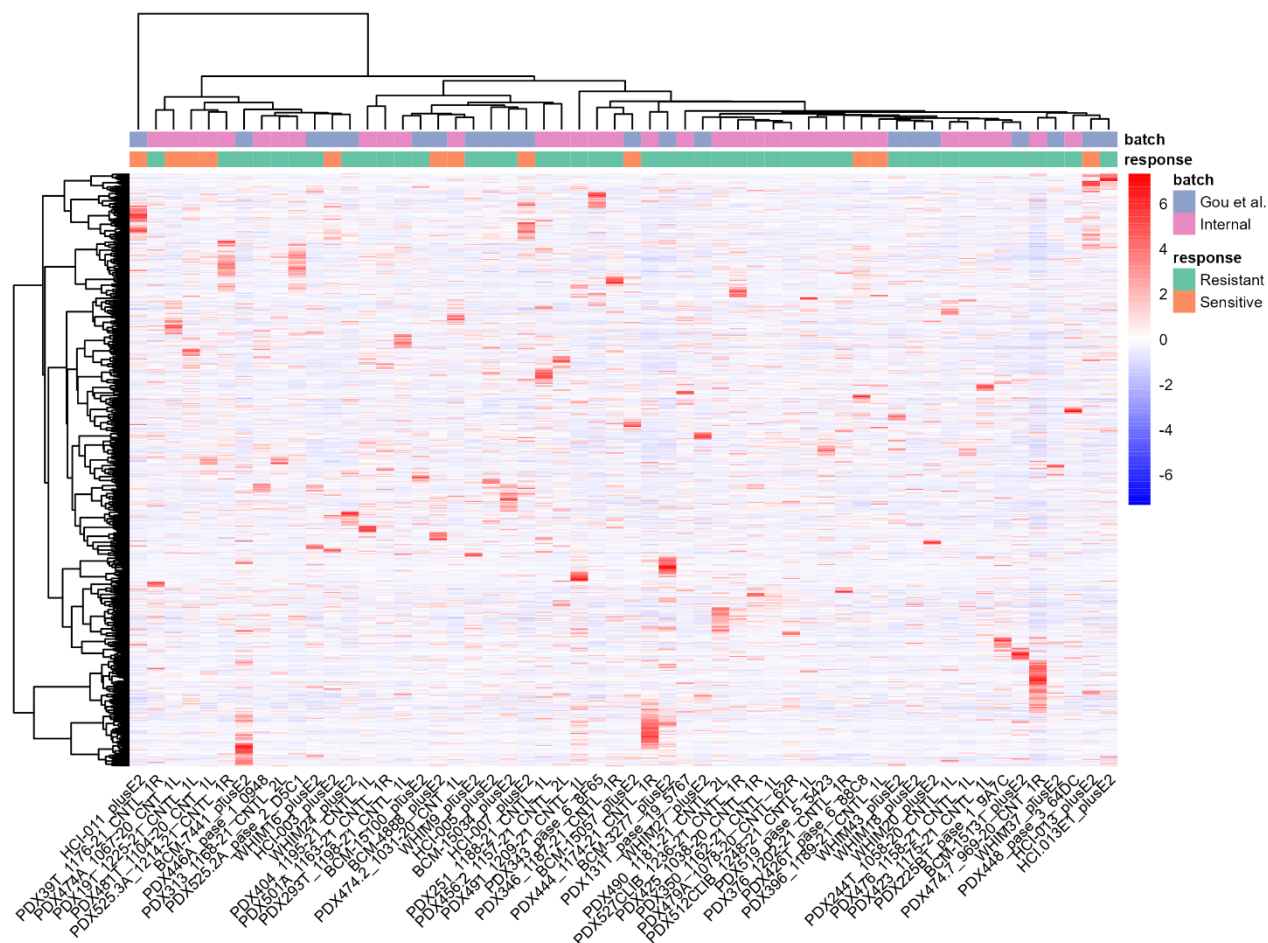
As the snRNAseq data will be used for validating the classification model trained on bulk RNA data, the creation of a compatible bulk data structure is needed. To address this, two distinct approaches were

undertaken. First, the standard `AverageExpression()` function was used to return a pseudobulk expression set containing the average expression of each patient across all cells. Like wise, a second pseudobulk validation set was made, via summing up the raw counts of a gene across the cells for each patient.

#### Differential expression analysis & feature selection

Both Limma (v3.54.0; Ritchie, M.E. et al., 2015) and DESeq2 (v1.38.2; Love, M.I. et al., 2014) were used for the DEA of the Internal data and a comparison was made based on the resulting differentially expressed genes (DEGs) between sensitive and resistant samples. Several models were tested consisting of unique combinations of clinical variables, but ultimately only patient identity (ID) was included in the model, since several PDXs originated from the same patient. It was then decided to proceed with DESeq2 for the DEA of the combined dataset, as a first step of the feature selection procedure. DESeq2 DEA was performed on the raw counts and the results were filtered for the FDR-adjusted p-value of 0.05 and log2 fold change (LFC) of 1.

Functional analysis was done using the `enricher()` function from the clusterProfiler package (v 4.8.3; Yu, G. et al. 2012). Gene set collections



**Figure 2:** Heatmap of 1072 DEGs in Combined Data: This heatmap displays the expression profiles of 1072 DEGs across samples, which include both the Internal and Gou et al. bulk RNAseq data. Rows represent genes, and columns represent samples. Hierarchical clustering (ward.D2 method) was applied to genes, and gene expression was scaled. The color codes indicate batch and response groups.

from hallmarks were obtained from the msigdb package (v1.8.0; Subramanian, A. et al. 2005).

The genes composing the signature were curated through distinct strategies. One approach mirrors the methodology outlined in the MOTERA study, involving the selection of DEGs linked to enriched hallmarks like ‘Estrogen response early’, ‘Estrogen response late’, and ‘Epithelial-mesenchymal transitions’, aligning with relevant biological contexts.

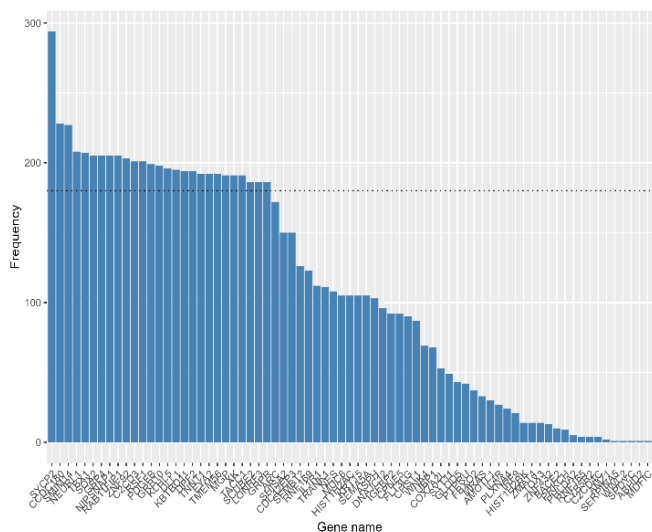
In parallel, an alternative model employed a penalized regression technique known as LASSO (Least Absolute Shrinkage and Selection Operator). LASSO automatically identifies significant variables by driving the coefficients of less impactful predictors toward zero. To address stochastic variations, the LASSO procedure underwent 1000 iterations, and the distribution of gene selection frequencies was visualized. As no outcomes emerged with  $\lambda_{min} + 1$  standard error (s.e.), we opted for  $\lambda_{min}$  as the penalty term. Next, a subset of features was generated, encompassing only those genes selected frequently in the LASSO iterations. This threshold, frequency > 180, was determined by examining the barplot depicting gene selection frequencies, selecting a point where a significant drop in frequencies was observed (Figure 3).

### Classification models

Different machine learning methods were used to construct the gene expression signature with the goal of correctly classifying patients as sensitive or resistant. Namely, SVMs, NNs, and RFs were employed.

The SVM is a powerful tool for classification tasks. It determines optimal class boundaries by transforming data into higher-dimensional space, aided by kernels. The linear kernel, simplifies this process by emphasizing relationships that resemble the product of observations. In addition, the SVM includes a parameter C that balances classification accuracy with model simplicity. By adjusting C, the SVM fine-tunes the classification method to avoid overfitting while still capturing the complex relationships within data. (Huang, S. et al., 2018; Hsu, C. et al., 2003). In R, the svm() function from the e1071 (v1.7-13; Meyer, D. et al., 2021) package was used to train the model on the test data.

NNs are a type of supervised learning algorithm that aims to find an optimal decision boundary between classes by employing nonlinear connections represented by interconnected nodes known as neurons. Each layer in the network, encompassing input, hidden, and output layers, processes data through weighted connections and activation functions (Trans, K. et al., 2021). For our analysis in R, we utilized the nnet() function from



**Figure 3:** Barplot of the LASSO-selected gene frequencies: Based on 1000 LASSO iterations (lambda.min) applied to the 1072 DEGs, this plot displays the frequencies of selected genes. The frequency represents the number of times a gene was not shrunken to zero after 1000 iterations. The dotted line ( $Y = 180$ ) indicates the threshold for the 'Filtered-Relaxed' feature set.

the `nnet()` package (v7.3-18; Venables, WN. et al., 2002) to fit single-hidden-layer NN models to our dataset.

RF is another popular supervised learning algorithm that builds multiple decision trees during training and combines their predictions to make more accurate and robust decisions. Each tree is built on a random subset of the training data and features, which reduces overfitting and increases generalization. By aggregating the results of individual trees, RF achieves high accuracy and can effectively handle high-dimensional data (Breiman, L. et al. 2001). In R, the `randomForest()` function from the `randomForest` package (v.7-1.1; Liaw, A. et al., 2002) was used for classification, that implemented Breiman's RF algorithm.

In essence, our methodology involved fourfold division using the `createFolds()` function sourced from the `caret` package (v6.0-94; Kuhn, M.; 2008). Subsequently, various models underwent training and assessment via `caret`'s `confusionMatrix()` function. F-scores (F1-scores) and accuracies were calculated using the following formulas:

$$\text{Accuracy} = (\text{True Positives} + \text{True Negatives}) / \text{Total Predictions}$$

	Reference	
Predicted	Sensitive	Resistant
Sensitive	A	B
Resistant	C	D

$$\text{Recall} = A / (A + C)$$

$$\text{Precision} = A / (A + B)$$

$$\text{F1-score} = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall})$$

Following the identification of the optimal model, a range of validation datasets were employed. These datasets included bulk RNAseq expression data from the Xia dataset, as well as snRNAseq data at the single-cell level and in the form of averaged and summed-up pseudobulks.

### 3 Results

In Figure 1, the process leading to the final model is outlined. It began with the merging of two raw-counts datasets, TMM normalization, and quality control to assess inter-batch sample relationships. During this phase, an outlier sample was removed. To facilitate DEA, DESeq2 was applied to the count data, resulting in the identification of DEGs between sensitive and resistant samples. Functional enrichment analysis of hallmarks provided biologically significant insights and the initial sets of features. Additionally, LASSO was employed to identify biomarkers for the models. After training and testing multiple models, the selection was based on the best performance achieved. External validation data were preprocessed and underwent quality control to eliminate outliers, to eventually evaluate the performances of the final models.

#### The analysis of combined bulk RNAseq data revealed 1072 DEGs distinguishing between resistant and sensitive samples.

Regarding the two separate bulk RNAseq datasets, namely Internal and Gou, an in-depth analysis was conducted. For the Internal dataset, DESeq2 analysis was employed with various models incorporating predictors derived from clinical data. However, issues arose due to collinearity between predictors like age, gender, and PAM50 subclass with patient identity. Consequently, only patient identity was integrated into the design formula, given its importance in accounting for inter-patient variability. In parallel, a Limma analysis was carried out, albeit without yielding genes surpassing the significance threshold (FDR of 0.05). Nevertheless, the top genes identified were consistent with those obtained through DESeq2, visualized in the volcano plots in Supplementary Figure 1. Ultimately, the DESeq2 approach was favored and applied to the Combined dataset, consequently unveiling contextually relevant DEGs.

Following the merging of the Internal and Gou bulk RNAseq datasets, a comprehensive analysis involving clustering and principal component analysis (PCA) of the TMM-normalized merged data unveiled an outlier among the resistant samples, namely 'PDX346\_1187-21\_CNTL\_1R' from the Internal dataset, observed after plotting the first two PCs (Supplementary Figure 3). Subsequently, DESeq2 analysis was executed on the Combined data, exclusively considering the response variable (resistant or sensitive sample) and patient identity. By applying a filter of  $|\log\text{FC}| > 1$  and adjusted p-value  $< 0.05$ , a total of 1072 DEGs were identified. The expression patterns of these 1072 genes are visualized in the heatmap (Figure 2), revealing no distinct separation based on response or batch categories. Notably, based on the expression of the DEGs, sensitive sample 'HCI-011\_plusE2' from the Gou data, was clustered separately from the others, but not excluded of further analysis. Remarkably, hallmark enrichment analysis performed on this DEG list yielded terms highly pertinent to context, including 'Estrogen Response Late', 'Estrogen Response Early', and 'Epithelial Mesenchymal Transition' (Supplementary Figure 2). These 1072 DEGs served as the foundation for subsequent feature selection.

#### Neural network models with LASSO-selected features demonstrated the highest levels of accuracy and best F1-scores.

In line with the methodology employed in the MOTERA paper, we initially adopted a similar approach, focusing on genes associated with hallmark terms related to estrogen response (early & late) and epithelial-mesenchymal transition (Supplementary Figure 2). Out of the initially identified 112 genes, reduced to 88 due to redundancy, only four genes (*RASGRP1*, *TFF1*, *VCAN*, and *TGM2*) were mutual with the 24-gene signature of the MOTERA paper, depicted in the Venn diagram in

## Gene Signature for ER+ Breast Cancer Classification

**Table 1:** Model Accuracies and F1-Scores for SVM, NN, and RF: This table presents the accuracies and F1-scores of SVM, NN, and RF models for three predictor sets: Hallmark genes, Relaxed, and Filtered-Relaxed LASSO-selected genes. These models were trained and tested using 4-fold cross-validation on a combination of internal and Gou et al. bulk RNAseq data.

		SVM	NN	RF
<b>Accuracy (%)</b> <b>F1-scores (0 – 1)</b>	Hallmark genes (88-genes)	78.16 0.000	72.80 0.545	62.52 0.000
	Relaxed (73-genes)	79.95 0.286	96.29 1.000	65.38 0.154
	Filtered-Relaxed (27-genes)	96.29 0.909	94.51 0.889	66.81 0.600

Supplementary Figure 6. The raw counts of these genes were extracted, TMM-normalized and scaled expression data from the merged dataset were chosen for training machine learning models, including NN, RF, and SVM. Initially, each NN was configured with a single neuron and a maximum of 2000 iterations. A decay rate of  $5e-4$  and a range of 0.1 were applied. The parameters for SVM models were left undefined, allowing the function to optimize them automatically. As for the RF model, it was initialized with 20 trees, as this configuration yielded the best performances after experimenting with different values. Additionally, the square root of the number of columns (features) was selected as the number of randomly chosen features during each split. Subsequent parameter optimization occurred after the initial selection process, but it did not lead to significant differences in performance. Consequently, the decision was made to continue using these parameters.

Although an epsilon-regulated radial SVM achieved the highest accuracy of 78.16%, its F-score was zero, rendering it unsuitable for our primary goal of accurately classifying sensitive patients. Therefore, the NN emerged as the second most accurate model, displaying an accuracy of 72.80% and a F-score of 0.545. The RF model achieved a similar accuracy of 62.52% but also had a F-score of zero. The suboptimal performance of these models raised concerns about the efficacy of the feature selection approach, and were no longer validated.

To address this issue, we turned to LASSO for feature selection. LASSO employed internal CV to optimize parameter values. The selection of the lambda value, which represents the stringency of LASSO, was crucial. Two LASSO iterations were conducted: one with the minimum lambda value (lambda.min) post CV, resulting in a more lenient predictor selection, and another with lambda.min plus one s.e., leading to a more rigorous LASSO.

The first iterative LASSO-selection with lambda.min yielded 73 genes (referred as the Relaxed feature set), each with coefficients unshrunk with a specific frequency over 1000 LASSO iterations, as seen in the barplot in Figure 3. In contrast, during the iterations when using lambda.min + 1 s.e., all genes were shrunk to zero. Nevertheless, a second gene set was generated by applying a threshold after observing a drop in gene frequency in the barplot. This set, referred to as the Filtered-Relaxed feature set, comprises a total of 27 genes with a frequency > 180 (dotted line in Figure 3).

Training models with the identified 73 features highlighted NNs as the superior choice. Its accuracy of 96.29% surpassed that of both RF (65.38%) and SVM (79.95%) models. Additionally, the NN flawlessly

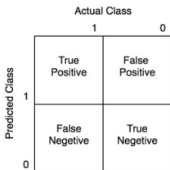
classified all sensitive samples, in contrast to the RF and SVM models with F-scores of 0.154 and 0.286, respectively. When trained on the 27 selected genes, the SVM model achieved the highest accuracy, reaching 96.29%, outperforming both NNs (94.51%) and RF (66.81%). In terms of F-scores, NNs achieved 0.889, while SVM excelled with a score of 0.909. Nevertheless, the NN model classified all 12 sensitive patients correctly (recall of 1), compared to SVM that classified 2 sensitive patients as resistant (recall of 0.83). All the results of the training and testing phase of the three different machine learning models (SVM, NN, and RF) on the three different features sets (hallmark genes, Relaxed genes, and Filtered-Relaxed genes), are summarized in Table 1.

### The 27-gene neural network model showed to the best performances across validation with diverse external data sets.

Subsequently, we conducted a comprehensive validation process for the NN models using both the Relaxed and Filtered-Relaxed LASSO-selected gene models. While the SVM model demonstrated slightly higher accuracy and F-score when trained on the Filtered-Relaxed 27-gene set, we chose to focus on the NN models for comparison of the number genes included in the models. Besides, missing values occurred in the validation data and would have to be imputed for the validation of the SVM model, which preferably is avoided. Additionally, the NN classified all sensitive patients correctly. The validation encompassed four distinct datasets: the Xia et al. external bulk RNAseq set, the snRNAseq data fed to the model cell-by-cell, and two variations of pseudobulk snRNAseq data (averaged and summed counts; See column names Table 2). Examination of the Xia et al. external bulk RNAseq dataset unveiled an outlying resistant sample named '015-1', which exhibited distinct clustering during ward.D2-based hierarchical clustering and PCA (Supplementary Figure 4). Next, to ensure data quality, the Seurat package was employed for in-depth analysis of the snRNAseq data, where cells with inadequate read counts, features, or elevated mitochondrial RNA content were filtered out. Moreover, due to a low number of cells, patients 15 and 16 were excluded from further analysis, as seen in the violin plots in Supplementary Figure 5. As previously mentioned, our analysis exclusively considered the tumoral fraction of cells, obviating the necessity for filtering based on cell type. Despite



**Table 2:** Confusion Matrices and Performance Metrics of NN-Based Models: This table displays the confusion matrices, accuracies, and F1-scores for two sets of predictors, namely, the Relaxed-LASSO selected and Filtered-Relaxed LASSO selected sets. These models were externally validated using four distinct datasets: Xia bulk RNAseq, snRNAseq analyzed at the single-cell level, and two pseudobulk datasets derived from the snRNAseq data, specifically summed and averaged pseudobulks.

		Xia bulk RNAseq	Single-cell wise snRNAseq	Summed pseudobulks nRNAseq	Averaged pseudobulk snRNAseq
<b>Confusion matrix</b> 	Relaxed (73-genes)	4   8 9   11	1784   2574 3942   11113	2   0 3   4	0   2 5   2
	Filtered-Relaxed (27-genes)	3   5 10   14	1616   2099 4110   11588	3   0 2   4	2   0 3   4
<b>Accuracy (%)</b> <b>F1-scores (0 – 1)</b>	Relaxed (73-genes)	46.88 0.308	66.43 0.356	60.00 0.571	20.00 0.000
	Filtered-Relaxed (27-genes)	53.12 0.308	68.02 0.345	63.64 0.750	50.00 0.571

this, upon performing dimensionality reduction, it is evident from the UMAP plot that cells naturally cluster according to the respective patients (Supplementary Figure 5).

In our validation using the Xia dataset, we evaluated two models: the Relaxed 73-gene model and the Filtered-Relaxed 27-gene model. The 73-gene model achieved an accuracy rate of 46.88%, while the 27-gene model achieved an accuracy rate of 53.12%. Both models exhibited an equal F-score of 0.308 (Table 2). These suboptimal results prompted us to conduct hierarchical clustering (ward.D2 method) of the Combined data (Internal and Gou data, utilized for training) alongside the Xia validation data, aiming to explore whether they exhibit separation based on their study origins. Notably, separation is indeed observed (Supplementary Figure 8).

Subsequently, we performed validation using snRNAseq data cell-by-cell. In this context, 73-gene model achieved an accuracy of 66.43%, and the 27-gene model achieved an accuracy of 68.02%. The corresponding F-scores were 0.356 and 0.345, respectively.

When validating with summed-up pseudo bulk data, the 73-gene model achieved an accuracy of 60.00%, while the 27-gene model achieved an accuracy of 63.64%. The F-scores for these models were 0.571 and 0.750, respectively (Table 2).

Finally, in the case of validation involving average pseudobulk data, the 73-gene model achieved an accuracy of 20.00%, whereas the 27-gene model achieved an accuracy of 50.00%. The corresponding F-scores were 0.000 and 0.571, respectively. A summary of all validation outcomes is provided in Table 2.

## 4 Discussion

In this study, we aimed to develop and compare various models for the classification of ER+ breast cancer patients based on their sensitivity to ET. To achieve this, we merged internally generated data with publicly available bulk RNAseq data from patients classified as either ET-sensitive

or -resistant. Initial DEA identified 1072 genes exhibiting significant expression differences between these groups, with biological relevance confirmed through hallmark enrichment analysis. Our comprehensive analysis encompassed a variety of models, each employing distinct features and machine learning algorithms. Ultimately, we arrived at two NN-based models: one derived from a 73-gene expression signature and another featuring a refined set of 27 genes, selected through the LASSO method for their predictive potential. External validation using four distinct methods, including the Xia et al. external bulk RNAseq dataset, snRNAseq data analyzed at the single-cell level, and two variations of pseudobulk snRNAseq data (averaged and summed counts), led to the construction of confusion matrices. Despite our efforts, these validation outcomes were less favorable than expected.

Both Limma and DESeq2 are valid tools for identifying DEGs between our sensitive and resistant samples. When applied to the Internal data, Limma didn't yield any DEGs meeting the stringent FDR threshold of 0.05. In contrast, DESeq2 identified several relevant DEGs. Notably, the top genes identified by Limma (before significance filtering) resembled those found by DESeq2. This aligns with a study by Tong Y. (2021) suggesting that DESeq2 tends to find more DEGs than Limma, but there's significant overlap, and Limma is generally considered more reliable, as it tends to produce less false positive results (Tong, Y., 2021). Nevertheless, our samples didn't cluster distinctly based on ET response in PCA plots. This suggests subtle discrepancies that conservative methods like Limma may not capture. Hence, we leaned towards using DESeq2 results as the foundation for subsequent feature selection, recognizing that, in our context, less stringent methods like DESeq2 may better reveal the nuances of gene expression patterns linked to ET response (Schurch, N. et al., 2016; Gauthier, M. et al., 2020).

Our choice to explore a feature selection approach akin to the one employed in the MOTERA paper stemmed from the shared purpose of both signatures. The MOTERA paper sought to differentiate patients with



activating *ESR1* gene fusions, linked to ET failure. This objective bears resemblance to our aim of identifying patients who are unlikely to respond to ET before initiating any treatment. As a result, it was reasonable to expect a gene overlap between these two sets, and our analysis confirmed the presence of ‘only’ four shared genes. However, the incorporation of these genes, and the other 84, into our signature led to suboptimal model performances. This outcome prompted us to adopt an alternative approach: LASSO shrinkage selection, aimed at identifying the most predictive genes aligned with our research objective.

While several gene signatures chose their candidate genes from a more biological perspective by selecting genes according to their known relevance in breast cancer, our choice to employ the LASSO algorithm reflects a more statistical and predictive-oriented approach. The key distinction lies in the emphasis placed on the predictive value of selected genes, without an explicit consideration of biological context (Paik, S., et al., 2004). However, it is important to note that the gene set on which LASSO operates is not devoid of context; it arises from a DEA designed to capture relevant genes associated with ET-response. This highlights that even within a statistical framework, the gene selection process remains inherently context-dependent. Notably, the LASSO approach has demonstrated its effectiveness in previous studies within the field of biomarker discovery and predictive modeling. Research conducted by Wang et al. (2022), Yang et al. (2021), and Tang et al. (2022) successfully employed the LASSO algorithm to identify robust gene signatures with significant predictive value in various contexts, further supporting the suitability of this method for our research (Wang, T. et al., 2022; Yang, Z. et al., 2021; Tang, Y. et al., 2022).

Throughout our model training process, a pattern emerged, revealing that RF consistently produced suboptimal results. This prompted us to reevaluate the applicability of RF to our dataset. Our dataset presented a distinct class imbalance, with only 12 sensitive samples compared to 43 resistant samples. As supported by previous literature, RF tends to encounter challenges when dealing with imbalanced datasets. Despite experimentation with various hyperparameters, RF consistently underperformed when compared to SVM and NN (Zhu, T., 2020).

The SVM model trained on the 27-gene feature set achieved good performance, even slightly higher than the ultimately chosen NN. The NN demonstrated perfect classification of the 12 sensitive samples with a recall of 1. Additionally, since some external validation sets lack expression data for certain genes, the SVM would require imputation of expression values during validation. Given the NN's similarly good performance and the avoidance of imputation, we opted to continue with it.

Despite achieving strong performance during the training of the NN models and robust results through CV, we encountered suboptimal performance during validation with external datasets. Several factors, biological, clinical and technical, could account for these discrepancies. From a biological perspective, breast cancer exhibits significant interpatient heterogeneity, and resistance to ET can arise through various pathways. As a result, patients classified as clinically resistant may employ diverse mechanisms to circumvent estrogen dependency, leading to distinct transcriptomic profiles detected through RNAseq. In addition, the approach for measuring estrogen response or ET resistance can differ significantly, such as variations in the classification criteria used. This heterogeneity becomes particularly pronounced when combining data from different studies with a limited number of samples for both model training and external validation (Sjöström, M. et al., 2018). Moreover, during our quality control procedure for the validation data, we noticed a separation in the PCA plot. This separation was evident between the training data (a combination of internally generated and Gou bulk RNAseq data) and the

external Xia bulk RNAseq validation data. This divergence in expression profiles, as indicated by the PCA plot, likely contributed to misclassification by the model during validation.

One challenging but innovative aspect of the study is our attempt to validate a model originally trained on bulk RNAseq data using snRNA-seq data. While the results were suboptimal, we were surprised by the feasibility of this approach, despite the questionability of averaging gene expression across different cells for each patient. To further align our validation process with the nature of bulk RNAseq data, we took an additional step. We aggregated the raw counts of specific genes across all cells for each patient, resulting in a single raw count value per gene per patient (summed pseudobulk validation set). This approach was guided by our hypothesis that it closely resembles the rationale of bulk RNA-seq. Upon evaluating the final accuracies and F-scores, we were gratified to observe that the summed pseudobulk method proved to be more effective in validating the model, bringing it closer in line with the characteristics of the training data.

Our results indicate that the Filtered-Relaxed 27-gene signature consistently outperforms the Filtered 73-gene signature in terms of classification performance after validation. This aligns with the fact that achieving comparable or better classification performance with a smaller number of genes is desirable, especially for clinical practice applications.

Notably, our study stands out by incorporating data from four different datasets derived from distinct studies, a less common approach in scientific literature. Despite the challenges and potential limitations observed during validation, we believe that this model, or similar ones, holds promise for future applications, provided more suitable validation data can be obtained. This pertains to both biological aspects, such as identifying samples with similar resistant pathways activated, and technical aspects, such as aligning with the characteristics of the training data.

In conclusion, our study paves the way for the development of a validated model that holds significant promise for clinical practice. Much like established signatures such as PAM50 or MOTERA, a successfully validated model could empower clinicians to make more precise treatment decisions. This would not only spare patients the harm and cost of unnecessary interventions, but also save the CDK4/6i-combination for those who do not respond, and extend the duration before resorting to chemotherapy.

## Acknowledgements

I would like to express my sincere gratitude to my supervisor, Lara Nonell, for all the guidance through this study. I am deeply appreciative of the knowledge and skills I have gained under her supervision. Special thanks for my academic tutor Malu Calle Rosingana for her willingness to help. I am also grateful to my colleagues, Pau, Irene, and Alba, for their support, insightful contributions, and advice during our collaborative lab meetings.

## References

- Duan, Y., Guo, D., Zhang, X., Lan, L., Meng, H., Wang, Y., Sui, C., Qu, Z., He, G., Wang, C., & Liu, X. (2023). Diagnostic accuracy of optical coherence tomography for margin assessment in breast-conserving surgery: A systematic review and meta-analysis. *Photodiagnosis and Photodynamic Therapy*, 103718. <https://doi.org/10.1016/J.PDPDT.2023.103718>
- Hanker, A. B., Sudhan, D. R., & Arteaga, C. L. (2020). Overcoming Endocrine Resistance in Breast Cancer. *Cancer Cell*, 37(4), 496–513. <https://doi.org/10.1016/J.CCELL.2020.03.009>
- Ma CX, Reinert T, Chmielewska I, Ellis MJ. Mechanisms of aromatase inhibitor resistance. *Nat Rev Cancer* 2015;15:261–75.
- Dustin D, Gu G, Fuqua SAW. ESR1 mutations in breast cancer. *Cancer* 2019;125: 3714–28.

- Fribbens C, O'Leary B, Kilburn L, Hrebien S, Garcia-Murillas I, Beaney M, et al. Plasma ESR1 mutations and the treatment of estrogen receptor-positive advanced breast cancer. *J Clin Oncol* 2016;34:2961–8.
- Spoerke JM, Gendreau S, Walter K, Qiu J, Wilson TR, Savage H, et al. Heterogeneity and clinical significance of ESR1 mutations in ER-positive metastatic breast cancer patients receiving fulvestrant. *Nat Commun* 2016;7:11579.
- Turner N. J., Slamon, D. J., Ro, J., Bondarenko, I., Im, S., Masuda, N., ... & Cristofanilli, M. (2018). Overall Survival With Palbociclib and Fulvestrant In Advanced Breast Cancer. *N Engl J Med*, 20(379), 1926-1936. <https://doi.org/10.1056/nejmoa1810527>
- Sonke, G. S., van Ommen - Nijhof, A., Wortelboer, N., van der Noort, V., Swinkels, A. C. P., Blommestein, H. M., Beeker, A., Beelen, K., Hamming, L. C., Heijns, J. B., Honkoop, A. H., de Jong, P. C., van Rossum-Schornagel, Q. C., van Schaik-van de Mheen, C., Tol, J., Tromp-Van Driel, C., Vrijaldenhoven, S., van Leeuwen-Stok, A. E., Konings, I., & Jager, A. (2023). Primary outcome analysis of the phase 3 SONIA trial (BOOG 2017-03) on selecting the optimal position of cyclin-dependent kinases 4 and 6 (CDK4/6) inhibitors for patients with hormone receptor-positive (HR+), HER2-negative (HER2-) advanced breast cancer (ABC). *Journal of Clinical Oncology*, 41(17\_suppl), LBA1000–LBA1000. [https://doi.org/10.1200/JCO.2023.41.17\\_suppl.LBA1000](https://doi.org/10.1200/JCO.2023.41.17_suppl.LBA1000)
- Al-qasem, A. J., Alves, C. L., & Ditzel, H. J. (2021). Resistance Mechanisms to Combined CDK4/6 Inhibitors and Endocrine Therapy in ER+/HER2- Advanced Breast Cancer: Biomarkers and Potential Novel Treatment Strategies. *Cancers*, 13(21). <https://doi.org/10.3390/CANCERS13215397>
- Sammons, S., Shastry, M., Dent, S., Anders, C., & Hamilton, E. (2020). Practical Treatment Strategies and Future Directions After Progression While Receiving CDK4/6 Inhibition and Endocrine Therapy in Advanced HR+/HER2- Breast Cancer. *Clinical breast cancer*, 20(1), 1–11. <https://doi.org/10.1016/j.clbc.2019.06.017>
- Sithara, S., Crowley, T. M., Walder, K., & Aston-Mourney, K. (2017). Gene expression signature: a powerful approach for drug discovery in diabetes. *The Journal of endocrinology*, 232(2), R131–R139. <https://doi.org/10.1530/JOE-16-0515>
- Qian, Y., Daza, J., Itzel, T., Betge, J., Zhan, T., Marmé, F., & Teufel, A. (2021). cells Prognostic Cancer Gene Expression Signatures: Current Status and Challenges. <https://doi.org/10.3390/cells10030648>
- Gou, X., Anurag, M., Lei, J. T., Kim, B. J., Singh, P., Seker, S., Fandino, D., Han, A., Rehman, S., Hu, J., Korchina, V., Doddapaneni, H., Dobrolecki, L. E., Mitsiades, N., Lewis, M. T., Welm, A. L., Li, S., Lee, A. v., Robinson, D. R., ... Ellis, M. J. (2021). Transcriptional Reprogramming Differentiates Active from Inactive ESR1 Fusions in Endocrine Therapy-Refractory Metastatic Breast Cancer. *Cancer Research*, 81(24), 6259–6272. <https://doi.org/10.1158/0008-5472.CAN-21-1256>
- Theilhaber, J., Chiron, M., Dreyman, J. et al. Construction and optimization of gene expression signatures for prediction of survival in two-arm clinical trials. *BMC Bioinformatics* 21, 333 (2020). <https://doi.org/10.1186/s12859-020-03655-7>
- Cantini, L., Calzone, L., Martignetti, L. et al. Classification of gene signatures for their information value and functional redundancy. *npj Syst Biol Appl* 4, 2 (2018). <https://doi.org/10.1038/s41540-017-0038-8>
- Chicco, D., Alameer, A., Rahmati, S., & Jurman, G. (2022). Towards a potential pan-cancer prognostic signature for gene expression based on probesets and ensemble machine learning. *BioData mining*, 15(1), 28. <https://doi.org/10.1186/s13040-022-00312-y>
- Yu, C., You, M., Zhang, P., Zhang, S., Yin, Y., & Zhang, X. (2021). A five-gene signature is a prognostic biomarker in pan-cancer and related with immunologically associated extracellular matrix. *Cancer medicine*, 10(13), 4629–4643. <https://doi.org/10.1002/cam4.3986>
- Xu, Q., Chen, J., Ni, S., Tan, C., Xu, M., Dong, L., Yuan, L., Wang, Q., & Du, X. (2016). Pan-cancer transcriptome analysis reveals a gene expression signature for the identification of tumor tissue origin. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc*, 29(6), 546–556. <https://doi.org/10.1038/modpathol.2016.60>
- Kalafi, E. Y., Nor, N. A. M., Taib, N. A., Ganggayah, M. D., Town, C., & Dhillon, S. K. (2019). Machine Learning and Deep Learning Approaches in Breast Cancer Survival Prediction Using Clinical Data. *Folia biologica*, 65(5-6), 212–220.
- Monserrat, L., Yañez, C., Sánchez-Guixé, M., Viaplana, C., Brasó-Maristany, F., Bellet, M., Chandarlapaty, S., Nonell, L., Vicent, G. P., Serra, V., ... Antitumor activity of the SARM RAD140 in hormone-independent estrogen receptor-positive breast cancer patient-derived xenografts. *EACR Congress 2022*. Seville, Spain
- Eisenhauer, E. A., Therasse, P., Bogaerts, J., Schwartz, L. H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M., Rubinstein, L., Shankar, L., Dodd, L., Kaplan, R., Lacombe, D., & Verweij, J. (2009). New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *European journal of cancer (Oxford, England : 1990)*, 45(2), 228–247. <https://doi.org/10.1016/j.ejca.2008.10.026>
- Harshil Patel, Phil Ewels, Alexander Peltzer, Rickard Hammarén, Olga Botvinnik, Gregor Sturm, Denis Moreno, Pranathi Vemuri, silviamorins, Lorena Pantano, Gavin Kelly, FriederikeHanssen, Maxime U. Garcia, nf-core bot, Chris Cheshire, rfenouil, marchoeppner, Peng Zhou, Gisela Gabernet, ... Alice Mayer. (2021). nf-core/maseq: nf-core/maseq v3.4 - Aluminium Aardvark (3.4). Zenodo. <https://doi.org/10.5281/zenodo.5550247>.
- Griffiths, J. I., Chen, J., Cosgrove, P. A., O'Dea, A., Sharma, P., Ma, C., Trivedi, M., Kalinsky, K., Wisinski, K. B., O'Regan, R., Makhoul, I., Spring, L. M., Bardia, A., Adler, F. R., Cohen, A. L., Chang, J. T., Khan, Q. J., & Bild, A. H. (2021). Serial single-cell genomics reveals convergent subclonal evolution of resistance as early-stage breast cancer patients progress on endocrine plus CDK4/6 therapy. *Nature Cancer*, 2(6), 658. <https://doi.org/10.1038/S43018-021-00215-7>
- Youli Xia, Xiaping He, Lorna Renshaw, Carlos Martínez-Perez, Charlene Kay, Mark Gray, James Meehan, Joel S. Parker, Charles M. Perou, Lisa A. Carey, J. Michael Dixon, Arran Turnbull; Integrated DNA and RNA Sequencing Reveals Drivers of Endocrine Resistance in Estrogen Receptor-Positive Breast Cancer. *Clin Cancer Res* 15 August 2022; 28 (16): 3618–3629. <https://doi.org/10.1158/1078-0432.CCR-21-3189>
- Love MI, Huber W, Anders S (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” *Genome Biology*, 15, 550. <https://doi.org/10.1186/s13059-014-0550-8>.
- Hao Y, Hao S, Andersen-Nissen E, III WMM, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zagar M, Hoffman P, Stoekius M, Papalexi E, Mimitou EP, Jain J, Srivastava A, Stuart T, Fleming LB, Yeung B, Rogers AJ, McElrath JM, Blish CA, Gottardo R, Smibert P, Satija R (2021). “Integrated analysis of multimodal single-cell data.” *Cell*. doi:10.1016/j.cell.2021.04.048. <https://doi.org/10.1016/j.cell.2021.04.048>.
- Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7), e47–e47.
- Yu G, Wang L, Han Y, He Q (2012). “clusterProfiler: an R package for comparing biological themes among gene clusters.” *OMICS: A Journal of Integrative Biology*, 16(5), 284–287. doi:10.1089/omi.2011.0118.
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), 15545–15550. [https://doi.org/10.1073/PNAS.0506580102/SUPPL\\_FILE/06580FIG7.JPG](https://doi.org/10.1073/PNAS.0506580102/SUPPL_FILE/06580FIG7.JPG)
- Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer genomics & proteomics*, 15(1), 41–51. <https://doi.org/10.21873/cgp.20063>
- Hsu, Chih-wei & Chang, Chih-chung & Lin, Chih-Jen. (2003). *A Practical Guide to Support Vector Classification*.
- Meyer, D., Dimitriadou, E., Hornik, K., Leisch, F., & Weingessel, A. (2021). e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.7-8. Retrieved from <https://CRAN.R-project.org/package=e1071>
- Tran, K. A., Kondrashova, O., Bradley, A., Williams, E. D., Pearson, J. V., & Waddell, N. (2021). Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome medicine*, 13(1), 152. <https://doi.org/10.1186/s13073-021-00968-x>
- Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*, Fourth edition. Springer, New York. ISBN 0-387-95457-0. <https://www.stats.ox.ac.uk/pub/MASS4/>.
- Breiman, L. *Random Forests*. *Machine Learning* 45, 5–32 (2001). <https://doi.org/10.1023/A:1010933404324>
- Liaw A, Wiener M (2002). “Classification and Regression by randomForest.” *R News*, 2(3), 18–22. <https://CRAN.R-project.org/doc/Rnews/>.
- Kuhn, Max (2008). “Building Predictive Models in R Using the caret Package.” *Journal of Statistical Software*, 28(5), 1–26. doi:10.18637/jss.v028.i05. <https://www.jstatsoft.org/index.php/jss/article/view/v028i05>.
- Tong, Y. (2021). The comparison of limma and DESeq2 in gene analysis. *E3S Web of Conferences*, 271, 3058. <https://doi.org/10.1051/e3sconf/202127103058>
- Schurch, N. J., Schofield, P., Gierlin'ski, M., Gierlin'ski, G., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G. G., Owen-Hughes, T., Blaxter, M., & Barton, G. J. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? <https://doi.org/10.1261/rna.053959.115>
- Gauthier, M., Agniel, D., Thiébaud, R., & Hejblum, B. P. (2020). dearseq: a variance component score test for RNA-seq differential analysis that effectively controls

## Gene Signature for ER+ Breast Cancer Classification

- the false discovery rate. *NAR Genomics and Bioinformatics*, 2(4), lqaa093. <https://doi.org/10.1093/nargab/lqaa093>
- Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F. L., Walker, M. G., Watson, D., Park, T., Hiller, W., Fisher, E. R., Wickerham, D. L., Bryant, J., & Wolmark, N. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *The New England journal of medicine*, 351(27), 2817–2826. <https://doi.org/10.1056/NEJMoa041588>
- Wang, T., Dai, L., Shen, S., Yang, Y., Yang, M., Yang, X., Qiu, Y., & Wang, W. (2022). Comprehensive Molecular Analyses of a Macrophage-Related Gene Signature With Regard to Prognosis, Immune Features, and Biomarkers for Immunotherapy in Hepatocellular Carcinoma Based on WGCNA and the LASSO Algorithm. *Frontiers in immunology*, 13, 843408. <https://doi.org/10.3389/fimmu.2022.843408>
- Yang, Z., Zi, Q., Xu, K., Wang, C., & Chi, Q. (2021). Development of a macrophages-related 4-gene signature and nomogram for the overall survival prediction of hepatocellular carcinoma based on WGCNA and LASSO algorithm. *International immunopharmacology*, 90, 107238. <https://doi.org/10.1016/j.intimp.2020.107238>
- Tang, Y., Tian, W., Xie, J., Zou, Y., Wang, Z., Li, N., Zeng, Y., Wu, L., Zhang, Y., Wu, S., Xie, X., & Yang, L. (2022). Prognosis and Dissection of Immunosuppressive Microenvironment in Breast Cancer Based on Fatty Acid Metabolism-Related Signature. *Frontiers in immunology*, 13, 843515. <https://doi.org/10.3389/fimmu.2022.843515>
- Zhu, T. (2020). Analysis on the Applicability of the Random Forest. *Journal of Physics: Conference Series*, 1607(1), 12123. <https://doi.org/10.1088/1742-6596/1607/1/012123>
- Sjöström, M., Staaf, J., Edén, P., Wärnberg, F., Bergh, J., Malmström, P., Fernö, M., Niméus, E., & Fredriksson, I. (2018). Identification and validation of single-sample breast cancer radiosensitivity gene expression predictors. *Breast cancer research: BCR*, 20(1), 64. <https://doi.org/10.1186/s13058-018-0978-y>