A WEB TOOL TO EXPLORE, ANNOTATE AND CLASSIFY THE ACIBADEM BREAST CANCER COHORT RNA-SEQ DATA WITH GENE SIGNATURES AND CLINICAL/MUTATION DATA, ACCORDING TO MOLECULAR SUBTYPES

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

MASTER OF SCIENCE

IN

MOLECULAR BIOLOGY AND GENETICS

By

Kübra Çalışır

June 2022

A WEB TOOL TO EXPLORE, ANNOTATE AND CLASSIFY THE ACIBADEM BREAST CANCER COHORT RNA-SEQ DATA WITH GENE SIGNATURES AND CLINICAL/MUTATION DATA, ACCORDING TO MOLECULAR SUBTYPES

By Kübra Çalışır

June 2022

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assoc. Prof. Özlen Konu Karakayalı(Advisor)

Asst. Prof. Onur Çizmecioğlu

Asst. Prof. Öznur Taştan

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

A WEB TOOL TO EXPLORE, ANNOTATE AND CLASSIFY THE ACIBADEM BREAST CANCER COHORT RNA-SEQ DATA WITH GENE SIGNATURES AND CLINICAL/MUTATION DATA, ACCORDING TO MOLECULAR SUBTYPES

Kübra Çalışır
M.S. in Molecular Biology and Genetics
Advisor: Assoc. Prof. Özlen Konu Karakayalı
June 2022

Transcriptomics-based approaches have revealed the molecular heterogeneity and distinct gene expression patterns across breast cancer subtypes since the early 2000s. This led to the usage of molecular subtypes in clinics and translational research in prognostic assessment, therapeutic efficacy prediction, and retrospective analysis of cohort studies. In this thesis, breast cancer subtypes of Acıbadem Breast Cancer Cohort (ABCC) RNA-seq data were classified with immunohistochemistry (IHC), PAM50, and SCMOD1 as molecular subtype predictors. The results revealed the moderate concordance of the methods across ABCC and selected five other public datasets. In addition, it was shown that the classification of ABCC and TCGA-BRCA RNA-seq data strongly depends on the gene signature selection. Further, a machine learning model trained with TCGA-BRCA RNA-seq data and PAM50 genes as predictors showed moderate results for ABCC and MATADOR due to the imbalanced nature of datasets where feature importance revealed a subset of PAM50 genes as predictors. Additionally, the R-Shiny-based classABCC app was developed to facilitate clustering of ABCC with six gene signatures, molecular subtyping of ABCC, and prediction of subtypes with TCGA-BRCA RNA-seq trained machine learning model.

Keywords: Breast Cancer Cohort, Clustering, Gene signature, Immunohistochemistry, Machine Learning, Molecular subtyping, PAM50, RNA-seq, R-shiny app, SCMOD1.

ÖZET

ACIBADEM MEME KANSERI KOHORTU RNA-SEQ VERILERINI GEN IMZALARI VE KLINIK/MUTASYON VERILERIYLE MOLEKULER ALT TIPLERE GORE KESFETMEK, BETIMLEMEK VE SINIFLANDIRMAK ICIN BIR WEB ARACI

Kübra Çalışır
Moleküler Biyoloji ve Genetik, Yüksek Lisans
Tez Danışmanı: Doçent Dr. Özlen Konu Karakayalı
Haziran 2022

Transkriptomik temelli yaklaşımlar, 2000'li yılların başından beri meme kanseri alt tipleri arasında moleküler heterojenliği ve farklı gen ekspresyon modellerini ortaya çıkarmıştır. Moleküler alt tiplerin kullanımı prognostik değerlendirme, terapötik etkinlik tahmini ve kohort çalışmalarının geriye dönük analizlerinde kliniklerde ve translasyonel araştırmalarda kullanımına yol açtı. Bu tezde, Acıbadem Meme Kanseri Kohortu (ABCC) RNA-seq verilerinin meme kanseri alt tipleri, moleküler alt tip belirleyicileri olan immünohistokimya (IHC), PAM50 ve SCMOD1 ile sınıflandırıldı. Sonuçlar, ABCC ve seçilen diğer beş halka açık veri genelinde yöntemlerin orta düzeyde uyumunu ortaya çıkardı. Ek olarak, ABCC ve TCGA-BRCA RNA-seq verilerinin sınıflandırılmasının güçlü bir şekilde gen imza seçimine bağlı olduğu gösterildi. Ayrıca, tahmin ediciler olarak PAM50 genleri ve TCGA-BRCA RNA-seq verileri ile eğitilmiş bir makine öğrenimi modeli veri kümelerinin dengesiz doğası nedeniyle ABCC ve MATADOR için orta düzeyde sonuçlar gösterdi. Tahmin ediciler olarak PAM50 genlerinin bir alt kümesini ortaya çıkartıldı. Ek olarak, R-Shiny tabanlı classABCC uygulaması, ABCC'nin altı gen imzasıyla kümelenmesini, ABCC ya da kullanıcı girdisi transkriptomik verilerinin moleküler alt tiplendirilmesini ve ve TCGA-BRCA RNA-seq eğitimli makine öğrenim modeliyle alt tiplerin tahminini kolaylaştırmak için geliştirildi.

Anahtar sözcükler: Meme Kanseri Kohortu, Kümeleme, Gen imzası, İmmünohistokimya, Makine Öğrenimi, Moleküler alt tipleme, PAM50, RNA-seq, R-shiny uygulaması, SCMOD1.

Acknowledgement

First of all, I would like to express my gratitude to my supervisor, Assoc. Prof. Ozlen Konu for giving me a chance to be a part of Konu Lab, her endless support, encouragement, and belief in me. I am very grateful to her for guiding me and giving me opportunities since my internship. It was a privilege to be a Master's student of Ozlen Hoca.

I would like to express my special thanks to my jury members Asst. Prof. Onur Cizmecioglu and Asst. Prof. Oznur Tastan for accepting to be my jury members in their rush time and for their valuable advices to improve my thesis.

I would like to thank Acıbadem University and Acıbadem Research Institue of Senology, Assoc. Dr. Fatma Tokat, Prof. Dr. Cengiz Yakicier, and Prof. Dr. Cihan Uraz for their collaboration and providing me with this opportunity. Furthermore, I would like to thank Sertac Kip (MD, PhD), Heather Milligan, Chris Dwan, Andrew Uzilov, Huan Wang, and SEMA4 Health Intelligence Company for providing the ABCC RNAseq data and their collaboration in this project.

I would like to thank Bilkent University Molecular Biology and Genetics Department for the scholarships and facilities which were provided. Then, I would like to thank TUBITAK (The Scientific and Technological Research Council of Turkey) for giving me the scholarship of BIDEB-2210/A scholarship.

I would like to thank all past and present Konu Lab members for creating a friendly and professional environment. I would like to thank Farid Ahadli and Moiz Aftab for helping in the project, supplying their ideas, and double-checking my codes. Finally, I want to thank Merve Vural, who taught me a lot in this lab, and she was like a big sister to me.

I would like thank my beautiful family, Ebru, Mehmet, Elif, and Metin Çalışır. I am very lucky to have you, and I love you so much. Your support and belief in me are everything to me. Thank you for being my family.

Acknowledgement

I want to start thanking my friends with Bugra Gungul. He is the best friend that someone can have. He is my dear who was always with me to cheer me up and make my life easier. I am very lucky to have you. Then, my sweetest friend Ozge Sengul always calmed me down and helped me on my worst days. Her help, support, and talks always mean a lot to me. Next, I would like to thank my beautiful friend Irmak Kaysudu who always helped me keep going on my worst days and motivated me. I very fortunate have you as my friend. Then I would like to thank Serdar Baysal and Nazlı Eda Yayvan for giving their full support and being always there for me whenever I needed them. Finally, I would like to thank Sultan Turan, who is like a mom and take care of us. I am very grateful to have you all in my life.

To my sister Elif and brother Metin,

Table of Contents

List of Figures 4

List of Tables 8

List of Appendices 11

Abbreviations 12

1. Introduction 13

1.1. Breast Cancer and its Subtypes 13

1.1.1. Luminal-like 13

1.1.1.1. Luminal A 13

1.1.1.2. Luminal B 14

1.1.2. HER2-positive / HER2-enriched 15

1.1.3. Basal-like 16

1.1.4. Normal-like 17

1.2. Breast Cancer Subtyping with Immunohistochemistry 17

1.3. Breast Cancer Subtyping with Transcriptomics 18

1.3.1. Single Sample Predictors (SSPs) and Subtype Classification Models (SCMs) 21

1.3.1.1. Prediction Analysis of Microarray 50 (PAM50) 22

1.3.1.2. Subtype Classification Model 1 (SCMOD1) 22

1.3.3.2. Other Available Gene Signatures 23

1.3.3.2.1. Abdelhafiz 17 Genes List/ 17-gene Signature 23

1.3.3.2.2. YMR16 Gene Signature/ 16-gene Signature 23

1.3.3.2.3. Mammaprint Gene Signature/ 70-gene Signature 24

1.3.3.2.4. OncotypeDx Gene Signature/ 21-gene Signature 24

1.3.3.2.5. EndoPredict Gene Signature/ 8-gene signature 25

1.4. Machine Learning (ML) 25

1.4.1. Unsupervised Learning 26

1.4.1.1. Clustering 26

1.4.1.2. Dimension Reduction 27

1.4.2. Supervised Learning 27

1.4.2.1. Classification 28

1.4.2.2. Regression 29

1.4.3. Preprocessing Techniques in Machine Learning 30

1.4.4. Overfitting and Underfitting 32

1.4.5. Imbalanced Data and Oversampling versus Undersampling 33

1.5. TCGA and Breast Cancer Cohort Studies 34

1.6. Shiny App 40

2. Aims and Rationale 41

3. Materials and Methods 42
3.1. Acibadem Breast Cancer Cohort (ABCC) Data Preparation 47
3.2. Molecular subtyping 48
3.3. Immunohistochemistry (IHC) subtyping 48
3.4. TCGA-BRCA Data Preparation 49
3.5. Batch Effect Control 50
3.6. Shiny App 50
3.6.1. Breast Cancer Subtype Differentiation (Heatmap) 50
3.6.2. Dimensional Reduction (PCA) 51
3.6.3. Machine Learning (ML) 51
3.7. GSE167977 (MATADOR) Public Data Preparation 53
3.8. Concordance Analysis using MAINZ, TRANSBIG, UPP, UNT, NKI 53
3.9. PI3K-AKT Signaling Pathway Genes as User Defined Signature 54

4. Results 55
4.1. Subtyping methods IHC, PAM50 and SCMOD1 do not always produce concordant results with each other 55
4.2. PAM50 and SCMOD1 subtyping predictors are 67% and 82% concordant across ABCC and the global populations, respectively 59
4.3. Molecular subtype assignment with transcriptomics data strongly depends on the gene signature selection for ABCC dataset. 64
4.4. Subtype structure of ABCC is comparable to that of TCGA-BRCA cohort 68
4.5. PAM50 signature based TCGA-BRCA machine learning model can moderately predict subtypes of ABCC and GSE167977 (MATADOR). 72
4.5.1. Model selection 73
4.5.2. Testing the model with the TCGA-BRCA validation data 74
4.5.3. Feature importance by ranking 76
4.5.4. Testing the model with the ABCC test data reveals moderate predictive ability, which is improved by undersampling 81
4.5.5. Testing the model with MATADOR RNA-seq data (GSE167977) reveals similar prediction power as in the case of ABCC 87
4.6. classABCC App 89
4.6.1. Clustering of ABCC with Signature and Input Annotation 92
4.6.2. Clustering of ABCC with Input Signature and Annotation 95
4.6.3. PCA 99
4.6.4. Molecular Subtyping with ABCC 100
4.6.5. TCGA 103
4.6.6. Predict 104

5. Discussion and Conclusion 106
5.1. Comparison of IHC, PAM50, and SCMOD1 methods 107
5.2. Comparison of PAM50 and SCMOD1 methods 111
5.3. Clustering of ABCC and TCGA-BRCA using selected gene signatures 112

5.4. Machine learning models to predict breast cancer subtypes **114**
5.5. A subset of PAM50 genes with feature importance **118**
5.6. Strengths and Limitations **120**
6. Future Perspectives **122**
7. References **124**
8. Appendix **139**

List of Figures

Figure 1. Diagram showing different approaches in machine learning [201]. 26

Figure 2. Machine Learning Classification and Regression Model difference modified from [202]. 28

Figure 3. The importance of accuracy in classification with class-imbalanced data. 29

Figure 4. Good fit, underfitting and overfitting machine learning models modified from [203]. 32

Figure 5. Error prediction as a function of the complexity of the model modified from [204]. 32

Figure 6. Machine Learning pipeline. 53

Figure 7. PC1 (32%) and PC2 (20%) scores from PCA of DESeq2 normalized PAM50 genes and subtypes were colored according to calls by IHC method *(the red circle indicates the two discordant patients)*. 58

Figure 8. PC1 (32%) and PC2 (20%) scores from PCA of DESeq2 normalized PAM50 genes colored by PAM50 *(top)* and by SCMOD1 *(bottom)* method predictions in the ABCC cohort *(Pink circles indicate patients which were identified as HER2-positive with SCMOD1 method but but different subtype with PAM50 method)*. 59

Figure 9. Heatmap showing probabilities of subtype predictions by SCMOD1 and PAM50 *(Rows show the patients and left annotation shows SCMOD1 and PAM50 subtypes with color code and legends are shown on the right side of the heatmap).* 60

Figure 10. Heatmap of logCPM values of A) 17-gene signature; B) PAM50 signature across ABCC data; C) annotated with ER, PR, HER2, KI67, Age *(KI67 Status; KI67< 20%: low proliferation, >20%: high proliferative).* 65

Figure 11. Heatmap of logCPM values of A) EndoPredict; B) Mammaprint signature across ABCC data; C) annotated with ER, PR, HER2, KI67, Age *(KI67 Status; KI67< 20%: low proliferation, >20%: high proliferative)*. 66

Figure 12. Heatmap of logCPM values of A) OncotypeDx; B) 16-gene signature across ABCC data; C) annotated with ER, PR, HER2, KI67, Age *(KI67 Status; KI67< 20%: low proliferation, >20%: high proliferative)*. 67

Figure 13. PC1 (46%) and PC2 (14%) scores from PCA of batch effect removed values of the 17-gene signature. 68

Figure 14. PCA of batch effect removed values of A) PC1 (29%) and PC2 (12%) scores from Mammaprint; and B) PC1 (45%) and PC2 (12%) scores from OncotypeDx. 69

Figure 15. PC1 (35%) and PC2 (29%) scores from PCA of batch effect removed values of PAM50. 70

Figure 16. PCA of batch effect removed values of A) PC1 (38%) and PC2 (27%) scores from EndoPredict; and B) PC1 (36%) and PC2 (16%) scores from 16-gene signature. 71

Figure 17. Accuracy and Kappa values of machine learning model trained with rf: random forest; svmLinear: support vector machine; regLogistic: regression logistic; nb: naive bayes; knn: k-nearest neighbors algorithms fed with TCGA-BRCA RNA-seq data. 74

Figure 18. Feature importance plot of PAM50 genes in random forest model fed with TCGA-BRCA train data *(x axis showing importance out of 100% and y axis showing PAM50 genes)*. 77

Figure 19. Model accuracy versus the number of randomly selected predictors/features *(PAM50 genes in this case)* of the Random Forest model fed with train data. 78

Figure 20. STRING network analysis of 26 genes selected with feature importance analysis. A) The genes from cellular composition organization Gene Ontology process labeled with

blue; and B) The genes from positive regulation of exit from mitosis pathway Gene Ontology process were labeled. 79

Figure 21. Bar plot showing F1 scores *(2*(Precision*Recall/Precision+Recall))* of each class/subtype by the null, oversampled and undersampled model *(x axis shows the subtypes; y-axis shows the F1 score)*. 86

Figure 22. The workflow of the app developed in this thesis *(ABCC: Acibadem Breast Cancer Cohort; TCGA-BRCA: The Cancer Genome Atlas- Breast Cancer; ML: Machine Learning)*. 89

Figure 23. The heatmap of logCPM values of 17-gene signature annotated with clinical data including status of ER, PR, HER2, and KI67, and age *(distance method: euclidean, clustering method: complete, number of column clusters: 5).* 93

Figure 24. The sidebar panel of the *Clustering with Signature* tab A) Gene signature selection bar; B) Normalization type selection bar; C) Slider of number of column clusters; D) Distance method selection bar; E) Clustering method selection bar; and F) Column annotation input bar. 94

Figure 25. The sidebar panel of the *Clustering with Input* tab A) The format of input gene signature CSV file; B) Normalization type selection bar; C) Slider of number of column clusters; D) Distance method selection bar; E) Clustering method selection bar; F) Column annotation input. 96

Figure 26. Heatmap of logTPM values of the genes involved in PI3K-AKT pathway annotated with ER, PR, HER2, KI67, and age status *(distance method: euclidean, clustering method: complete, number of column clusters: 2).* 97

Figure 27. Heatmap of logTPM values of BRCA1, BRCA2, PTEN, TP53, CDH1, STK11, CHEK2, BRIP1, ATM, PALB2 annotated with ER, PR, HER2, KI67, and age status *(distance method: euclidean, clustering method: complete, number of column clusters: 2).* 98

Figure 28. Screenshots of the *PCA* tab of classABCC. (A) Sidebar panel of *PCA* tab that enables gene signature and grouping selection; B) The grouping selection bar of *PCA* tab presents grouping by patients' id, IHC, PAM50, and SCMOD1 subtypes; and C) The classABCC output of PCA of PAM50 genes and colored by IHC subtypes *(the same result discussed in Figure 7)*. 99

Figure 29. Screenshots of the *Molecular Subtyping with ABCC* tab of the classABCC. A) *Clinical Data* subtab containing clinical data and (IHC) subtypes; B) *PAM50* subtab showing the table of subtypes by PAM50 method along with probabilities; and C) *SCMOD1* subtab showing the table of subtypes by SCMOD1 method along with probabilities. 100

Figure 30. The *Compare* subtab of the *Molecular Subtyping of ABCC* tab of the classABCC. 101

Figure 31. *Compare with Scatter Plot* and *Heatmap (the same result discussed in Figure 9)* subtabs of the *Molecular Subtyping with ABCC* tab. 102

Figure 32. Screenshots of the *TCGA* tab of classABCC. A) Sidebar panel of the *TCGA* tab that enables gene signature selection containing six different gene signatures; B) A heatmap of PAM50 genes across TCGA-BRCA and ABCC datasets; and C) the visualization output as PCA of PAM50 genes across TCGA-BRCA and ABCC datasets *(the same result discussed in Figure 15)*. 103

Figure 33. Screenshots of the *Predict* tab of classABCC. A) Confusion matrix of subtype prediction of ABCC data with TCGA-BRCA trained model; and B) Comparison of observed *(subtypes by PAM50 method)* and predicted subtypes for each patient. 105

List of Tables

Table 1. Some IHC common markers and their purposes [46]. 18

Table 2. Some studies that used intrinsic breast cancer subtyping. Gray boxes present the identified subtype and the numbers in the boxes represent the number of subtypes in each subtype. 20

Table 3. The comparison of supervised learning problems, regression and classification. 30

Table 4. Recently published TCGA studies. 34

Table 5. Some breast cancer cohort studies. *(BC: Breast Cancer; MGPs: Multiple Gene Predictors; GGI: Gene Expression Grade; SET: Sensitivity to Endocrine Therapy; TNBC: Triple Negative Breast Cancer; GEP: Gene Expression Profiles; *: shows related studies with this thesis).* 36

Table 6. Recently published R-Shiny Apps. 40

Table 7. Key resources table. 42

Table 8. Criteria of IHC categorization of ABCC. 49

Table 9. Used filters in downloading of TCGA-BRCA files from NIH GDC Data Portal. 50

Table 10. The number of subtypes in ABCC using PAM50, SCMOD1, and IHC. 55

Table 11. Expression values of ERBB2, ESR1, and PGR along with PAM50 and SCMOD1 probabilities obtained by genefu package and subtypes for selected patients *(exp: expression in logCPM; P(): Probability calculated by PAM50 or SMOD1 method; The probability of other subtypes showed in parentheses with italic)*. 56

Table 12. Subtype probabilities of selected patients by PAM50 *(top)* and SCMOD1 *(bottom)*. 61

Table 13. Overview of datasets used in the concordance analysis of PAM50 and SCMOD1.62

Table 14. Confusion matrix of SCMOD1 versus PAM50 subtype predictions. 63

Table 15. Subtype predictions of PAM50 versus SCMOD1 predictors. A) Overall statistics; B) Statistics by class. 63

Table 16. The number of each subtype in TCGA-BRCA (train and test) and ABCC datasets. 72

Table 17. The class distribution (percentage, %) of subtypes in the train, validation and test data. 73

Table 18. The confusion matrix that shows the number of predicted and observed subtypes by the model fed with TCGA-BRCA train data and tested with the validation data.. 75

Table 19. Statistics calculated for the model fed with TCGA-BRCA train data and tested with the TCGA-BRCA validation data. A) overall statistics; and B) statistics by class. 75

Table 20. The confusion matrix showing the number of predicted and observed subtypes by the model obtained with the TCGA-BRCA train data and tested with the ABCC test data. 81

Table 21. Statistics calculated for the model fed with TCGA-BRCA train data and tested with the ABCC test data. A) overall statistics; and B) statistics by class. 82

Table 22. The confusion matrix shows the number of predicted and observed subtypes by the oversampled model fed with TCGA-BRCA train data and tested with the ABCC test data. 83

Table 23. Statistics calculated for the oversampled model fed with TCGA-BRCA train data and tested with the ABCC test data. A) overall statistics; and B) statistics by class. 83

Table 24. The confusion matrix shows the number of predicted and observed subtypes by the undersampled model fed with TCGA-BRCA train data and tested with the ABCC test data.85

Table 25. Statistics calculated for the undersampled model fed with TCGA-BRCA train data and tested with the ABCC test data . A) overall statistics; and B) statistics by class. 85

Table 26. The confusion matrix that shows the number of predicted and observed subtypes by the undersampled model fed with TCGA-BRCA train data and tested with the MATADOR test data. 87

Table 27. Statistics calculated for the model fed with TCGA-BRCA train data and tested with the MATADOR test data with undersampling. A) overall statistics; and B) statistics by class. 88

Table 28. An overview of classABCC presenting the name of the tab and subtabs, the job, data used, options and the type of available visualizations. 90

List of Appendices

Appendix Table 1. The list of PAM50 genes and the centroids. 139

Appendix Table 2. The concordance of subtypes from each method. *(IHC_PAM50: (IHC versus PAM50 subtypes; IHC_SCMOD1: IHC versus SCMOD1 subtypes; PAM50 versus SCMOD1 subtypes.) (if the pair gives the same subtype it is annotated as 1 and if the pair gives the different subtypes it is annotated as 0. And rowtotal is the sum of all columns).* 141

Appendix Table 3. The list of genes selected with feature importance sorted by the importance. 144

Appendix Table 4. Genes found in each signature used in this thesis and classABCC. 145

Appendix Figure 1. Scatter plot showing subtype probabilities calculated by PAM50 *(top)* and SCMOD1 *(bottom)*. 147

Appendix Figure 2. Heatmap of ComBat adjusted values of A) 17-gene signature; B) EndoPredict; C) Mammaprint; D) OncotypeDx; E) PAM50; F) 16-gene signatures across TCGA-BRCA and ABCC data; and G) annotated with the PAM50 subtypes and data source. 149

Appendix Figure 3. Distribution of PAM50 genes in TCGA. 151

Appendix Figure 4. Distribution of PAM50 genes in ABCC. 153

Appendix Figure 5. Box plot of each feature *(PAM50 genes)* in train data. 155

Abbreviations

Abbreviations	Meaning
ABCC	Acibadem Breast Cancer Cohort
BC	Breast Cancer
ER	Estrogen Receptor
HER2	Human Epidermal Growth Factor Receptor 2
KI67	Proliferation Marker Protein KI67
PR	Progesterone Receptor
PAM50	Prediction Analysis of Microarray 50
RNA-seq	RNA-Sequencing
SCM	Subtype Classification Method
SCMOD1	Subtype Classification Method 1
TNBC	Triple Negative Breast Cancer
TCGA-BRCA	The Cancer Genome Atlas - Breast Cancer

1. Introduction

1.1. Breast Cancer and its Subtypes

Breast cancer is a cancer type which is the most common worldwide [1]. It is a heterogeneous complex disease with intertumor and intratumor molecular heterogeneity that results in distinct responses to treatments and clinical outcomes [2], [3]. Breast cancer is divided into four main molecular subtypes: Luminal A, Luminal B, HER2-enriched, and basal-like based on immunohistochemistry categorization [4].

1.1.1. Luminal-like

Luminal-like breast cancer is a subtype characterized by ER and/or PR expression that confers two-thirds of all breast cancer insides [5]. Luminal breast cancer expresses ER, PR, ER-responsive genes, and proteins encoded by normal luminal breast cells [6]. Luminal breast cancer is also divided into subtypes regarding the expression profile, proliferative potential, aggressiveness, and response to hormone therapy [6]. However, the characterization of subtypes of luminal breast cancer depends on the usage of sets of gene signatures for clustering. Thus, Hu et al. [7] combined gene signatures from three different studies [8]–[10] and identified Luminal A and Luminal B as subtypes, supported by subsequent studies [9], [11], [12].

1.1.1.1. Luminal A

Luminal A is the most common type of breast cancer, 50-60% of all breast cancers [13]. Luminal A is characterized by the expression of cytokeratins including CK 8 and CK18, and ER, PR, and ER activator genes such as LIV1, and CCND1, with negative HER2, histologically low grade, and low Ki67 mitotic index (the protein encoded by the MKi67 gene used for a prognostic indicator to assess how many cells are dividing) [6], [9], [10].

The primary therapy for Luminal A patients is hormone therapy since Luminal A is more responsive to anti-hormone therapy than chemotherapy [14], [15]. Kennecke et al. [16] studied 3726 early-stage breast cancer patients, 44% of patients are Luminal A breast cancer. Luminal A patients showed the highest survival (70%) in ten-year estimates than Luminal B, basal-like, and HER2-positive subtypes. In addition, Luminal A patients achieved the highest duration for distant metastasis, which is 2.2 years of survival. Further, the 15-years of distant relapse rate was lower in Luminal A patients than in other subtypes. Although the recurrence in bone is common, the rate of metastasis is less than 10% in the liver, lung, and brain [16]. Therefore, patients with Luminal A breast cancer have the best prognosis in addition to lower recurrence when compared to Luminal B and other subtypes [9], [10], [17], [18].

1.1.1.2. Luminal B

Luminal B accounts for 15-20% of all breast cancers, which is more poor in prognosis, higher in grade histologically, and is proliferative (higher Ki67 mitotic index) than Luminal A. Luminal B subtype has a higher relapse rate (42.9% in 15-years) and lower survival rate (46.1% in ten-years survival estimate) [16]. Although Luminal B is an ER-positive tumor, the expression of hormone receptors is not very high [19]. Thus, Luminal B cancers are less responsive to hormone therapy compared to Luminal A [20]. Further, it can be responsive to HER2-inhibiting if the sample is HER2-positive; however, it is less sensitive compared to HER2-positive and basal-like subtypes [20], [21, p. 21], [22]. It was found that untreated Luminal B are generally classified as high-risk tumors with basal-like or HER2-positive subtypes [13].

Although proliferation-related genes such as v-MYB, GGH, LAPTM4B, NSEP1, and CCNE1 create the main distinction between luminal subtypes [23], the expression distribution of the proliferative genes is not always distinctive. For instance, 30% of HER2-positive breast cancer is classified as Luminal B using immunohistochemistry since proliferative-related genes activate growth receptor signaling genes in Luminal B [23], [24]. Therefore, there is a need for criteria or methods that can be applied to differentiate luminal subtypes.

KI67 mitotic index is a candidate. Cheang et al. analyzed 375 breast cancer microarray data with KI67 index, and HER2 status with immunohistochemistry revealed that 14% of KI67 cutoff differentiate Luminal A and Luminal B [25, p. 67]. However, the method has limitations, including reproducibility, antibody selection, and cell counting (both manually and automated). Thus, recent studies focus on the identification of new molecular biomarkers. Fibroblast growth factor signaling pathway and PI3K pathway appear promising, which are related to poor diagnosis, insensitivity to endocrine therapy, and high proliferation [26], [27].

1.1.2. HER2-positive / HER2-enriched

HER2 is a member of the human epidermal growth factor receptor family, which are membrane tyrosine kinases [28]. Although HER proteins undergo dimerization or transphosphorylation upon ligand binding in the activation process, HER2 does not have a ligand and requires ligand binding for activation [28]. HER2 is activated with homodimerization or heterodimerization with other family members when it is over-expressed [29], [30]. Upon activation of HER2, downstream crosstalks with various signaling pathways promote cell growth, proliferation, survival, angiogenesis, invasion, and

metastasis [29], [30]. HER2-positive tumors account for almost 20% of breast cancers and express high levels of HER2 and HER2-related genes and/or HER2 amplification in the genome [13]. HER2-positive tumors are highly aggressive, histologically high grade, and almost 50% of them are ER-positive (but with lower expression) and P53 mutation [31]. HER2-positive subtypes can be targeted with the HER2-inhibiting drug, trastuzumab, with a significant reduction in relapse time observed with a combination of chemotherapy [32], [33]. PTEN loss [34] and CXCR4 upregulation [35] result in resistance to trastuzumab.

The identification of HER2-positive breast cancer subtypes does not match perfectly using immunohistochemistry and molecular subtyping since only 70% of HER2-positive tumors, according to molecular subtyping, shows high HER2 protein level [36], [37]. Contrarily, the HER2 amplification or overexpression samples were not classified as HER2-positive in microarray analysis [36], [37]. Thus, Staff et al. identified different subtypes of HER2-positive subtype using HER2-derived prognostic predictor (HDPP) gene analysis [38]. HDPP does not only use HER2 genes and proliferative-related genes but also genes related to the response of immunity to metastasis and invasion [38].

1.1.3. Basal-like

Basal-like breast cancers express high levels of cytokeratins and myoepithelial markers, including CK5, CK14, CK17, and laminin, and do not express ER, PGR, and HER2 refers to triple-negative breast cancer (TNBC) [39], [40]. Further, it is found that integrin expression is deregulated. However, basal-like and triple-negative are not completely synonymous with each other. Not all triple-negative subtypes are basal-like [41], [42], and not all basal-like subtypes are triple-negative [41], [43]. Triple-negative corresponds to tumors lacking ER, PR,

and HER2 at protein level, which can be easily identified pathologically [39]. However, triple-negative is a more heterogeneous subtype than basal-like [41].

Multiple studies showed that basal-like subtypes have a poor prognosis [10], [17], [19]. However, a higher response was observed in basal-like subtypes than non-basal-like subtypes with anthracycline or anthracycline and taxane combined neoadjuvant chemotherapy [15].

1.1.4. Normal-like

Normal-like is a breast cancer subtype identified with molecular subtyping that accounts for 7.8% of all breast cancer cases in the lymph node-negative cohort [44]. It is a poorly characterized subtype that groups with normal breast cells [1]. Normal-like subtype ensembles like Luminal A subtype immunohistochemically, which is ER-positive and/or PR-positive and HER2-negative with low KI67 levels [44]. However, the expression profile of normal-like subtypes is similar to normal breast with poor outcome and grade III [4], [44].

1.2. Breast Cancer Subtyping with Immunohistochemistry

Immunohistochemistry (IHC) is a commonly used immunostaining method that uses monoclonal or polyclonal antibodies to detect antigens in and on cells of any tissue section [45], [46]. IHC has been used in clinics and research for diagnostic and prognostic purposes, including identifying breast cancer subtypes, guiding treatment decisions, and assessing proliferative, invasive, and apoptotic potential [45], [46] that gives greater insight than basic histological examination [47], [48]. There are various markers used in a broad sense shown in **Table 1**.

Table 1. Some IHC common markers and their purposes [46].

Marker(s)	Purpose(s)
ER, PR, HER2	Distinguish subtypes
KRT14, KRT5	Distinguish in-situ from invasive carcinomas
GCDFP15, CEA, Mammaglobin A	Determine metastatic carcinoma origin
KI67, BCL2, TP53	Assess tumor proliferation and apoptosis
SMA, Calponin, p63, SMMHC	Myoepithelial markers
E-cadherin, CK8	Distinguish lobular or ductal carcinoma

These markers are very convenient for assessing the subtype and some tumor characteristics, including the presence of hormone receptors such as ER, PR, and HER2. Thus, it can be known whether the patient will respond to hormone therapy or chemotherapy [49]. Although IHC is very descriptive about breast cancer subtypes, recent studies have been indicating that IHC is not always sufficient to identify the complexity of breast cancer by itself [50].

1.3. Breast Cancer Subtyping with Transcriptomics

High-throughput technologies, including microarray studies, next-generation sequencing, and RNA-seq technology, have changed the understanding of diseases, especially cancer. Genome-wide studies revealed that breast cancer is a very heterogeneous cancer type dividing into subtypes, showing striking differences in molecular level, response to treatments, and clinical outcomes. Gene expression data provides information about molecular heterogeneity, environmental status, cell identity, and subtypes.

The traditional classification of breast cancer, including tumor size, lymph node involvement, age of the patient, histological appearance, hormone receptor status (ER, PR, HER2), is not

sufficient for patient-oriented treatment methods, recurrence prediction and survival, and do not reflect the heterogeneity of breast cancer [25, p. 67], [50], [54], [55].

Therefore, recent studies focus on breast cancer subtype identification with Global Gene Expression Profiling (GEP) and genetic knowledge to investigate cell identity, cell environment, and distinct molecular biological classes [8], [56] (**Table 2**). Perou et al. analyzed gene expression of 65 paired tumor samples from 42 individuals using a microarray representing 8102 human genes [6]. The number of genes was reduced to 496 based on the low variance between repeating sampling from the same sample and high variance among different individuals called intrinsic genes. Then the samples were clustered according to their similarity to each other using unsupervised methods. A dendrogram with two main branches was identified according to the clustering, showing ER-positive and ER-negative samples. The ER-positive branch represented luminal groups, and the ER-negative branch was further classified into three sub-branches representing basal-like, normal-like, and HER2-positive groups. As a result, Perou et al. identified four main subtypes of breast cancer using microarray data [6].

Further, Sørlie et al. investigated 85 cDNA expression data from 78 patients with breast cancer [10]. The novel finding of the study was that the previously identified ER-positive luminal group is further divided into at least two subgroups with distinct expression patterns. Moreover, it was shown that survival significantly differs between subtypes of breast cancer in which basal and HER2-positive subtypes were associated with the lowest survival. In addition, Luminal A showed the best survival among other luminal groups [10].

Finally, Van't Veer et al. analyzed DNA microarray data of 117 primary breast cancer patients and clustered samples with supervised methods to identify markers of metastasis in lymph node-negative patients. 231 genes were identified, significantly associated with disease outcome and distant metastasis in the 5th year [57]. As a result, these studies showed that expression data is a great tool to identify breast cancer subtypes and predict survival, the risk of recurrence, and metastasis.

Table 2. Some studies that used intrinsic breast cancer subtyping. Gray boxes present the identified subtype and the numbers in the boxes represent the number of subtypes in each subtype.

Study	Data Type	Luminal		HER2	TNP		Normal -like
		Lum A	Lum B	over-expression	Basal	Non-basal	
Sørlie [10]	Gene Expression Profiling	1	1	1	1		1
Hu [7]	Gene Expression Profiling	1	1	1	1		1
PAM50 [58]	Gene Expression Profiling	1	1	1	1		1
Abd El-Rehim [11]	Tissue Microarray	1	1	1	1		1
Sotiriou [9]	Gene Expression Profiling	3		1	2		
Fan [59]	Gene Expression Profiling	1	1	1	1		
Lehmann [60]	Gene Expression				2	4	

	Profiling						
Gatza [61]	Pathway –based subtyping	2	5	2	3		
		4					
		1					
METAFABRIC [62]	Integrative view	3	3	1	1		
		1					
		1					
TCGA [63]	Integrative view	1	1	1	1		
Dai [64]	Integrative view	1	1	1	1		

1.3.1. Single Sample Predictors (SSPs) and Subtype Classification Models (SCMs)

Breast cancer subtyping was introduced by Perou et al.. They showed that breast cancer was divided into four groups HER2-positive, ER-positive/HER2-negative (Luminal A and Luminal B), basal, and normal-like subtypes using 84 cDNA microarray samples [6]. This subtyping was identified using a group of intrinsic genes that showed low variance within repeating sampling of the same tumor but high variance intertumors. Then, patients were clustered using hierarchical clustering methods into transcriptional subtypes [6]. This approach is not always applicable since it is not always possible to have repeated samples taken from the same patients [65]. Thus, the authors developed a new method called Single Sample Predictors (SSPs), which uses a simple nearest centroid model [7], [17], [66]. The first SSP model, SSP2003 [17], was further refined to SSP2006 and PAM50 [7], [66], which were applied to different breast cancer data [7], [36]. Each method has different intrinsic gene lists with differing numbers of genes (500, 306, and 50 genes, respectively). However, some limitations of SSPs were revealed, including the stability of the method [67], [68], the concordance of the intrinsic gene list [69], and the interobserver agreement [70].

To overcome the downsides of SSP models, Subtype Classification Models including SCMOD1, SCMOD2, and SCMGENE were developed (726, 663, 3 genes respectively) [65], [71]–[73]. In contrast to SSP models, SCMs use a Gaussian distribution mixture in two-dimensional space with the dimensions specified by ER gene and HER2 genes with the AURKA gene to differentiate high and low proliferative tumors [65]. SCMs use intrinsic gene lists and ER/HER2/AURKA genes called prototype genes where intrinsic genes are correlated with prototype genes [65].

1.3.1.1. Prediction Analysis of Microarray 50 (PAM50)

In 2009, Parker et al. introduced the PAM50 (Prediction Analysis of Microarray 50) method, which tests tumors based on 50 genes to predict sample subtypes [66]. As explained before, PAM50 is a member of SSPs, which assigns samples to the nearest centroids based on the intrinsic genes. A training data composed of 189 breast cancer samples analyzed by hierarchical clustering using 1906 intrinsic genes showed subtypes of Luminal A, Luminal B, HER2-enriched, normal-like, and basal subtypes [65], [66]. Then, the intrinsic genes were reduced from 1906 genes to 50 genes to identify the genes that differentiate the subtypes [66]. The 50 genes and centroids list is shown in **Appendix Table 1**.

1.3.1.2. Subtype Classification Model 1 (SCMOD1)

Desmedt et al. developed a gene expression module that uses tumor invasion (PLAU), immune response (STAT1), angiogenesis (VEGF), apoptosis (CASP3), proliferation (AURKA), and ER and HER2 signaling [71]. However, AURKA revealed the importance of proliferative genes in classification since it showed the highest percentage of specific association [71]. Therefore, SCMOD1 was simplified as a subtype classification method that

uses a mixture of three normal distributions in two dimensions by ER and HER2 genes and the AURKA gene for proliferative differentiation as low and high [65], [71].

1.3.3.2. Other Available Gene Signatures

SSPs and SCMs are methods used commonly in the clinic and research. However, there is a need for new biomarkers to make more accurate subtyping of breast cancer and prediction of prognosis and diagnosis since there is a controversy on the complexity of models and the uncertainty about the robustness and clinical relevance [67]–[69], [74]. Therefore, various studies have been conducted on new intrinsic gene signatures in breast cancer subtypes coupled with clinical variables [65], [75]–[80]. Moreover, there are commercially available multi-gene signatures *(Mammaprint, EndoPredict, OncotypeDx)* for breast cancer patients to predict clinical outcomes [44].

1.3.3.2.1. Abdelhafiz 17 Genes List/ 17-gene Signature

Abdelhafiz et al. identified 17 breast cancer classification genes that differentiate luminal breast cancer subtypes. ER, PR, HER2, and KI67 genes were assessed immunohistochemically [75]. Based on differential expression, breast cancer classification genes were selected with RT-PCR between tumor and normal samples [75].

1.3.3.2.2. YMR16 Gene Signature/ 16-gene Signature

Huang et al. used available gene signatures to construct a better signature with a smaller set of genes for prognostic purposes [80]. For these purposes, the authors overlapped 33 gene signatures (shown in Table 1 of the paper) [80] and compared them with KEGG and Gene Ontology (GO) terms. Moreover, the overlapped genes were further refined with METAFABRIC and two available datasets (GSE1456 and GSE2034) [80].

1.3.3.2.3. Mammaprint Gene Signature/ 70-gene Signature

The Mammaprint gene signature, which consists of 70 genes, is an FDA-approved commercially available gene signature [81, p. 70], [82]. The selection of 70 genes was made from a randomized study of 6693 patients with early-stage breast cancer [83]. Mammaprint gene signature showed that the survival rate of patients with a high clinical risk of recurrence is lower without chemotherapy than with chemotherapy receiving ones [83]. Thus, a 70-gene signature provides valuable information for deciding whether a patient should receive chemotherapy or not by classifying it as high and low risk [81]–[83].

1.3.3.2.4. OncotypeDx Gene Signature/ 21-gene Signature

OncotypeDx is a 21 gene expression signature containing 16 known cancer genes and five reference genes, which quantifies the 10-year distant recurrence, chemotherapy, and hormone therapy response of breast cancer patients with ER-positive and axillary lymph node-negative [84]. The commercially available signature is commonly used for prognostic purposes by classifying it as low, intermediate, and high risk [85], [86].

1.3.3.2.5. EndoPredict Gene Signature/ 8-gene signature

Endopredict gene signature is used to predict the metastases risk of ER-positive, HER2-negative breast cancer patients exposed to endocrine therapy [87]. The signature is used to determine treatment by incorporating clinical and pathological information, including the size of tumors and nodal status [87]. The signature has eight genes, BIRC5, UBE2C, DHCR7, RBBP8, IL6ST, AZGP1, MGP, and STC2, with reference genes [87].

1.4. Machine Learning (ML)

Machine learning is a concept in which computer algorithms can learn from data, identify patterns, and make decisions or predictions based on the mathematical model constructed from data learning [88], [89]. The ML process starts with feeding training data into a computer algorithm. By using the training data, machine learning models construct valuable and meaningful patterns in the data to further make decisions or predictions on the relevant test event [89], [90].

There are different approaches to machine learning based on the feedback mechanism for learning [91] shown in **Figure 1**. Before explaining the different approaches, there are some terms to be explained. In ML, the target is called a label named a dependent variable in statistics. Moreover, variables in statistics are called features in ML [89].

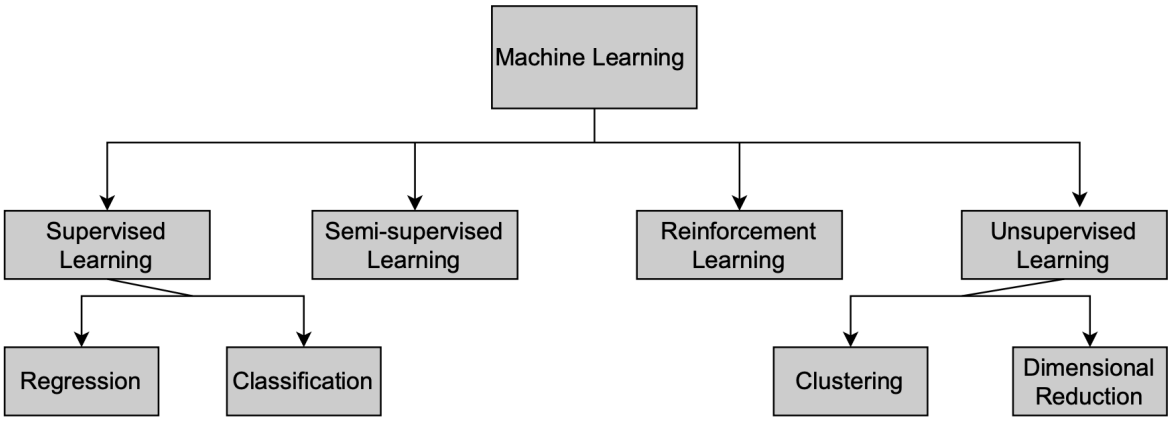


Figure 1. Diagram showing different approaches in machine learning modified from [201].

1.4.1. Unsupervised Learning

The first approach is unsupervised learning which deals with unlabeled data, unlike supervised learning [91]. Since there are no labels and input-output pairs to make decisions, there is no feedback mechanism for learning [92]. Unsupervised learning is generally used to extract features, identify trends and patterns, and group the data [91], [92]. There are two main unsupervised learning problems which are clustering and dimension reduction.

1.4.1.1. Clustering

Cluster analysis or clustering is a grouping method that groups members so that members in the same group are more similar to each other than the members of the other group(s) [93], which means that the analysis minimizes distance of intra-cluster and maximizes distance of inter-cluster .

Clustering is a statistical method that relies on similarity and dissimilarity between data points. Therefore, the distance method selection is one of the key clustering points [94]. The recommended distance method depends on the data types, which can be binary categorical

data with more than two levels, continuous data, and mixed data. Jaccard coefficient [95], Rogers and Tanimoto measure [96], Sokal and Sneath measure [97], and Gower and Legendre measures for binary [98], genetic similarity measure for categorical data [99], Euclidean, city block, Minkowski, Canberra, Pearson correlation for continuous data, and Gower similarity matrix for mixed data [100] are some of the options. Furthermore, the choice of algorithm is another parameter that affects the clustering analysis [94]. There are different methods: partitioning methods (K-means [101], K-Medoids [102]) ; density-based methods (DBSCAN [103], OPTICS [104]); hierarchical-based methods (Agglomerative [105], top-down [106]); model-based methods (GMM [107], SOM [108]).

1.4.1.2. Dimension Reduction

The high number of dimensions is a challenging task in data processing. Dimension reduction is an unsupervised learning technique that transforms multi-dimensional data into low-dimensional space by preserving relations between data points [91], [93]. Dimensions that correspond to the number of features in data are either selected, i.e., feature selection; (ANOVA, Pearson's correlation), or generating new features from the existing ones, i.e., feature extraction; (PCA) in dimension reduction [109], [110].

1.4.2. Supervised Learning

The second approach is supervised learning which can be adopted if data has labels where the data is called train data or labeled data [92]. In this concept, the feedback mechanism of learning is direct. Supervised learning learns a function (f) by mapping input (x) directly to the output (y) based on the input-output pairs presented in the training data, which means that algorithms learn the relationship between input variables and designated dependent variables

to predict new instances [111]. The most common supervised learning tasks are classification and regression [93].

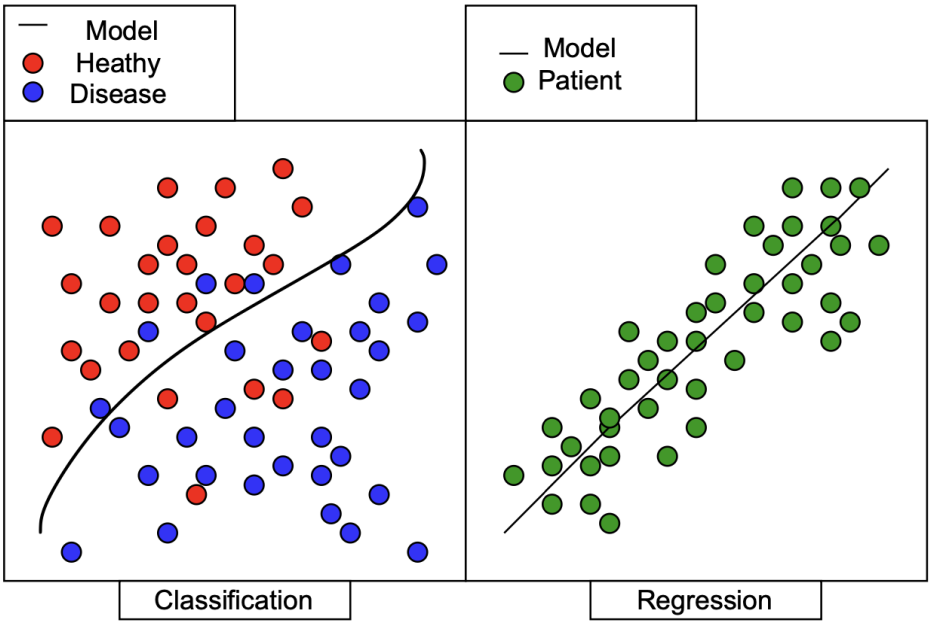
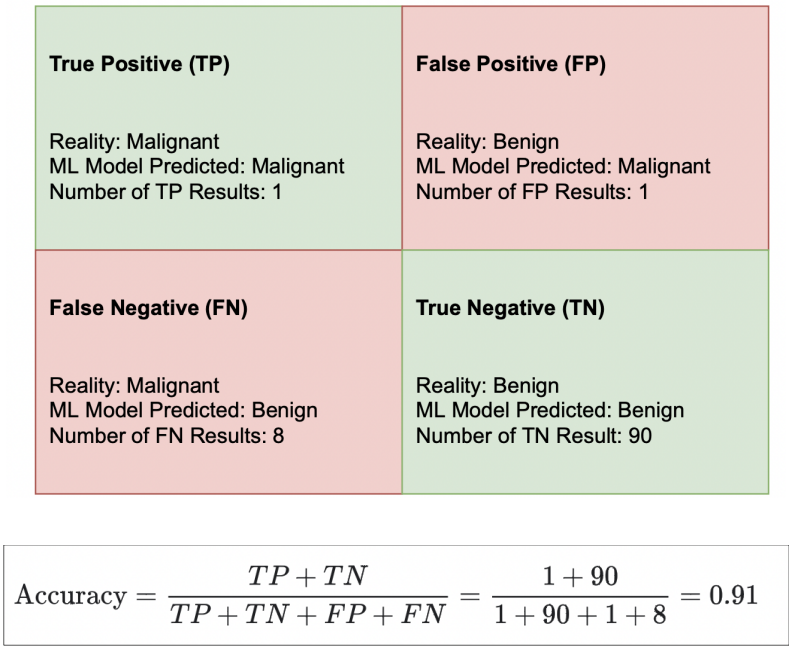


Figure 2. Machine Learning Classification and Regression Model difference modified from [202].

1.4.2.1. Classification

Classification is a supervised learning problem that can be applied to data used to predict a label or category, a discrete class label [92], [94]. A model trained to decide if a cluster of cells shows disease phenotype or not is an example of a classification problem that separates data based on the labels [112] shown in **Figure 2**. The common classification problems are binary classification (two-class), multiclass classification (multiple-class) and multilabel (multiple labels) classification [91]. Artificial Neural Networks (ANN), Bayesian Network (BN), Decision Trees (DT), k-Nearest Neighbors (k-NN), Logistic Regression (LR), Random Forests (RF), and Support Vector Machine (SVM) are some algorithms that can be used in the classification of supervised learning [112].

The performance of classification models is generally evaluated by the percentage of true classifications called accuracy [113]. Although accuracy is a metric used to assess the performance of classification models, it is not enough to conclude that the model is reliable, especially with class-imbalanced data [114], [115]. **Figure 3** shows that although 91% is a high accuracy, the model failed to identify eight malignant out of 9. In other words, the model has almost zero predictive ability in distinguishing malignant tumors from benign tumors. Therefore, precision, recall, and F1 score should be considered for the final decision on the model.



$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{1 + 90}{1 + 90 + 1 + 8} = 0.91$$

Figure 3. The importance of accuracy in classification with class-imbalanced data.

1.4.2.2. Regression

Regression is another supervised problem that can be applied to predict a continuous quantity, unlike a discrete variable in classification by fitting the data [91], [92]. A model trained to predict a baby's height at 20 years is an example of a regression problem since the answer is a

continuous quantity. Therefore, the performance of regression models is evaluated with root mean squared error (RMSE) or mean absolute error (MAE) [116]. **Table 3** explains the main differences between classification and regression problems in supervised learning.

Table 3. The comparison of supervised learning problems, regression and classification.

	Regression Algorithm	**Classification Algorithm**
Output	Continuous (numerical) value	Discrete (categorical) value
Task	Map input value to continuous value	Map input value to discrete value
Data	Continuous labeled data	Discrete labeled data
Prediction Method	Draws best line	Finds boundaries among different classes
Subcategories	Linear and Nonlinear Regression	Binary and Multi-class Classifier
Type of algorithm	Supervised learning algorithm	Supervised learning algorithm
How to assess model fit	Root Mean Square Error (RMSE)	Percentage of current classifications
Classes/inputs	Multiple inputs correspond to multivariate regression problems.	Multiple classes correspond to multi-class classification.

1.4.3. Preprocessing Techniques in Machine Learning

Data preprocessing is the fundamental step in machine learning to improve data quality which directly affects the ability of the selected model to learn [93], [94]. Thus, it is a critical and integral step that should be completed before feeding the model.

Generally, data contains some null (NA) values for various reasons. Machine learning algorithms cannot handle NA values. Thus it should be removed prior to feeding the model. There are different ways to handle NA values. The first is dropping rows/columns containing null variables [121]. Although it is an easy choice, it is not the best since it may lead to

important information loss. Instead of dropping the null values, what can be done is imputation, which means replacing missing values of the data [122]. NA values can be imputed with the mean or median of the remaining values [121].

Scaling is another integral step of data preprocessing. Generally, there are two reasons to transform the train data [121]. The first reason is that the distribution of each feature may not be skewed distributions. Linear models tend to normalize the data, which improves the model's performance [123]. The second reason is that features of the data may have values in different units. For example, one feature can be age in years, and the other one can be RNA level in counts per million (CPM). Indeed, the units are different. Algorithms using distances among points and calculating coefficients will create very different results due to the unit differences. Scaling will solve that problem by putting them on equal footing [124].

Encoding categorical values prior to feeding models is essential since machine learning algorithms cannot handle categorical values. Thus, they are converted to numerical values as 0 and 1 [121], [122]. For example, if there is a disease phenotype for an individual, it should be presented as 1. This is called dummy variable creation with one-hot encoding, where 0 and 1 represent the absence or presence of the categorical value. However, the process should be carried out carefully to avoid dummy variable traps. The dummy variable trap is a case related to multicollinear in which one variable can be enough to predict others [125]. Thus, if there is an m number of categories, the model should use an m-1 number of categories [125].

1.4.4. Overfitting and Underfitting

The purpose of the training data is to discover the true relation between variables so that the machine learning model has the power to predict the unseen/un-trained data. However, overfitting and underfitting can be experienced, resulting in the machine learning algorithm [93].

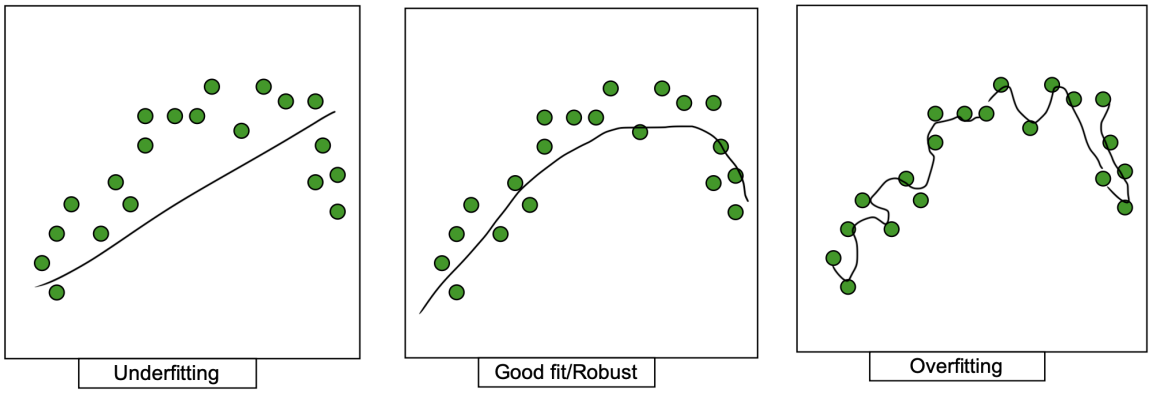


Figure 4. Good fit, underfitting and overfitting machine learning models modified from [203].

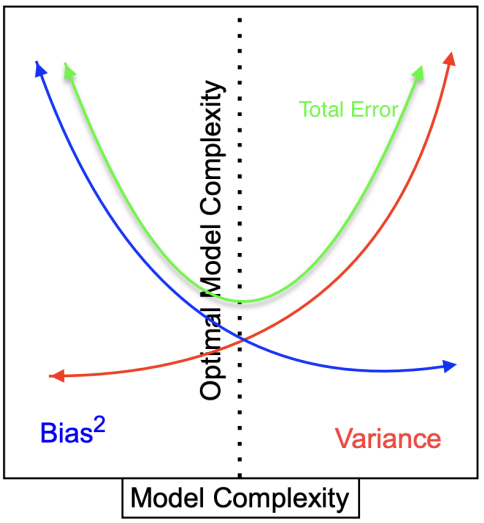


Figure 5. Error prediction as a function of the complexity of the model modified from [204].

Underfitting is a concept when a model has not learned or generalized the patterns well in training data which results in unreliable results with test data [93], [126] (**Figure 4**). Underfitting is high bias and low data variance, insufficient data size, and a straightforward model (**Figure 5**). Cleaning data from noise, increasing model complexity, the number of features, training duration, and controlling bias-variance trade-off can overcome the underfitting problem [93]. Overfitting is a concept when model performance is good with training data but poor with test data. In this case, the model learns the noise and details of the train data so well that it results in poor performance with test data and leads to false-positive discovery [111] (**Figure 4**). In this concept, noise corresponds to data points that do not reveal the true pattern of the data but random chance [89]. There can be various reasons for overfitting: low bias, high variance, insufficient size of data, or complex model (**Figure 5**). Using k-fold cross-validation [127], regularization methods [128], decreasing size of data [94], tuning parameters [89] and adopting resembling techniques [129] are some ways to overcome the overfitting problem [89], [90].

1.4.5. Imbalanced Data and Oversampling versus Undersampling

Imbalanced data means that the distribution of classes found in a data is not equal. Therefore, imbalanced data contain majority class(es): and minority class(es) where data is dominated by the majority class [186]. Therefore, machine learning models do perform poorly with the imbalanced trained data since minority classes are also predicted as majority class by the model [186]. To overcome imbalanced data problems, undersampling and oversampling can be applied. The logic of oversampling is to increase the number of minority classes to have more balanced data [186-188]. On the other hand, undersampling decreases the number of majority classes. As a result, imbalance data can be handled with resampling techniques [186-188].

1.5. TCGA and Breast Cancer Cohort Studies

The Cancer Genome Atlas (TCGA) is a cancer genomics program [130]. Under the program, primary cancers were matched with normal samples covering 33 cancer types. TCGA generated genomic, transcriptomic, proteomic, and genomic data, which led to an understanding of cancer diagnosis, prognosis, and prevention [130]. The data is currently over 2.5 petabytes publicly available for scientific purposes. TCGA deepened the understanding of cancer subtypes using molecular characterization, established rich genomic data for a broad research area, helped advancement of health and science technologies, and changed the diagnosis and treatment of cancer patients in clinics [131] (**Table 4**).

Table 4. Recently published TCGA studies.

Study	Title of study
Zhang [132]	Comprehensive Analysis of Genetics Ancestry and Its Molecular Correlates in Cancer
Gao [133]	Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data
Wang [134]	Comprehensive Molecular Characterization of the Hippo Signaling Pathway in Cancer
Corces [135]	The Chromatin Accessibility Landscape of Primary Human Cancers

Cohort studies are observational studies in which a group of individuals is followed over time by measuring the outcomes one or more times. Cohort studies are divided into retrospective and prospective [136]. A prospective study is the starting of the study with a subject before the outcome of interest is developed. On the other hand, the outcome is already present in subjects when the study is enrolled in a retrospective study [136].

Cohort studies have been used commonly in research (**Table 5**). Many publicly available cohort data were subjected to the identification of subtypes To understand the complexity of breast cancer [137], markers [138] and intrinsic gene lists [139], biological processes among subtypes in breast cancer [140], response and sensitivity to chemotherapy [141] and hormone therapy [142], the risk of metastasis [143] and recurrence [144]. The cohort studies generally used multiple array data or RNA-seq data to increase the sample size to employ supervised or unsupervised clustering. K-means, hierarchical, and consensus clustering methods are common for unsupervised clustering. On the other hand, training and test data split with classification learning methods are generally used in supervised clustering.

Table 5. Some breast cancer cohort studies. *(BC: Breast Cancer; MGPs: Multiple Gene Predictors; GGI: Gene Expression Grade; SET: Sensitivity to Endocrine Therapy; TNBC: Triple Negative Breast Cancer; GEP: Gene Expression Profiles; *: shows related studies with this thesis).*

Study	Focus of the study	The result	Train/ Test Split	TCGA data	Type of Data	# of Sample	Data Accessibility	Preservation Method
Wang [144]	Distant metastasis of lymph-node-negative BC.	Signature for distant recurrence	+	-	Array	286	GSE2034	Fresh frozen
Desmedt [143]	Validation of signature for distant metastasis.	Strong time-dependence of signature	-	-	Array	198	GSE7390	Fresh frozen
Jindal [145]*	Comparing ER+ Postpartum and Nulliparous BC	Poor prognostic phenotype of postpartum BC	-	-	RNA-seq	16	GSE158854	FFPE
Varley [146]	Fusion transcripts in BC	*SCNN1A-TNFRSF1A* and *CTSD-IFITM10* as fusion transcripts	-	-	RNA-seq	168	GSE58135	Fresh frozen
Kester [147]*	Microenvironment in BC	Survival and therapy benefit by populations of epithelial and immune cells.	+	-	RNA-seq	528	GSE167977	FFPE
					scRNA-seq	47	GSE168410	
Brueffer [138]*	Clinical value of RNA-seq-based classifiers	RNA-seq-based classifiers	+	-	RNA-seq	405	GSE81538	FFPE/ Fresh frozen

	for five biomarkers	complementary diagnostic for BC						
Popovici [148]*	Accuracy of genomic prediction	Effect of training-sample size and classification difficulty on accuracy	+	-	Array	278	GSE20194	Fresh frozen
Pawitan [149]*	Adjuvant BC therapy	Differential gene signatures among treated and untreated BC patients	+	-	Array	318	GSE1456	Fresh frozen
Iwamoto [140]	Chemotherapy sensitivity in BC	Different biological process for chemotherapy among ER+ and ER- BC	-	-	Array Array	103 61	GSE22093 GSE23988	Cell lines were used.
Shen [150]	Performance of MGPs	Consistent prediction of patient clinical outcomes by the cell line-derived MGPs	+	-	Array Array	91	GSE42822	Public data were used.
Shen [151]	Performance of MGPs in combination chemotherapy	Good performance with 291-probeset genomic predictor	-	-	Array	502	GSE3494	Fresh frozen
Miller [152, p. 53]*	p53 Status	Prediction of BC outcome by p53 status	+	-	Array	136	GSE12093	Fresh frozen
Zhang [139]*	Tamoxifen treatment	Assessment of benefit from tamoxifen by	-	-	Array	741	GSE12093	Fresh frozen

		76-gene signature						
Loi [77]	Molecular subtypes with GGI	Identification of 2 ER+ subtypes with GGI	-	-	Array	200	GSE6532	Fresh frozen
Schmidt [153]	Role of humoral immunity in node-negative BC	Association of B-cells in metastasis-free survival	-	-	Array	200	GSE6532	Public data
Symmans [142]	Genomic index for SET	Prediction of survival benefit from SET index of ER-related transcripts	-	-	Array	298	GSE11121	Fresh frozen
Burstein [137]*	Subtypes of TNBC	Identification of four TNBC subtypes	-	-	Array	265	GSE17705	Fresh frozen
Hu [7]*	Identification of intrinsic subtype	The conservation of subtypes across studies	+	-	Array	170	GSE76275	Fresh frozen
Anders [154]	Age-dependency of BC	Higher aggressiveness in younger ages	-	-	Array	344	GSE1992	Fresh frozen
Sabatier [155]*	Subtypes basal BC	Two basal subtypes with 368-gene model	+	-	Array	266	GSE20624	-
Sotiriou [156]	Relapse-free survival	Associated with gene expression grade	+	-	Array	189	GSE2990	Public data
Ivshina [157]	Subtyping with histological grade	Genetic grade signature in prognosis and diagnosis	-	-	Array	578	GSE4922	FFPE
Yang [141]	Chemotherapy resistance and recurrence in BC	Associated with overexpression of	-	-	Array	115	GSE19615	-

		LAMTM4B and YWHAZ						
Hatzis [158]*	Survival and response to TAX-ANT chemotherapy	Development of genomic predictor combined with ER status	+	-	Array	198	GSE25065	Fresh frozen

1.6. Shiny App

Shiny is an R package that enables the development of web apps in R without knowledge of HTML, CSS, or JavaScript by automating "reactive" binding between inputs and outputs [159], [160]. Shiny apps have been used in various niches, including network analysis, prognostic studies, and multi-omics data analysis (**Table 6**).

Table 6. Recently published R-Shiny Apps.

Study	**Focus of the app**
Manouchehri [161]	PAR(1) machine learning model
Syed [162]	Simulation of multi-omics data
Fanidis [163]	Multi-omics database and mining tool
Jagla [164]	Single-cell analysis
Evrenoglu [165]	Meta-analyzes of COVID-19 trials
Germani [166]	Tissue image visualizer

2. Aims and Rationale

Breast cancer is a complex and heterogeneous disease. The diversity among tumors, cancer-bearing individuals (intertumor heterogeneity), and even within the tumor (intratumor heterogeneity) is high. The diversity determines the progression of the tumor, therapeutic resistance, and treatment. To assess the subtype of breast cancer, molecular subtyping using transcriptomics data (and other types of omics data) has become a popular option.Various methods and gene signatures are available in the literature among which concordance issues have been observed. Thus, I aimed to

1. compare the concordance between IHC, PAM50, and SCMOD1 methods regarding breast cancer subtyping across ABCC, and publicly available array datasets Mainz, Transbig, UPP, UNT, and NKI.
2. explore six selected gene signatures including prognostic gene signatures and intrinsic gene lists across ABCC and TCGA-BRCA RNA-seq datasets.
3. predict breast cancer subtypes of ABCC and MATADOR RNA-seq data comparatively with a machine learning model trained with TCGA-BRCA RNA-seq data.
4. develop an R Shiny-based web-tool to facilitate a user-friendly environment to explore and predict subtypes of ABCC data and demonstrate use of the tool's modules.

3. Materials and Methods

Table 7. Key resources table.

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
TCGA-BRCA RNA-seq data	NCI GDC	https://portal.gdc.cancer.gov
MATADOR RNA-seq [147]	NCBI GEO	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE167977
MAINZ	breastCancerMAINZ	https://bioconductor.org/packages/release/data/experiment/html/breastCancerMAINZ.html
TRANSBIG	breastCancerTRANSBIG	https://bioconductor.org/packages/release/data/experiment/html/breastCancerTRANSBIG.html
UPP	breastCancerUPP	https://bioconductor.org/packages/release/data/experiment/html/breastCancerUPP.html
UNT	breastCancerUNT	http://bioconductor.org/packages/release/data/experiment/html/breastCancerUNT.html
NKI	breastCancerNKI	https://bioconductor.org/packages/release/data/experiment/html/breastCancerNKI.html
ABCC RNA-seq data	Acıbadem University	This thesis
Software and Packages		
R (version 4.1.2)	R Development Core Team	https://www.r-project.org
R Studio (version 1.4.1717)	R Studio Development Core Team	https://www.rstudio.com
Gene Expression Omnibus	NCBI	https://www.ncbi.nlm.nih.gov/geo/
Shiny App		
caret	Kuhn (2021)	https://cran.r-project.org/web/packages/caret/index.html

ComplexHeatmap	Gu (2016)	https://www.bioconductor.org/packages/release/bioc/html/ComplexHeatmap.html
DESeq2	Love et al. (2014)	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
DT	Xie et al. (2021)	https://cran.r-project.org/web/packages/DT/index.html
edgeR	Robinson et al. (2010)	https://bioconductor.org/packages/release/bioc/html/edgeR.html
genefu	Gendoo et al. (2021)	https://www.bioconductor.org/packages/release/bioc/html/genefu.html
ggplot2	Wickham (2016)	https://cran.r-project.org/web/packages/ggplot2/index.html
gplots	Warnes et al. (2020)	https://cran.r-project.org/web/packages/gplots/index.html
heatmaply	Galili et al. (2021)	https://cran.r-project.org/web/packages/heatmaply/index.html
InteractiveComplexHeatmap	Gu (2021)	https://www.bioconductor.org/packages/release/bioc/html/InteractiveComplexHeatmap.html
limma	Ritchie (2015)	https://bioconductor.org/packages/release/bioc/html/limma.html
plotly	Sievert (2020)	https://cran.r-project.org/web/packages/plotly/index.html
png	Urbanek (2013)	https://cran.r-project.org/web/packages/png/index.html
RColorBrewer	Neuwirth (2014)	https://cran.r-project.org/web/packages/RColorBrewer/index.html
shiny	Chang et al. (2021)	https://cran.r-project.org/web/packages/shiny/index.html
shinycssloaders	Sali et al. (2020)	https://cran.r-project.org/web/packages/shinycssloaders/index.html
shinydashboard	Chang et al. (2021)	https://cran.r-project.org/web/packages/shinydashboard/index.html

tidyverse	Wickham et al. (2019)	https://cran.r-project.org/web/packages/tidyverse/index.html
TCGA-BRCA Data Preparation		
biomaRt	Durinck et al. (2009)	https://bioconductor.org/packages/release/bioc/html/biomaRt.html
dplyr	Wickham et al. (2022)	https://cran.r-project.org/web/packages/dplyr/index.html
jsonlite	Ooms (2014)	https://cran.r-project.org/web/packages/jsonlite/index.html
TCGAbiolinks	Silva et al. (2016)	https://bioconductor.org/packages/release/bioc/html/TCGAbiolinks.html
ABCC Data Preparation		
biomaRt	Durinck et al. (2009)	https://bioconductor.org/packages/release/bioc/html/biomaRt.html
DESeq2	Love et al. (2014)	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
dplyr	Wickham et al. (2022)	https://cran.r-project.org/web/packages/dplyr/index.html
Molecular Subtyping		
Biobase	Huber et al. (2015)	https://bioconductor.org/packages/release/bioc/html/Biobase.html
biomaRt	Durinck et al. (2009)	https://bioconductor.org/packages/release/bioc/html/biomaRt.html
caret	Kuhn (2021)	https://cran.r-project.org/web/packages/caret/index.html
genefu	Gendoo et al. (2021)	https://www.bioconductor.org/packages/release/bioc/html/genefu.html
rmeta	Lumley (2018)	https://cran.r-project.org/web/packages/rmeta/index.html
survival	Therneau et al. (2000)	https://cran.r-project.org/web/packages/survival/index.html

xtable	Dahl et al. (2019)	https://cran.r-project.org/web/packages/xtable/index.html
Batch Effect Control		
edgeR	Robinson et al. (2010)	https://bioconductor.org/packages/release/bioc/html/edgeR.html
stringr	Wickham (2019)	https://cran.r-project.org/web/packages/stringr/index.html
sva	Leek et al. (2021)	https://bioconductor.org/packages/release/bioc/html/sva.html
Mutation Data Preparation		
vcf2maf	Kandoth (2018)	https://github.com/mskcc/vcf2maf
Machine Learning		
caret	Kuhn (2021)	https://cran.r-project.org/web/packages/caret/index.html
genefu	Gendoo et al. (2021)	https://www.bioconductor.org/packages/release/bioc/html/genefu.html
Hmisc	Harrell (2022)	https://cran.r-project.org/web/packages/Hmisc/index.html
stringr	Wickham (2019)	https://cran.r-project.org/web/packages/stringr/index.html
Concordance Analysis		
breastCancerMAINZ	Schroeder (2008)	https://bioconductor.org/packages/release/data/experiment/html/breastCancerMAINZ.html
breastCancerTRANSBIG	Schroeder (2007)	https://bioconductor.org/packages/release/data/experiment/html/breastCancerTRANSBIG.html
breastCancerUPP	Schroeder (2005)	https://bioconductor.org/packages/release/data/experiment/html/breastCancerUPP.html
breastCancerUNT	Schroeder (2007)	http://bioconductor.org/packages/release/data/experiment/html/breastCancerUNT.html

breastCancerNKI	Schroeder (2002)	https://bioconductor.org/packages/release/data/experiment/html/breastCancerNKI.html
genefu	Gendoo et al. (2021)	https://www.bioconductor.org/packages/release/bioc/html/genefu.html

3.1. Acibadem Breast Cancer Cohort (ABCC) Data Preparation

RSEM RNAseq expression data files were generated by Mount Sinai Genomics, Inc. d/b/a Sema4 Health Intelligence Company (CT, USA; Nefize Sertac Kip, MD PhD) from the tumors of retrospectively collected formalin fixed paraffin embedded (FFPE) blocks of De-Identified breast cancer patients of Acıbadem University, Research Institute of Senology, Departments of Pathology, Surgery and Molecular Biology and Genetics (Prof. Dr. Cengiz Yakicier, Prof. Cihan Uras, and Assoc. Dr. Fatma Tokat); and were provided to Prof. Cengiz Yakicier by Sema4. The Acibadem University Medical Ethics Committee (ATADEK Decision Number: 2022-09/22) approved bioinformatics analyses of these RNAseq data, which belonged to 91 unique patients with clinical/mutation data available, in the context of this thesis and any associated publication. One patient, for which the molecular subtyping with the genefu package [170] gave NA subtype, was excluded from analyses.

RSEM files supplied by Sema4 were with HGNC ids (HGNC: 100), which were then converted to HGNC symbols (ASIC1) with the biomaRt package [167]. TPM (transcript per million) and the number of reads (rounded version of raw count data) values that were supplied in each patient's RSEM file were used in the downstream analyses.

In the TPM filtration step, the genes that do not have at least one patient with one TPM count were removed. For heatmap visualization, TPM data was transformed into logarithmic scale (log2 base). In determining the number of reads passing filtration, the genes with zero count in every patient were removed. The number of reads was converted to logarithmic scale (log2 base) CPM values (Counts per Million) with the edgeR package [168] or normalized with the DeSEq2 package [169, p. 2] for visualization by heatmap and PCA, respectively.

3.2. Molecular subtyping

Before molecular subtyping, HGNC symbols were converted to Entrez gene ids with the biomaRt package [167], and variance stabilizing transformation was applied to the number of reads with equal variance and normally distributed data using the DeSEq2 package [169, p. 2]. Molecular subtyping of the 90 patients was done with the genefu package [170] using SCMOD1 and PAM50 methods.

3.3. Immunohistochemistry (IHC) subtyping

Prof. Dr. Cengiz Yakicier in collaboration with Acibadem University Hospital Pathologist and surgeon, Assoc. Prof. Dr. Fatma Tokat and Prof. Dr. Cihan Uras, respectively, have supplied the clinical data for each patient. The clinical data contained the status of each biomarker, namely, ER, PR, HER2, KI67, PPH3, and MET, as well as the age, stage of cancer, and pathological diagnosis of each patient. In this thesis the IHC subtyping was done considering the ER, PR, HER2 status, and MKI67 percentage for each patient. Categorization of IHC scores was performed as shown in **Table 8**. 20% was used as the threshold for proliferation (< 20%: low proliferation, >20%: high proliferative) [171].

Table 8. Criteria of IHC categorization of ABCC.

Molecular Subtypes	Immunohistochemical categorization
Luminal A	ER+ and/or PR+, Her2-, low proliferation (KI67)
Luminal B	ER+ and/or PR+, and Her2+
	ER+ and/or PR+, and Her2-, high proliferation (KI67)
Her2-positive	ER-, PR-, Her2+
Basal-like	ER-, PR-, Her2-

3.4. TCGA-BRCA Data Preparation

1164 TCGA-BRCA RNA-seq data files were downloaded on the 21st of January, 2022, from the National Cancer Institute (NIH) GDC Data Portal Repository [172] based on the listed filter shown in **Table 9**. Moreover, metadata were downloaded for these files to match file names (21feac5d-641f-4802-a910-51b73431de3c.htseq.count) and the entity submitter id (TCGA-EW-A1J6-01A-11R-A13Q-07). Ensembl gene ids were converted to HGNC symbols using the biomaRt package [167].

Subtypes of TCGA-BRCA files were pulled from PanCancerAtlas [130] using the TCGAbiolinks package [173]. Molecular subtypes of TCGA-BRCA RNA-seq data were predicted with the PAM50 method by The Cancer Genome Atlas Network [63] using original centroids from Parker et al. [66]. The number of files dropped from 1164 to 1149 since subtypes of some patients were not supplied.

Table 9. Used filters in downloading of TCGA-BRCA files from NIH GDC Data Portal.

Filter Category	Selected filter category
Gender	Female
Disease type	Ductal and Lobular Neoplasm
Primary site	Breast
Program name	TCGA
Project id	TCGA-BRCA
Workflow type	HTSeq-Counts
Data category	Transcriptome profiling
Data type	Gene expression quantification
Experimental strategy	RNA-seq

3.5. Batch Effect Control

TCGA-BRCA and ABCC data were logCPM normalized with the edgeR package [168], separately, prior to batch effect control. Next, TCGA-BRCA and ABCC data were merged, and phenotype data were prepared that indicate samples and batches (Batch 1: TCGA-BRCA; Batch 2: ABCC). These two batches were detected and were variance adjusted with the *ComBat* function from the sva package [174]. The ComBat adjusted data were used to generate a heatmap with TCGA-BRCA and ABCC data together [174].

3.6. Shiny App

classABCC is an R-based Shiny app I developed to perform classification of ABCC patients and predict subtypes in an user facilitative manner for exploration of ABCC RNA-seq data. The methods and data types used in the app were described in the detail below:

3.6.1. Breast Cancer Subtype Differentiation (Heatmap)

Heatmap of logCPM or logTPM values of 17-gene [75], EndoPredict [87], Mammaprint [74, p. 70], OncotypeDx [84], PAM50 [66], YMR16 Signature (16-gene signature) [80] across

ABCC RNA-seq data was constructed with the ComplexHeatmap package [176]. ABCC data heatmap has ER, PR, Her2, KI67 percentage, age status. Heatmap of logCPM values of the same signatures across ABCC and TCGA-BRCA together were constructed with the same package indicating PAM50 subtypes and the data source as column annotations.

3.6.2. Dimensional Reduction (PCA)

Principal Component Analysis (PCA), an unsupervised method used in dimension reduction, of 17-gene [75], EndoPredict [87], Mammaprint [74, p. 70], OncotypeDx [84], PAM50 [66], YMR16 Signature (16-gene signature) [80] across ABCC with/without TCGA-BRCA RNA-seq data was constructed with the DESeq2 package [169, p. 2]. Raw count reads were used in the DESeq2 package [169, p. 2] pipeline with variance stabilizing transformation (because the rlog method slows down the process). Moreover, the batch effect was removed with *removeBatchEffect* function from the limma package [178]. Moreover, the "intgroup" argument was specified to group by patients, PAM50, SCMOD1, and IHC subtypes.

3.6.3. Machine Learning (ML)

TCGA-BRCA and ABCC RNA-seq data were subset for only PAM50 genes before the model training and test data prediction. TCGA data containing 1149 samples were first shuffled with the *sample* function from R Base, and then the data were divided as a train (80%) and test data (20%) with the *createDataPartition* function from the caret package [177]. Train data contained 921 samples, and test data contained 228 samples. Train data were preprocessed with the *preProcess* function using the scale and center method from the caret package [177].

Before feeding the model, the control of the train model was set using the *trainControl* function from the caret package [177] with repeated cross-validation as the method using ten iterations, and ten times cross-validation. Random forest (rf), k-nearest neighbors (knn), naive Bayes (nb), regression logistic (regLogistic), and linear model of support vector machine (svmLinear) algorithms were used to train the data. Random forest was selected for further processing since its higher accuracy and the kappa value was obtained with random forest. The model was tested with the test data, and a confusion matrix was constructed to test the model. Undersampling and oversampling was done by specifying the *sampling* argument of the *trainControl* function in the caret package as "up" or "down", respectively [177]. Moreover, xgboost algorithm was tested with the undersampled TCGA-BRCA model before deciding on the random forest model.

Trained TCGA-BRCA RNA-seq data with the random forest model was saved as a rda file. The model was tested with ABCC RNA-seq as well as public data. I also used GSE167977 public datasets for comparative purposes. The workflow is presented in **Figure 6**. The performance of the model was assessed with an accuracy value in the model selection. However, precision, recall, and F1 scores were also considered for the performance of the model with test data since they represented more reliable metrics for imbalanced data.

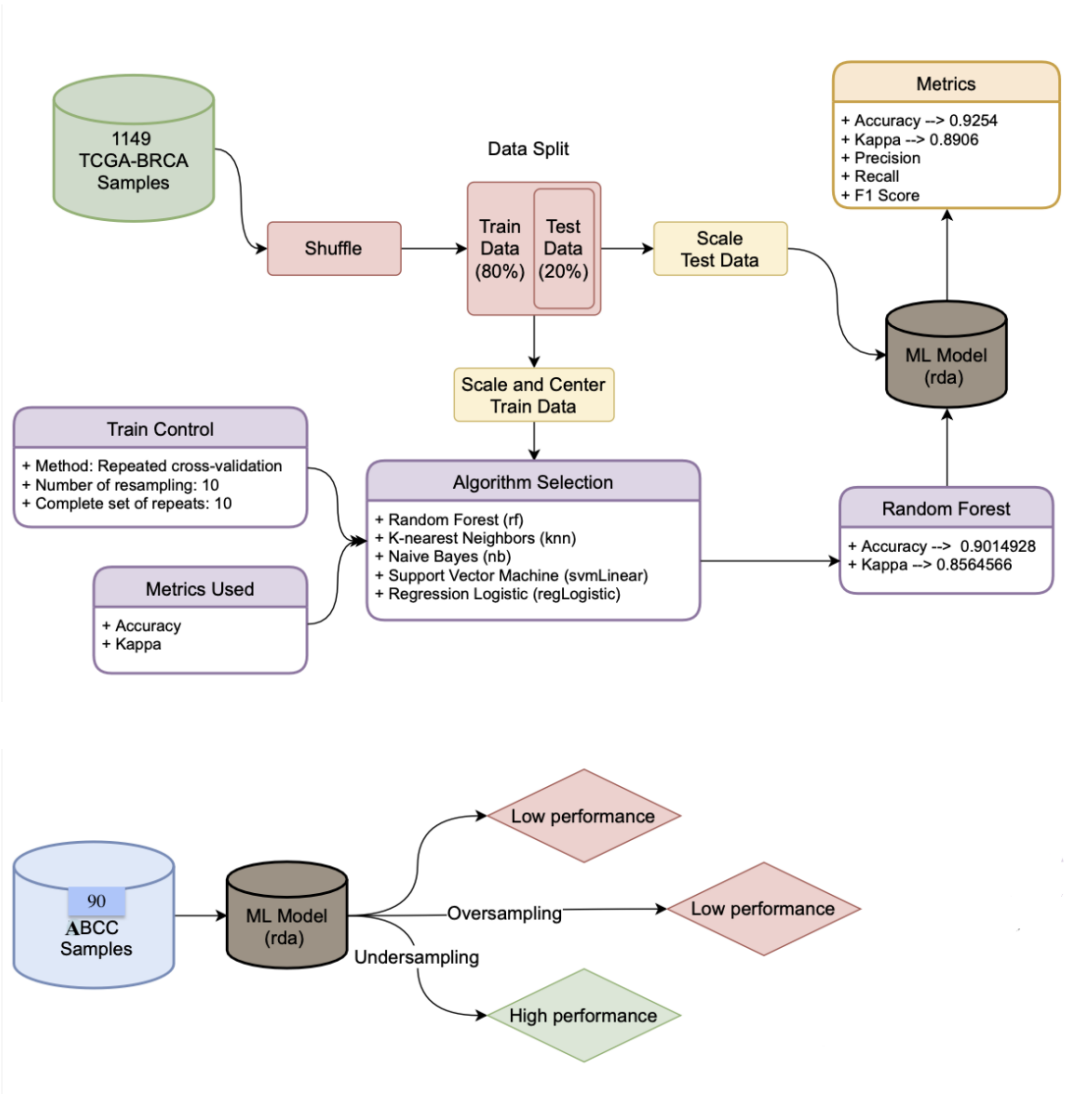


Figure 6. Machine Learning pipeline.

3.7. GSE167977 (MATADOR) Public Data Preparation

GSE167977 [180] is an RNA-seq data consisting of 528 breast cancer patient samples. The count data was directly downloaded from GEO and Ensembl gene ids were converted to gene symbols using the biomart package [167]. Molecular subtypes of GSE167977 were predicted with the genefu package using PAM50 method [170]. The GSE167977 dataset was further used as test data to test with the machine learning model developed for classABCC to test the machine learning model in comparison with the ABCC dataset.

3.8. Concordance Analysis using MAINZ, TRANSBIG, UPP, UNT, NKI

Public datasets were collected to test the concordance between PAM50 and SCMOD1 prediction methods in comparison with the ABCC dataset. Mainz, Transbig, UPP, and UNT datasets with GEO ids of GSE11121 [153], GSE7390 [143], GSE3494 [151], and GSE2990 [156], respectively, were used. Moreover, NKI dataset [181] is available in the supplementary of the publication. Each dataset was obtained using the breastCancerMAINZ [190], breastCancerTRANSBIG [191], breastCancerUPP [192], breastCancerUNT [193], breastCancerNKI [194] packages, respectively. Duplicate patients were removed before performing molecular subtyping with PAM50 and SCMOD1 using the genefu package [170]. The concordance analysis conducted by the caret package [177] using the *confusionMatrix* function gave a measurement of agreement between PAM50 and SCMOD1 methods showing the confusion matrix of PAM50 versus SCMOD1.

3.9. PI3K-AKT Signaling Pathway Genes as User Defined Signature

PI3K-AKT Signaling pathway genes were obtained from the KEGG Database [231]. The gene list containing 354 genes was saved as a .csv file to use in the classABCC to cluster ABCC data. The format of .csv file for input gene list to cluster ABCC has been shown in **Figure 25A**.

4. Results

4.1. Subtyping methods IHC, PAM50 and SCMOD1 do not always produce concordant results with each other

Molecular subtyping of breast cancer was obtained from clinical data with commonly used IHC markers **(Table 8)**, and from RNA-seq data with PAM50 and SCMOD1. The number of patients belonging to each molecular subtype and the percentages obtained from each subtype exhibited differences among the scoring systems. In addition, the normal-like subtype was assessed with only the PAM50 method **(Table 10)**.

Table 10. The number of subtypes in ABCC using PAM50, SCMOD1, and IHC.

	Basal	**Her2**	**LumA**	**LumB**	**Normal-like**
PAM50	9 **(10%)**	7 **(7.78%)**	42 **(46.7)**	21 **(23.3%)**	11 **(12.2%)**
SCMOD1	7 **(7.78%)**	16 **(17.8%)**	41 **(45.6%)**	26 **(28.9%)**	NA
IHC	8 **(8.89%)**	4 **(4.44%)**	38 **(42.2%)**	41 **(44.4%)**	NA

Subtyping with IHC, PAM50, and SCMOD1 was concordant among some patients such that 48 out of 90 patients were the same subtype for all three methods. However, there were disconcordant results for some patients: subtypes of 36 patients were the same only for two methods and 6 patients' subtypes did not match at all in any method **(Appendix Table 2)**. Molecular subtypes of patients with ids of RSM521517 and RSM521661 were found to be discordant between IHC and PAM50 or SCMOD1. According to the IHC subtyping, the patients were classified as Luminal B. On the other hand, PAM50 and SCMOD1 classified these patients as basal with high probabilities (PAM50: 0.95, 0.93; 1, 1: SCMOD1 respectively) **(Table 11)**. RSM521517 identifies as ER and PR-positive, and HER2-negative whereas RSM521661 was identified as ER-positive, PR-negative, and HER2 negative.

Table 11. Expression values of ERBB2, ESR1, and PGR along with PAM50 and SCMOD1 probabilities obtained by genefu package and subtypes for selected patients *(exp: expression in logCPM; P(): Probability calculated by PAM50 or SCMOD1 method; The probability of other subtypes showed in parentheses with italic)*.

Patient ids	exp ERBB2	exp ESR1	exp PGR	IHC Subtype	P(Basal) PAM50	Pam50 Subtype	P(Basal) SCMOD1	SCMOD1 Subtype
RSM487411	9.81	3.2	3.47	Her2	0.54 *(0.46 HER2)*	Basal	0	Her2 *(0.99)*
RSM487446	6.33	-0.83	-2.55	Basal	0.94 *(0.06 HER2)*	Basal	1	Basal
RSM521505	-5.38	-5.38	-5.38	Basal	0.62 *(0.17 HER2, 0.21 Lum B)*	Basal	1	Basal
RSM521517	**6.15**	**6.18**	**-3.77**	**LumB**	**0.95** ***(0.05 HER2)***	**Basal**	**1**	**Basal**
RSM521557	6.63	2.15	1.18	Basal	0.87 *(0.13 Normal)*	Basal	0.01	Her2 *(0.98)*
RSM521573	6.52	4.41	-1.6	Basal	0.76 *(0.17 HER2, 0.07 Normal)*	Basal	1	Basal
RSM521579	7.38	3.02	2.27	Basal	1	Basal	1	Basal

RSM521631	6.74	2.83	1.58	Basal	0.64 *(0.36 Normal)*	Basal	1	Basal
RSM521661	**6.63**	**3.39**	**-1.84**	**LumB**	**0.93** ***(0.07 HER2)***	**Basal**	**1**	**Basal**

The dimension reduction method PCA was applied to the data with PAM50 genes through which PC1 and PC2 explained more than 50% of the variance. The discordance for the two patients could also be observed in the PCA as well **(Figure 7)** where they were clustered in the basal group when using the PAM50 genes' expression profiles but they were grouped as Luminal B based on IHC subtyping.

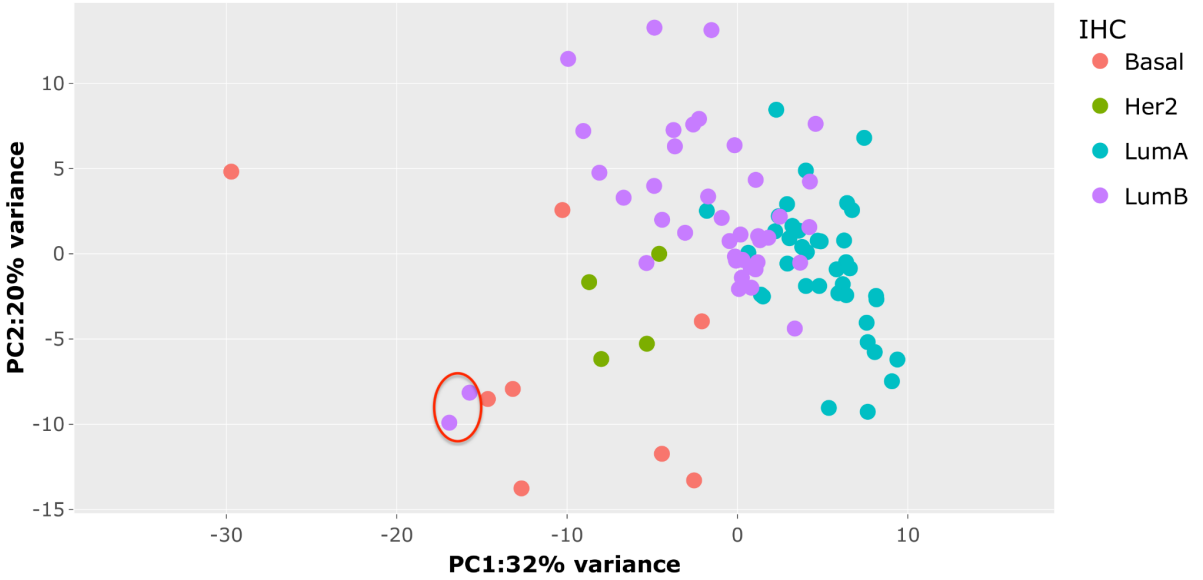


Figure 7. PC1 (32%) and PC2 (20%) scores from PCA of DESeq2 normalized PAM50 genes and subtypes were colored according to calls by IHC method *(the red circle indicates the two discordant patients)*.

4.2. PAM50 and SCMOD1 subtyping predictors are 67% and 82% concordant across ABCC and the global populations, respectively

PAM50 and SCMOD1 subtyping provided the same subtype for 60 patients but not for the remaining 30 patients **(Appendix Table 2)**. The differences could be also observed in the PCA plot **(Figure 8)** where only PAM50 genes were used to show the difference. The subtyping discrepancy between PAM50 and SCMOD1 showed itself especially in the *HER2-enriched* subtype. However, there is no predisposition of subtyping the HER2-positive as any other subtype.

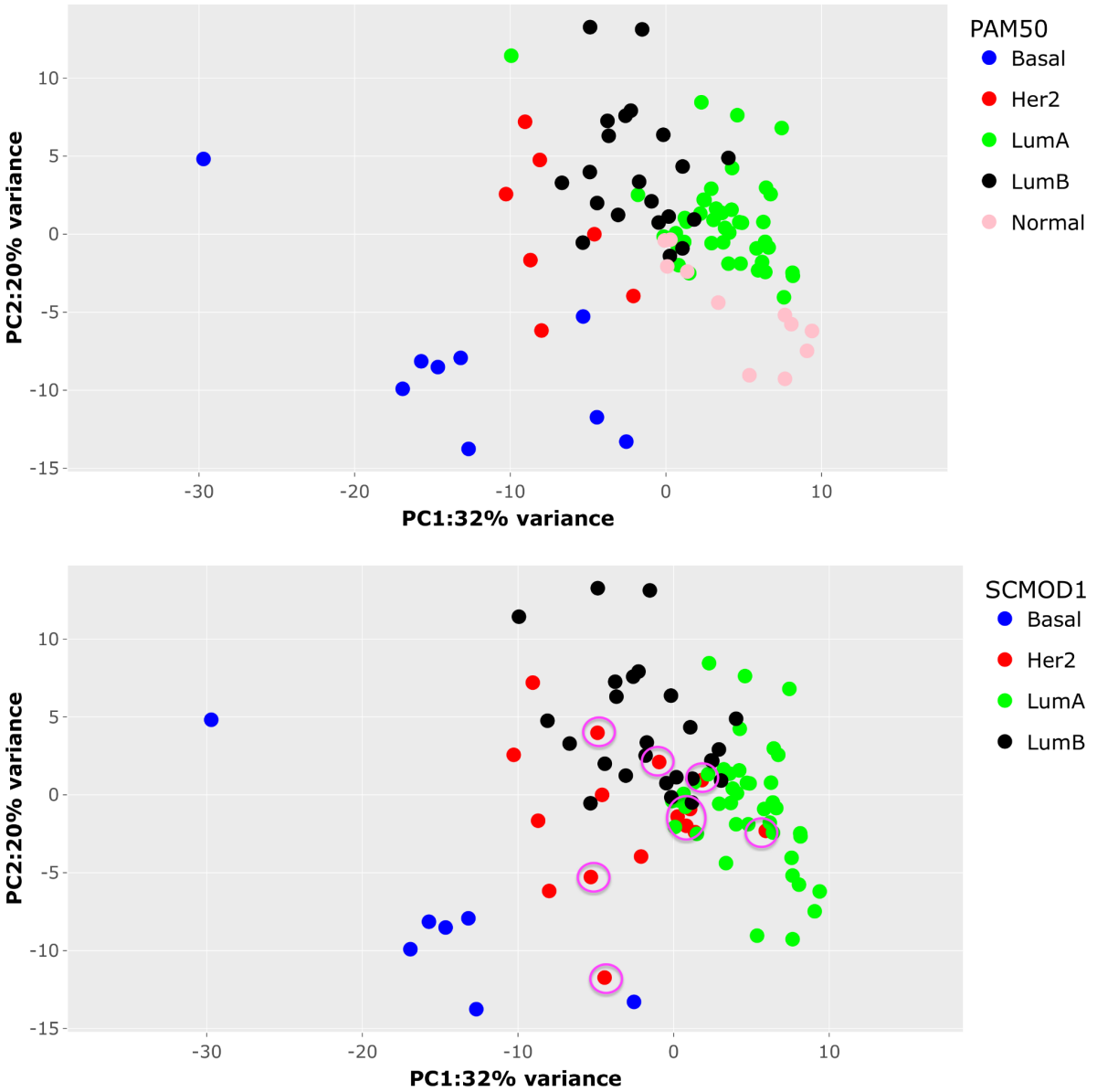


Figure 8. PC1 (32%) and PC2 (20%) scores from PCA of DESeq2 normalized PAM50 genes colored by PAM50 *(top)* and by SCMOD1 *(bottom)* method predictions in the ABCC cohort *(Pink circles indicate patients which were identified as HER2-positive with SCMOD1 method but but different subtype with PAM50 method)*.

Both PAM50 and SCMOD1 methods calculated a probability for each patient. For example, a patient with an id of RSM487431 was labeled Luminal A with probability of 0.8 but there can be other subtypes contributing with the remaining probability of 0.2 by PAM50. The range of probability values for the PAM50 method was 0.4-1.0; on the other hand, the range for SCMOD1 method was 0.52-1.0 **(Appendix Figure 1)**. SCMOD1 performed subtype prediction with higher probabilities compared to PAM50 **(Figure 9)**.

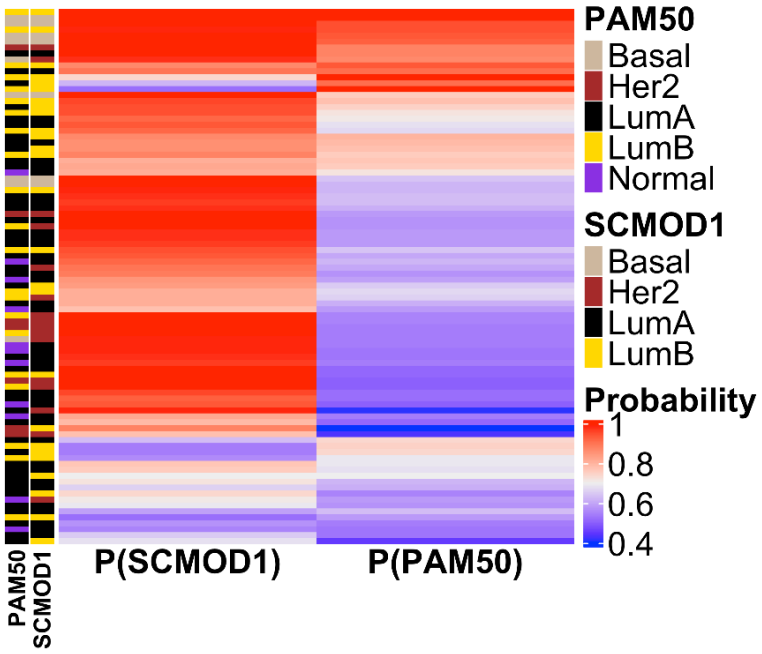


Figure 9. Heatmap showing probabilities of subtype predictions by SCMOD1 and PAM50 *(Rows show the patients and left annotation shows SCMOD1 and PAM50 subtypes with color code and legends are shown on the right side of the heatmap)*.

For some patients, the discordance between PAM50 and SCMOD1 was more obvious. **Table 12** shows selected patients which had different subtypes predicted by PAM50 and SCMOD1. In addition, for these patients the prediction probabilities were not as high towards a single subtype especially in PAM50. Part of the disconcordancy observed could arise from the fact that SCMOD1 inherently did not provided "normal-like" subytpe while PAM50 did. The disconcordancy rate was recalculated by removing samples labeled "normal-like" by PAM50 and then we found that the subtypes of 60 patients out of 79 were classified into the same subtype with PAM50 and SCMOD1 methods. Therefore, the concordance percentage was increased from 67% to 76% after normal-like subtype removal.

Table 12. Subtype probabilities of selected patients by PAM50 *(top)* and SCMOD1 *(bottom)*.

Patient ids	**P(Basal) PAM50**	**P(Her2) PAM50**	**P(LumA) PAM50**	**P(LumB) PAM50**	**P(Normal -like) PAM50**	**PAM50 Subtype**
RSM487372	0	0	0.41	0	0.59	Normal-like
RSM487373	0	0	0.74	0.26	0	LumA
RSM487387	0	0.34	0	0.66	0	LumB

Patient ids	**P(Basal) SCMOD1**	**P(Her2) SCMOD1**	**P(LumA) SCMOD1**	**P(LumB) SCMOD1**	**P(Normal -like) SCMOD1**	**SCMOD1 Subtype**
RSM487372	0	0.13	0.85	0.01	NA	LumA
RSM487373	0	0.1	0.36	0.54	NA	LumB
RSM487387	0	0.81	0	0.19	NA	Her2

To further determine how PAM50 and SCMOD1 were concordant as subtype predictors, Mainz, Transbig, UPP, and UNT expression (microarray) datasets with GEO ids of GSE11121 [153], GSE7390 [143], GSE3494 [151], and GSE2990 [156], respectively, were

used. Also, NKI dataset [181] was obtained from the supplementary data of the publication (**Table 13**).

Table 13. Overview of datasets used in the concordance analysis of PAM50 and SCMOD1.

Dataset	**Patient [#]**	**ER+ [#]**	**HER2+ [#]**	**Age**	**Grade [1/2/3]**	**Platform**
MAINZ	200	155	23	25/90	29/136/35	HGU133A
TRANSBIG	198	123	35	24-60	30/83/83	HGU133A
UPP	251	175	46	28-93	67/128/54	HGU133AB
UNT	137	94	21	24-73	32/51/29	HGU133AB
NKI	337	212	53	26-62	79/109/149	Agilent

Each dataset was classified using PAM50 as well as SCMOD1 subtyping algorithms separately after duplicate patients removal. The performance of both molecular predictors was compared with the confusion matrix without the normal-like subtype patients since SCMOD1 does not predict this subtype (**Table 14**). The results revealed that the concordance of PAM50 and SCMOD1 predictors was around 82 percent (**Table 15A**). The predictors showed the best concordance in Luminal A and basal subtypes. However, the concordance was worse in Luminal B and HER-positive subtypes where SCMOD1 was prone to identify HER2-positive subtypes (**Table 15B**). Kappa coefficient is a metric that indicates the degree of agreement between two evaluators, known also as inter-rater reliability. Kappa score of 0.74 corresponds to a substantial agreement [234]. In addition, balanced accuracy is another score to evaluate the concordance between PAM50 and SCMOD1 which basically corresponds to (Sensitivity+Specificity)/2.

Table 14. Confusion matrix of SCMOD1 versus PAM50 subtype predictions.

	SCMOD1			
PAM50	Basal	Her2	LumA	LumB
Basal	148	2	1	3
Her2	0	86	1	31
LumA	5	4	334	36
LumB	8	24	52	236
Normal-like	0	0	0	0

Table 15. Subtype predictions of PAM50 versus SCMOD1 predictors. A) Overall statistics; B) Statistics by class.

A

Accuracy	0.8196
95% CI	(0.7941, 0.8432)
No Information Rate	0.4057
P-Value [Acc > NIR]	<2e-16
Kappa	0.7429

B

	Basal	**Her2**	**LumA**	**LumB**
Sensitivity	0.9193	0.74138	0.8392	0.7712
Specificity	0.9927	0.94566	0.9314	0.8756
Balanced Accuracy	0.9560	0.84352	0.8853	0.8234

4.3. Molecular subtype assignment with transcriptomics data strongly depends on the gene signature selection for ABCC dataset.

Heatmaps of logTPM values of genes of the six different signatures using ABCC expression data were constructed. Each cluster was then annotated with clinical data available with regards to receptor status, i.e., ER, PR, HER2; proliferation marker KI67, and the age. The number of column clusters was set as two to differentiate luminal subtypes from non-luminal subtypes; the distance method for clustering was set as Euclidean; the agglomeration method was set as complete.

Heatmaps of 17-gene and PAM50 signatures showed the best separation of luminal and non-luminal subtypes. The two main clusters appeared to discriminate the ER and/or PR-positive (Luminal) patients from the ER, PR, and HER2-negative subtypes (Basal), although some patients were clustered indiscriminately, especially more with PAM50 (**Figure 10A, B**).

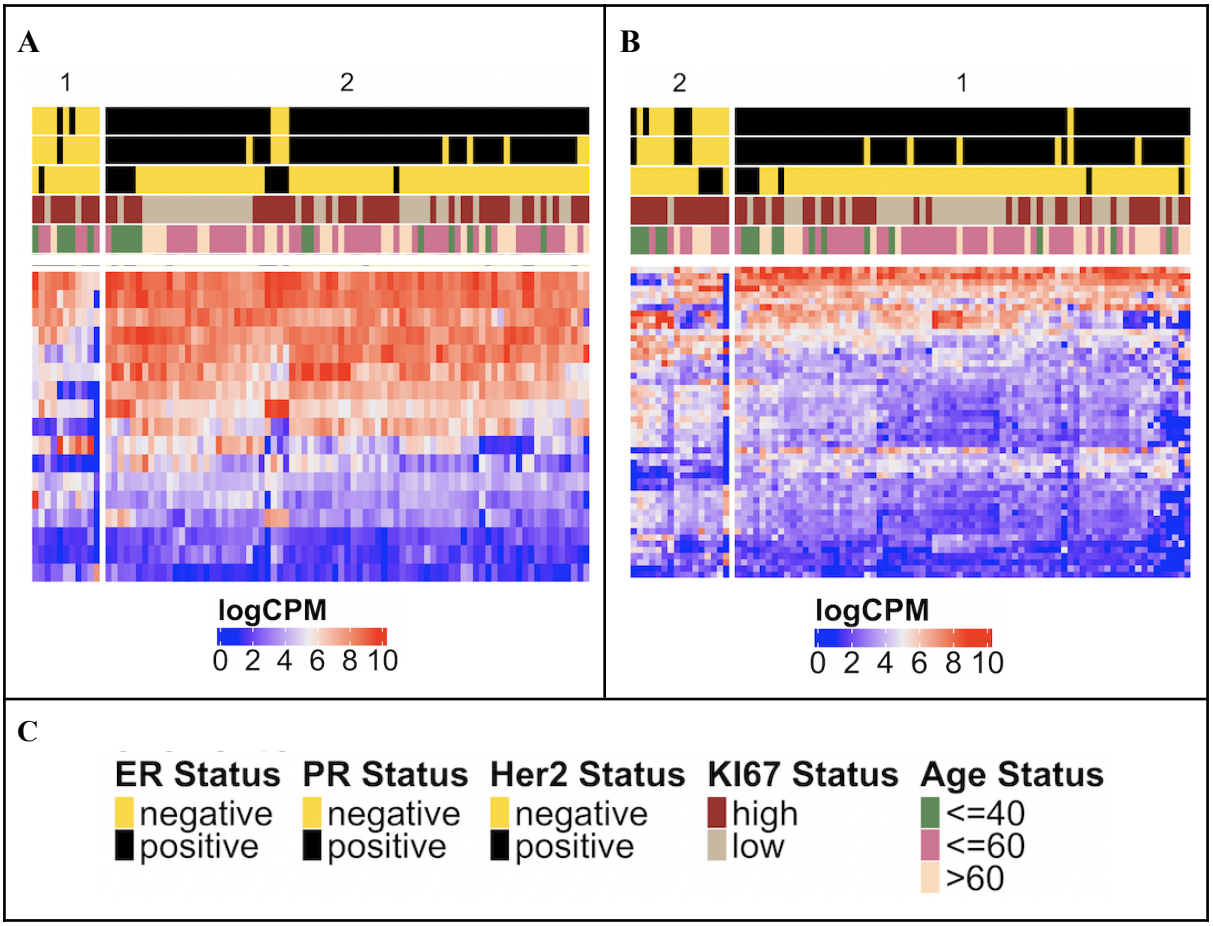


Figure 10. Heatmap of logCPM values of A) 17-gene signature; B) PAM50 signature across ABCC data; C) annotated with ER, PR, HER2, KI67, Age *(KI67 Status; KI67< 20%: low proliferation, >20%: high proliferative).*

Heatmaps of EndoPredict and Mammaprint signatures showed moderate differentiation of luminal subtypes from non-luminal subtypes. However, both of these two signatures had a high power of clustering according to KI67 status as high and low (**Figure 11A,B**).

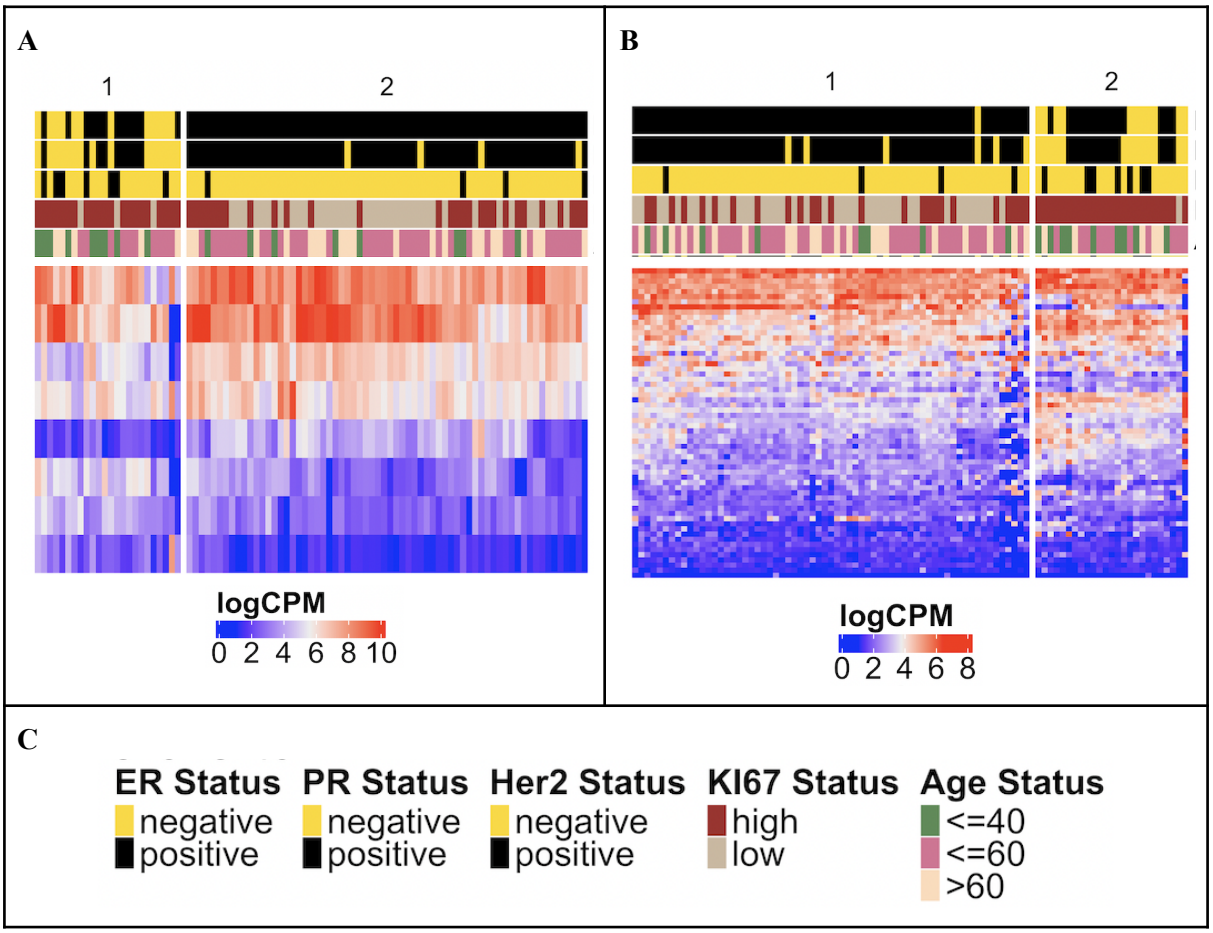


Figure 11. Heatmap of logCPM values of A) EndoPredict; B) Mammaprint signature across ABCC data; C) annotated with ER, PR, HER2, KI67, Age *(KI67 Status; KI67< 20%: low proliferation, >20%: high proliferative)*.

However, heatmaps of OncotypeDx and 16-gene signatures did not show apparent clustering of the clinical and molecular biomarkers plotted and could not differentiate the luminal from non-luminal subtypes. Therefore, KI67 clustering observed in EndoPredict and Mammaprint were not observed with these signatures (**Figure 12A,B**). In conclusion, each signature

exhibited a different pattern to cluster ABCC RNAseq data along with the clinical and mutation data.

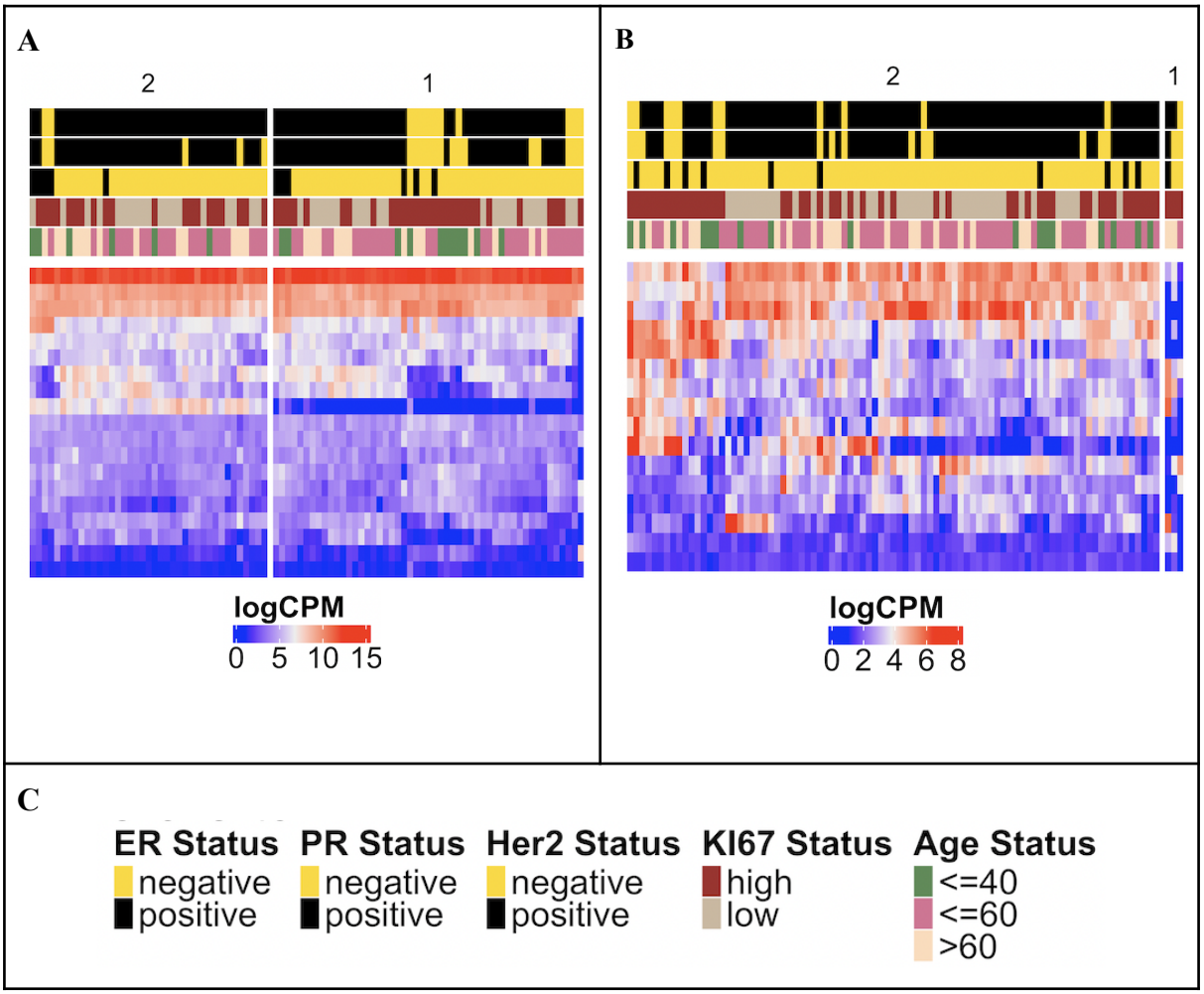


Figure 12. Heatmap of logCPM values of A) OncotypeDx; B) 16-gene signature across ABCC data; C) annotated with ER, PR, HER2, KI67, Age *(KI67 Status; KI67< 20%: low proliferation, >20%: high proliferative)*.

4.4. Subtype structure of ABCC is comparable to that of TCGA-BRCA cohort

ABCC RNAseq data were clustered together with TCGA-BRCA data to investigate the differences between these two cohorts and test the power of signatures in differentiating breast cancer subtypes also in each of the datasets as well as in the combined dataset. For this purpose, the heatmaps and PCA results were plotted. ABCC data were scattered within the TCGA-BRCA data in both the heatmap and PCA plots with all gene signatures, implying the similarity of the two data regardless of gene signature used, after the batch effect removal (**Figure 13-16; Appendix Figure 2**). However, heatmap and especially PCA showed that the different gene signatures have different power to differentiate breast cancer molecular subtypes implying the results obtained in the previous section were not specific to ABCC data.

The 17-gene signature exhibited the best differentiation of luminal groups from basal, normal-like, and HER2-positive subtypes, which were differentiated along the PC1 axis (**Figure 13**).

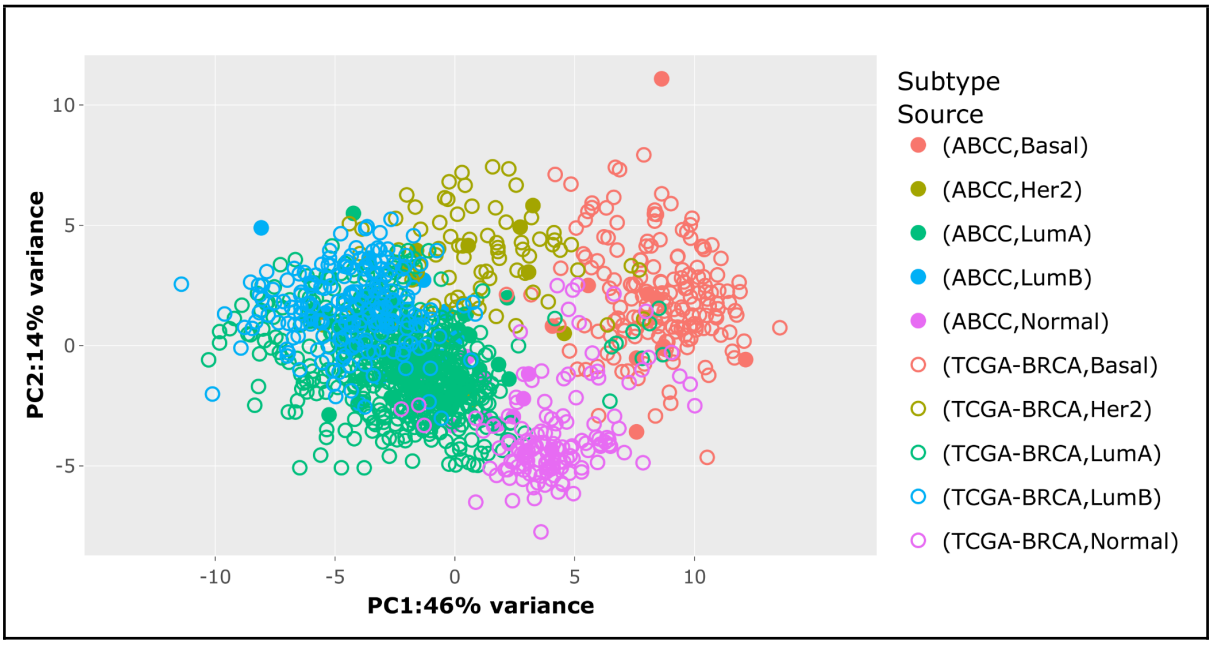


Figure 13. PC1 (46%) and PC2 (14%) scores from PCA of batch effect removed values of the 17-gene signature.

Mammaprint and OncotypeDx signatures had moderate power in differentiating PAM50 subtypes. Both signatures clustered HER2-positive and basal subtypes together but apart from luminal type and clustered the normal-like patients distinctly in the PC2 axis (**Figure 14A, B**).

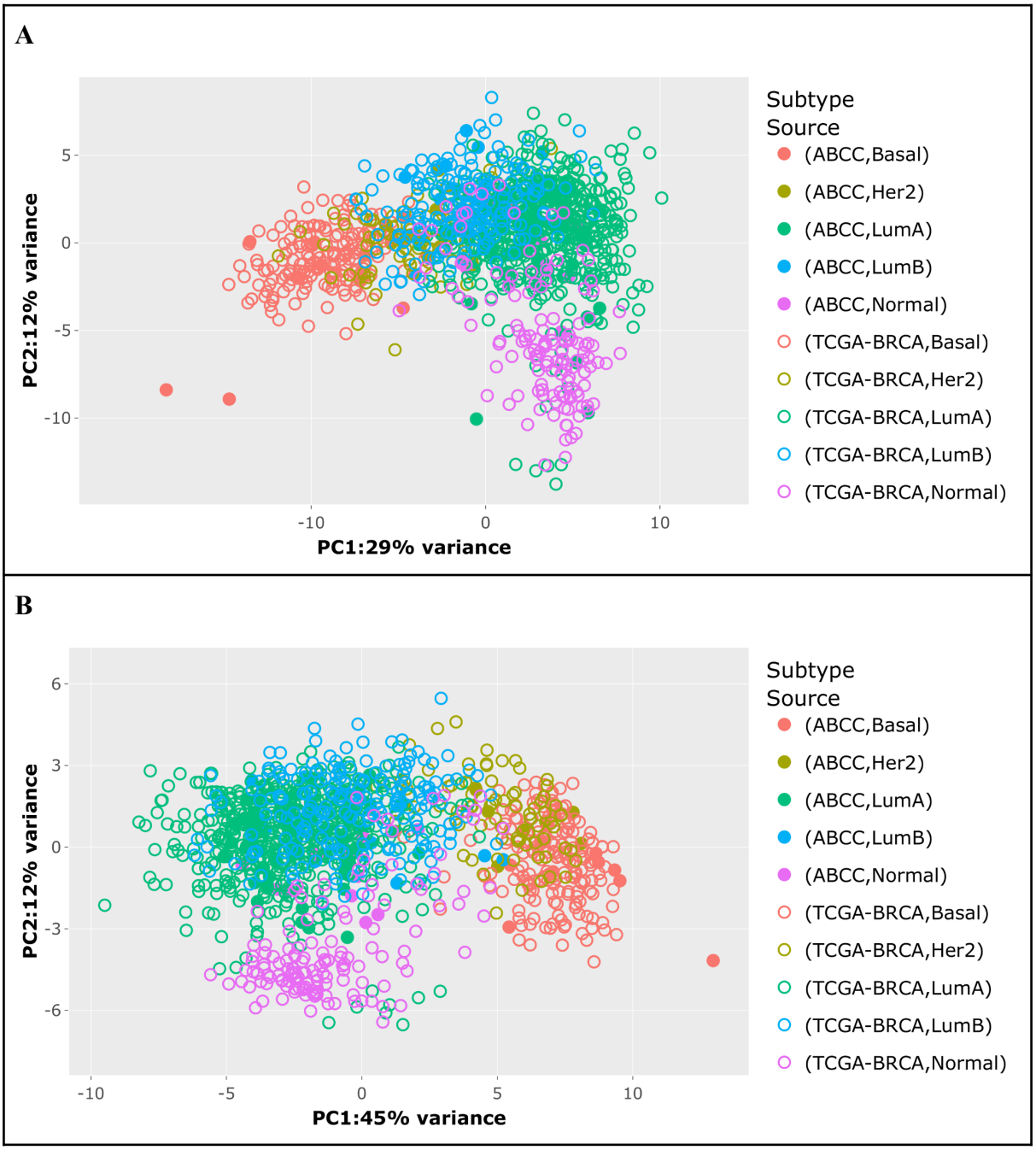


Figure 14. PCA of batch effect removed values of A) PC1 (29%) and PC2 (12%) scores from Mammaprint; and B) PC1 (45%) and PC2 (12%) scores from OncotypeDx.

On the other hand, PAM50 led to reasonable differentiation between the Luminal A, Luminal B, and normal-like subtypes. However, the HER2-positive subtype cluster was distinctly apart from basal yet partially mixed with luminal B subtype on the PC2 axis (**Figure 15**).

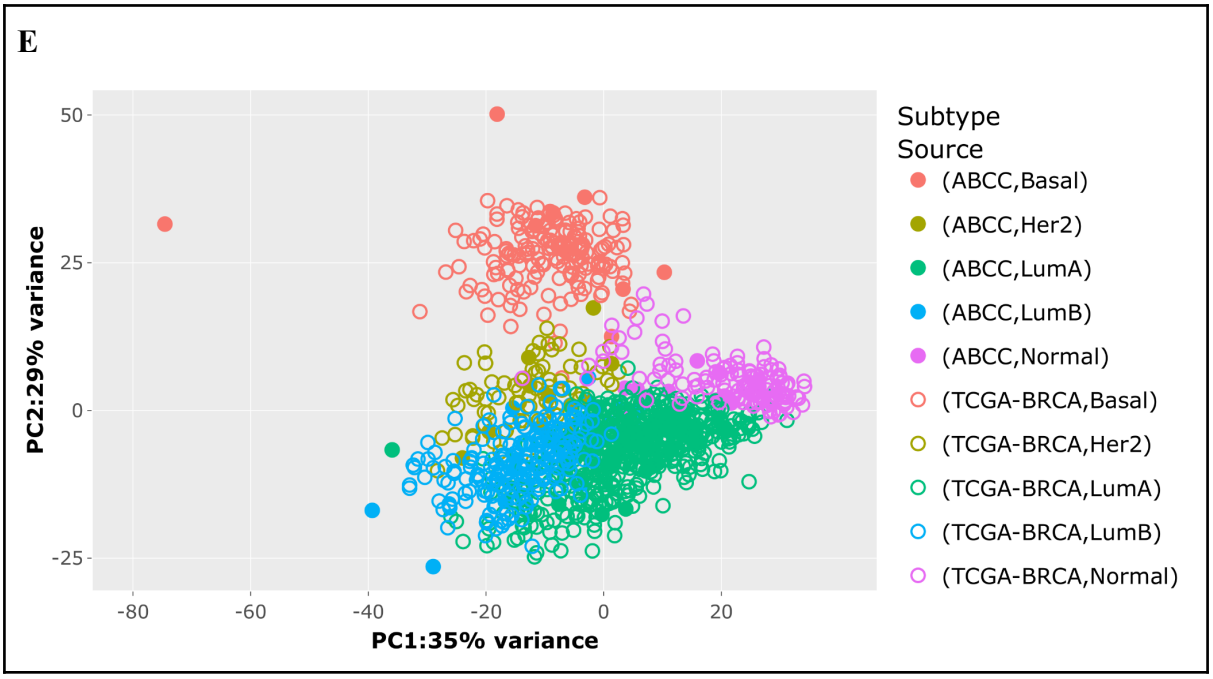


Figure 15. PC1 (35%) and PC2 (29%) scores from PCA of batch effect removed values of PAM50.

Finally, EndoPredict and 16-gene signatures showed the worst strength in differential clustering of the subtypes, which were grouped together in the PC1 axis when using 16-gene signatures (**Figure 16A, B**).

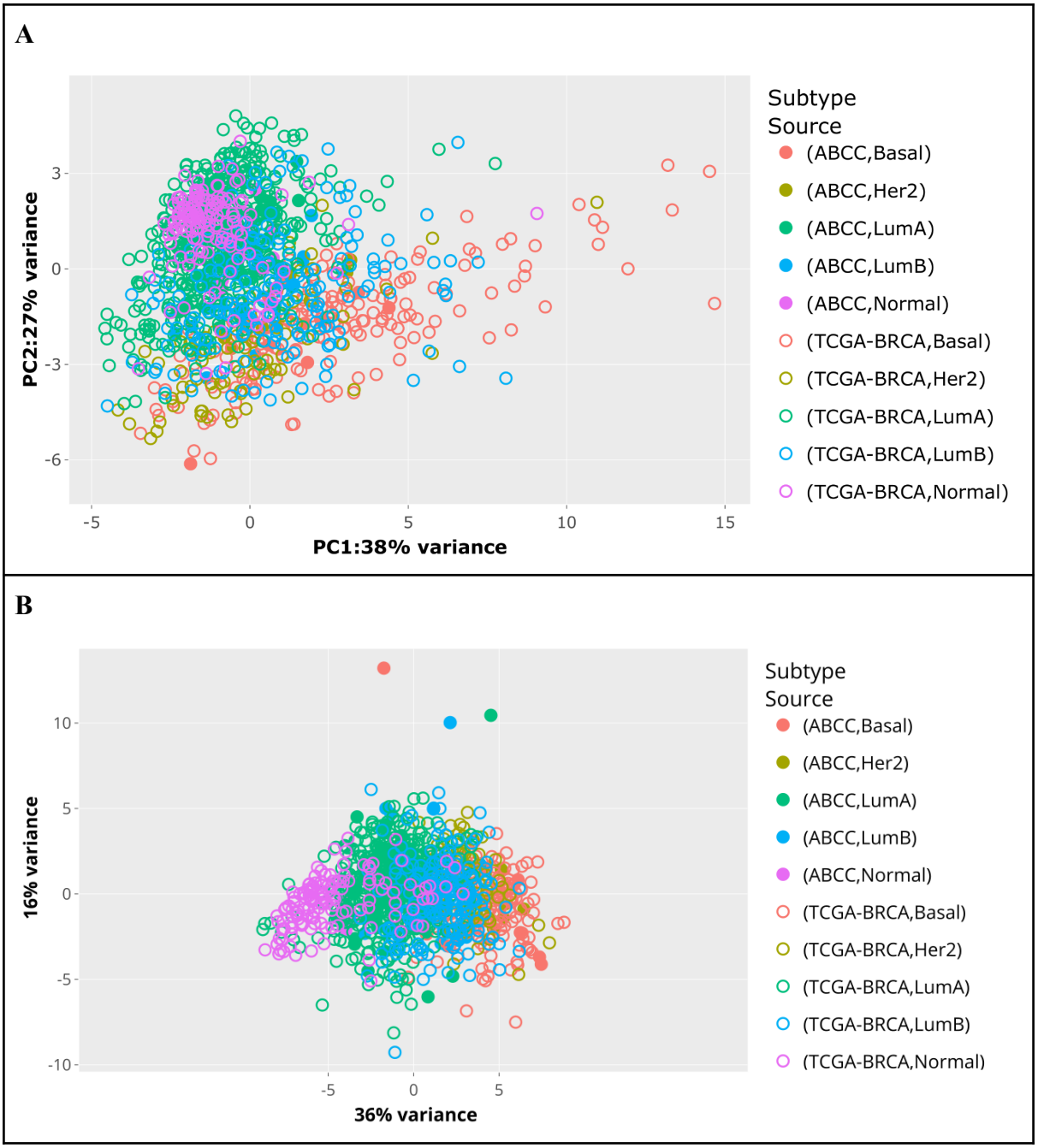


Figure 16. PCA of batch effect removed values of A) PC1 (38%) and PC2 (27%) scores from EndoPredict; and B) PC1 (36%) and PC2 (16%) scores from 16-gene signature .

4.5. PAM50 signature based TCGA-BRCA machine learning model can moderately predict subtypes of ABCC and GSE167977 (MATADOR).

PAM50 genes of TCGA-BRCA RNA-seq data were used as the train data while PAM50 genes of ABCC RNA-seq data were used as the test data. In this process, TCGA-BRCA RNA-seq and ABCC RNA-seq datasets were not normalized with ComBat meaning that the test data was completely independent of the train data. The labels which corresponded to five classes of breast cancer subtypes were obtained from PanCancerAtlas for TCGA-BRCA dataset while the PAM50 method present in genefu package was used for the ABCC dataset. The number of subtypes found in each dataset including the train, validation and test data were presented in **Table 16**.

Table 16. The number of each subtype in TCGA-BRCA (train and test) and ABCC datasets.

	Basal	**Her2**	**LumA**	**LumB**	**Normal-like**	**Total**
TCGA-BRCA	175	80	553	203	138	1149
Train Data	140	64	443	163	111	921
Validation Data	35	16	110	40	27	228
ABCC (Test Data)	9	7	42	21	11	90

After train and validation data split, the class distribution (percentage) of each subtype in train, validation and test data were also checked (**Table 17**). The subtype distribution across train and validation data did not show a significant difference according to Pearson's Chi-squared test (p-value= 0.2202).

Table 17. The class distribution (percentage, %) of subtypes in the train, validation and test data.

	Basal	Her2	LumA	LumB	Normal
Percentage (validation data)	15.35	7.01	48.24	17.54	11.84
Percentage (train data)	15.20	6.94	48.09	17.69	12.005
Percentage (test data)	10.0	7.77	46.66	23.33	12.22
Chi-squared test	X-squared= 20, df=16, p-value= 0.2202				

4.5.1. Model selection

Since the prediction of breast cancer subtype is a classification problem of machine learning not a regression problem, random forest, regression logistic, naive bayes, k-nearest neighbors, and support vector machine algorithms were selected to choose the best performing model. According to the accuracy and kappa values, the random forest model showed the best performance (**Figure 17**). Thus, in the downstream analyses I continued with the random forest model.

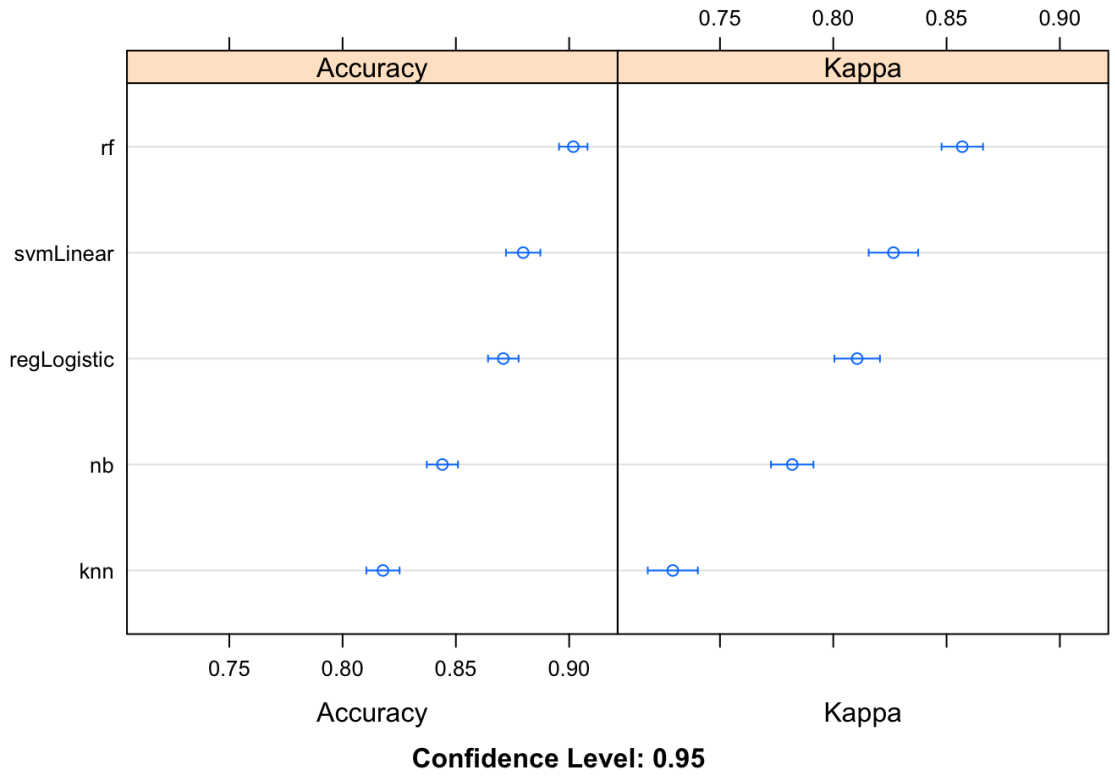


Figure 17. Accuracy and Kappa values of machine learning model trained with rf: random forest; svmLinear: support vector machine; regLogistic: regression logistic; nb: naive bayes; knn: k-nearest neighbors algorithms fed with TCGA-BRCA RNA-seq data.

4.5.2. Testing the model with the TCGA-BRCA validation data

The random forest model fed with train data was tested first with the validation data. The confusion matrix showed the number of reference subtypes versus predicted subtypes for each patient (**Table 18**). The accuracy of this prediction was 0.91 (**Table 19A**). Furthermore, statistics were calculated for prediction of each subtype where F1 scores for subtypes were: basal, 0.99; Her2, 0.87; Luminal A, 0.93; Luminal B, 0.86; and normal-like, 0.84 (**Table 19B**).

Table 18. The confusion matrix that shows the number of predicted and observed subtypes by the model fed with TCGA-BRCA train data and tested with the validation data.

	Observed				
Predicted	Basal	Her2	LumA	LumB	Normal-like
Basal	35	0	0	0	1
Her2	0	13	0	0	1
LumA	0	1	106	7	4
LumB	0	2	2	33	0
Normal-like	0	0	2	0	21

Table 19. Statistics calculated for the model fed with TCGA-BRCA train data and tested with the TCGA-BRCA validation data. A) overall statistics; and B) statistics by class.

A

Accuracy	0.9123
95% CI	(0.8678, 0.9456)
No Information Rate	0.4825
P-Value [Acc > NIR]	<2.2e-16
Kappa	0.8713

B

	Basal	**Her2**	**LumA**	**LumB**	**Normal-like**
Sensitivity	1	0.81250	0.9636	0.8250	0.77778
Specificity	0.9948	0.99528	0.8983	0.9787	0.99005
Precision	0.9722	0.92857	0.8983	0.8919	0.91304
Recall	1	0.81250	0.9636	0.8250	0.77778
F1	0.9859	0.86667	0.9298	0.8571	0.84000
Balanced Accuracy	0.9974	0.90389	0.9310	0.9019	0.88391

4.5.3. Feature importance by ranking

PAM50 genes were the features/predictors of the model. Although there would not be a feature selection for this case, a feature importance plot was constructed which showed that more than half of the PAM50 genes did not influence the prediction of breast cancer subtypes (**Figure 18**). This result was supported with the plot of model accuracy versus the number randomly selected predictors showing that 26 features gave the maximum accuracy where the higher number of features made the model more complex and decreased the prediction accuracy (**Figure 19; Appendix Table 3**). The 26 genes were enriched in positive regulation of exit from mitosis, i.e., Gene Ontology process (GO:0031536), by strength and cellular component organization, i.e., Gene Ontology process (GO:0016043), by count in the network using STRING analysis (**Figure 20**). Moreover, the feature plots (box plot) were constructed for each PAM50 gene and subtypes were set as factors showing relation of each gene and subtype (**Appendix Figure 3-5**).

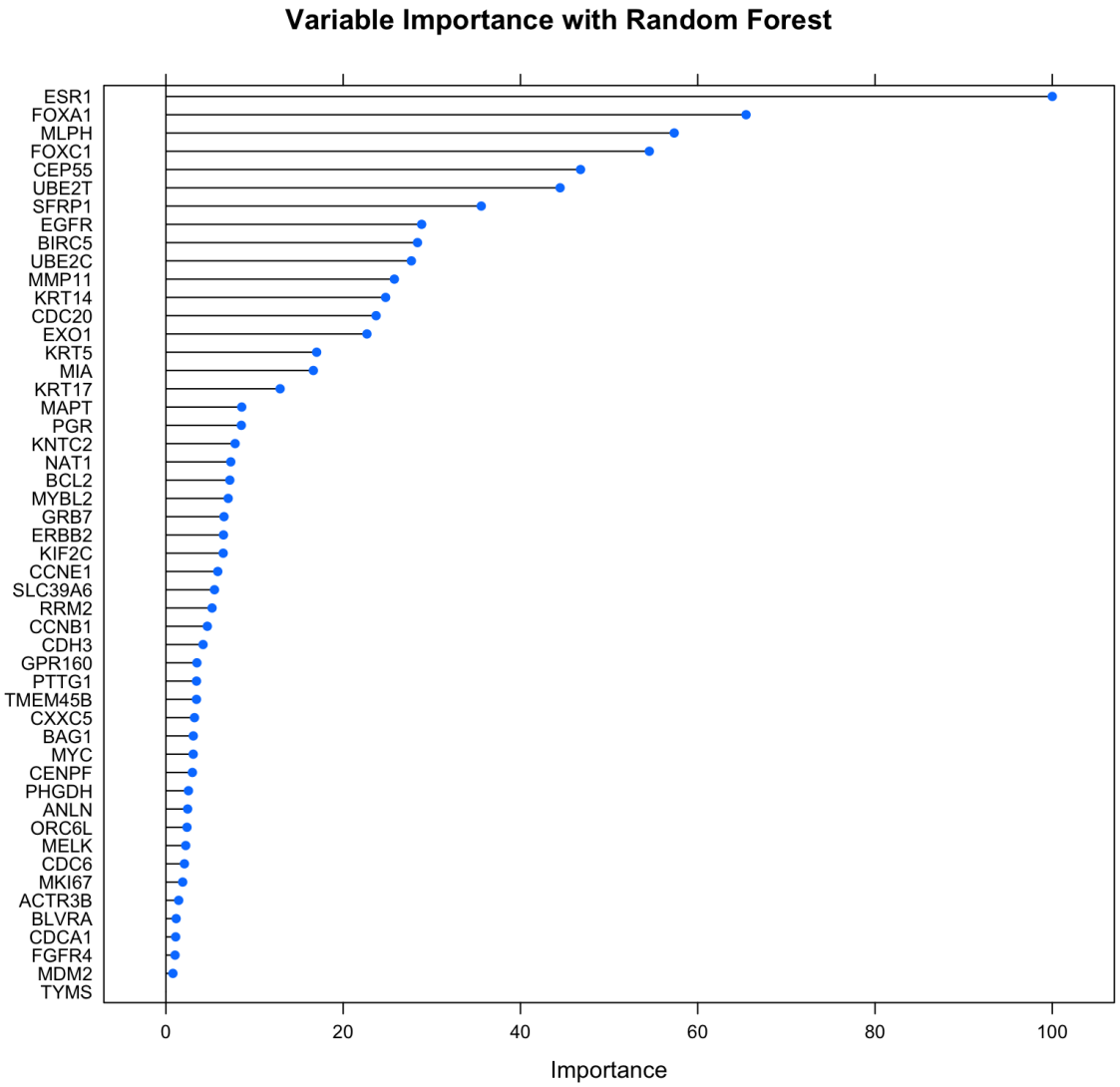


Figure 18. Feature importance plot of PAM50 genes in random forest model fed with TCGA-BRCA train data *(x axis showing importance out of 100% and y axis showing PAM50 genes)*.

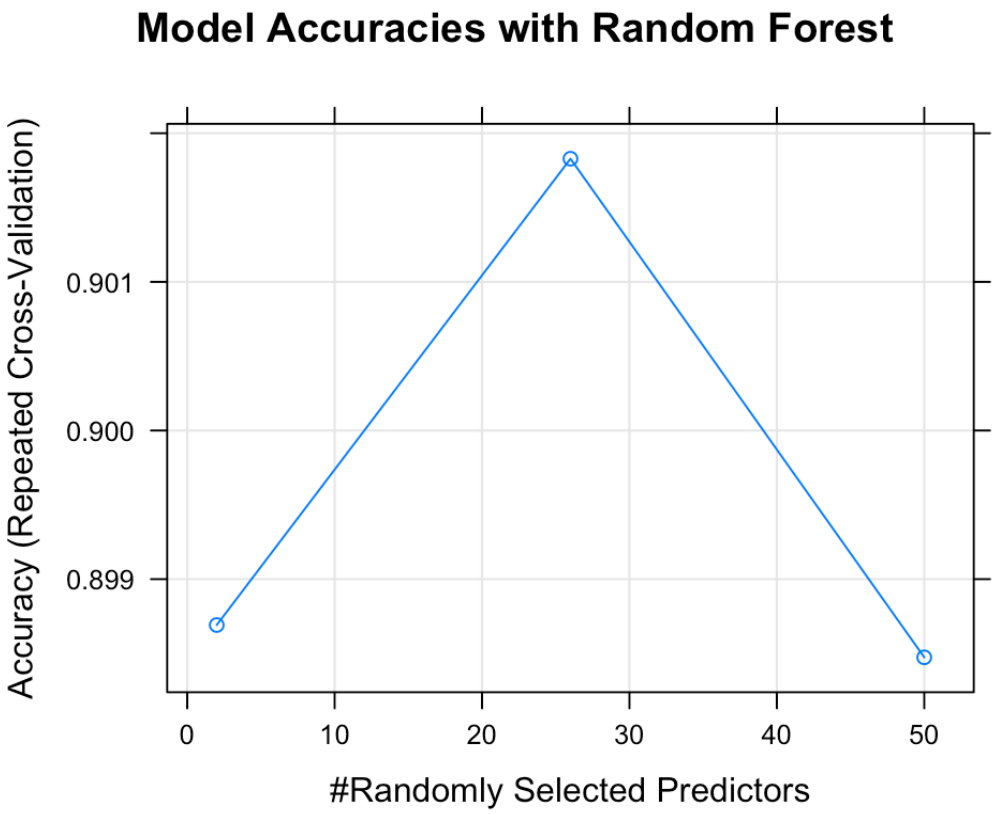
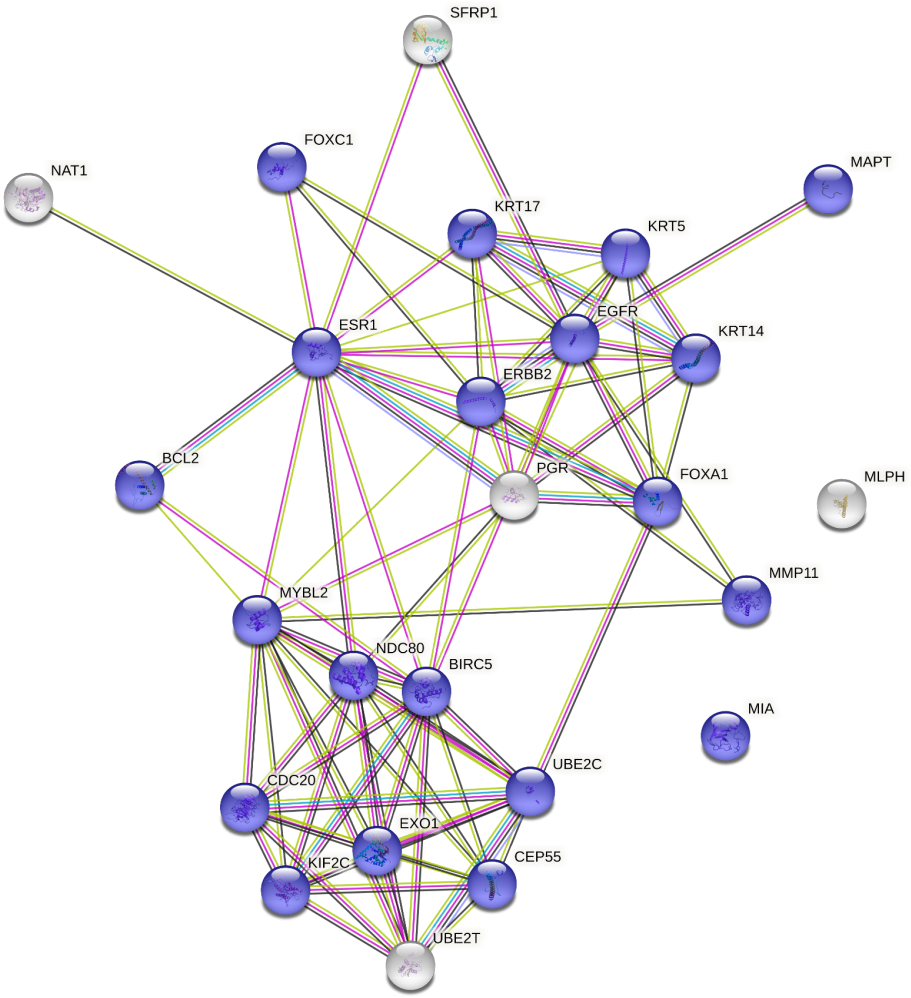


Figure 19. Model accuracy versus the number of randomly selected predictors/features *(PAM50 genes in this case)* of the Random Forest model fed with train data.

A


SFRP1
FOXC1
NAT1
MAPT
KRT17
KRT5
EGFR
ESR1
KRT14
ERBB2
BCL2
PGR
FOXA1
MLPH
MMP11
MYBL2
NDC80
BIRC5
MIA
UBE2C
CDC20
EXO1
KIF2C
CEP55
UBE2T

B

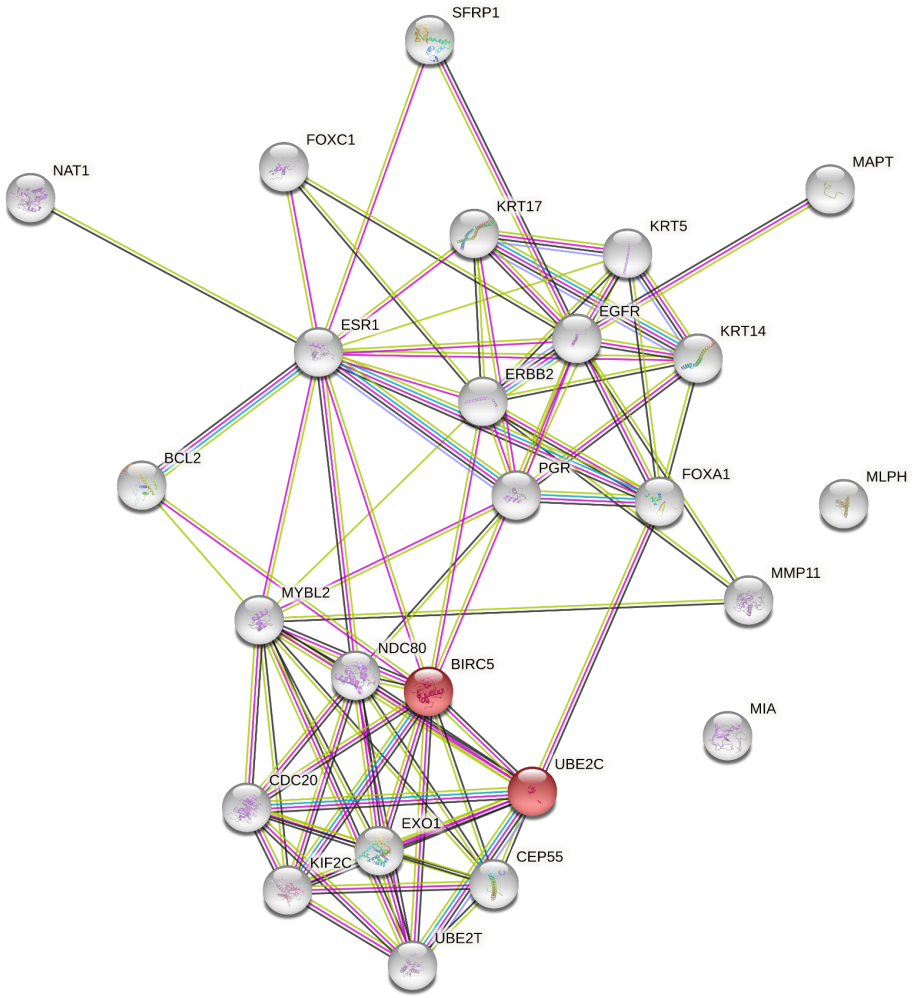


Figure 20. STRING network analysis of 26 genes selected with feature importance analysis. A) The genes from cellular composition organization Gene Ontology process labeled with blue; and B) The genes from positive regulation of exit from mitosis pathway Gene Ontology process were labeled.

4.5.4. Testing the model with the ABCC test data reveals moderate predictive ability, which is improved by undersampling

The model was tested with the validation data which were split from TCGA-BRCA RNA-seq data before training (20%). Since the model showed good performance using the validation data, it allowed testing of the model with ABCC RNA-seq data, representing a totally independent test data. ABCC data as the test data were labeled by the genefu package and each patient was assigned a subtype using the PAM50 method. Therefore, this enabled the comparison of subtypes predicted with the TCGA-BRCA model and PAM50 method by the genefu method, originally designed for microarray datasets. The observed (PAM50 subtypes) and predicted (model predicted subtypes) number of subtypes by the model showed poor performance (**Table 20**) where the model's accuracy was 0.68 (**Table 21**A). It was apparent that the model was an overfit for the Luminal A subtype that predicted Luminal B and Normal-like subtypes as Luminal A since the data were imbalanced towards the Luminal A subtype (**Table 21**B). To increase the performance of the model, oversampling and undersampling were performed comparatively since these were the suggested approaches for imbalanced data [186], [187].

Table 20. The confusion matrix showing the number of predicted and observed subtypes by the model obtained with the TCGA-BRCA train data and tested with the ABCC test data.

	Observed				
Prediction	Basal	Her2	LumA	LumB	Normal-like
Basal	7	1	0	0	0
Her2	2	6	0	0	0
LumA	0	0	42	16	10
LumB	0	0	0	5	0
Normal-like	0	0	0	0	1

Table 21. Statistics calculated for the model fed with TCGA-BRCA train data and tested with the ABCC test data. A) overall statistics; and B) statistics by class.

A

Accuracy	0.6778
95% CI	(0.571, 0.7725)
No Information Rate	0.4667
P-Value [Acc > NIR]	4.283e-05
Kappa	0.478

B

	Basal	**Her2**	**LumA**	**LumB**	**Normal-like**
Sensitivity	0.77778	0.85714	1	0.23810	0.09091
Specificity	0.98765	0.97590	0.4583	1	1
Precision	0.87500	0.75000	0.6176	1	1
Recall	0.77778	0.85714	1	0.23810	0.09091
F1	0.82353	0.80000	0.7636	0.38462	0.16667
Balanced Accuracy	0.88272	0.91652	0.7292	0.61905	0.54545

To improve the performance of the model, oversampling was applied to the model. Oversampling involves random sampling from the minority classes and adding them to the train data to increase the prediction performance of the minority classes [186], [187]. In ABCC data, the minority classes were basal, HER2-positive and normal-like subtypes since they were less in number when compared to Luminal A and B. However, oversampling showed its power only on the basal subtype where all the basal groups were predicted correctly.

On the other hand, oversampling did not affect the prediction performance of the model on the other minority classes (**Table 22**). Moreover, the accuracy of the model decreased to 0.67

(**Table 23A**). The ineffectiveness of the oversampling can be seen also in the F1 scores of HER2-positive and normal-like subtypes (**Table 23B**).

Table 22. The confusion matrix shows the number of predicted and observed subtypes by the oversampled model fed with TCGA-BRCA train data and tested with the ABCC test data.

	Observed				
Prediction	Basal	Her2	LumA	LumB	Normal-like
Basal	9	1	0	0	0
Her2	0	3	0	0	0
LumA	0	1	42	16	10
LumB	0	2	0	5	0
Normal-like	0	0	0	0	1

Table 23. Statistics calculated for the oversampled model fed with TCGA-BRCA train data and tested with the ABCC test data. A) overall statistics; and B) statistics by class.

A

Accuracy	0.6667
95% CI	(0.5595, 0.7626)
No Information Rate	0.4667
P-Value [Acc > NIR]	0.0001029
Kappa	0.4527

B

	Basal	Her2	LumA	LumB	Normal-like
Sensitivity	1	0.42857	1	0.23810	0.09091
Specificity	0.9877	1	0.4375	0.97101	1
Precision	0.9000	1	0.6087	0.71429	1
Recall	1	0.42857	1	0.23810	0.09091
F1	0.9474	0.6000	0.7568	0.35714	0.16667
Balanced Accuracy	0.9938	0.71429	0.7188	0.60455	0.54545

Since oversampling did not improve performance of the model, undersampling was applied to the model. Undersampling balances the imbalanced data by decreasing the size of the majority class and by keeping the size of minority classes [186], [187]. In ABCC data, Luminal A was the majority class, thus the size of Luminal A subtype was decreased. Undersampling improved the performance by increasing the accuracy of the model up to 0.78, representing a considerable increase (**Table 25A**). The prediction performance of the model improved greatly especially for the basal, HER2-positive and Luminal B groups in which F1 scores became 0.82, 0.80, 072 respectively (**Table 25B**).

An improvement in the prediction of the normal-like subtype was also an issue where 10 patients out of 11 were predicted as Luminal A suggesting overfitting. However, undersampling led to prediction of four patients as the normal-like subtype (**Table 24**). Since the number of this subtype was low to begin with, this drastically affected the general accuracy of the model.

Table 24. The confusion matrix shows the number of predicted and observed subtypes by the undersampled model fed with TCGA-BRCA train data and tested with the ABCC test data.

	Observed				
Prediction	Basal	Her2	LumA	LumB	Normal-like
Basal	7	1	0	0	0
Her2	1	6	1	0	0
LumA	0	0	41	9	7
LumB	0	0	0	12	0
Normal-like	1	0	0	0	4

Table 25. Statistics calculated for the undersampled model fed with TCGA-BRCA train data and tested with the ABCC test data . A) overall statistics; and B) statistics by class.

A

Accuracy	0.7778
95% CI	(0.6779, 0.8587)
No Information Rate	0.4667
P-Value [Acc > NIR]	1.578e-09
Kappa	0.6585

B

	Basal	**Her2**	**LumA**	**LumB**	**Normal-like**
Sensitivity	0.77778	0.85714	0.9762	0.5714	0.36364
Specificity	0.98765	0.97590	0.6667	1	0.98734
Precision	0.87500	0.75000	0.7193	1	0.80000
Recall	0.77778	0.85714	0.9762	0.5714	0.36364
F1	0.82353	0.80000	0.8283	0.7273	0.50000
Balanced Accuracy	0.88272	0.91652	0.8214	0.7857	0.67549

The TCGA-BRCA trained random forest model using only the PAM50 genes showed moderate performance in predicting subtypes of ABCC data. The overfitting was one of the problems in which undersampling improved the metrics for Luminal A, Luminal B and normal subtype (**Figure 21**) and the overall accuracy improved from 0.68 to 0.78.

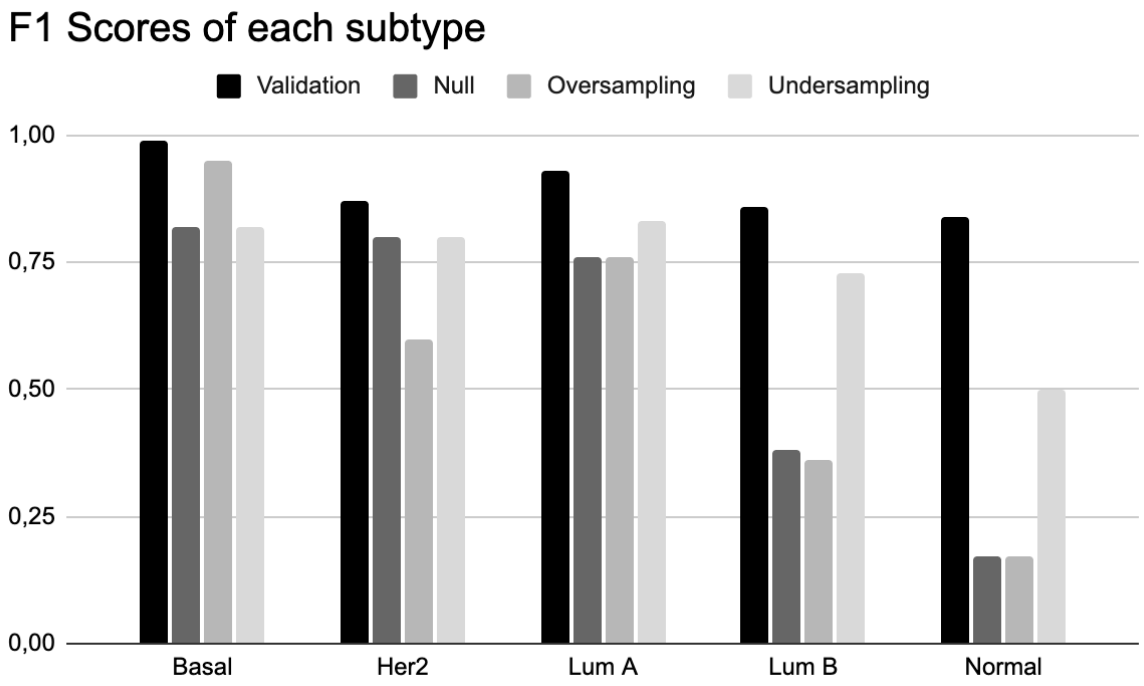


Figure 21. Bar plot showing F1 scores *(2*(Precision*Recall/Precision+Recall))* of each class/subtype by the null, oversampled and undersampled model *(x axis shows the subtypes; y-axis shows the F1 score)*.

4.5.5. Testing the model with MATADOR RNA-seq data (GSE167977) reveals similar prediction power as in the case of ABCC

TCGA-BRCA RNA-seq trained RF model with undersampling was tested with MATADOR public data, which consisted of 528 patient samples, to test whether all PAM50 genes are good predictors. Feature importance analysis showed in ABCC data that all PAM50 genes made the model more complex and decreased the power of the model to predict subtypes. MATADOR data confirmed that the model with all PAM50 genes as predictors had moderate prediction power except for basal and luminal subtypes (**Table 26**). The accuracy of the prediction was 0.67 (**Table 27A**). When the prediction results of MATADOR data were compared with ABCC data, 20% of decrease was observed in each subtype in F1 scores except basal subtype (**Table 26-27**). The drop may reflect the differences in class distribution of basal, HER2, Luminal A, Luminal B, and normal-like: 15, 12, 28, 44, 1% for MATADOR; 9, 7, 42, 21, 11% for ABCC data, respectively. The majority class were Luminal A for ABCC whereas Luminal B for MATADOR. This difference may require undersampling to be done for Luminal B for MATADOR.

Table 26. The confusion matrix that shows the number of predicted and observed subtypes by the undersampled model fed with TCGA-BRCA train data and tested with the MATADOR test data.

	Observed				
Prediction	Basal	Her2	LumA	LumB	Normal-like
Basal	75	1	0	0	0
Her2	0	28	0	3	0
LumA	1	25	149	124	5
LumB	0	8	0	104	0
Normal-like	1	0	2	0	2

Table 27. Statistics calculated for the model fed with TCGA-BRCA train data and tested with the MATADOR test data with undersampling. A) overall statistics; and B) statistics by class.

A

Accuracy	0.678
95% CI	(0.6363, 0.7177)
No Information Rate	0.4375
P-Value [Acc > NIR]	< 2.2e-16
Kappa	0.5494

B

	Basal	**Her2**	**LumA**	**LumB**	**Normal-like**
Sensitivity	0.9740	0.45161	0.9868	0.4502	0.285714
Specificity	0.9978	0.99356	0.5889	0.9731	0.994242
Precision	0.9868	0.90323	0.4901	0.9286	0.400000
Recall	0.9740	0.45161	0.9868	0.4502	0.285714
F1	0.9804	0.60215	0.6549	0.6064	0.333333
Balanced Accuracy	0.9859	0.72259	0.7878	0.7116	0.639978

4.6. classABCC App

classABCC is an R-based Shiny App developed to facilitate the unsupervised and supervised clustering of ABCC and TCGA-BRCA with the given gene signatures of known or user-defined subtyping gene sets (**Figure 22**). Various analyses are presented in different tabs: Clustering with Signature, Clustering with Input, PCA, Molecular Subtyping with ABCC, Molecular subtyping with Input, TCGA and Predict. The details of the tabs are presented in **Table 28**.

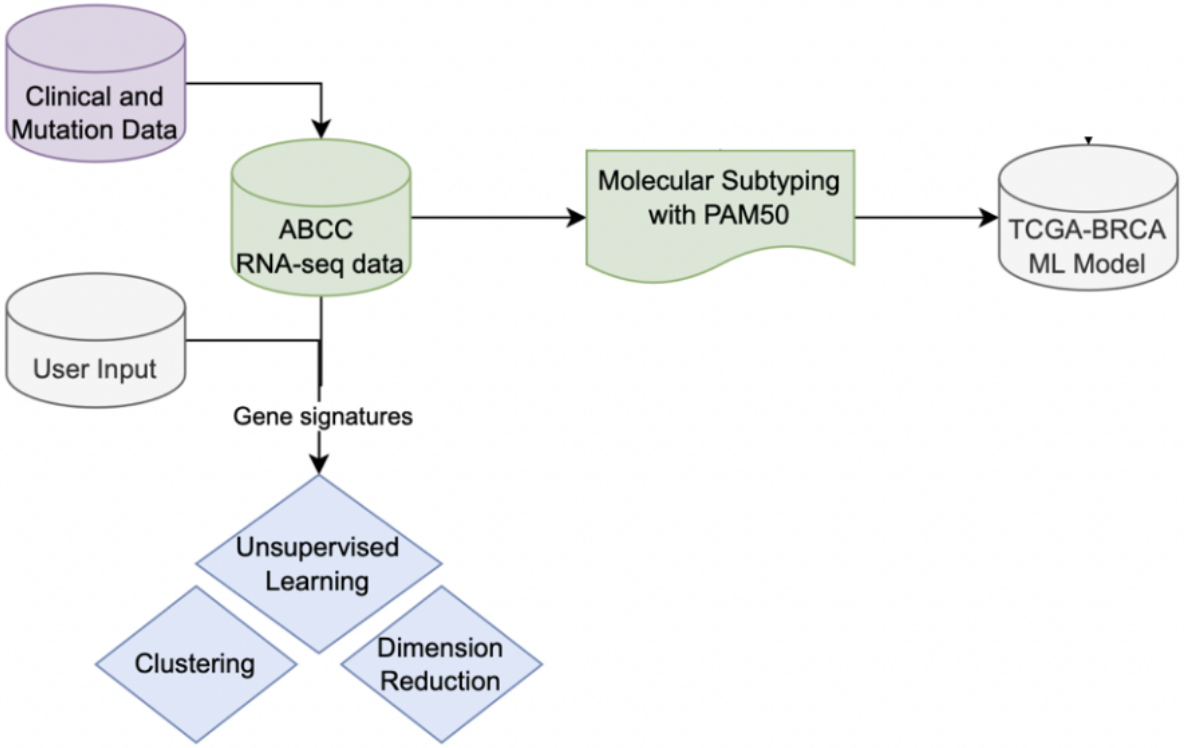


Figure 22. The workflow of the app developed in this thesis *(ABCC: Acibadem Breast Cancer Cohort; TCGA-BRCA: The Cancer Genome Atlas- Breast Cancer; ML: Machine Learning)*.

Table 28. An overview of classABCC presenting the name of the tab and subtabs, the job, data used, options and the type of available visualizations.

Name of the tab	Subtabs	The job	Used data	Available options / Content	Visualization
Clustering with Signature	**1.** Table **2.** Heatmap **3.** Heatmap with Annotation	Unsupervised clustering with heatmap	ABCC	• The selection of gene signature • The selection of normalization type • The selection of distance method • The selection of agglomeration methods • The adjustment of number of clusters • User input of clinical/mutation data annotation	Static
Clustering with Input	**1.** Table **2.** Heatmap **3.** Heatmap with Annotation	Unsupervised clustering with heatmap	ABCC	• The gene signature input by the user • The selection of normalization type • The selection of agglomeration method • The selection of distance method • The adjustment of number of clusters • User input of clinical/mutation data annotation	Static
PCA	-	Dimension reduction	ABCC	• The selection of gene signature • The grouping option with patients, IHC, PAM50, and SCMOD1 results	Interactive
Molecular Subtyping with ABCC	**1.** Clinical Data **2.** PAM50 **3.** SCMOD1 **4.** Compare **5.** Compare with Scatter Plot **6.** Compare with Heatmap	Molecular Subtyping results obtained by IHC, PAM50, and SCMOD1	ABCC	• Clinical data • PAM50 results • SCMOD1 results • Comparison of IHC, PAM50, and SCMOD1 • Heatmap and scatter plot representing the probability of PAM50 and SCMOD1 subtyping	Static Interactive
TCGA	**1.** Heatmap **2.** PCA	Unsupervised clustering with heatmap and PCA	ABCC TCGA	• The selection of gene signature	Static Interactive

Predict	**1.** Predict ABCC	Subtyping prediction with TCGA-BRCA trained model with PAM50 genes	TCGA ABCC	• Subtype prediction of ABCC	Static

4.6.1. Clustering of ABCC with Signature and Input Annotation

Clustering with Signature is the first tab of the classABCC in which ABCC data can be clustered using unsupervised methods. In this tab, a sidebar panel is available to specify clustering features. The heatmap of selected gene signature genes can be constructed with clinical and mutation data as annotations if available (**Figure 23**). The integrated gene signatures include a 17-gene list from Abdelhafiz et al. 2021, EndoPredict, Mammaprint, OncotypeDx, 16-gene list (YMR16), and PAM50 (**Appendix Table 4**; **Figure 24A**). In addition, the normalization type can be selected among logCPM and logTPM (**Figure 24B**). The number of column clusters can be adjusted to visualize subtype specifications (**Figure 24C**). The distance method can be specified from the list of Binary, Canberra, Euclidian, Kendall, Manhattan, Maximum, Minkowski, Pearson, and Spearman (**Figure 24D**). Moreover, the agglomeration method (clustering method) can be chosen from the list containing Ward.D, Ward.D2, Single, Complete, Average (UPGMA), Mcquitty (WPGMA), Median (WPGMC), and Centroid (UPGMC) methods (**Figure 24E**). Finally, an annotation file in a CSV format can be uploaded to add column annotations in clustering of the ABCC data. This annotation file can include clinical and/or mutation data (**Figure 24F**).

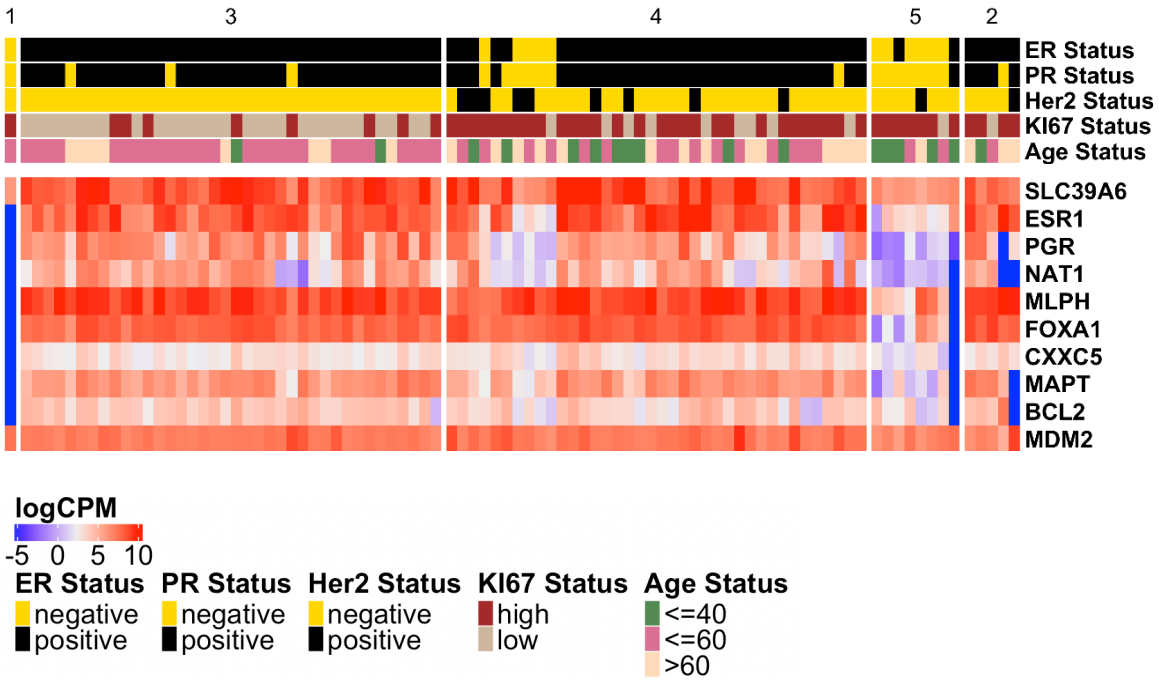


Figure 23. The heatmap of logCPM values of 17-gene signature annotated with clinical data including status of ER, PR, HER2, and KI67, and age *(distance method: euclidean, clustering method: complete, number of column clusters: 5).*

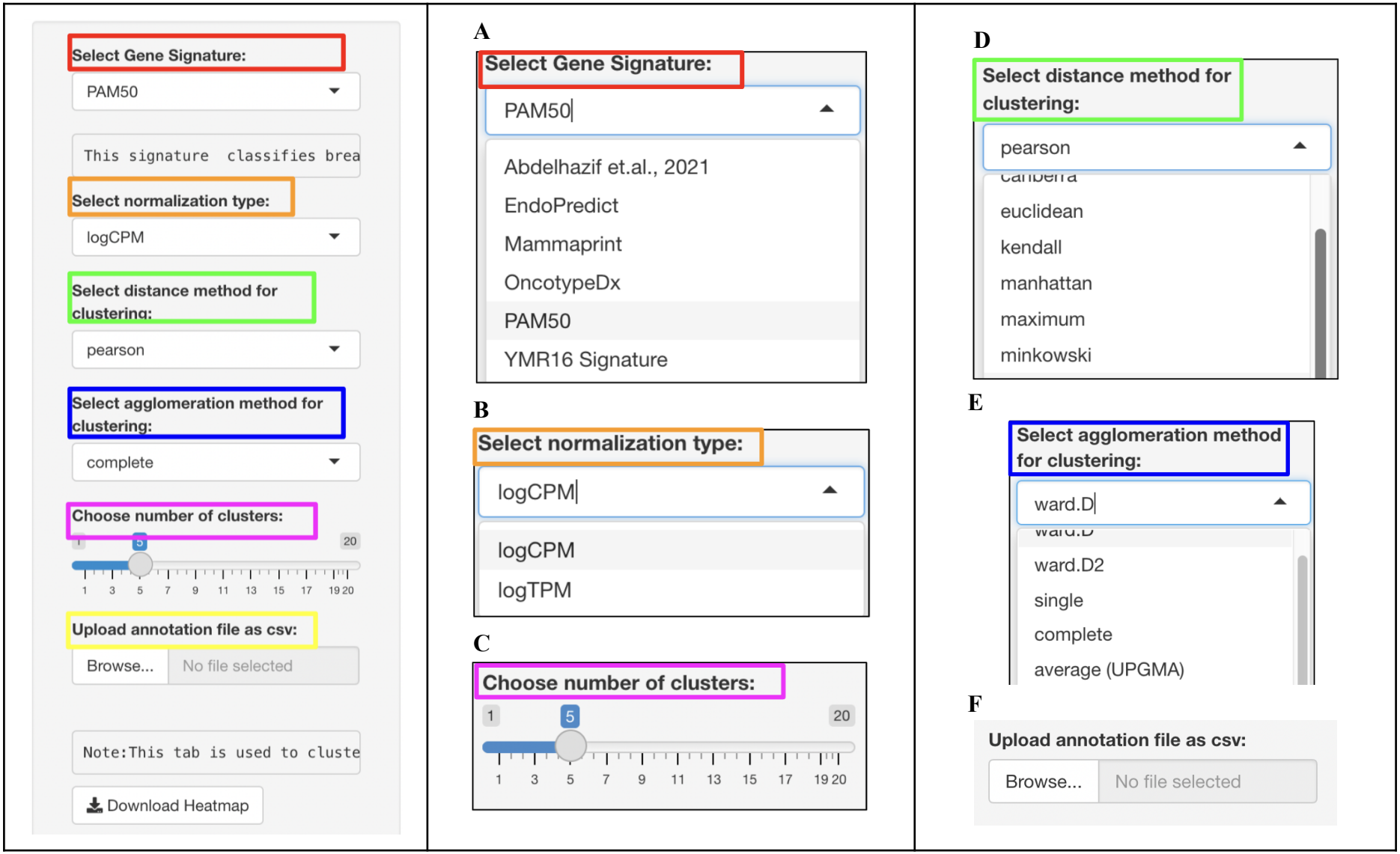


Figure 24. The sidebar panel of the *Clustering with Signature* tab A) Gene signature selection bar; B) Normalization type selection bar; C) Slider of number of column clusters; D) Distance method selection bar; E) Clustering method selection bar; and F) Column annotation input bar.

4.6.2. Clustering of ABCC with Input Signature and Annotation

Clustering with Input is the second tab of the classABCC in which ABCC data is clustered with unsupervised methods given a user input gene list. The normalization type (**Figure 25B**), the number of column clusters (**Figure 25C**), distance method (**Figure 25D**), clustering methods (**Figure 25E**) can be specified and an annotation file as CSV format can be integrated as column annotations (**Figure 25F**) (as previously defined for the Clustering with Signature tab. This tab allows users to cluster ABCC RNA-seq data with user input gene lists which can be a pathway entry (e.g. core WNT signaling genes [216]), gene family (e.g. N-myc downstream-regulated gene (NDRG) family [216]) or other available gene signatures (e.g. Breast Cancer Index, 8-gene signature [215]) and intrinsic gene lists (e.g. 76-gene signature [176]). Further, an annotation file can be added to heatmaps of ABCC as column annotation in the classABCC. This feature allows annotation of clinical data and/or mutation and copy-number variation (cnv) features of patients. Thus, classABCC creates a flexible environment to cluster ABCC data by gene list selection or user gene list input and with annotation options.

The user can input their list of genes as a CSV file and get the heatmap of those respective genes against the samples present in the ABCC data. The input file should have a specific format (**Figure 25A**). In the input file, there should be a column named “gene_ids”. Under the ‘gene_ids’ title, you can list your HGNC gene names. The number of genes as input is not limited however maximum 50 genes are recommended for better visualization of gene names as row names.

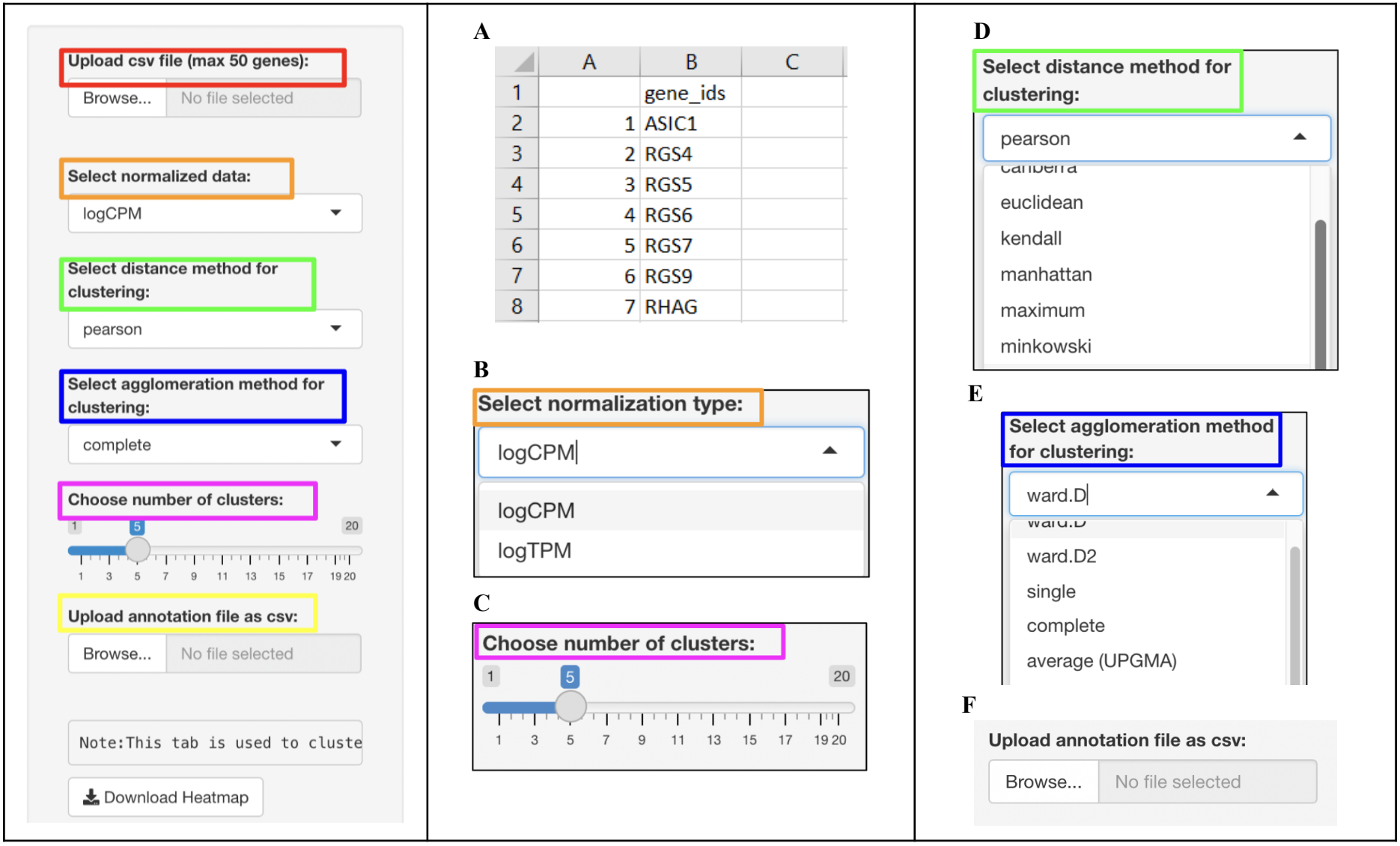


Figure 25. The sidebar panel of the *Clustering with Input* tab A) The format of input gene signature CSV file; B) Normalization type selection bar; C) Slider of number of column clusters; D) Distance method selection bar; E) Clustering method selection bar; F) Column annotation input.

The *Clustering with Input tab* allows the gene signature input to cluster ABCC data which has great importance since different gene lists can be used to classify subtypes of ABCC data. As an example, the PI3K-AKT pathway including 354 genes were used as a gene list (**Figure 26**). The pathway leads to cell growth and tumor proliferation and plays a significant role in endocrine resistance in breast cancer [232] and found association of the genes with breast cancer subtypes regarding mutation status [233]. Therefore, the genes were used to test whether the genes involved in the PI3K-AKT pathway are good signatures of breast cancer subtypes at expression level combined with clinical data. The heatmap showed moderate classification of luminal and nonluminal subtypes.

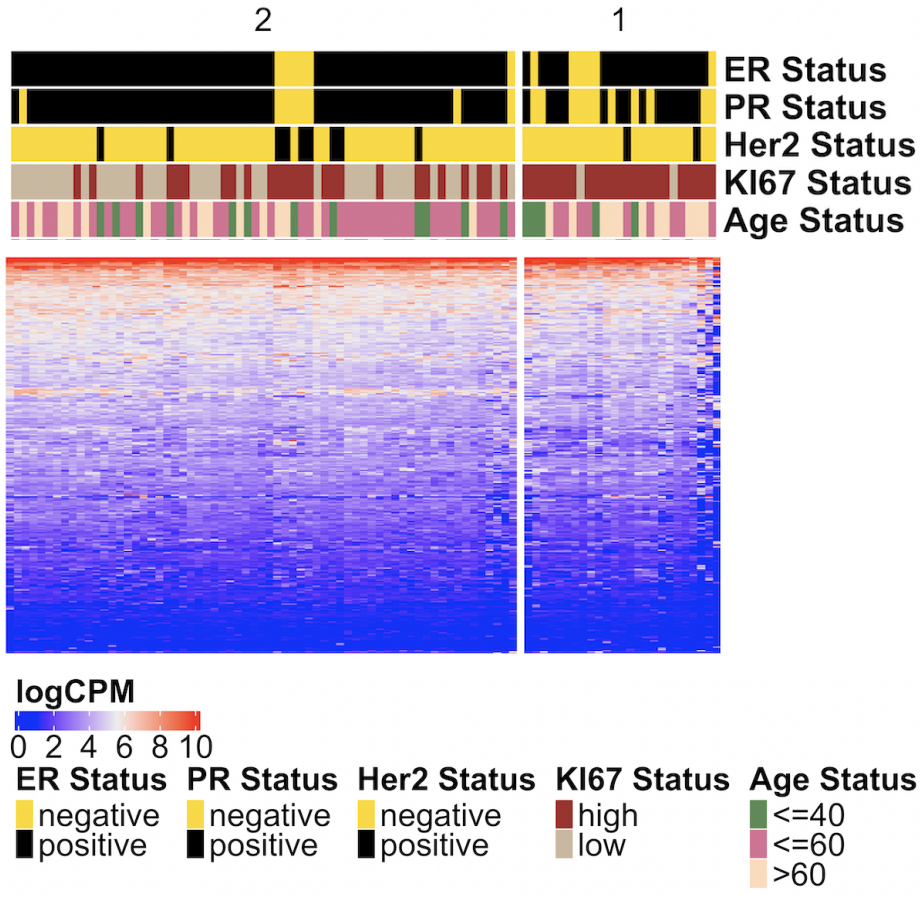


Figure 26. Heatmap of logTPM values of the genes involved in PI3K-AKT pathway annotated with ER, PR, HER2, KI67, and age status *(distance method: euclidean, clustering method: complete, number of column clusters: 2).*

Moreover, another list of genes containing BRCA1, BRCA2, PTEN, TP53, CDH1, STK11, CHEK2, BRIP1, ATM, PALB2 were used to cluster ABCC data to observe classification of clinical data. The selected genes are associated with hereditary cases where *BRCA1*, *BRCA2*, *PTEN*, *TP53*, *CDH1*, and *STK11* are penetrant genes conferring up to an 80% lifetime risk of BC whereas *CHEK2*, *BRIP1*, *ATM*, and *PALB2* are moderate-penetrance genes associated with a twofold increase in risk [234]. The list showed poor classification of clinical data as luminal and non-luminal subtype compared to PI3K-AKT Signaling pathway genes (**Figure 27**) although a statistical assessment is needed.

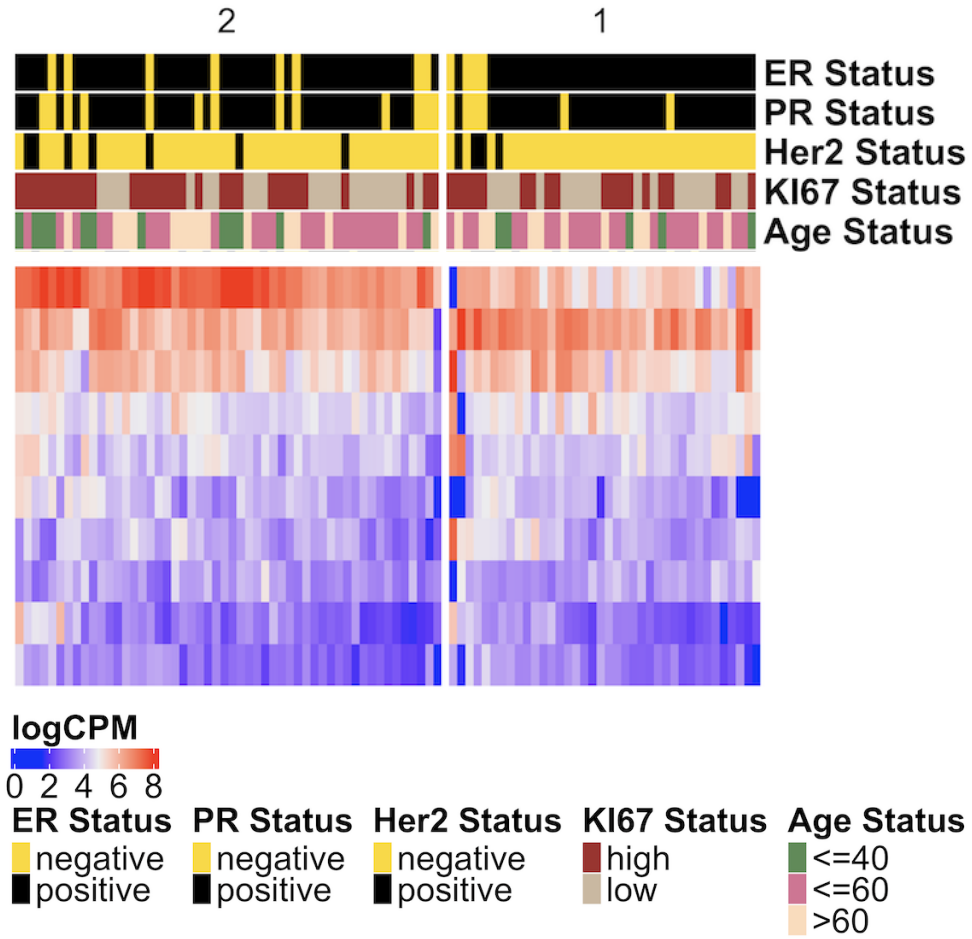


Figure 27. Heatmap of logTPM values of BRCA1, BRCA2, PTEN, TP53, CDH1, STK11, CHEK2, BRIP1, ATM, PALB2 annotated with ER, PR, HER2, KI67, and age status *(distance method: euclidean, clustering method: complete, number of column clusters: 2).*

4.6.3. PCA

PCA tab is the third tab of the classABCC that constructs PCA of selected gene signatures out of six available (**Figure 28A**), and the PCA can be grouped by patients' id, IHC, PAM50, or SCMOD1 subtypes (**Figure 28B**). An interactive PCA is presented (**Figure 28C**).

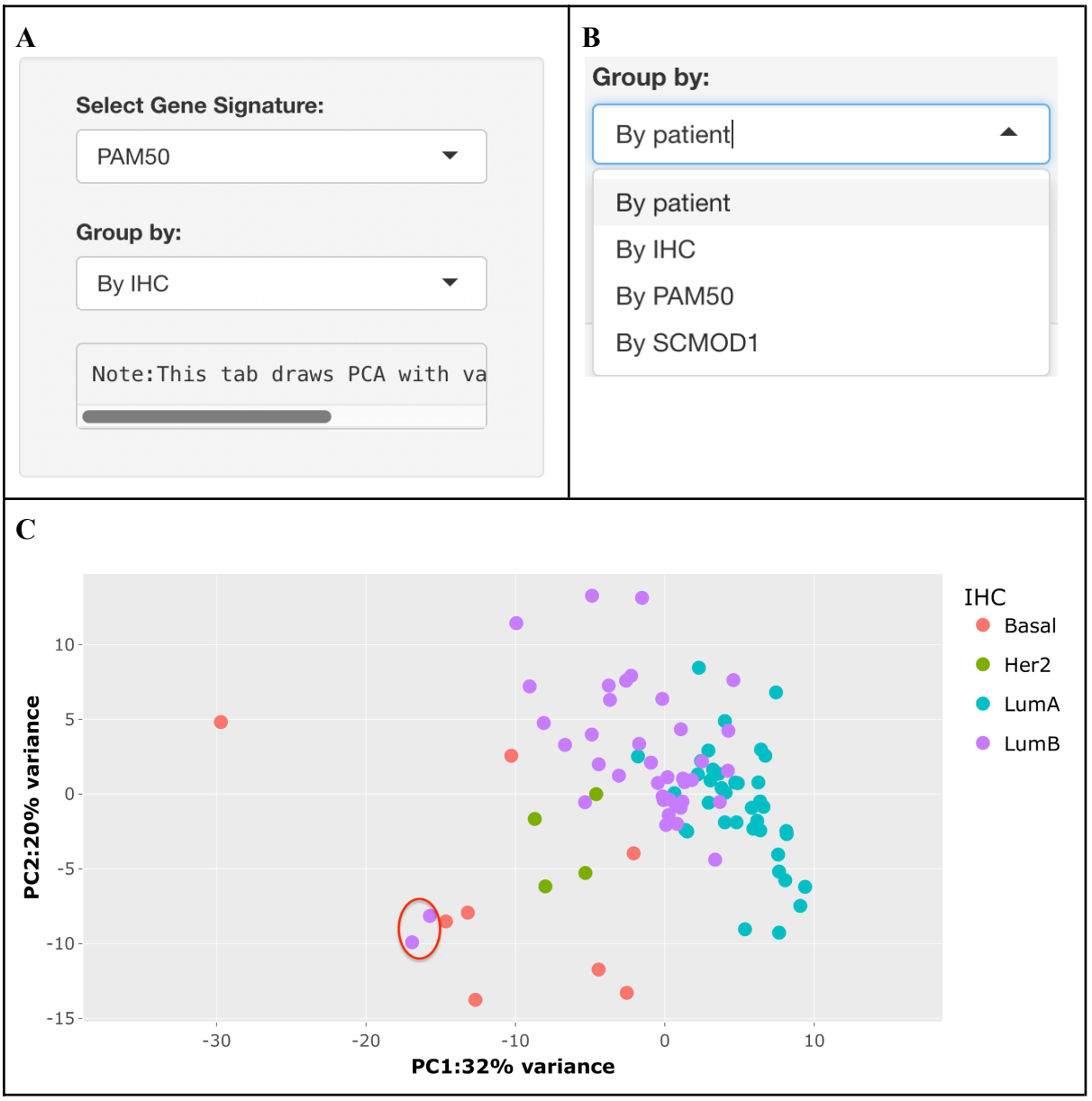


Figure 28. Screenshots of the *PCA* tab of classABCC. (A) Sidebar panel of *PCA* tab that enables gene signature and grouping selection; B) The grouping selection bar of *PCA* tab presents grouping by patients' id, IHC, PAM50, and SCMOD1 subtypes; and C) The classABCC output of PCA of PAM50 genes and colored by IHC subtypes *(the same result discussed in Figure 7)*.

4.6.4. Molecular Subtyping with ABCC

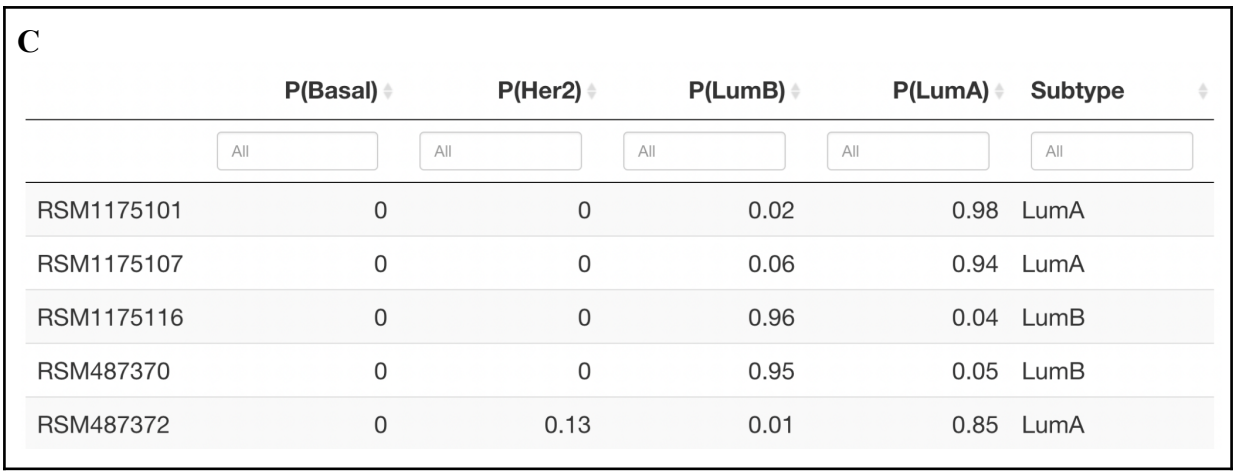
In the *Molecular Subtyping with ABCC* tab, there are subtabs: *Clinical Data, PAM50, SCMOD1, Compare, Compare with Scatter Plot and Compare with Heatmap*. In the Clinical Data subtab (**Figure 29A**), the table of clinical data with IHC subtypes and IHC characterization criteria are available. Moreover, a pie chart showing the percentage of each subtype is presented. In the *PAM50* subtab, a table of PAM50 subtypes with probabilities and a pie chart are presented (**Figure 29B**). Similarly, a table of SCMOD1 subtypes with probabilities and a pie chart are presented in the *SCMOD1* subtab (**Figure 29C**).

A

	ER	PR	HER2	%KI67	KI67 Status	Age	Age Status	Type	Subtype
									All
RSM1175101	positive	positive	negative	10	low	69	>60	ER+	LumA
RSM1175107	positive	positive	negative	18	low	68	>60	ER+	LumA
RSM1175116	positive	positive	negative	25	high	82	>60	ER+	LumB
RSM487370	positive	positive	negative	35	high	74	>60	ER+	LumB
RSM487372	positive	positive	negative	12	low	42	<=60	ER+	LumA

B

	Basal_prob	Her2_prob	LumA_prob	LumB_prob	Normal_prob	Subtype
	All	All	All	All	All	All
RSM1175101	0	0	0.61	0.39	0	LumA
RSM1175107	0	0	0.68	0	0.32	LumA
RSM1175116	0	0.01	0.2	0.78	0	LumB
RSM487370	0	0	0.36	0.64	0	LumB
RSM487372	0	0	0.41	0	0.59	Normal

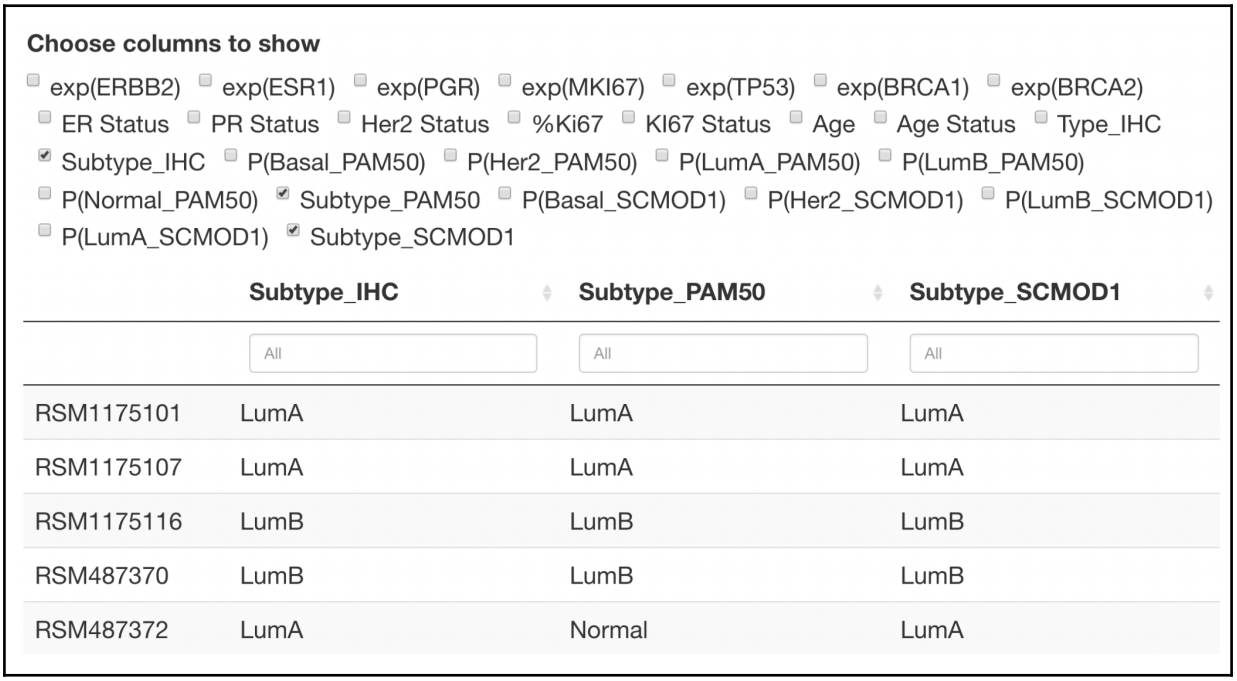


C

	P(Basal)	P(Her2)	P(LumB)	P(LumA)	Subtype
	All	All	All	All	All
RSM1175101	0	0	0.02	0.98	LumA
RSM1175107	0	0	0.06	0.94	LumA
RSM1175116	0	0	0.96	0.04	LumB
RSM487370	0	0	0.95	0.05	LumB
RSM487372	0	0.13	0.01	0.85	LumA

Figure 29. Screenshots of the *Molecular Subtyping with ABCC* tab of the classABCC. A) *Clinical Data* subtab containing clinical data and (IHC) subtypes; B) *PAM50* subtab showing the table of subtypes by PAM50 method along with probabilities; and C) *SCMOD1* subtab showing the table of subtypes by SCMOD1 method along with probabilities.

The results from the previous tab and the expression values of selected breast cancer-related genes were available to compare in one table in the *Compare* tab (**Figure 30**). The subtype probabilities calculated by PAM50 and SMOCD1 are presented as visualizations in the *Compare with Scatter Plot* and *Heatmap* tab (**Figure 31A, B**).



Choose columns to show

exp(ERBB2) exp(ESR1) exp(PGR) exp(MKI67) exp(TP53) exp(BRCA1) exp(BRCA2)
ER Status PR Status Her2 Status %Ki67 KI67 Status Age Age Status Type_IHC
Subtype_IHC P(Basal_PAM50) P(Her2_PAM50) P(LumA_PAM50) P(LumB_PAM50)
P(Normal_PAM50) Subtype_PAM50 P(Basal_SCMOD1) P(Her2_SCMOD1) P(LumB_SCMOD1)
P(LumA_SCMOD1) Subtype_SCMOD1

	Subtype_IHC	Subtype_PAM50	Subtype_SCMOD1
	All	All	All
RSM1175101	LumA	LumA	LumA
RSM1175107	LumA	LumA	LumA
RSM1175116	LumB	LumB	LumB
RSM487370	LumB	LumB	LumB
RSM487372	LumA	Normal	LumA

Figure 30. The *Compare* subtab of the *Molecular Subtyping of ABCC* tab of the classABCC.

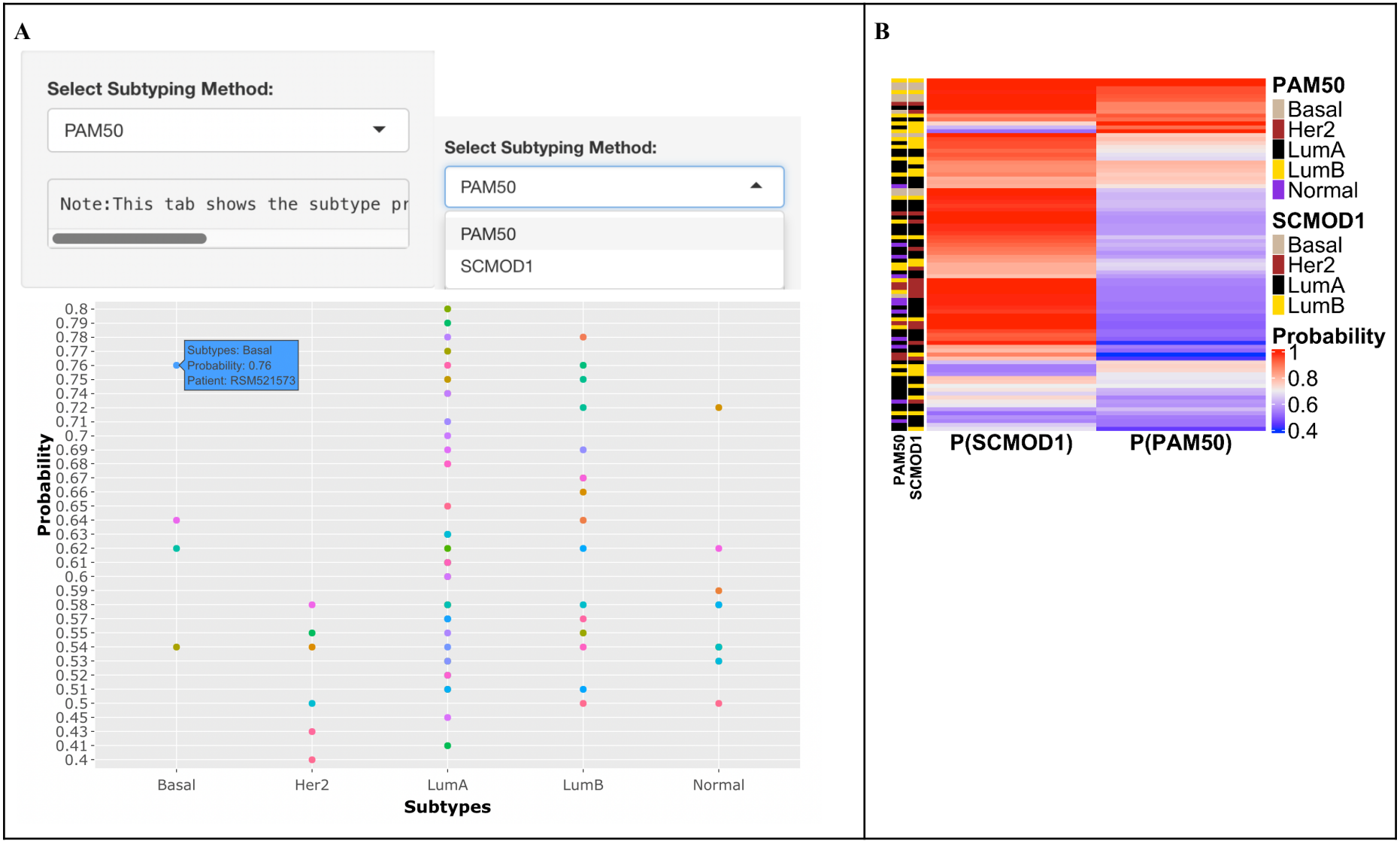
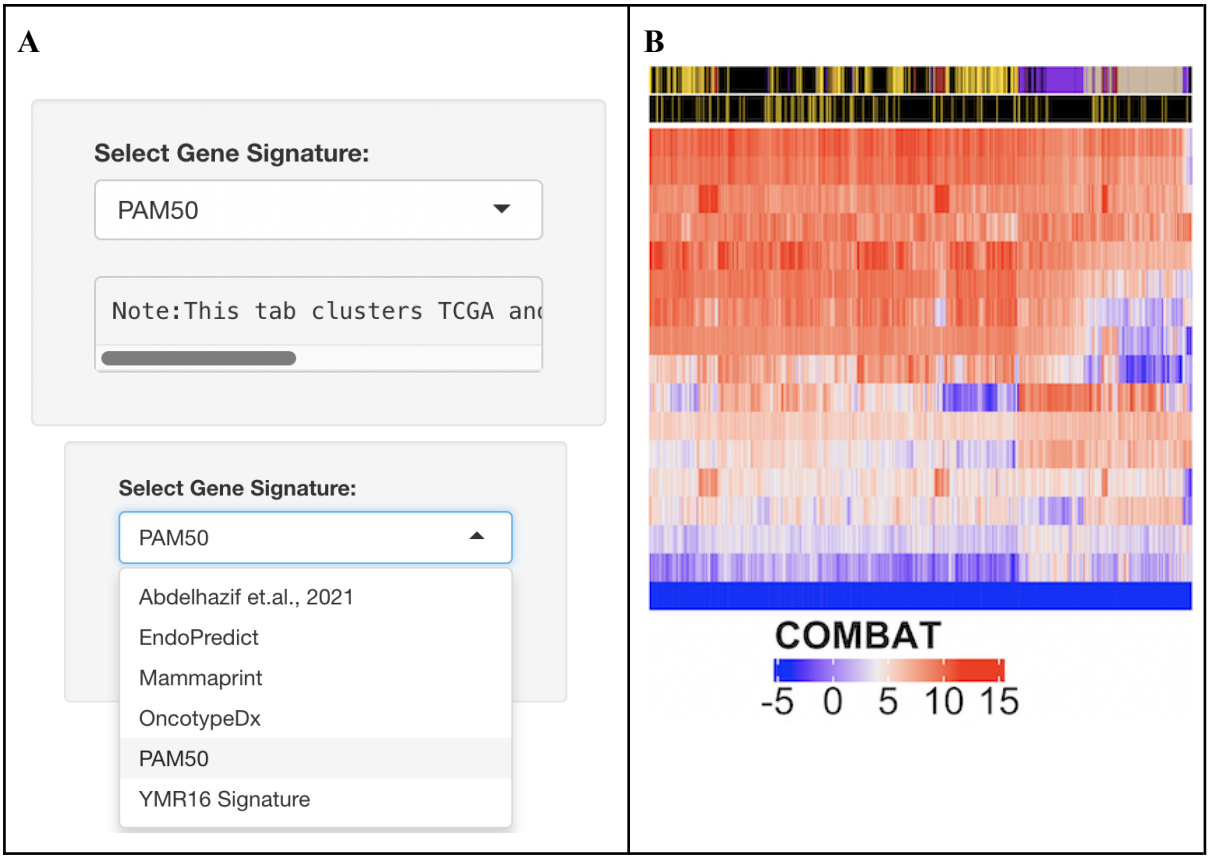


Figure 31. *Compare with Scatter Plot* and *Heatmap (the same result discussed in Figure 9)* subtabs of the *Molecular Subtyping with ABCC* tab.

4.6.5. TCGA

In the *TCGA* tab, ABCC data can be visualized together with TCGA-BRCA data based on the gene signature selection among six signatures. Since TCGA-BRCA and ABCC data from different sources, they clustered separately without batch effect normalization. Therefore, these two datasets were batch normalized prior to clustering. Moreover, PAM50 subtypes which were obtained from PanCancerAtlas for TCGA-BRCA RNA-seq data and from the genefu package for ABCC RNA-seq data were used as annotation (**Figure 32A**). There are subtabs: *Heatmap* and *PCA*. In the *Heatmap* subtab, heatmaps of genes of selected signatures across TCGA-BRCA and ABCC data are clustered along with PAM50 subtypes and the source of the data as annotations (**Figure 32B**). In the *PCA* subtab, PCA plotted with signatures and colors indicate PAM50 subtypes; shapes indicate the source of the data (**Figure 32C**).



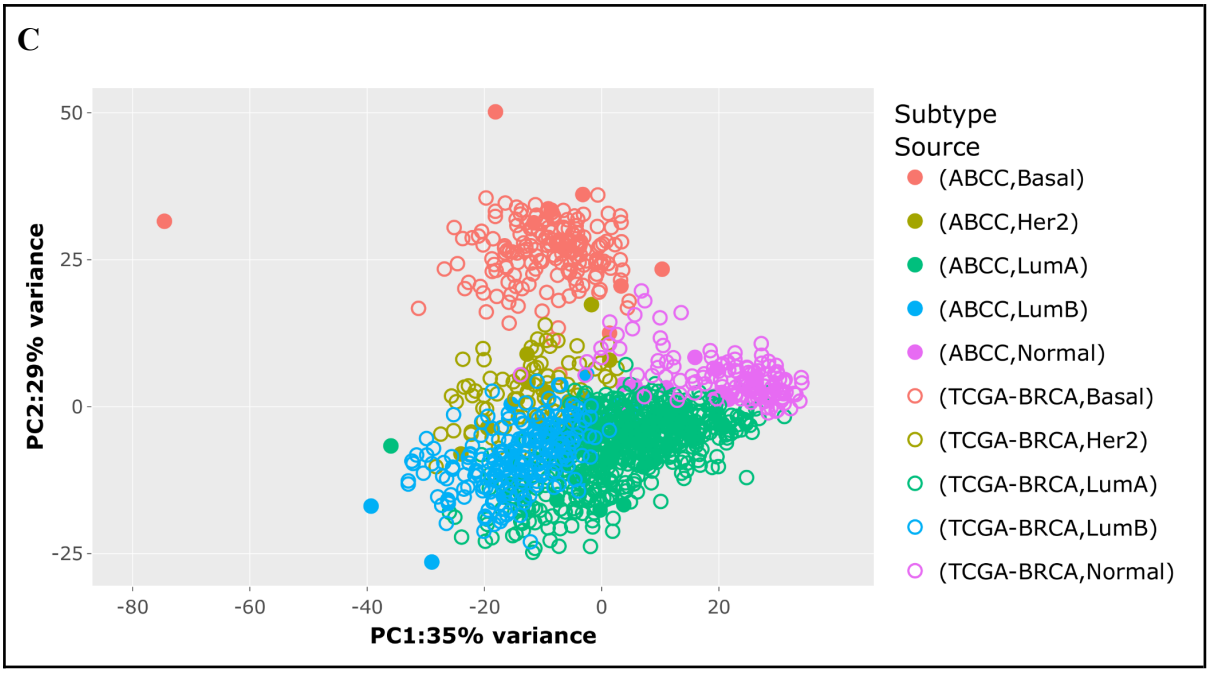


Figure 32. Screenshots of the *TCGA* tab of classABCC. A) Sidebar panel of the *TCGA* tab that enables gene signature selection containing six different gene signatures; B) A heatmap of PAM50 genes across TCGA-BRCA and ABCC datasets; and C) the visualization output as PCA of PAM50 genes across TCGA-BRCA and ABCC datasets *(the same result discussed in Figure 15)*.

4.6.6. Predict

In the *Predict* tab, TCGA-BRCA trained model was integrated into the app in which ABCC data were used to predict subtypes of ABCC patients. The confusion matrix (**Figure 33A**) and observed-predicted subtypes (**Figure 33B**) are presented as a table in the tab.

A

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall	F1
	Al	Al			A		
Class: Basal	0.78	0.99	0.87	0.98	0.88	0.78	0.82
Class: Her2	0.86	0.98	0.75	0.99	0.75	0.86	0.8
Class: LumA	0.98	0.67	0.72	0.97	0.72	0.98	0.83
Class: LumB	0.57	1	1	0.88	1	0.57	0.73
Class: Normal	0.36	0.99	0.8	0.92	0.8	0.36	0.5

B

	Prediction	Reference	Freq
	All	All	All
1	Basal	Basal	7
2	Her2	Basal	1
3	LumA	Basal	0
4	LumB	Basal	0
5	Normal	Basal	1

Figure 33. Screenshots of the *Predict* tab of classABCC. A) Confusion matrix of subtype prediction of ABCC data with TCGA-BRCA trained model; and B) Comparison of observed *(subtypes by PAM50 method)* and predicted subtypes for each patient.

5. Discussion and Conclusion

Transcriptomics-based approaches have revealed the molecular heterogeneity and distinct gene expression patterns across breast cancer subtypes since the early 2000s. This has led to the usage of molecular subtypes in clinics and translational research in prognostic assessment, therapeutic efficacy prediction, and retrospective analysis of cohort studies [50], [54], [55]. Molecular subtyping classification emerged from unsupervised clustering of gene-expression data to identify molecular subsets of similar gene expression patterns in extensive cohort studies. Then, this analysis evolved more towards prediction by supervised methods by comparing outcome groups to identify differentially expressed genes to develop multi-gene outcome predictors. Single sample predictors (SSPs), including PAM50, are gene expression-based classifiers used to classify breast cancer subtypes with hierarchical clustering followed by nearest centroid classification [66]. However, SSPs have several limitations, including stability and concordance across different signatures/methods and the requirement of a large retrospective study [7], [17], [66]. Subtype Classification Models (SCMs), including SCMOD1, classify subtypes with a mixture of Gaussian distribution based on the genes that are correlated with ER, PR, and AURKA [7], [36], [66], is only one among molecular subtyping algorithms, published intrinsic gene lists, and prognostic gene signatures that are also used for molecular subtyping purposes [60], [80] [82], [84], [139], [155]. As a result, there are various methods, gene lists, and signatures to identify breast cancer subtypes, however the concordance among them, stability of these methods, and their applicability to each dataset from different genetic backgrounds are not well known and may present limitations.

5.1. Comparison of IHC, PAM50, and SCMOD1 methods

Subtype prediction using IHC, PAM50, and SCMOD1 methods showed varying results in predicting subtypes across ABCC. In IHC subtyping, the predictors were ER, PR, HER2, and KI67 proteins predicting Luminal A, Luminal B, HER2-positive and basal subtypes. PAM50 uses a 50-intrinsic gene list to predict Luminal A, Luminal B, HER2-positive, basal and normal-like subtypes. Finally, SCMOD1 uses 726 genes to classify data into Luminal A, Luminal B, HER2-positive and basal subtypes. The main difference between the three predictor methods was observed in the Luminal B subtype, where the IHC method was more prone to overpredicting it (PAM50: 23.3%; SCMOD1: 28.9%; IHC: 44.4%). Since both IHC and SCMOD1 methods do not involve the classification of normal-like subtypes, I was expecting to see a similar distribution of subtypes among them. However, SCMOD1 was more concordant with PAM50 than IHC regarding Luminal B subtyping. This can be related to KI67 percent, and HER2 status since both are the predictors of Luminal B in the IHC method, which raises a question about the reliability or adequacy of the IHC method [23], [182], still being widely used in pathology labs.

The reliability of IHC results could be questioned especially for a subset of samples giving discordant results. For example, analyses showed that some patients were identified as basal by both PAM50 and SCMOD1 predictors. However, the same patients were classified as Luminal B by IHC. The patients with RSM521517 and RSM521661 ids had low expression of ER, PR, and HER2 at the expression level according to RNA-seq results. Therefore, they were identified as the basal subtype by both PAM50 and SCMOD1 as expected. However, according to the clinical data, RSM521517 identified as ER- and PR-positive and HER2-negative whereas RSM521661 was identified as ER-positive and PR- and

HER2-negative at protein level. Furthermore, the KI67 percent were supplied as 62% and 66% for RSM521517 and RSM521661, respectively, which was classified as high. In the cohort, there were no other patients which showed such discordance. The molecular subtype prediction probability of both predictors was high, implying that the prediction stating basal subtype was more likely than Luminal B with high probability. Moreover, PAM50 and SCMOD1 revealed that both predictors had high concordance for basal subtype prediction. Therefore, it could be considered less likely to observe wrong predictions of the basal class by PAM50 or SCMOD1. Therefore, these patients could wrongly be classified as Luminal B, although they were basal subtypes. This could be because the molecular subtyping uses more genes and correlated gene expression patterns while IHC relies on hormone receptors' and proliferation markers' protein expression patterns [65], which can be prone to technical problems associated with the technique [51], [52]as well as presence of high tumor heterogeneity. These results show that IHC may not always agree with molecular subtyping and may be inadequate in identifying subtypes. Moreover, one of the reasons is that the success of IHC depends on various parameters such as type and duration of fixation, quality of fresh or frozen sectioning, and antibody specificity, which makes the subtyping more complex [51]. Moreover, the differences between subtypes are not always observable pathologically, anatomically, or at the protein level [18], [52]. While IHC is a protein based assay, the PAM50 and SCMOD1 methods rely on expression data and they may not be comparable with each other in the presence of gene mutations or copy number variations as well as extensive post transcriptional modifications, that are common in some cancers [235], [236]. Furthermore, there is no consensus on determining the scoring cut-offs of several markers [224]. For example, high expression of KI67 is defined as 10%, 15%, 30% in different studies [25, p. 67], [46], [53]. Finally, scoring subjectivity and intra-tumor

heterogeneity creates variation in the IHC method [224]. A potential solution could be to perform follow up IHC measurements with these tumors to address intra-tumoral differences.

The concordance issue between IHC and PAM50 methods has been shown before by Bastien et al [220]. They showed that breast cancer subtyping using ER, PR, and HER2 status based on IHC and PAM50 methods were adequately concordant with each other. IHC strongly assigned the ER-positive samples as Luminal A and Luminal B. However, only 75% ER-negative samples were associated with HER-positive and basal groups. Authors concluded that although the IHC method is a great tool and gene expression of ER and HER2 showed good concordance with the protein level, IHC has more prognostic purposes than PAM50 [220]. Furthermore, the discordant results were obtained where 38.4% of patients were subtyped differently by IHC and PAM50 concluding that IHC could nevertheless mislead the decision for treatment and hence may result in poor outcome [222]. Our results also support this view with a totally independent and novel BRCA dataset from Acibadem University in Turkey.

Furthermore, Luminal A and Luminal B subtypes make a continuum since both subtypes are ER-responsive [214], [215]. Therefore, ER and ER-related genes have been successfully used to classify them. However, Luminal B is a more aggressive and proliferative subtype in which proliferative genes and HER2 may play a role in its correct classification where Luminal B also makes a continuum with the HER2-positive subtype [214]. Therefore, Luminal A, Luminal B, and HER2-positive subtypes can frequently get mixed up encouraging me to provide comparative means for more accurate subtyping by integrating with IHC with at least two of the most used subtyping methods, PAM50 and SCMOD1. Moreover, correct and robust subtyping could be critical since basal, HER2, and Luminal B

subtypes are very distinct regarding their treatment, survival, recurrence, and metastasis. The basal subtype is more aggressive with a high risk of early metastatic relapse [183]. Moreover, there is no available targeted therapy for the basal subtype; only chemotherapy is validated [183]. On the other hand, Luminal B is a hormone-therapy responsive subtype. However, it is not sensitive as Luminal A since ER expression is generally lower in Luminal B [23], [182]. Moreover, Luminal B is insensitive to chemotherapy compared to basal and HER2-positive subtypes [182]. Thus, the accurate classification of breast cancer into subtypes matters significantly due to its effects on the treatments and prognosis decisions [225]. In a novel patient cohort as in the case of ABCC, the use of multiple prediction methodologies may provide clinicians with a better diagnosis and possible prognosis via use of specialized gene signatures of their patients.

Although the concordance issues have been revealed for IHC in association with other predictor methods, IHC is still a useful and informative approach especially in clinics. Therefore, it would not be realistic to claim that transcriptomics and mutation status should not be used in clinics for prognostic purposes. Therefore, one solution can be improvement of IHC approaches. The fist approach can be using IHC not only for prognostic purposes but also subtyping purposes since subtypes are more informative about the heterogeneity. The second approach can be addition of new markers which are good predictors in transcriptomic level including AURKA [65], BRCA1, and BRCA2 [221]. The third approach can be reconsideration of available IHC markers. For example, Daemen et al. claimed that HER2 is not a subtype but it is an event where HER2 amplification and overexpression is observed. They said that HER2-positive subtype is independent from HER2-amplification but that reflects androgen receptor signaling as replacement for estrogen receptor (ER)-driven tumorigenesis [223].

5.2. Comparison of PAM50 and SCMOD1 methods

The discordant results between PAM50 and SCMOD1 methods were observed across ABCC where the two methods were only 67% concordant. This result can be confirmed with the robustness analysis of all SSPs and SCMs methods across 5715 breast cancer patients' data collected from different publicly available data, where robustness is defined as the score of the assignment of the same tumors to the same subtypes regardless of the gene expression data found in the classifier. [65]. While SCMs yielded prediction strength over 0.8, SPPs yielded much lower strength in 0.45-0.59. SCMs are statistically significantly more robust than SSP models [65]. Furthermore, the limitations of PAM50 molecular subtyping due to ER-status balancing and gene centering approach affects the prediction strength of PAM50. Therefore, Kumar et al. developed a new approach called PCA-PAM50 to perform molecular subtyping with ER-unbalanced data showing 6-9.3% improvement inconsistency with IHC, especially for Luminal B subtype (25-49% improvement) [210].

Furthermore, I have checked whether PAM50 and SCMOD1, which are two of the commonly used gene-expression predictors, showed discrepant results for the patient data across the globe to remove the effect of genetic background on the subtyping methods. Five publicly available array datasets (Mainz, Transbig, UPP, UNT, and NKI) were used containing patients from Germany, France, the United Kingdom, Sweden, and Norway. The concordance analysis of the predictions between PAM50 and SCMOD1 was found to be around 82%, suggesting a moderate to a high concordance. The normal-like subtype was removed from the concordance analysis since SCMOD1 does not make normal-like subtype predictions. The results showed that the concordance of the two methods was high especially for the basal and Luminal A subtypes since Luminal A and basal subtypes were relatively easy to predict due to their well-known molecular nature consistent with a previous study [65]. However,

HER2-positive and Luminal B concordance were low. The low concordance of the HER2-positive subtype can be associated with the heterogeneous nature of this subtype [211]. The HER2-positive subtype was classified based on the negative hormone receptors and positive HER2 receptors. However, HER2-positive subtypes generally show varying molecular backgrounds with low clustering abilities since they may carry transcriptional subtypes of both Luminal A and basal [184], [211]. Moreover, the concordance of six predictor models (three SCMs and three SSPs) was analyzed in a study that showed that the concordance score of PAM50 and SCMOD1 was 0.67 (0.65-0.68). This concordance is classified as substantial (0.60-0.80) by Mackay et al. [65] [70].

5.3. Clustering of ABCC and TCGA-BRCA using selected gene signatures

The normalization type used was logTPMin the unsupervised clustering since TPM was based on length-normalization, whereas CPM included only by depth normalization. Unsupervised clustering of ABCC data with six multi-gene signatures showed varying levels of clustering power for each signature. 17-gene signature and PAM50 showed the best separation of luminal subtypes from non-luminal subtypes , as shown before [65], [185]. Mammaprint and EndoPredict were designed to classify breast cancer samples based on the response to chemotherapy by dividing samples as high risk and low risk [82], [87]. Hence, these two signatures moderately differentiated subtypes as luminal and non-luminal, both of which may contain high as well as low risk patients. Since these two signatures' foci were the risk of recurrence, they also helped cluster patients' KI67 status as low and high [82], [87]. Moreover, EndoPredict and Mammaprint are predictors for prognosis and risk predictors, which have been used for clinical purposes. However, I showed that they also behaved as moderately effective signatures that could differentiate luminal subtypes for ABCC data where PAM50 and 17-gene signatures have been developed specifically for subtyping

purposes [80]. On the other hand, OncotypeDx and 16-gene signature showed relatively low power to differentiate luminal from non-luminal tumors in ABCCs, although it was shown previously that they had the power to differentiate subtypes [80], [84]. Acordingly, this implies that the clustering power of signatures is also dependent on the data selected. It is also interesting that the 16-gene signature did not show any clustering given the clinical variables provided, although it is a gene signature containing a minimal overlap of 33 available gene signatures [80]. Therefore, it can be concluded that the overlapped genes found in several signatures did not reflect the PAM50 classes of ABCC and TCGA-BRCA RNA-seq data and may provide a residual effect that has not been stated before. However, the differential clustering power of signatures for hormone status, age or proliferation status, defined as clustering the samples according to the given clinical or any molecular measurement, is best understood by supervised analyses; and clustering could only be helpful as an exploratory tool rather than a prediction. The robustness of clusters generated by these methods can further be studied by existing methods and the number of clusters that best describe the data given the signatures can be studied in the future [213]. For example contingency tests for binary scores and Wilcoxon tests for continuous or ranked values split into two clusters can be performed to test significant differences between clusters.

The clustering of ABCC with gene signatures showed differing molecular subtyping abilities. When ABCC data were ComBat normalized with TCGA-BRCA, RNA-seq data showed similar gene expression patterns, which was observed with the heatmap and dimension reduction method. This is an important finding and can be a prototype for novel BRCA cohorts as exemplified for ABCC. Thus, it allows further analysis of ABCC regarding TCGA-BRCA, including TCGA-BRCA fed machine learning model to predict ABCC.

Furthermore, the clustering of ABCC data using user selected input gene lists was demonstrated with a use case gene list, PI3K-AKT signaling pathway genes from KEGG and highly mutated genes in BRCA. PI3K-AKT signaling pathway genes and mutated genes that are related with high risk of breast cancer showed moderate and poor classification of clinical data, respectively. The results showed that highly mutated genes in breast cancer may not be the best indicators for subtype classification at least for ABCC data which showed different mutation profiles across subtypes [232-234]. However the clustering by these lists may identify a different clinical association then the hormone status of discretized KI67 status and may warrant future studies. However, the classABCC user entered gene list capability offers a useful option so that any gene list can be used to cluster ABCC data to check the clustering of breast tumors from this novel cohort..

5.4. Machine learning models to predict breast cancer subtypes

A model was constructed with a random forest algorithm and fed with TCGA-BRCA data performed best. Since TCGA-BRCA was a large, reliable breast cancer cohort it was selected to test the predictive power of PAM50 genes for training and cross validation. Since it was commonly used for clinical and research purposes and described as the gold standard [199], NIH uses PAM50 subtyping in TCGA-BRCA data [66], [189]. The model showed excellent performance with validation data; however, the metrics decreased with test data, including the ABCC. The first reason for this poorer performance was identified as overfitting, which is generally observed when data are imbalanced. All breast cancer cohort data, including ABCC and TCGA-BRCA have been dominated by Luminal A, the most common breast cancer subtype in the literature. Therefore, the model could be prone to predict non-Luminal A subtypes as Luminal A due to its imbalanced nature [188]. There are different approaches to overcome the overfitting problem due to imbalanced data; oversampling and undersampling

[186-188]. Oversampling did not improve the performance except for the basal subtype, whereas undersampling did improve the model's performance by 10%. The undersampling improved F1 scores of Luminal A, Luminal B, and normal-like subtypes, whereas oversampling only affected the basal subtype. This result implies that train dataset is already complex enough, so increasing the number of minority classes with oversampling did not improve the performance. The need for the model seems to be the simplification by the reduction of the majority class, which is Luminal A. The imbalanced data problem of breast cancer data can be solved by integrating several datasets and providing a balanced number of each class by decreasing the major classes and increasing the minor classes [195]. The undersampling model showed good performance in predicting basal and Luminal A; moderate performance in predicting HER2-positive and Luminal B; low performance in normal-like subtype. The issue of Luminal A, Luminal B, and normal-like subtype differentiation was also been brought up by Mackay et al. [185], in which Luminal A and B differentiation problems can be due to the continuum expression levels of proliferation-related genes [65], [69–72]. Furthermore, List et al., who also used TCGA-BRCA data as train data and PAM50 genes as predictors, after removal of the 20% of the least importance genes, showed that this model performed well in the differentiation of basal and Luminal A but poorly in the differentiation of HER2 and normal-like subtypes [199], confirming the findings presented in this thesis.

Another dataset was used to check the performance of the model in addition to ABCC as a test data; thus MATADOR data was used since it is a well-known public data with high sample size. The model performance was relatively poorer with MATADOR data which had not been used so far for the concept of molecular subtyping with machine learning. The decrease observed in F1 scores by about 20% may reflect the difference in class distribution

between ABCC and MATADOR, where ABCC was dominated by Luminal A as in the case of TCA-BRCA cohort. Therefore, undersampling considered Luminal A as the majority group and decreased the number of samples. However, the majority group is Luminal B in MATADOR data, so the model performance is expectedly lower; hence undersampling using the Luminal B as the majority group may be needed for a better classification performance. . The varying performance of the model with TCGA-BRCA RNA-seq data as train data was shown across ABCC and MATADOR data. There can be other reasons for that in addition to the imbalanced nature of all selected data. The first reason can be related to genetic differences of breast cancer cohorts which are from different populations. It was shown that breast cancer and its subtypes show varying biological features, prognosis, and risk factors regarding population-based differences, etiological or demographic differences [226], [227]. Therefore, the results may reflect the more genetic similarity of TCGA-BRCA RNA-seq data with ABCC than MATADOR. Therefore, datasets used in machine learning pipeline could be adjusted based on genetic background considering etiological and demographic information. Moreover, the demographic information can be included in the train data to remove the effect of population-based differences which can be used in predicting risk factors and early diagnosis purposes in BC using machine learning [228], [229].

Another reason for the disconcordancy may be related with the source of predictors. TCGA-BRCA data was subtyped with the PAM50 method by The Cancer Genome Atlas Network in 2012 [63] where the subtypes were available in the TCGABioLinks package. On the other hand, molecular subtyping was done using the PAM50 method with the genefu package for ABCC and MATADOR data since the subtypes were not available for these datasets. Although PAM50 was the method applied to all datasets, TCGA-BRCA RNA-seq subtyping with the genefu package showed differences with the results of The Cancer Genome Atlas Network. The concordance between them was found as 75% for

TCGA-BRCA data where the majority class changed from Luminal A to Luminal B. This can be related with the genefu package which was stated as a package for breast cancer gene expression analysis allowing analysis of microarray and RNA-seq data [170]. However, a contributor of the package stated that an adjustment for RNA-seq data can be needed for microarray centroids where there is no established method that performs best for that adjustment (personal communication). The genefu package is a useful package to predict molecular subtypes of gene expression data, however, more adjustments may be needed for RNAseq data.

Furthermore, the discordant results may be related to the type of omics data, bulk RNA-seq which has poor power to reflect the true heterogeneity of samples [197]. To improve the ability of breast cancer classification, methylation data for deeper insight on epigenetic level [199], single-cell RNA-seq data [147], hematoxylin-and-eosin stained slides for visual signals [198], copy number variation data [200] can be integrated into the train data. Daement et al. showed that data subtyped with copy number aberrations, mutations, gene and isoform expression, promoter methylation, and protein expression information improved the prediction by 65% compared to using only the expression data [200].

Finally, the model always showed the worst performance in predicting normal-like subtype regardless of the selected dataset which also affected the performance of the model significantly. Thus, removing the normal-like subtype can be an option and has been applied. There is a controversy on the normal-like subtype that can be an artifact resulting from stromal contamination [70]. Normal-like samples described as an artifact where the number of tumor cells are low with a large background of breast cells in the sample [220].

5.5. A subset of PAM50 genes with feature importance

Interestingly, it was shown that more than half of the 50 predictors of the PAM50 gene signature had almost no impact on the subtype prediction by the feature importance analysis. The refinement of PAM50 can be an issue that is already discussed by Ektefaie et al., who dropped 20% of the least important PAM50 genes [198], indicating a subset of PAM50 reflected cancer morphology. Moreover, the top 15 important genes were selected as a subset of PAM50 genes, which are FOXA1, ESR1, MLPH, FOXC1, CEP55, NDC80, EGFR, NUF2, SFRP1, BIRC5, MIA, MYBL2, UBE2T, EXO1, SLC39A6 [199]. As mentioned, I identified a 26-gene list with feature selection analysis with TCGA-BRCA RNA-seq random forest model. The 26-gene list that I identified includes the 15-gene list identified by List et al. [199] except NUF2, SFRP1, and SLC39A6 genes. The feature importance analysis selected 26 genes mainly related to cell composition and cell division processes confirming cancer hallmarks [196]. Moreover, KRT17, KRT14, EGFR, ERBB2, PGR, KRT5, and ESR1 genes have been used as basal subtype markers, which explains the model's power in differentiating basal subtype [205-208]. The concept of basal-positive and basal-negative Luminal A may explain the power of basal markers to differentiate Luminal A subtypes implying Luminal A subtypes of ABCC are basal-negative [209]. The results have indicated that a subset of PAM50 can have enough power to differentiate subtypes confirming the need for new signatures that only involve genes with high prediction strength of breast cancer subtypes.

The split of large breast cancer cohort as train and test data to conduct a supervised analysis is not a new concept; actually, it is commonly used to predict the cancer recurrence, the sensitivity to hormone therapy and chemotherapy, metastasis, and to identify gene signatures for subtype differentiation [138], [144], [147-150], [152,p.53]. However, no available publication focuses on a TCGA-BRCA trained random forest model that is integrated into a

web tool to facilitate exploration of a novel dataset combined with comparison of IHC, PAM50, and SCMOD1 methods and gene signatures.

In conclusion, the thesis has shown the heterogeneous nature of breast cancer across ABCC, TCGA-BRCA, and other publicly available datasets. These findings supported the existing notion of heterogeneity as in the previous literature [2-7]. The thesis also confirms that histology and multi-omics status should be systematically investigated further. Moreover, it has shown that there is still a need for novel or comprehensive gene signatures, standardized methods, or algorithms to identify the breast cancer subtypes accurately, regardless of the population-based differences to understand the true differences between breast cancer subtypes to allow more personalized treatment and better outcomes in future.

5.6. Strengths and Limitations

One of the strengths of this thesis is that it uses ABCC dataset which is based on a novel breast cancer cohort, mainly consisting of the patients from Turkey/Turkiye with available clinical data of each patient that support the clustering and help confirm/compare with PAM50 and SCMOD1 subtype predictions. The second strength is that concordance issues between IHC, PAM50, and SCMOD1 have been shown also with ABCC and global datasets confirming previous results in literature. The third strength is that some commonly used gene signatures were compared across ABCC and TCGA-BRCA data, showing clustering of breast cancer subtypes and clinical data depends on selected gene signatures and data; this demonstrates further the need for discovering new gene signatures or approaches for classifying breast cancer subtypes regardless of genetic background of data. The fourth strength is that a machine learning model was developed with TCGA-BRCA data to test both ABCC and publicly available data showing the limitations of molecular subtyping with very heterogeneous, complex, and imbalanced data. Finally, a user-friendly pipeline was constructed and integrated into an R Shiny-based web tool that facilitates the investigation of ABCC and TCGA-BRCA data.

In the field of machine learning model development, TCGA-BRCA RNA-seq data has been previously and herein used; however, other available data types including mutation, clinical data and demographic information can be included to improve the analysis's strength and have a deeper understanding of molecular subtypes. Machine learning is a broad area where various methods and approaches are available for each case. To improve the power and performance of the model, preprocessing, tuning of parameters, oversampling, undersampling, and feature selection were tried. However, different solutions can be additionally tried, including regularization, adding more datasets, splitting datasets with different percentages (i.e., 60% - 20% - 20%), applying different preprocessing, and feature

extraction for minority classes. Furthermore, xgboost algorithm did perform poorer than random forest (data not shown) however it was not checked with undersampling and oversampling where an improvement could be observed. Finally, classABCC is a user-friendly web-tool specifically developed for ABCC RNA-seq data. Therefore, the app can be developed into a more generic pipeline to facilitate user input transcriptomics data for a more broad usage. Moreover, the app focuses on molecular subtyping of breast cancer including Luminal A, Luminal B, normal-like, basal, and HER2-positive. However, there are more subtypes of BC including immune subtypes [230] and future studies should focus on incorporating those as well.

6. Future Perspectives

- The training model can be improved with additional datasets including copy number variation, mutation data, clinical data, methylation data which are available for TCGA-BRCA.
- Train and test data can be adjusted based on demographic information if they are available for the data. Since there is genetic diversity among populations which should be considered before the model is tested and trained.
- Single-cell RNA-seq can be used for breast cancer molecular subtyping in breast cancer instead of bulk RNA-seq to discover molecular differences in better resolution. Or deconvolution of bulk RNA-seq can be another option in identifying the true number of subtypes.
- Chromatin conformation of the genome can be investigated if molecular subtyping is observable in the 3D spatial formation of DNA since it can give a better understanding toward the heterogeneity and diversity of breast cancer subtypes.
- For unsupervised clustering, the data were investigated as two clusters for luminal and non-luminal subtypes. For a more advanced approach, the number of clusters can be predicted with the Elbow method, Silhouette method, Self Organizing Maps (SOM) ot consensus clustering [212], [213]. Clustering quality can be assessed with intrinsic and extrinsic measures: Silhouette Coefficient [218] and Adjusted Rand index [219], respectively. Moreover, clustering tendency can be measured with Hopkin's Test [217]. Thus, clustering based on gene signatures can be evaluated more confidently across ABCC and TCGA-BRCA RNA-seq data.
- The luminal subtype is the dominant subtype in breast cancer; therefore, a gene signature specific for luminal subtype differentiation can be constructed. Moreover, triple-negative and basal subtype differentiation can be another perspective.

Moreover, the difference between these subtypes can be also studied by discovering specific markers.

- Training was only performed with TCGA-BRCA RNA-seq data for subtype prediction with machine learning. Therefore, other datasets such as METABRIC can be used as training data so that the imbalanced nature of the can be balanced with the additional datasets.
- Features of machine learning prediction are limited to PAM50 genes and it was shown that a subset of PAM50 could be enough to classify the subsets. Therefore, PAM50 genes identified with feature importance can be used to train the model. Feature selection or user input option can be added to the app. Moreover ROC/AUC may be used as performance measurements.

7. References

[1] "Cancer." https://www.who.int/news-room/fact-sheets/detail/cancer (accessed Apr. 23, 2022).

[2] K. Polyak, "Heterogeneity in breast cancer," *J. Clin. Invest.*, vol. 121, no. 10, pp. 3786–3788, Oct. 2011, doi: 10.1172/JCI60534.

[3] G. Turashvili and E. Brogi, "Tumor Heterogeneity in Breast Cancer," *Front. Med.*, vol. 4, p. 227, 2017, doi: 10.3389/fmed.2017.00227.

[4] N. Kumar, P. Patni, A. Agarwal, M. A. Khan, and N. Parashar, "Prevalence of molecular subtypes of invasive breast cancer: A retrospective study," *Med. J. Armed Forces India*, vol. 71, no. 3, pp. 254–258, Jul. 2015, doi: 10.1016/j.mjafi.2015.04.006.

[5] M. Ignatiadis and C. Sotiriou, "Luminal breast cancer: from biology to treatment," *Nat. Rev. Clin. Oncol.*, vol. 10, no. 9, pp. 494–506, Sep. 2013, doi: 10.1038/nrclinonc.2013.124.

[6] C. M. Perou *et al.*, "Molecular portraits of human breast tumours," *Nature*, vol. 406, no. 6797, pp. 747–752, Aug. 2000, doi: 10.1038/35021093.

[7] Z. Hu *et al.*, "The molecular portraits of breast tumors are conserved across microarray platforms," *BMC Genomics*, vol. 7, p. 96, Apr. 2006, doi: 10.1186/1471-2164-7-96.

[8] L. J. van 't Veer *et al.*, "Expression profiling predicts outcome in breast cancer," *Breast Cancer Res.*, vol. 5, no. 1, pp. 57–58, 2003, doi: 10.1186/bcr562.

[9] C. Sotiriou *et al.*, "Breast cancer classification and prognosis based on gene expression profiles from a population-based study," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, no. 18, pp. 10393–10398, Sep. 2003, doi: 10.1073/pnas.1732912100.

[10] T. Sørlie *et al.*, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 98, no. 19, pp. 10869–10874, Sep. 2001, doi: 10.1073/pnas.191367098.

[11] D. M. Abd El-Rehim *et al.*, "High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses," *Int. J. Cancer*, vol. 116, no. 3, pp. 340–350, Sep. 2005, doi: 10.1002/ijc.21004.

[12] H. O. Habashy *et al.*, "A review of the biological and clinical characteristics of luminal-like oestrogen receptor-positive breast cancer," *Histopathology*, vol. 60, no. 6, pp. 854–863, May 2012, doi: 10.1111/j.1365-2559.2011.03912.x.

[13] O. Yersal and S. Barutca, "Biological subtypes of breast cancer: Prognostic and therapeutic implications," *World J. Clin. Oncol.*, vol. 5, no. 3, pp. 412–424, Aug. 2014, doi: 10.5306/wjco.v5.i3.412.

[14] R. Rouzier *et al.*, "Breast cancer molecular subtypes respond differently to preoperative chemotherapy," *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.*, vol. 11, no. 16, pp. 5678–5685, Aug. 2005, doi: 10.1158/1078-0432.CCR-04-2421.

[15] L. Carey *et al.*, "The triple negative paradox: primary tumor chemosensitivity of the basal-like breast cancer (BBC) phenotype.," Jan. 2004, pp. S48–S49.

[16] H. Kennecke *et al.*, "Metastatic behavior of breast cancer subtypes," *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.*, vol. 28, no. 20, pp. 3271–3277, Jul. 2010, doi: 10.1200/JCO.2009.25.9820.

[17] T. Sorlie *et al.*, "Repeated observation of breast tumor subtypes in independent gene expression data sets," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, no. 14, pp. 8418–8423, Jul. 2003, doi: 10.1073/pnas.0932692100.

[18] B. Weigelt, F. L. Baehner, and J. S. Reis-Filho, "The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective

of the last decade," *J. Pathol.*, vol. 220, no. 2, pp. 263–280, Jan. 2010, doi: 10.1002/path.2648.
[19] C. Sotiriou and L. Pusztai, "Gene-expression signatures in breast cancer," *N. Engl. J. Med.*, vol. 360, no. 8, pp. 790–800, Feb. 2009, doi: 10.1056/NEJMra0801289.
[20] E. F. Attiyeh *et al.*, "Chromosome arm 11q deletion predicts for neuroblastoma outcome: A Children's Oncology Group study," *J. Clin. Oncol.*, vol. 23, no. 16_suppl, pp. 6–6, Jun. 2005, doi: 10.1200/jco.2005.23.16_suppl.6.
[21] S. Paik *et al.*, "Expression of the 21 genes in the Recurrence Score assay and tamoxifen clinical benefit in the NSABP study B-14 of node negative, estrogen receptor positive breast cancer," *J. Clin. Oncol.*, vol. 23, no. 16_suppl, pp. 510–510, Jun. 2005, doi: 10.1200/jco.2005.23.16_suppl.510.
[22] L. Gianni *et al.*, "Gene expression profiles of paraffin-embedded core biopsy tissue predict response to chemotherapy in patients with locally advanced breast cancer," *J. Clin. Oncol.*, vol. 22, no. 14_suppl, pp. 501–501, Jul. 2004, doi: 10.1200/jco.2004.22.90140.501.
[23] J. S. Reis-Filho, B. Weigelt, D. Fumagalli, and C. Sotiriou, "Molecular profiling: moving away from tumor philately," *Sci. Transl. Med.*, vol. 2, no. 47, p. 47ps43, Sep. 2010, doi: 10.1126/scitranslmed.3001329.
[24] S. Loi *et al.*, "Gene expression profiling identifies activated growth factor signaling in poor prognosis (Luminal-B) estrogen receptor positive breast cancer," *BMC Med. Genomics*, vol. 2, p. 37, Jun. 2009, doi: 10.1186/1755-8794-2-37.
[25] M. C. U. Cheang *et al.*, "Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer," *J. Natl. Cancer Inst.*, vol. 101, no. 10, pp. 736–750, May 2009, doi: 10.1093/jnci/djp082.
[26] N. Turner and R. Grose, "Fibroblast growth factor signalling: from development to cancer," *Nat. Rev. Cancer*, vol. 10, no. 2, pp. 116–129, Feb. 2010, doi: 10.1038/nrc2780.
[27] X. Fu, C. K. Osborne, and R. Schiff, "Biology and therapeutic potential of PI3K signaling in ER+/HER2-negative breast cancer," *Breast Edinb. Scotl.*, vol. 22 Suppl 2, pp. S12-18, Aug. 2013, doi: 10.1016/j.breast.2013.08.001.
[28] R. Callahan and S. Hurvitz, "HER2-Positive Breast Cancer: Current Management of Early, Advanced, and Recurrent Disease," *Curr. Opin. Obstet. Gynecol.*, vol. 23, no. 1, pp. 37–43, Feb. 2011.
[29] L. A. Bazley and W. J. Gullick, "The epidermal growth factor receptor family," *Endocr. Relat. Cancer*, vol. 12, no. Supplement_1, pp. S17–S27, Jul. 2005, doi: 10.1677/erc.1.01032.
[30] M. J. Wieduwilt and M. M. Moasser, "The epidermal growth factor receptor family: biology driving targeted therapeutics," *Cell. Mol. Life Sci. CMLS*, vol. 65, no. 10, pp. 1566–1584, May 2008, doi: 10.1007/s00018-008-7440-8.
[31] S. Tsutsui, S. Ohno, S. Murakami, A. Kataoka, J. Kinoshita, and Y. Hachitanda, "Prognostic significance of the coexpression of p53 protein and c-erbB2 in breast cancer," *Am. J. Surg.*, vol. 185, no. 2, pp. 165–167, Feb. 2003, doi: 10.1016/s0002-9610(02)01203-5.
[32] E. H. Romond *et al.*, "Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer," *N. Engl. J. Med.*, vol. 353, no. 16, pp. 1673–1684, Oct. 2005, doi: 10.1056/NEJMoa052122.
[33] F. Montemurro *et al.*, "The 41st Annual Meeting of The American Society of Clinical Oncology," *Expert Opin. Pharmacother.*, vol. 6, no. 16, pp. 2907–2921, Dec. 2005, doi: 10.1517/14656566.6.16.2907.
[34] Y. Nagata *et al.*, "PTEN activation contributes to tumor inhibition by trastuzumab, and loss of PTEN predicts trastuzumab resistance in patients," *Cancer Cell*, vol. 6, no. 2, pp.

117–127, Aug. 2004, doi: 10.1016/j.ccr.2004.06.022.
[35] D. Tripathy, S. Hassan, U. Verma, P. Gurnani, A. Nandi, and K. Rosenblatt, “Phenotypic and proteomic alterations of acquired trastuzumab resistance,” *J. Clin. Oncol.*, vol. 23, no. 16_suppl, pp. 3121–3121, Jun. 2005, doi: 10.1200/jco.2005.23.16_suppl.3121.
[36] J. J. de Ronde *et al.*, “Concordance of clinical and molecular breast cancer subtyping in the context of preoperative chemotherapy response,” *Breast Cancer Res. Treat.*, vol. 119, no. 1, pp. 119–126, Jan. 2010, doi: 10.1007/s10549-009-0499-6.
[37] A. Prat and C. M. Perou, “Deconstructing the molecular portraits of breast cancer,” *Mol. Oncol.*, vol. 5, no. 1, pp. 5–23, Feb. 2011, doi: 10.1016/j.molonc.2010.11.003.
[38] J. Staaf *et al.*, “Identification of subtypes in human epidermal growth factor receptor 2--positive breast cancer reveals a gene signature prognostic of outcome,” *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.*, vol. 28, no. 11, pp. 1813–1820, Apr. 2010, doi: 10.1200/JCO.2009.22.8775.
[39] E. A. Rakha *et al.*, “Triple-Negative Breast Cancer: Distinguishing between Basal and Nonbasal Subtypes,” *Clin. Cancer Res.*, vol. 15, no. 7, pp. 2302–2310, Apr. 2009, doi: 10.1158/1078-0432.CCR-08-2132.
[40] M. van de Rijn *et al.*, “Expression of cytokeratins 17 and 5 identifies a group of breast carcinomas with poor clinical outcome,” *Am. J. Pathol.*, vol. 161, no. 6, pp. 1991–1996, Dec. 2002, doi: 10.1016/S0002-9440(10)64476-8.
[41] F. Bertucci *et al.*, “How basal are triple-negative breast cancers?,” *Int. J. Cancer*, vol. 123, no. 1, pp. 236–240, Jul. 2008, doi: 10.1002/ijc.23518.
[42] B. Kreike *et al.*, “Gene expression profiling and histopathological characterization of triple-negative/basal-like breast carcinomas,” *Breast Cancer Res. BCR*, vol. 9, no. 5, p. R65, 2007, doi: 10.1186/bcr1771.
[43] M. Jumppanen *et al.*, “Basal-like phenotype is not associated with patient survival in estrogen-receptor-negative breast cancers,” *Breast Cancer Res.*, vol. 9, no. 1, p. R16, 2007, doi: 10.1186/bcr1649.
[44] X. Dai *et al.*, “Breast cancer intrinsic subtype classification, clinical use and future trends,” *Am. J. Cancer Res.*, vol. 5, no. 10, pp. 2929–2943, Sep. 2015.
[45] J. Duraiyan, R. Govindarajan, K. Kaliyappan, and M. Palanisamy, “Applications of immunohistochemistry,” *J. Pharm. Bioallied Sci.*, vol. 4, no. Suppl 2, pp. S307–S309, Aug. 2012, doi: 10.4103/0975-7406.100281.
[46] D. C. Zaha, “Significance of immunohistochemistry in breast cancer,” *World J. Clin. Oncol.*, vol. 5, no. 3, pp. 382–392, Aug. 2014, doi: 10.5306/wjco.v5.i3.382.
[47] J. Roulson, E. W. Benbow, and P. S. Hasleton, “Discrepancies between clinical and autopsy diagnosis and the value of post mortem histology; a meta-analysis and review,” *Histopathology*, vol. 47, no. 6, pp. 551–559, Dec. 2005, doi: 10.1111/j.1365-2559.2005.02243.x.
[48] F. D. C. Bernardi, P. H. N. Saldiva, and T. Mauad, “Histological examination has a major impact on macroscopic necropsy diagnoses,” *J. Clin. Pathol.*, vol. 58, no. 12, pp. 1261–1264, Dec. 2005, doi: 10.1136/jcp.2005.027953.
[49] A. B. Effi, N. A. Aman, B. S. Koui, K. D. Koffi, Z. C. Traoré, and M. Kouyate, “Immunohistochemical determination of estrogen and progesterone receptors in breast cancer: relationship with clinicopathologic factors in 302 patients in Ivory Coast,” *BMC Cancer*, vol. 17, p. 115, Feb. 2017, doi: 10.1186/s12885-017-3105-z.
[50] A. Prat *et al.*, “Clinical implications of the intrinsic molecular subtypes of breast cancer,” *The Breast*, vol. 24, pp. S26–S35, Nov. 2015, doi: 10.1016/j.breast.2015.07.008.

[51] J. T. Painter, N. P. Clayton, and R. A. Herbert, “Useful Immunohistochemical Markers of Tumor Differentiation,” *Toxicol. Pathol.*, vol. 38, no. 1, pp. 131–141, Jan. 2010, doi: 10.1177/0192623309356449.
[52] T. Iwamoto and L. Pusztai, “Predicting prognosis of breast cancer with gene signatures: are we lost in a sea of data?,” *Genome Med.*, vol. 2, no. 11, p. 81, Nov. 2010, doi: 10.1186/gm202.
[53] X. Zhu *et al.*, “The prognostic and predictive potential of Ki-67 in triple-negative breast cancer,” *Sci. Rep.*, vol. 10, no. 1, Art. no. 1, Jan. 2020, doi: 10.1038/s41598-019-57094-3.
[54] C. S. Vallejos *et al.*, “Breast cancer classification according to immunohistochemistry markers: subtypes and association with clinicopathologic variables in a peruvian hospital database,” *Clin. Breast Cancer*, vol. 10, no. 4, pp. 294–300, Aug. 2010, doi: 10.3816/CBC.2010.n.038.
[55] S. Masood, “Breast Cancer Subtypes: Morphologic and Biologic Characterization,” *Womens Health*, vol. 12, no. 1, pp. 103–119, Jan. 2016, doi: 10.2217/whe.15.99.
[56] “Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer - PubMed.” https://pubmed.ncbi.nlm.nih.gov/16720680/ (accessed Apr. 23, 2022).
[57] “Gene expression profiling predicts clinical outcome of breast cancer | Nature.” https://www.nature.com/articles/415530a (accessed Apr. 23, 2022).
[58] M. Gnant *et al.*, “Predicting distant recurrence in receptor-positive breast cancer patients with limited clinicopathological risk: using the PAM50 Risk of Recurrence score in 1478 postmenopausal patients of the ABCSG-8 trial treated with adjuvant endocrine therapy alone,” *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.*, vol. 25, no. 2, pp. 339–345, Feb. 2014, doi: 10.1093/annonc/mdt494.
[59] C. Fan *et al.*, “Concordance among gene-expression-based predictors for breast cancer,” *N. Engl. J. Med.*, vol. 355, no. 6, pp. 560–569, Aug. 2006, doi: 10.1056/NEJMoa052933.
[60] B. D. Lehmann *et al.*, “Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies,” *J. Clin. Invest.*, vol. 121, no. 7, pp. 2750–2767, Jul. 2011, doi: 10.1172/JCI45014.
[61] M. L. Gatza *et al.*, “A pathway-based classification of human breast cancer,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 15, pp. 6994–6999, Apr. 2010, doi: 10.1073/pnas.0912708107.
[62] C. Curtis *et al.*, “The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups,” *Nature*, vol. 486, no. 7403, pp. 346–352, Apr. 2012, doi: 10.1038/nature10983.
[63] Cancer Genome Atlas Network, “Comprehensive molecular portraits of human breast tumours,” *Nature*, vol. 490, no. 7418, pp. 61–70, Oct. 2012, doi: 10.1038/nature11412.
[64] X. Dai, A. Chen, and Z. Bai, “Integrative investigation on breast cancer in ER, PR and HER2-defined subgroups using mRNA and miRNA expression profiling,” *Sci. Rep.*, vol. 4, p. 6566, Oct. 2014, doi: 10.1038/srep06566.
[65] B. Haibe-Kains *et al.*, “A Three-Gene Model to Robustly Identify Breast Cancer Molecular Subtypes,” *JNCI J. Natl. Cancer Inst.*, vol. 104, no. 4, pp. 311–325, Feb. 2012, doi: 10.1093/jnci/djr545.
[66] J. S. Parker *et al.*, “Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes,” *J. Clin. Oncol.*, vol. 27, no. 8, pp. 1160–1167, Mar. 2009, doi: 10.1200/JCO.2008.18.1370.

[67] L. Pusztai, C. Mazouni, K. Anderson, Y. Wu, and W. F. Symmans, “Molecular classification of breast cancer: limitations and potential,” *The Oncologist*, vol. 11, no. 8, pp. 868–877, Sep. 2006, doi: 10.1634/theoncologist.11-8-868.
[68] F. Andre and L. Pusztai, “Molecular classification of breast cancer: implications for selection of adjuvant chemotherapy,” *Nat. Clin. Pract. Oncol.*, vol. 3, no. 11, pp. 621–632, Nov. 2006, doi: 10.1038/ncponc0636.
[69] A. V. Kapp *et al.*, “Discovery and validation of breast cancer subtypes,” *BMC Genomics*, vol. 7, p. 231, Sep. 2006, doi: 10.1186/1471-2164-7-231.
[70] A. Mackay *et al.*, “Microarray-based class discovery for molecular classification of breast cancer: analysis of interobserver agreement,” *J. Natl. Cancer Inst.*, vol. 103, no. 8, pp. 662–673, Apr. 2011, doi: 10.1093/jnci/djr071.
[71] C. Desmedt *et al.*, “Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes,” *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.*, vol. 14, no. 16, pp. 5158–5165, Aug. 2008, doi: 10.1158/1078-0432.CCR-07-4756.
[72] P. Wirapati *et al.*, “Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures,” *Breast Cancer Res. BCR*, vol. 10, no. 4, p. R65, 2008, doi: 10.1186/bcr2124.
[73] B. Haibe-Kains, C. Desmedt, F. Rothé, M. Piccart, C. Sotiriou, and G. Bontempi, “A fuzzy gene expression-based computational approach improves breast cancer prognostication,” *Genome Biol.*, vol. 11, no. 2, p. R18, 2010, doi: 10.1186/gb-2010-11-2-r18.
[74] B. Weigelt *et al.*, “Breast cancer molecular profiling with single sample predictors: a retrospective analysis,” *Lancet Oncol.*, vol. 11, no. 4, pp. 339–349, Apr. 2010, doi: 10.1016/S1470-2045(10)70008-5.
[75] A. S. Abdelhafiz *et al.*, “Gene expression analysis of invasive breast carcinoma yields differential patterns in luminal subtypes of breast cancer,” *Ann. Diagn. Pathol.*, vol. 55, p. 151814, Dec. 2021, doi: 10.1016/j.anndiagpath.2021.151814.
[76] K. S. Albain *et al.*, “Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: a retrospective analysis of a randomised trial,” *Lancet Oncol.*, vol. 11, no. 1, pp. 55–65, Jan. 2010, doi: 10.1016/S1470-2045(09)70314-6.
[77] S. Loi *et al.*, “Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade,” *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.*, vol. 25, no. 10, pp. 1239–1246, Apr. 2007, doi: 10.1200/JCO.2006.07.1522.
[78] P.-L. Jerevall *et al.*, “Prognostic utility of HOXB13:IL17BR and molecular grade index in early-stage breast cancer patients from the Stockholm trial,” *Br. J. Cancer*, vol. 104, no. 11, pp. 1762–1769, May 2011, doi: 10.1038/bjc.2011.145.
[79] “A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer | NEJM.” https://www.nejm.org/doi/full/10.1056/nejmoa041588 (accessed Apr. 26, 2022).
[80] S. Huang, L. Murphy, and W. Xu, “Genes and functions from breast cancer signatures,” *BMC Cancer*, vol. 18, no. 1, p. 473, Apr. 2018, doi: 10.1186/s12885-018-4388-4.
[81] M. Tsai *et al.*, “Association of 70-Gene Signature Assay Findings With Physicians’ Treatment Guidance for Patients With Early Breast Cancer Classified as Intermediate Risk by the 21-Gene Assay,” *JAMA Oncol.*, vol. 4, no. 1, p. e173470, Jan. 2018, doi: 10.1001/jamaoncol.2017.3470.
[82] E. A. Slodkowska and J. S. Ross, “MammaPrint 70-gene signature: another milestone in personalized medical care for breast cancer patients,” *Expert Rev. Mol. Diagn.*, vol. 9, no. 5, pp. 417–422, Jul. 2009, doi: 10.1586/erm.09.32.

[83] F. Cardoso *et al.*, “70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer,” *N. Engl. J. Med.*, vol. 375, no. 8, pp. 717–729, Aug. 2016, doi: 10.1056/NEJMoa1602253.
[84] V. Kaklamani, “A genetic signature can predict prognosis and response to therapy in breast cancer: Oncotype DX,” *Expert Rev. Mol. Diagn.*, vol. 6, no. 6, pp. 803–809, Nov. 2006, doi: 10.1586/14737159.6.6.803.
[85] M. Toi *et al.*, “Clinical significance of the 21-gene signature (Oncotype DX) in hormone receptor-positive early stage primary breast cancer in the Japanese population,” *Cancer*, vol. 116, no. 13, pp. 3112–3118, 2010, doi: 10.1002/cncr.25206.
[86] K. Jonsdottir *et al.*, “Prognostic Value of Gene Signatures and Proliferation in Lymph-Node-Negative Breast Cancer,” *PLOS ONE*, vol. 9, no. 3, p. e90642, Mar. 2014, doi: 10.1371/journal.pone.0090642.
[87] B. M. Müller *et al.*, “The EndoPredict Gene-Expression Assay in Clinical Practice - Performance and Impact on Clinical Decisions,” *PLoS ONE*, vol. 8, no. 6, p. e68252, Jun. 2013, doi: 10.1371/journal.pone.0068252.
[88] A. C. Müller and S. Guido, *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O’Reilly Media, Inc., 2016.
[89] Y. Liu, P.-H. C. Chen, J. Krause, and L. Peng, “How to Read Articles That Use Machine Learning: Users’ Guides to the Medical Literature,” *JAMA*, vol. 322, no. 18, pp. 1806–1816, Nov. 2019, doi: 10.1001/jama.2019.16489.
[90] J. Arbet, C. Brokamp, J. Meinzen-Derr, K. E. Trinkley, and H. M. Spratt, “Lessons and tips for designing a machine learning study using EHR data,” *J. Clin. Transl. Sci.*, vol. 5, no. 1, p. e21, doi: 10.1017/cts.2020.513.
[91] I. H. Sarker, “Machine Learning: Algorithms, Real-World Applications and Research Directions,” *SN Comput. Sci.*, vol. 2, no. 3, p. 160, Mar. 2021, doi: 10.1007/s42979-021-00592-x.
[92] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Elsevier, 2011.
[93] J. G. Greener, S. M. Kandathil, L. Moffat, and D. T. Jones, “A guide to machine learning for biologists,” *Nat. Rev. Mol. Cell Biol.*, vol. 23, no. 1, Art. no. 1, Jan. 2022, doi: 10.1038/s41580-021-00407-0.
[94] H. Bhaskar, D. C. Hoyle, and S. Singh, “Machine learning in bioinformatics: A brief survey and recommendations for practitioners,” *Comput. Biol. Med.*, vol. 36, no. 10, pp. 1104–1125, Oct. 2006, doi: 10.1016/j.compbiomed.2005.09.002.
[95] S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu, “Using of Jaccard Coefficient for Keywords Similarity,” *Hong Kong*, p. 5, 2013.
[96] D. J. Rogers and T. T. Tanimoto, “A Computer Program for Classifying Plants,” vol. 132, p. 5, 1960.
[97] R. R. Sokal, “The Principles and Practice of Numerical Taxonomy,” *Taxon*, vol. 12, no. 5, pp. 190–199, 1963, doi: 10.2307/1217562.
[98] “Metric and Euclidean properties of dissimilarity coefficients | SpringerLink.” https://link.springer.com/article/10.1007/BF01896809 (accessed Apr. 26, 2022).
[99] H. N. Munro, *Mammalian Protein Metabolism: Volume III*. Elsevier, 2013.
[100] “A General Coefficient of Similarity and Some of Its Properties on JSTOR.” https://www.jstor.org/stable/2528823?seq=1 (accessed Apr. 26, 2022).
[101] “Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics.” https://projecteuclid.org/proceedings/berkeley-symposium-on-mathematical-statistics-and-probability/proceedings-of-the-fifth-berkeley-symposium-on-mathematical-statistics-and-probability-volume-1-statistics/toc/bsmsp/1200512974 (accessed Apr. 26, 2022).

[102] X. Jin and J. Han, "K-Medoids Clustering," in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, 2010, pp. 564–565. doi: 10.1007/978-0-387-30164-8_426.
[103] M. Ester, H.-P. Kriegel, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," p. 6.
[104] "OPTICS: ordering points to identify the clustering structure: ACM SIGMOD Record: Vol 28, No 2." https://dl.acm.org/doi/abs/10.1145/304181.304187 (accessed Apr. 26, 2022).
[105] S. Patel, S. Sihmar, and A. Jatain, "A study of hierarchical clustering algorithms," in *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, Mar. 2015, pp. 537–541.
[106] S. Takumi and S. Miyamoto, "Top-down vs bottom-up methods of linkage for asymmetric agglomerative hierarchical clustering," in *2012 IEEE International Conference on Granular Computing*, Aug. 2012, pp. 459–464. doi: 10.1109/GrC.2012.6468689.
[107] C. Rasmussen, "The Infinite Gaussian Mixture Model," in *Advances in Neural Information Processing Systems*, 1999, vol. 12. Accessed: Apr. 26, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/1999/hash/97d98119037c5b8a9663cb21fb8ebf47-Abstract.html
[108] G. A. Carpenter and S. Grossberg, "A massively parallel architecture for a self-organizing neural pattern recognition machine," *Comput. Vis. Graph. Image Process.*, vol. 37, no. 1, pp. 54–115, Jan. 1987, doi: 10.1016/S0734-189X(87)80014-2.
[109] I. H. Sarker, H. Alqahtani, F. Alsolami, A. I. Khan, Y. B. Abushark, and M. K. Siddiqui, "Context pre-modeling: an empirical analysis for classification based user-centric context-aware predictive modeling," *J. Big Data*, vol. 7, no. 1, p. 51, Jul. 2020, doi: 10.1186/s40537-020-00328-3.
[110] H. Liu, "Inst. of Sci. & Indus. Res. Osaka University Ibaraki, Osaka, Japan 567-0047," p. 6.
[111]C. Xu and S. A. Jackson, "Machine learning and complex biological data," *Genome Biol.*, vol. 20, no. 1, p. 76, Apr. 2019, doi: 10.1186/s13059-019-1689-0.
[112] A. Singh, N. Thakur, and A. Sharma, "A review of supervised machine learning algorithms," in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, Mar. 2016, pp. 1310–1315.
[113] M. Hossin and M. N. Sulaiman, "A Review on Evaluation Metrics for Data Classification Evaluations," *Int. J. Data Min. Knowl. Manag. Process IJDKP*, vol. 5, no. 2, pp. 1–11, Nov. 2019, doi: 10.5281/zenodo.3557376.
[114] A. Kulkarni, D. Chong, and F. A. Batarseh, "Foundations of data imbalance and solutions for a data democracy," *ArXiv210800071 Cs Stat*, Jul. 2021, Accessed: Apr. 26, 2022. [Online]. Available: http://arxiv.org/abs/2108.00071
[115] M. Grandini, E. Bagli, and G. Visani, "Metrics for Multi-Class Classification: an Overview," *ArXiv200805756 Cs Stat*, Aug. 2020, Accessed: Apr. 26, 2022. [Online]. Available: http://arxiv.org/abs/2008.05756
[116] A. Botchkarev, "Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology," *Interdiscip. J. Inf. Knowl. Manag.*, vol. 14, pp. 045–076, 2019, doi: 10.28945/4184.
[117] M. Mohammed, M. B. Khan, and E. B. M. Bashier, *Machine Learning: Algorithms and Applications*. Boca Raton: CRC Press, 2016. doi: 10.1201/9781315371658.

[118] "Cybersecurity data science: an overview from machine learning perspective | SpringerLink." https://link.springer.com/article/10.1186/s40537-020-00318-5 (accessed Apr. 26, 2022).
[119] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement Learning: A Survey," *J. Artif. Intell. Res.*, vol. 4, pp. 237–285, May 1996, doi: 10.1613/jair.301.

[120] J. D. Kelleher, B. M. Namee, and A. D'Arcy, *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. MIT Press, 2015.
[121] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data," *Front. Energy Res.*, vol. 9, 2021, Accessed: Apr. 26, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fenrg.2021.652801
[122] H. Wang, J. Tang, M. Wu, X. Wang, and T. Zhang, "Application of machine learning missing data imputation techniques in clinical decision making: taking the discharge assessment of patients with spontaneous supratentorial intracerebral hemorrhage as an example," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, p. 13, Jan. 2022, doi: 10.1186/s12911-022-01752-6.
[123] M. Ribeiro, K. Grolinger, H. F. ElYamany, W. A. Higashino, and M. A. M. Capretz, "Transfer learning with seasonal and trend adjustment for cross-building energy forecasting," *Energy Build.*, vol. 165, pp. 352–363, Apr. 2018, doi: 10.1016/j.enbuild.2018.01.034.
[124] X. Yu, S. Ergan, and G. Dedemen, "A data-driven approach to extract operational signatures of HVAC systems and analyze impact on electricity consumption," *Appl. Energy*, vol. 253, p. 113497, Nov. 2019, doi: 10.1016/j.apenergy.2019.113497.
[125] C. Fan, Y. Sun, Y. Zhao, M. Song, and J. Wang, "Deep learning-based feature engineering methods for improved building energy prediction," *Appl. Energy*, vol. 240, pp. 35–45, Apr. 2019, doi: 10.1016/j.apenergy.2019.02.052.
[126] A. Ghasemian, H. Hosseinmardi, and A. Clauset, "Evaluating Overfit and Underfit in Models of Network Community Structure," *IEEE Trans. Knowl. Data Eng.*, pp. 1–1, 2019, doi: 10.1109/TKDE.2019.2911585.
[127] B. Ghojogh and M. Crowley, "The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial," *ArXiv190512787 Cs Stat*, May 2019, Accessed: Apr. 26, 2022. [Online]. Available: http://arxiv.org/abs/1905.12787
[128] X. Ying, "An Overview of Overfitting and its Solutions," *J. Phys. Conf. Ser.*, vol. 1168, p. 022022, Feb. 2019, doi: 10.1088/1742-6596/1168/2/022022.
[129] "What is Overfitting?," Mar. 06, 2021. https://www.ibm.com/cloud/learn/overfitting (accessed Apr. 26, 2022).
[130] "The Cancer Genome Atlas Program - National Cancer Institute," Jun. 13, 2018. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga (accessed Apr. 27, 2022).
[131] J. N. Weinstein *et al.*, "The Cancer Genome Atlas Pan-Cancer Analysis Project," *Nat. Genet.*, vol. 45, no. 10, pp. 1113–1120, Oct. 2013, doi: 10.1038/ng.2764.
[132] J. Carrot-Zhang *et al.*, "Comprehensive Analysis of Genetic Ancestry and Its Molecular Correlates in Cancer," *Cancer Cell*, vol. 37, no. 5, pp. 639-654.e6, May 2020, doi: 10.1016/j.ccell.2020.04.012.
[133] G. F. Gao *et al.*, "Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data," *Cell Syst.*, vol. 9, no. 1, pp. 24-34.e10, Jul. 2019, doi: 10.1016/j.cels.2019.06.006.

[134] Y. Wang *et al.*, “Comprehensive Molecular Characterization of the Hippo Signaling Pathway in Cancer,” *Cell Rep.*, vol. 25, no. 5, pp. 1304-1317.e5, Oct. 2018, doi: 10.1016/j.celrep.2018.10.001.
[135] “The chromatin accessibility landscape of primary human cancers.” https://www.science.org/doi/10.1126/science.aav1898 (accessed Apr. 28, 2022).
[136] X. Wang and M. W. Kattan, “Cohort Studies: Design, Analysis, and Reporting,” *Chest*, vol. 158, no. 1, Supplement, pp. S72–S78, Jul. 2020, doi: 10.1016/j.chest.2020.03.014.
[137] M. D. Burstein *et al.*, “Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer,” *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.*, vol. 21, no. 7, pp. 1688–1698, Apr. 2015, doi: 10.1158/1078-0432.CCR-14-0432.
[138] C. Brueffer *et al.*, “Clinical Value of RNA Sequencing–Based Classifiers for Prediction of the Five Conventional Breast Cancer Biomarkers: A Report From the Population-Based Multicenter Sweden Cancerome Analysis Network—Breast Initiative,” *JCO Precis. Oncol.*, vol. 2, p. PO.17.00135, Mar. 2018, doi: 10.1200/PO.17.00135.
[139] Y. Zhang *et al.*, “The 76-gene signature defines high-risk patients that benefit from adjuvant tamoxifen therapy,” *Breast Cancer Res. Treat.*, vol. 116, no. 2, pp. 303–309, Jul. 2009, doi: 10.1007/s10549-008-0183-2.
[140] T. Iwamoto *et al.*, “Gene pathways associated with prognosis and chemotherapy sensitivity in molecular subtypes of breast cancer,” *J. Natl. Cancer Inst.*, vol. 103, no. 3, pp. 264–272, Feb. 2011, doi: 10.1093/jnci/djq524.
[141] Y. Li *et al.*, “Amplification of LAPTM4B and YWHAZ contributes to chemotherapy resistance and recurrence of breast cancer,” *Nat. Med.*, vol. 16, no. 2, pp. 214–218, Feb. 2010, doi: 10.1038/nm.2090.
[142] W. F. Symmans *et al.*, “Genomic index of sensitivity to endocrine therapy for breast cancer,” *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.*, vol. 28, no. 27, pp. 4111–4119, Sep. 2010, doi: 10.1200/JCO.2010.28.4273.
[143] C. Desmedt *et al.*, “Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series,” *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.*, vol. 13, no. 11, pp. 3207–3214, Jun. 2007, doi: 10.1158/1078-0432.CCR-06-2765.
[144] Y. Wang *et al.*, “Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer,” *Lancet Lond. Engl.*, vol. 365, no. 9460, pp. 671–679, Feb. 2005, doi: 10.1016/S0140-6736(05)17947-1.
[145] S. Jindal *et al.*, “Postpartum breast cancer has a distinct molecular profile that predicts poor outcomes,” *Nat. Commun.*, vol. 12, no. 1, p. 6341, Nov. 2021, doi: 10.1038/s41467-021-26505-3.
[146] K. E. Varley *et al.*, “Recurrent read-through fusion transcripts in breast cancer,” *Breast Cancer Res. Treat.*, vol. 146, no. 2, pp. 287–297, Jul. 2014, doi: 10.1007/s10549-014-3019-2.
[147] L. Kester *et al.*, “Differential Survival and Therapy Benefit of Patients with Breast Cancer Are Characterized by Distinct Epithelial and Immune Cell Microenvironments,” *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.*, vol. 28, no. 5, pp. 960–971, Mar. 2022, doi: 10.1158/1078-0432.CCR-21-1442.
[148] V. Popovici *et al.*, “Effect of training-sample size and classification difficulty on the accuracy of genomic predictors,” *Breast Cancer Res. BCR*, vol. 12, no. 1, p. R5, 2010, doi: 10.1186/bcr2468.

[149] Y. Pawitan *et al.*, “Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts,” *Breast Cancer Res. BCR*, vol. 7, no. 6, pp. R953-964, 2005, doi: 10.1186/bcr1325.
[150] K. Shen *et al.*, “A Systematic Evaluation of Multi-Gene Predictors for the Pathological Response of Breast Cancer Patients to Chemotherapy,” *PLOS ONE*, vol. 7, no. 11, p. e49529, Nov. 2012, doi: 10.1371/journal.pone.0049529.

[151] K. Shen *et al.*, “Cell line derived multi-gene predictor of pathologic response to neoadjuvant chemotherapy in breast cancer: a validation study on US Oncology 02-103 clinical trial,” *BMC Med. Genomics*, vol. 5, p. 51, Nov. 2012, doi: 10.1186/1755-8794-5-51.
[152] L. D. Miller *et al.*, “An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 38, pp. 13550–13555, Sep. 2005, doi: 10.1073/pnas.0506230102.
[153] M. Schmidt *et al.*, “The humoral immune system has a key prognostic impact in node-negative breast cancer,” *Cancer Res.*, vol. 68, no. 13, pp. 5405–5413, Jul. 2008, doi: 10.1158/0008-5472.CAN-07-5206.
[154] C. K. Anders *et al.*, “Breast Carcinomas Arising at a Young Age: Unique Biology or a Surrogate for Aggressive Intrinsic Subtypes?,” *J. Clin. Oncol.*, vol. 29, no. 1, pp. e18–e20, Jan. 2011, doi: 10.1200/JCO.2010.28.9199.
[155] R. Sabatier *et al.*, “A gene expression signature identifies two prognostic subgroups of basal breast cancer,” *Breast Cancer Res. Treat.*, vol. 126, no. 2, pp. 407–420, Apr. 2011, doi: 10.1007/s10549-010-0897-9.
[156] C. Sotiriou *et al.*, “Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis,” *J. Natl. Cancer Inst.*, vol. 98, no. 4, pp. 262–272, Feb. 2006, doi: 10.1093/jnci/djj052.
[157] A. V. Ivshina *et al.*, “Genetic Reclassification of Histologic Grade Delineates New Clinical Subtypes of Breast Cancer,” *Cancer Res.*, vol. 66, no. 21, pp. 10292–10301, Nov. 2006, doi: 10.1158/0008-5472.CAN-05-4414.
[158] C. Hatzis *et al.*, “A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer,” *JAMA*, vol. 305, no. 18, pp. 1873–1881, May 2011, doi: 10.1001/jama.2011.593.
[159] W. Chang *et al.*, *shiny: Web Application Framework for R*. 2021. Accessed: Apr. 27, 2022. [Online]. Available: https://CRAN.R-project.org/package=shiny
[160] “Shiny.” https://shiny.rstudio.com/ (accessed Apr. 27, 2022).
[161] T. Manouchehri and A. R. Nematollahi, “PAR(1) model analysis: a web-based shiny application for analysing periodic autoregressive models,” *J. Stat. Comput. Simul.*, vol. 0, no. 0, pp. 1–22, Jan. 2022, doi: 10.1080/00949655.2021.2021528.
[162] H. Syed, “MOPower: an R-shiny application for the simulation and power calculation of multi-omics studies,” p. 14.
[163] D. Fanidis, P. Moulos, and V. Aidinis, “Fibromine is a multi-omics database and mining tool for target discovery in pulmonary fibrosis,” *Sci. Rep.*, vol. 11, no. 1, Art. no. 1, Nov. 2021, doi: 10.1038/s41598-021-01069-w.
[164] B. Jagla *et al.*, “SCHNAPPs - Single Cell sHiNy APPlication(s),” *J. Immunol. Methods*, vol. 499, p. 113176, Dec. 2021, doi: 10.1016/j.jim.2021.113176.
[165] T. Evrenoglou, I. Boutron, and A. Chaimani, “metaCOVID: An R-Shiny application for living meta-analyses of COVID-19 trials.” medRxiv, p. 2021.09.07.21263207, Sep. 10, 2021. doi: 10.1101/2021.09.07.21263207.

[166] E. Germani, H. Lelouard, and M. Fallet, “SAPHIR: a Shiny application to analyze tissue section images.” F1000Research, Apr. 08, 2021. doi: 10.12688/f1000research.27062.2.
[167] S. Durinck, W. Huber, S. Davis, F. Pepin, V. S. Buffalo, and M. Smith, *biomaRt: Interface to BioMart databases (i.e. Ensembl)*. Bioconductor version: Release (3.14), 2022. doi: 10.18129/B9.bioc.biomaRt.
[168] Y. Chen *et al.*, *edgeR: Empirical Analysis of Digital Gene Expression Data in R*. Bioconductor version: Release (3.14), 2022. doi: 10.18129/B9.bioc.edgeR.
[169] M. Love *et al.*, *DESeq2: Differential gene expression analysis based on the negative binomial distribution*. Bioconductor version: Release (3.14), 2022. doi: 10.18129/B9.bioc.DESeq2.
[170] D. M. A. Gendoo *et al.*, *genefu: Computation of Gene Expression-Based Signatures in Breast Cancer*. Bioconductor version: Release (3.14), 2022. doi: 10.18129/B9.bioc.genefu.
[171] E. Luporsi *et al.*, “Ki-67: level of evidence and methodological considerations for its role in the clinical management of breast cancer: analytical and critical review,” *Breast Cancer Res. Treat.*, vol. 132, no. 3, pp. 895–915, Apr. 2012, doi: 10.1007/s10549-011-1837-z.
[172] R. L. Grossman *et al.*, “Toward a Shared Vision for Cancer Genomic Data,” *N. Engl. J. Med.*, vol. 375, no. 12, pp. 1109–1112, Sep. 2016, doi: 10.1056/NEJMp1607591.
[173] A. Colaprico *et al.*, *TCGAbiolinks: TCGAbiolinks: An R/Bioconductor package for integrative analysis with GDC data*. Bioconductor version: Release (3.14), 2022. doi: 10.18129/B9.bioc.TCGAbiolinks.
[174] J. T. L. <jtleek at gmail.com> *et al.*, *sva: Surrogate Variable Analysis*. Bioconductor version: Release (3.14), 2022. doi: 10.18129/B9.bioc.sva.
[175] C. Kandoth *et al.*, *mskcc/vcf2maf: vcf2maf v1.6.16*. Zenodo, 2018. doi: 10.5281/zenodo.1185418.
[176] Z. Gu, *ComplexHeatmap: Make Complex Heatmaps*. Bioconductor version: Release (3.14), 2022. doi: 10.18129/B9.bioc.ComplexHeatmap.
[177] M. Kuhn [aut *et al.*, *caret: Classification and Regression Training*. 2022. Accessed: Apr. 27, 2022. [Online]. Available: https://CRAN.R-project.org/package=caret
[178] Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). *limma powers differential expression analyses for RNA-sequencing and microarray studies*. Nucleic Acids Research 43(7), e47.
[179] Wang Y, Klijn JG, Zhang Y, Sieuwerts AM et al. *Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.* Lancet 2005 Feb 19-25;365(9460):671-9. PMID: 15721472.
[180] Kester L, Seinstra D, van Rossum AGJ, Vennin C et al. *Differential Survival and Therapy Benefit of Patients with Breast Cancer Are Characterized by Distinct Epithelial and Immune Cell Microenvironments.* Clin Cancer Res 2022 Mar 1;28(5):960-971. PMID: 34965952.
[181] Schroeder M, Haibe-Kains B, Culhane A, Sotiriou C, Bontempi G, Quackenbush J (2022). *breastCancerNKI: Genexpression dataset published by van't Veer et al. [2002] and van de Vijver et al. [2002] (NKI).*. R package version 1.34.0, http://compbio.dfci.harvard.edu/.
[182] B. Tran and P. Bedard, "Luminal-B breast cancer and novel therapeutic targets", 2022.

[183] H. S, "Basal-Like Breast Cancer: Main Features of a Very Particular Entity of Breast Cancer", *Clinmedjournals.org*, 2022. [Online]. Available: https://clinmedjournals.org/articles/cmrcr/clinical-medical-reviews-and-case-reports-cmrcr-6-272.php?jid=cmrcr#:~:text=This%20subtype%20of%20breast%20cancer,validated%20systemic%20therapy%20is%20chemotherapy. [Accessed: 17- May- 2022].

[184] F. Schettiniab and A. Prat, "Dissecting the biological heterogeneity of HER2-positive breast cancer", *Elsevier*, 2021. [Online]. Available: http://sciencedirect.com/science/article/pii/S0960977621004276. [Accessed: 17- May- 2022].

[185] Weigelt B, Mackay A, A'Hern R, et al. Breast cancer molecular profiling with single sample predictors: a retrospective analysis. *Lancet Oncol.* 2010;11(4):339–349.

[186] R. Mohammed, J. Rawashdeh and M. Abdullah, "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results," *2020 11th International Conference on Information and Communication Systems (ICICS)*, 2020, pp. 243-248, doi: 10.1109/ICICS49469.2020.239556.

[187] Lauron, Maureen Lyndel C., and Jaderick P. Pabico. "Improved sampling techniques for learning an imbalanced data set." *arXiv preprint arXiv:1601.04756* (2016).

[188] Bach, Małgorzata, Aleksandra Werner, and Mateusz Palt. "The proposal of undersampling method for learning from imbalanced datasets." *Procedia Computer Science* 159 (2019): 125-134.

[189] Compilation of TCGA molecular subtypes", *Bioconductor.org*, 2022. [Online]. Available:https://bioconductor.org/packages/devel/bioc/vignettes/TCGAbiolinks/inst/doc/subtypes.html. [Accessed: 18- May- 2022].

[190] Markus Schroeder, Benjamin Haibe-Kains, Aedin Culhane, Christos Sotiriou, Gianluca Bontempi and John Quackenbush (2021). breastCancerMAINZ: Gene expression dataset published by Schmidt et al. [2008] (MAINZ).. R package version 1.32.0. http://compbio.dfci.harvard.edu/

[191] Markus Schroeder, Benjamin Haibe-Kains, Aedin Culhane, Christos Sotiriou, Gianluca Bontempi and John Quackenbush (2021). breastCancerTRANSBIG: Gene expression dataset published by Desmedt et al. [2007] (TRANSBIG).. R package version 1.32.0. http://compbio.dfci.harvard.edu/

[192] Markus Schroeder, Benjamin Haibe-Kains, Aedin Culhane, Christos Sotiriou, Gianluca Bontempi and John Quackenbush (2021). breastCancerUPP: Gene expression dataset published by Miller et al. [2005] (UPP).. R package version 1.32.0. http://compbio.dfci.harvard.edu/

[193] Markus Schroeder, Benjamin Haibe-Kains, Aedin Culhane, Christos Sotiriou, Gianluca Bontempi and John Quackenbush (2021). breastCancerUNT: Gene expression dataset published by Sotiriou et al. [2007] (UNT).. R package version 1.32.0. http://compbio.dfci.harvard.edu/

[194] Markus Schroeder, Benjamin Haibe-Kains, Aedin Culhane, Christos Sotiriou, Gianluca Bontempi and John Quackenbush (2021). breastCancerNKI: Genexpression dataset published by van't Veer et al. [2002] and van de Vijver et al. [2002] (NKI).. R package version 1.32.0. http://compbio.dfci.harvard.edu/

[195] Ying, Xue. "An overview of overfitting and its solutions." *Journal of Physics: Conference Series*. Vol. 1168. No. 2. IOP Publishing, 2019.

[196] Hanahan, Douglas, and Robert A. Weinberg. "The hallmarks of cancer." *cell* 100.1 (2000): 57-70.

[197] Jew, Brandon, et al. "Accurate estimation of cell composition in bulk expression through robust integration of single-cell information." *Nature communications* 11.1 (2020): 1-11.
[198] Ektefaie, Yasha, et al. "Integrative multiomics-histopathology analysis for breast cancer classification." *NPJ breast cancer* 7.1 (2021): 1-6.
[199] List, Markus, et al. "Classification of breast cancer subtypes by combining gene expression and DNA methylation data." *Journal of integrative bioinformatics* 11.2 (2014): 1-14.
[200] Daemen, Anneleen, et al. "Modeling precision treatment of breast cancer." *Genome biology* 14.10 (2013): 1-14.
[201] Tangirala, Suryakanthi. "Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm." *International Journal of Advanced Computer Science and Applications* 11.2 (2020): 612-619.

[202] Ngoma, Yves Matanga. *Analysis of Control Attainment in Endogenous Electroencephalogram Based Brain Computer Interfaces*. Diss. Tshwane University of Technology, 2017.
[203] McCoubrey, Laura E., et al. "Harnessing machine learning for development of microbiome therapeutics." *Gut Microbes* 13.1 (2021): 1872323.
[204] Belkin, Mikhail, et al. "Reconciling modern machine-learning practice and the classical bias–variance trade-off." *Proceedings of the National Academy of Sciences* 116.32 (2019): 15849-15854.
[205] Sasa M, Bando Y, Takahashi M, Hirose T, Nagao T. Screening for basal marker expression is necessary for decision of therapeutic strategy for triple-negative breast cancer. J Surg Oncol. 2008 Jan 1;97(1):30-4. doi: 10.1002/jso.20906. PMID: 17929254.
[206] Razzaghi H, Troester MA, Gierach GL, Olshan AF, Yankaskas BC, Millikan RC. Association between mammographic density and basal-like and luminal A breast cancer subtypes. Breast Cancer Res. 2013;15(5):R76. doi: 10.1186/bcr3470. PMID: 24008056; PMCID: PMC3978452.
[207] Alshareeda AT, Soria D, Garibaldi JM, Rakha E, Nolan C, Ellis IO, Green AR. Characteristics of basal cytokeratin expression in breast cancer. Breast Cancer Res Treat. 2013 May;139(1):23-37. doi: 10.1007/s10549-013-2518-x. Epub 2013 Apr 16. PMID: 23588953.
[208] Kim NH, Lim HY, Im KS, Kim JH, Sur JH. Identification of triple-negative and basal-like canine mammary carcinomas using four basal markers. J Comp Pathol. 2013 May;148(4):298-306. doi: 10.1016/j.jcpa.2012.08.009. Epub 2012 Oct 15. PMID: 23079102.
[209] Sung, H., Garcia-Closas, M., Chang-Claude, J., Blows, F. M., Ali, H. R., Figueroa, J., Nevanlinna, H., Fagerholm, R., Heikkilä, P., Blomqvist, C., Giles, G. G., Milne, R. L., Southey, M. C., McLean, C., Mannermaa, A., Kosma, V. M., Kataja, V., Sironen, R., Couch, F. J., Olson, J. E., … Yang, X. R. (2016). Heterogeneity of luminal breast cancer characterised by immunohistochemical expression of basal markers. *British journal of cancer*, *114*(3), 298–304. https://doi.org/10.1038/bjc.2015.437
[210] Raj-Kumar, Praveen-Kumar et al. “PCA-PAM50 improves consistency between breast cancer intrinsic and clinical subtyping reclassifying a subset of luminal A tumors as luminal B.” *Scientific reports* vol. 9,1 7956. 28 May. 2019, doi:10.1038/s41598-019-44339-4
[211] A. Godoy-Ortiz *et al.*, “Deciphering HER2 Breast Cancer Disease: Biological and Clinical Implications,” *Frontiers in Oncology*, vol. 9, 2019, doi: 10.3389/fonc.2019.01124.

[212] Zait, Mohamed, and Hammou Messatfa. "A comparative study of clustering methods." *Future Generation Computer Systems* 13.2-3 (1997): 149-159.
[213] Fraley, Chris, and Adrian E. Raftery. "How many clusters? Which clustering method? Answers via model-based cluster analysis." *The computer journal* 41.8 (1998): 578-588.
[214] Tran B, Bedard PL. Luminal-B breast cancer and novel therapeutic targets. *Breast Cancer Res*. 2011;13(6):221. doi:10.1186/bcr2904
[215] Feeley, L., Mulligan, A., Pinnaduwage, D. *et al.* Distinguishing luminal breast cancer subtypes by Ki67, progesterone receptor or TP53 status provides prognostic information. *Mod Pathol* 27, 554–561 (2014). https://doi.org/10.1038/modpathol.2013.153
[215] Vieira, André Filipe, and Fernando Schmitt. "An update on breast cancer multigene prognostic tests—emergent clinical biomarkers." *Frontiers in medicine* (2018): 248.
[216] Xie, Longxiang, et al. "Systematic review of prognostic gene signature in gastric cancer patients." *Frontiers in Bioengineering and Biotechnology* (2020): 805.
[217] Cross, G. R., and A. K. Jain. "Measurement of clustering tendency." *Theory and Application of Digital Control*. Pergamon, 1982. 315-320.
[218] Starczewski, Artur, and Adam Krzyżak. "A modification of the silhouette index for the improvement of cluster validity assessment." *International Conference on Artificial Intelligence and Soft Computing*. Springer, Cham, 2016.
[219] Golalipour, Keyvan, et al. "From clustering to clustering ensemble selection: A review." *Engineering Applications of Artificial Intelligence* 104 (2021): 104388.
[220] Bastien, R.R., Rodríguez-Lescure, Á., Ebbert, M.T. *et al.* PAM50 Breast Cancer Subtyping by RT-qPCR and Concordance with Standard Clinical Molecular Markers. *BMC Med Genomics* 5, 44 (2012). https://doi.org/10.1186/1755-8794-5-44.
[221] Larsen, Martin J et al. “Classifications within molecular subtypes enables identification of BRCA1/BRCA2 mutation carriers by RNA tumor profiling.” *PloS one* vol. 8,5 e64268. 21 May. 2013, doi:10.1371/journal.pone.0064268.
[222] Kim, H. K., Park, K. H., Kim, Y., Park, S. E., Lee, H. S., Lim, S. W., Cho, J. H., Kim, J. Y., Lee, J. E., Ahn, J. S., Im, Y. H., Yu, J. H., & Park, Y. H. (2019). Discordance of the PAM50 Intrinsic Subtypes Compared with Immunohistochemistry-Based Surrogate in Breast Cancer Patients: Potential Implication of Genomic Alterations of Discordance.” *Cancer research and treatment* vol. 51,2 (2019): 737-747. doi:10.4143/crt.2018.342.
[223] Daemen, A., Manning, G. HER2 is not a cancer subtype but rather a pan-cancer event and is highly enriched in AR-driven breast tumors. *Breast Cancer Res* 20, 8 (2018). https://doi.org/10.1186/s13058-018-0933-y.
[224] Szymiczek, Agata, Amna Lone, and Mohammad R. Akbari. "Molecular intrinsic versus clinical subtyping in breast cancer: A comprehensive review." *Clinical Genetics* 99.5 (2021): 613-637.
[225] Viale, G., de Snoo, F.A., Slaets, L. *et al.* Immunohistochemical versus molecular (BluePrint and MammaPrint) subtyping of breast carcinoma. Outcome results from the EORTC 10041/BIG 3-04 MINDACT trial. *Breast Cancer Res Treat* 167, 123–131 (2018). https://doi.org/10.1007/s10549-017-4509-9.
[226] Yang XR, Sherman ME, Rimm DL, Lissowska J, Brinton LA, Peplonska B, Hewitt SM, Anderson WF, Szeszenia-Dabrowska N, Bardin-Mikolajczak A, Zatonski W, Cartun R, Mandich D, Rymkiewicz G, Ligaj M, Lukaszek S, Kordek R, García-Closas M. Differences in risk factors for breast cancer molecular subtypes in a population-based study. Cancer Epidemiol Biomarkers Prev. 2007 Mar;16(3):439-43. doi: 10.1158/1055-9965.EPI-06-0806. PMID: 17372238.

[227] Del Casar JM, Martín A, García C, Corte MD, Alvarez A, Junquera S, González LO, Bongera M, García-Muñiz JL, Allende MT, Vizoso F. Characterization of breast cancer subtypes by quantitative assessment of biological parameters: relationship with clinicopathological characteristics, biological features and prognosis. Eur J Obstet Gynecol Reprod Biol. 2008 Dec;141(2):147-52. doi: 10.1016/j.ejogrb.2008.07.021. Epub 2008 Sep 2. PMID: 18768247.

[228] Behravan, H., Hartikainen, J.M., Tengström, M. *et al.* Predicting breast cancer risk using interacting genetic and demographic factors and machine learning. *Sci Rep* 10, 11044 (2020). https://doi.org/10.1038/s41598-020-66907-9

[229] Reza Rabiei; Seyed Mohammad Ayyoubzadeh; Solmaz Sohrabei; Marzieh Esmaeili; Alireza Atashi. "Prediction of Breast Cancer using Machine Learning Approaches". *Journal of Biomedical Physics and Engineering*, 12, 3, 2022, 297-308. doi: 10.31661/jbpe.v0i0.2109-1403

[230] Yao, Jia, Shengwei Li, and Xiaosheng Wang. "Identification of Breast Cancer Immune Subtypes by Analyzing Bulk Tumor and Single Cell Transcriptomes." *Frontiers in Cell and Developmental Biology* 9 (2021): 781848-781848.

[231] Kanehisa, M. and Goto, S.; KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30 (2000).

[232] Paplomata, Elisavet, and Ruth O'Regan. “The PI3K/AKT/mTOR pathway in breast cancer: targets, trials and biomarkers.” *Therapeutic advances in medical oncology* vol. 6,4 (2014): 154-66. doi:10.1177/1758834014530023

[233] Khoury, Katia, et al. "Prevalence of Phosphatidylinositol-3-Kinase (PI3K) pathway alterations and co-alteration of other molecular markers in breast cancer." *Frontiers in oncology*(2020): 1475.

[234] Shiovitz, S, and L A Korde. “Genetics of breast cancer: a topic in evolution.” *Annals of oncology : official journal of the European Society for Medical Oncology* vol. 26,7 (2015): 1291-9. doi:10.1093/annonc/mdv022

[234] McHugh, Mary L. “Interrater reliability: the kappa statistic.” *Biochemia medica* vol. 22,3 (2012): 276-82.

[235] Chen, L., Liu, S. & Tao, Y. Regulating tumor suppressor genes: post-translational modifications. *Sig Transduct Target Ther* 5, 90 (2020). https://doi.org/10.1038/s41392-020-0196-9

[236] Holstein, Elisa et al. “The Burden of Post-Translational Modification (PTM)-Disrupting Mutations in the Tumor Matrisome.” *Cancers*vol. 13,5 1081. 3 Mar. 2021, doi:10.3390/cancers13051081

8. Appendix

Appendix Table 1. The list of PAM50 genes and the centroids.

	Basal	Her2	LumA	LumB	Normal-like
ACTR3B	0.718331891	-0.481665675	0.00998107	-0.190551328	0.465722871
ANLN	0.537372301	0.266931609	-0.579245716	0.098804179	-0.836939593
BAG1	-0.574506867	-0.476072868	0.758221161	-0.405458622	0.316552973
BCL2	-0.11876043	-0.157913959	0.28748744	-0.441339498	0.533978871
BIRC5	0.300488641	0.405733099	-0.881434366	0.603850777	-0.876636424
BLVRA	-0.642677513	0.335336041	0.042042017	0.691204962	-0.163412812
CCNB1	0.191208143	0.135476652	-0.491662114	0.503176358	-0.545269312
CCNE1	0.560271028	0.066872232	-0.430291227	-0.01666143	-0.255476058
CDC20	0.399695242	0.00835552	-0.46904401	-0.070412466	-0.04550481
CDC6	0.159418279	0.58900682	-0.612824305	0.510895969	-0.595752175
CDCA1	0.472400168	-0.023819207	-0.712520819	0.589626883	-0.370533365
CDH3	0.508362012	0.210889692	-0.513649344	-1.419134437	0.757920624
CENPF	0.482976288	-0.02926616	-0.543740234	0.278228556	-0.070583075
CEP55	0.567748894	0.276381022	-0.746721735	0.460015762	-1.162374186
CXXC5	-0.920385813	-0.241550612	0.467411571	0.32133502	0.050901436
EGFR	-0.030416849	-0.096382621	0.009162963	-0.412401259	0.341637082
ERBB2	-0.80835398	1.759844231	0.608191264	0.159651874	-0.870238456
ESR1	-2.746513086	-1.513111253	2.161411882	1.605899914	-0.418282349
EXO1	0.428090356	0.049297194	-0.567474505	0.141241282	-0.450780537
FGFR4	-0.271238025	0.821778152	0.170811925	-0.247036038	0.857472777
FOXA1	-2.626946721	0.022827151	1.017457421	0.360757795	-0.782812106
FOXC1	1.490451472	-0.947174192	-0.17495796	-1.564854964	1.111547864
GPR160	-1.054974672	0.583194826	0.685489973	0.714407601	-0.423568467
GRB7	-0.276128586	1.03065778	0.041568986	0.087750887	0.241710991
KIF2C	0.20357258	-0.165102048	-0.505394668	-0.182890713	-0.390014484
KNTC2	0.600356167	0.042546792	-0.588220989	0.386706844	-1.06962886
KRT14	0.096826723	-0.443646142	0.368375943	-0.639446966	1.73568631
KRT17	0.482565528	-0.337837101	0.014209862	-1.463742934	1.759598437
KRT5	0.506640416	-0.428261778	0.215320068	-0.91160727	1.785116895

MAPT	-0.425829273	-0.357506541	0.700622718	-0.190340574	0.117828498
MDM2	-0.251366205	-0.106728681	0.14195743	-0.133779037	0.274214011
MELK	0.523033872	0.198013115	-0.582088108	0.44793463	-0.743764676
MIA	1.578276368	-0.904898621	-0.165258584	-1.422926267	2.038859556
MKI67	0.476537448	0.06566236	-0.501871622	-0.145217868	-0.166004063
MLPH	-0.339972459	-0.195228658	0.339304418	-0.456149915	0.750758365
MMP11	-0.556037672	0.50675876	-0.00625509	0.334199309	-2.326985119
MYBL2	0.389893452	0.20526358	-0.843569931	0.46728199	-0.601704754
MYC	0.178763812	-1.046832832	-0.090830821	0.015264397	1.029176203
NAT1	-0.936848946	-0.089988492	2.922786792	0.470788042	-0.363273764
ORC6L	0.216304797	0.204402449	-0.352220667	0.11062765	-0.255879493
PGR	-0.429133389	-0.279409916	0.445785003	-0.448839844	0.126011482
PHGDH	0.63451887	-0.186625862	-0.398682234	-1.030139318	0.660437753
PTTG1	0.264131894	0.055809895	-0.63446827	0.249725281	-0.54978126
RRM2	0.156204676	0.682724889	-0.9507602	0.35066384	-1.12105493
SFRP1	0.987988459	-1.048202667	0.131566364	-1.720458262	2.436288668
SLC39A6	-1.051125052	-0.695736457	2.061459075	1.653303022	0.116889694
TMEM45B	-1.109458181	1.330636172	0.446242045	0.375688226	0.036208914
TYMS	0.449800897	0.052944897	-0.644602075	0.492606521	-0.726989454
UBE2C	0.218534147	0.061080598	-0.519818399	0.292799306	-0.408894685
UBE2T	0.389908898	0.284536813	-0.539259391	0.738952133	-0.952381005

Appendix Table 2. The concordance of subtypes from each method. *(IHC_PAM50: (IHC versus PAM50 subtypes; IHC_SCMOD1: IHC versus SCMOD1 subtypes; PAM50 versus SCMOD1 subtypes.) (if the pair gives the same subtype it is annotated as 1 and if the pair gives the different subtypes it is annotated as 0. And rowtotal is the sum of all columns).*

	IHC_PAM50	IHC_SCMOD1	PAM50_SCMOD1	rowtotal
RSM1175101	1	1	1	3
RSM1175107	1	1	1	3
RSM1175116	1	1	1	3
RSM487370	1	1	1	3
RSM487372	0	1	0	1
RSM487373	0	1	0	1
RSM487376	1	1	1	3
RSM487379	0	0	1	1
RSM487385	0	0	1	1
RSM487387	1	0	0	1
RSM487392	0	0	0	0
RSM487393	0	0	0	0
RSM487395	1	1	1	3
RSM487402	0	1	0	1
RSM487404	0	0	1	1
RSM487405	1	1	1	3
RSM487409	1	1	1	3
RSM487411	0	1	0	1
RSM487416	0	0	1	1
RSM487419	1	0	0	1
RSM487424	0	0	0	0
RSM487427	1	1	1	3
RSM487431	1	0	0	1
RSM487434	1	1	1	3
RSM487438	0	1	0	1
RSM487439	0	0	1	1
RSM487444	0	0	1	1

RSM487446	1	1	1	3
RSM487449	1	1	1	3
RSM487451	0	0	0	0
RSM487453	1	1	1	3
RSM487455	1	1	1	3
RSM521481	0	0	0	0
RSM521483	1	1	1	3
RSM521485	1	1	1	3
RSM521487	1	1	1	3
RSM521489	1	1	1	3
RSM521491	1	1	1	3
RSM521493	1	1	1	3
RSM521501	1	1	1	3
RSM521503	1	1	1	3
RSM521505	1	1	1	3
RSM521517	0	0	1	1
RSM521521	1	1	1	3
RSM521527	1	1	1	3
RSM521529	0	1	0	1
RSM521531	0	0	1	1
RSM521535	0	0	1	1
RSM521539	1	1	1	3
RSM521541	0	1	0	1
RSM521543	1	1	1	3
RSM521545	1	1	1	3
RSM521547	1	1	1	3
RSM521549	0	0	0	0
RSM521551	1	1	1	3
RSM521557	1	0	0	1
RSM521559	1	1	1	3
RSM521561	1	1	1	3
RSM521565	1	1	1	3
RSM521573	1	1	1	3

RSM521579	1	1	1	3
RSM521583	1	1	1	3
RSM521585	1	1	1	3
RSM521587	1	1	1	3
RSM521595	1	1	1	3
RSM521597	0	1	0	1
RSM521599	1	0	0	1
RSM521605	1	1	1	3
RSM521617	1	0	0	1
RSM521619	1	0	0	1
RSM521625	0	1	0	1
RSM521627	1	1	1	3
RSM521629	1	1	1	3
RSM521631	1	1	1	3
RSM521633	1	1	1	3
RSM521635	0	1	0	1
RSM521641	1	1	1	3
RSM521643	1	0	0	1
RSM521645	1	1	1	3
RSM521647	0	0	1	1
RSM521651	1	0	0	1
RSM521657	1	1	1	3
RSM521661	0	0	1	1
RSM521665	1	1	1	3
RSM521669	1	0	0	1
RSM521671	0	0	1	1
RSM521673	0	1	0	1
RSM521675	0	1	0	1
RSM521679	0	1	0	1
RSM521687	1	0	0	1

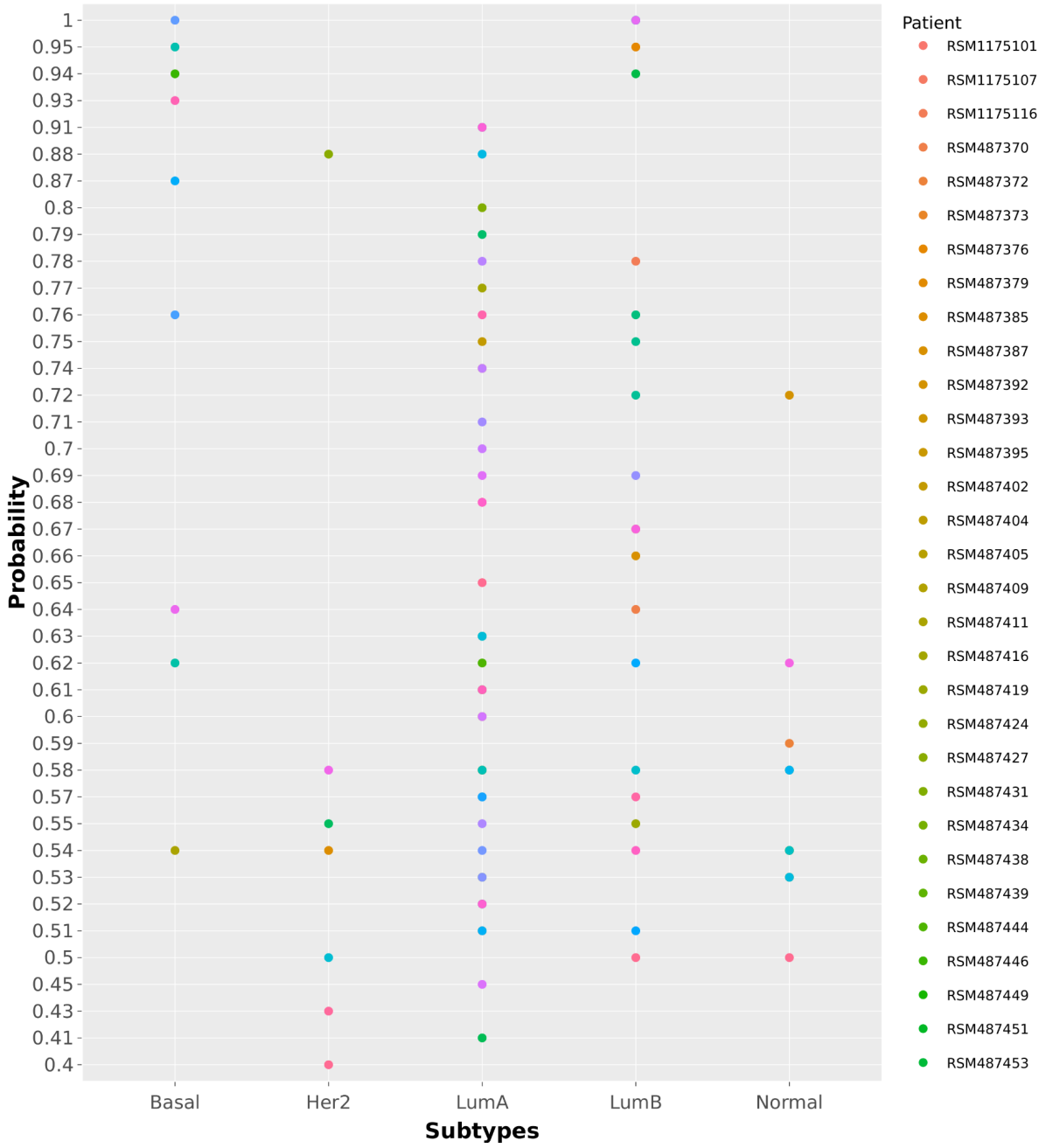
Appendix Table 3. The list of genes selected with feature importance sorted by the importance.

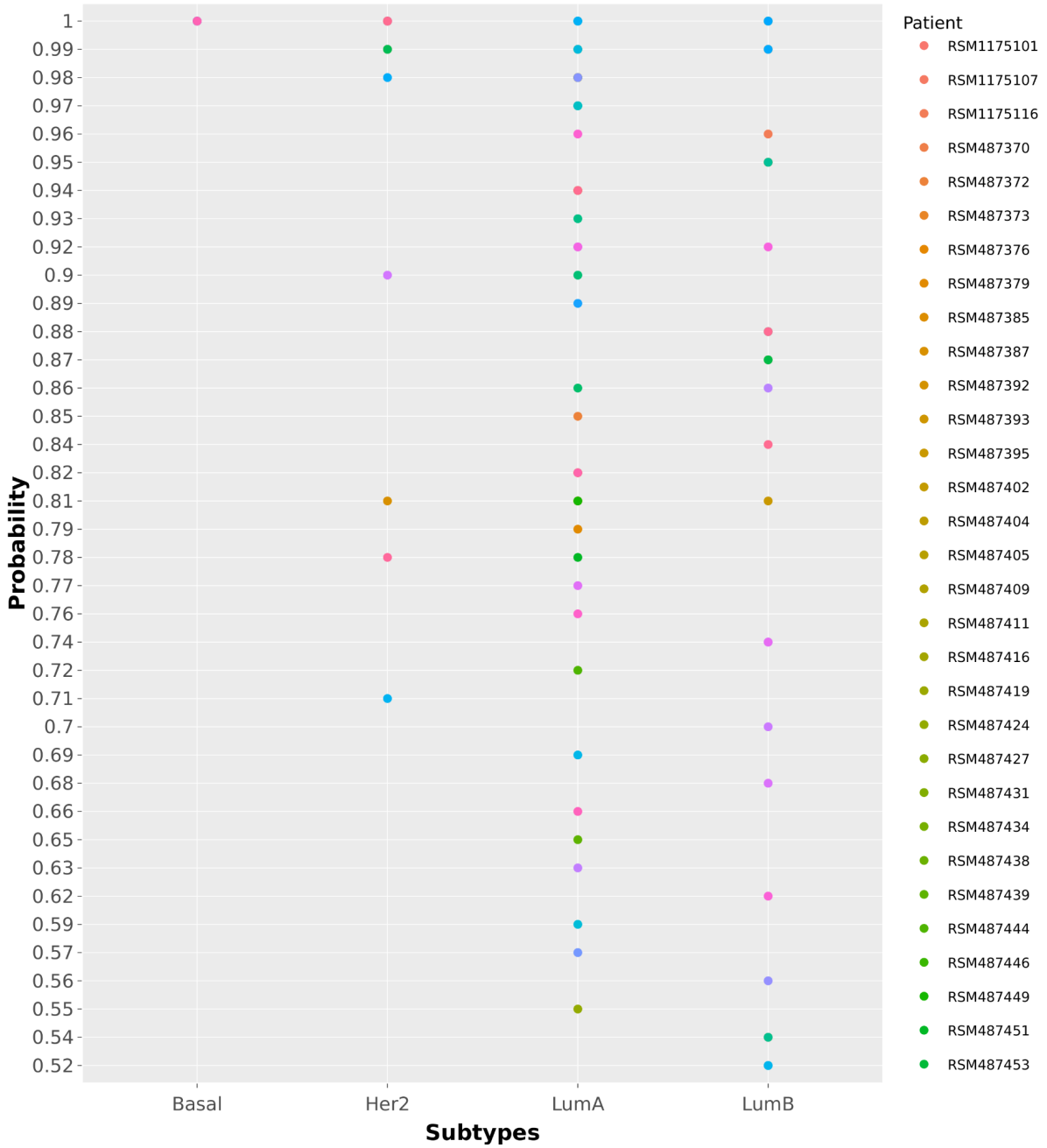
ESR1
FOXA1
MLPH
FOXC1
CEP55
UBE2T
SFRP1
EGFR
BIRC5
UBE2C
MMP11
KRT14
CDC20
EXO1
KRT5
MIA
KRT17
MAPT
PGR
KNTC2
NAT1
MYBL2
GBR7
ERBB2
KIF2C

Appendix Table 4. Genes found in each signature used in this thesis and classABCC.

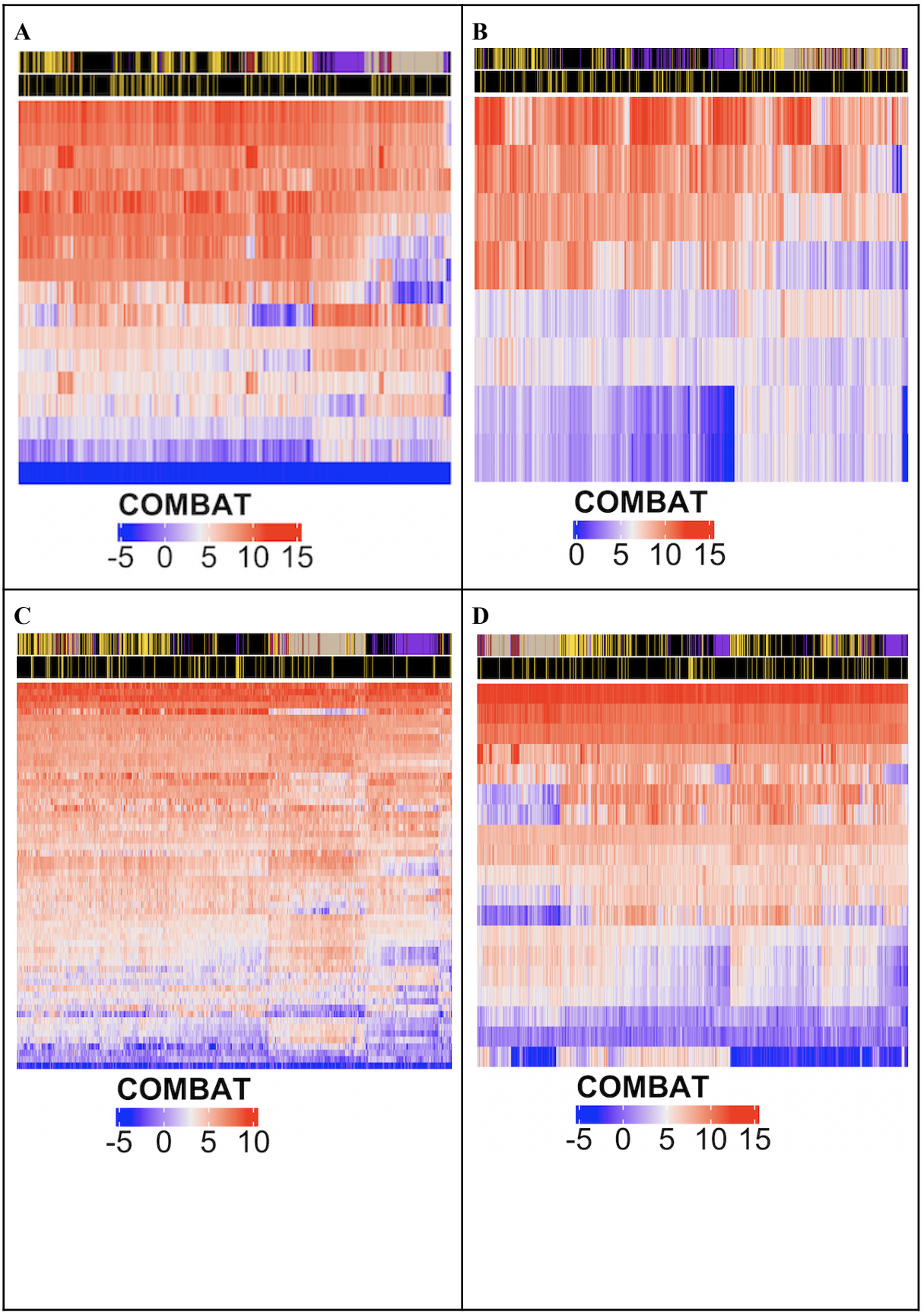
Mammaprint	**16-gene**	**OncotypeDx**	**Pam50**	**EndoPredict**	**17-gene**
ADGRG6	AURKA	ACTB	ACTR3B	BIRC5	ESR1
ALDH4A1	CBX7	AURKA	ANLN	UBE2C	KRT8
AP2B1	CDC20	BAG1	BAG1	DHCR7	KRT18
BBC3	DDIT4	BCL2	BCL2	RBBP8	SLC39A6
CCN4	GSTM3	BIRC5	BIRC5	IL6ST	ERBB2
CCNE2	IGFBP4	CCNB1	BLVRA	AZGP1	TFF3
CDC42BPA	PITX1	CD68	CCNB1	MGP	XBP1
CDCA7	PTHLH	CTSV	CCNE1	STC2	GRB7
CENPA	QPRT	ERBB2	CDC20		BIRC5
CMC2	RAI2	ESR1	CDC6		EGFR
COL4A2	RGS5	GAPDH	CDH3		KRT5
DCK	S100P	GRB2	CENPF		NOTCH1
DHX58	TGFBR3	GSTM1	CEP55		ID1
DIAPH3	TMEM132A	MMP11	CXXC5		MMP2
DTL	UBE2C	MYBL2	EGFR		PTGS2
EBF4	WDR19	PGR	ERBB2		FOXA1
ECI2		RPLP0	ESR1		GATA3
ECT2		SCUBE2	EXO1		
EGLN1		TFRC	FGFR4		
ESM1			FOXA1		
EXT1			FOXC1		
FGF18			GPR160		
FLT1			GRB7		
GMPS			KIF2C		
GNAZ			KRT14		
GPR180			KRT17		
GSTM3			KRT5		
IGFBP5			MAPT		
KDM7A			MDM2		

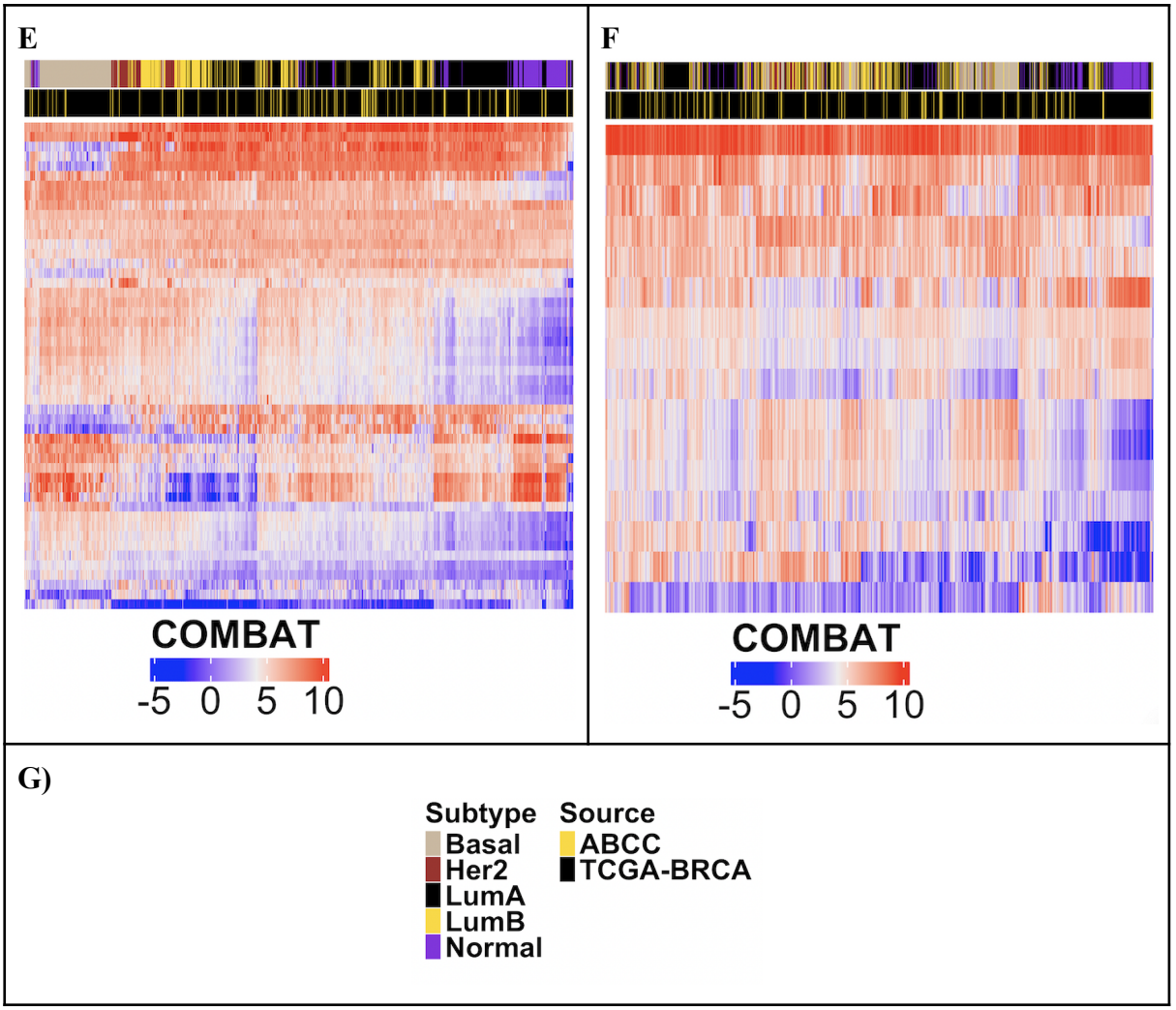
LIN9			MELK		
LPCAT1			MIA		
MCM6			MKI67		
MELK			MLPH		
MMP9			MMP11		
MS4A7			MYBL2		
MSANTD3			MYC		
MTDH			NAT1		
NDC80			PGR		
NUSAP1			PHGDH		
ORC6			PTTG1		
OXCT1			RRM2		
PALM2			SFRP1		
PITRM1			SLC39A6		
PLAAT1			TMEM45B		
PRC1			TYMS		
QSOX2			UBE2C		
RAB6B			UBE2T		
RASSF7					
RECQL5					
RFC4					
RTN4RL1					
RUNDC1					
SCUBE2					
SERF1A					
SLC2A3					
STK32B					
TGFB3					
TMEM74B					
TSPYL5					
UCHL5					
ZNF385B					


Probability
1
0.95
0.94
0.93
0.91
0.88
0.87
0.8
0.79
0.78
0.77
0.76
0.75
0.74
0.72
0.71
0.7
0.69
0.68
0.67
0.66
0.65
0.64
0.63
0.62
0.61
0.6
0.59
0.58
0.57
0.55
0.54
0.53
0.52
0.51
0.5
0.45
0.43
0.41
0.4
Basal
Her2
LumA
LumB
Normal
Subtypes
Patient
RSM1175101
RSM1175107
RSM1175116
RSM487370
RSM487372
RSM487373
RSM487376
RSM487379
RSM487385
RSM487387
RSM487392
RSM487393
RSM487395
RSM487402
RSM487404
RSM487405
RSM487409
RSM487411
RSM487416
RSM487419
RSM487424
RSM487427
RSM487431
RSM487434
RSM487438
RSM487439
RSM487444
RSM487446
RSM487449
RSM487451
RSM487453

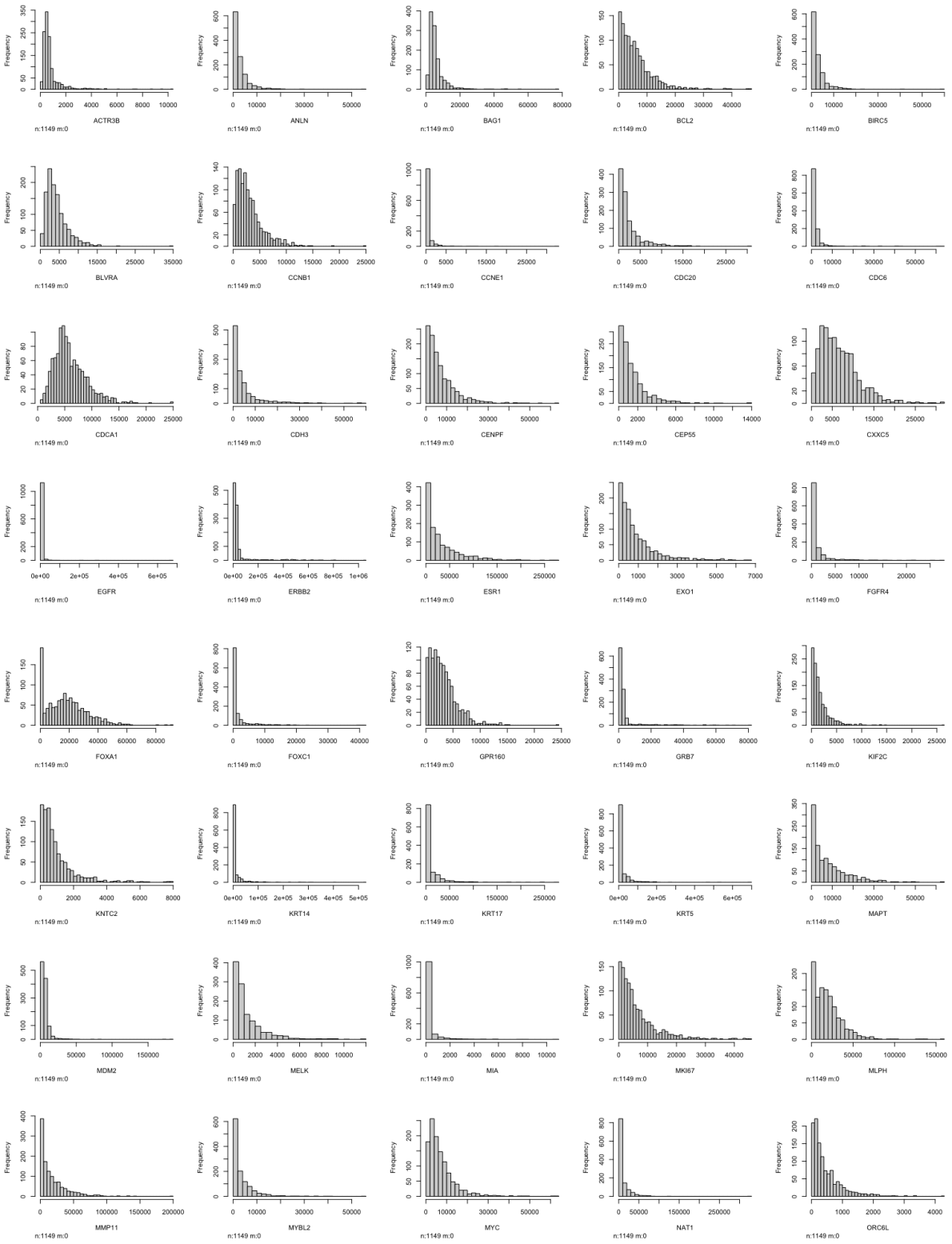


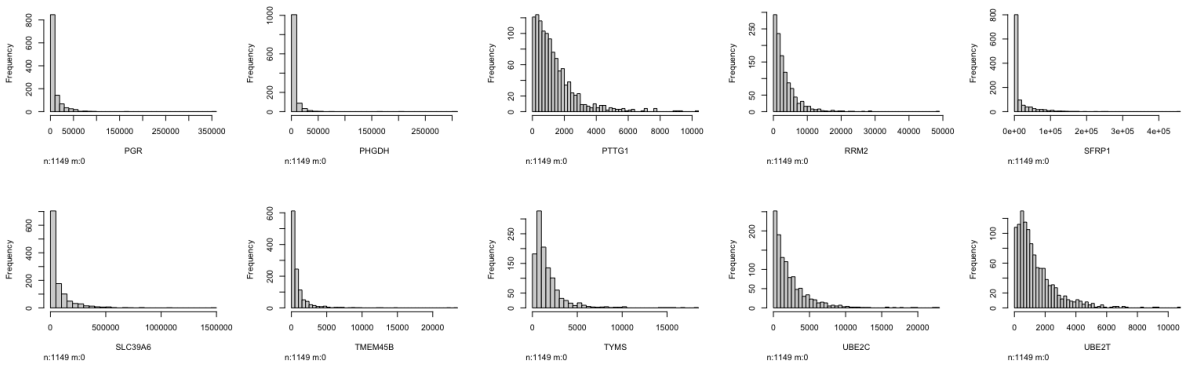
Appendix Figure 1. Scatter plot showing subtype probabilities calculated by PAM50 *(top)* and SCMOD1 *(bottom)*.


A
COMBAT
-5 0 5 10 15
B
COMBAT
0 5 10 15
C
COMBAT
-5 0 5 10
D
COMBAT
-5 0 5 10 15

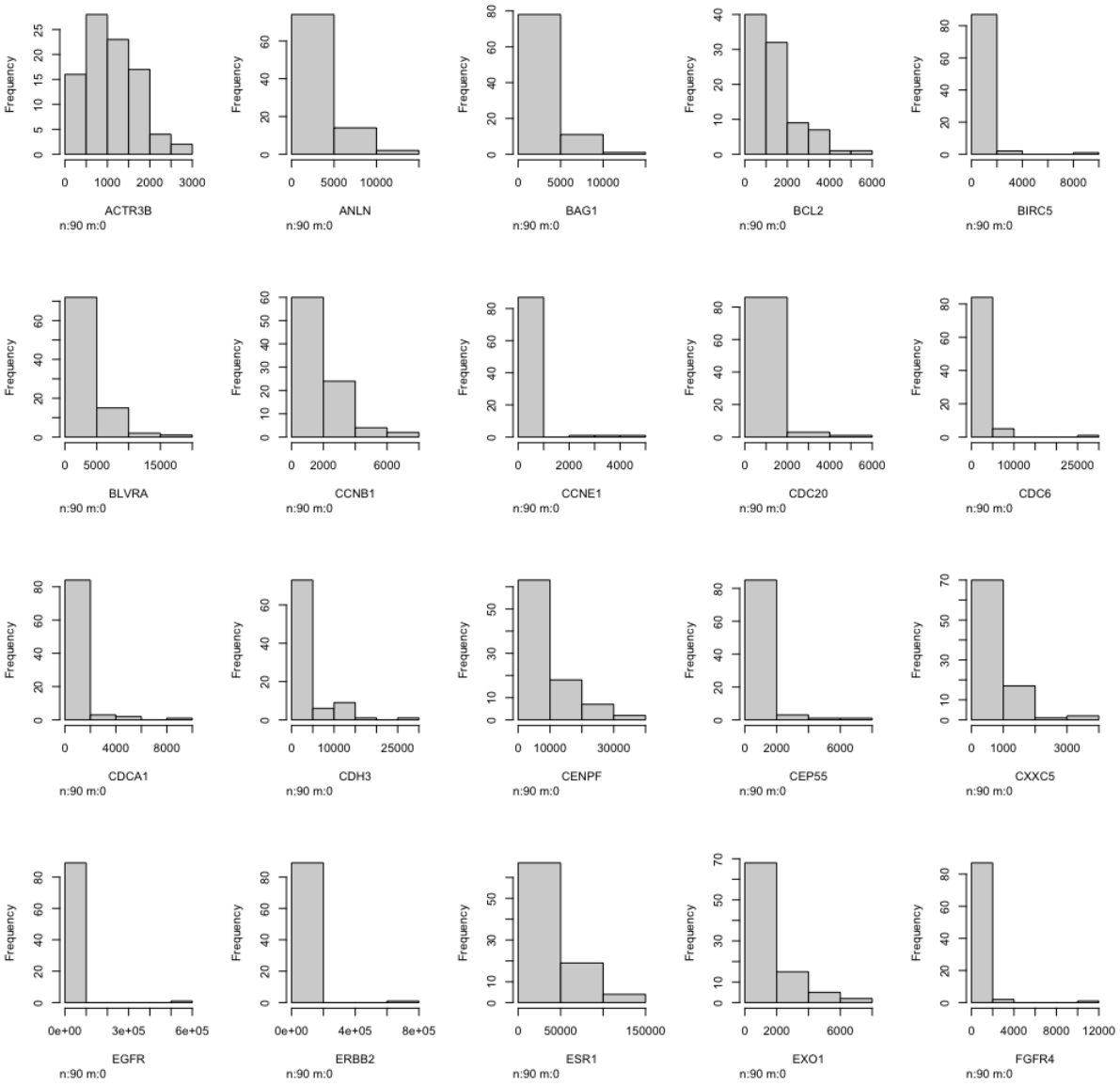


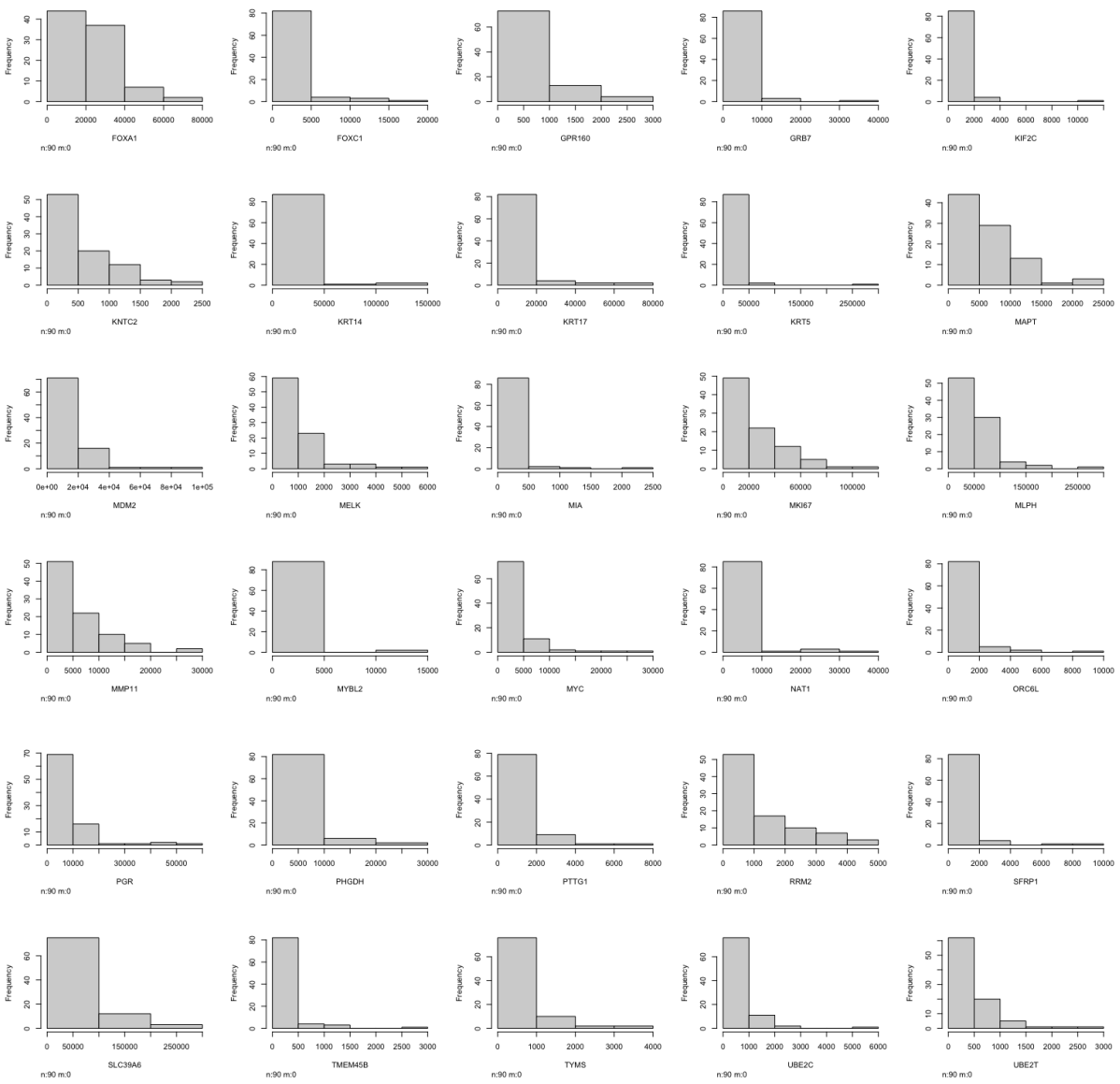
Appendix Figure 2. Heatmap of ComBat adjusted values of A) 17-gene signature; B) EndoPredict; C) Mammaprint; D) OncotypeDx; E) PAM50; F) 16-gene signatures across TCGA-BRCA and ABCC data; and G) annotated with the PAM50 subtypes and data source.


Frequency
ACTR3B
n:1149 m:0
ANLN
n:1149 m:0
BAG1
n:1149 m:0
BCL2
n:1149 m:0
BIRC5
n:1149 m:0
BLVRA
n:1149 m:0
CCNB1
n:1149 m:0
CCNE1
n:1149 m:0
CDC20
n:1149 m:0
CDC6
n:1149 m:0
CDCA1
n:1149 m:0
CDH3
n:1149 m:0
CENPF
n:1149 m:0
CEP55
n:1149 m:0
CXXC5
n:1149 m:0
EGFR
n:1149 m:0
ERBB2
n:1149 m:0
ESR1
n:1149 m:0
EXO1
n:1149 m:0
FGFR4
n:1149 m:0
FOXA1
n:1149 m:0
FOXC1
n:1149 m:0
GPR160
n:1149 m:0
GRB7
n:1149 m:0
KIF2C
n:1149 m:0
KNTC2
n:1149 m:0
KRT14
n:1149 m:0
KRT17
n:1149 m:0
KRT5
n:1149 m:0
MAPT
n:1149 m:0
MDM2
n:1149 m:0
MELK
n:1149 m:0
MIA
n:1149 m:0
MKI67
n:1149 m:0
MLPH
n:1149 m:0
MMP11
n:1149 m:0
MYBL2
n:1149 m:0
MYC
n:1149 m:0
NAT1
n:1149 m:0
ORC6L
n:1149 m:0

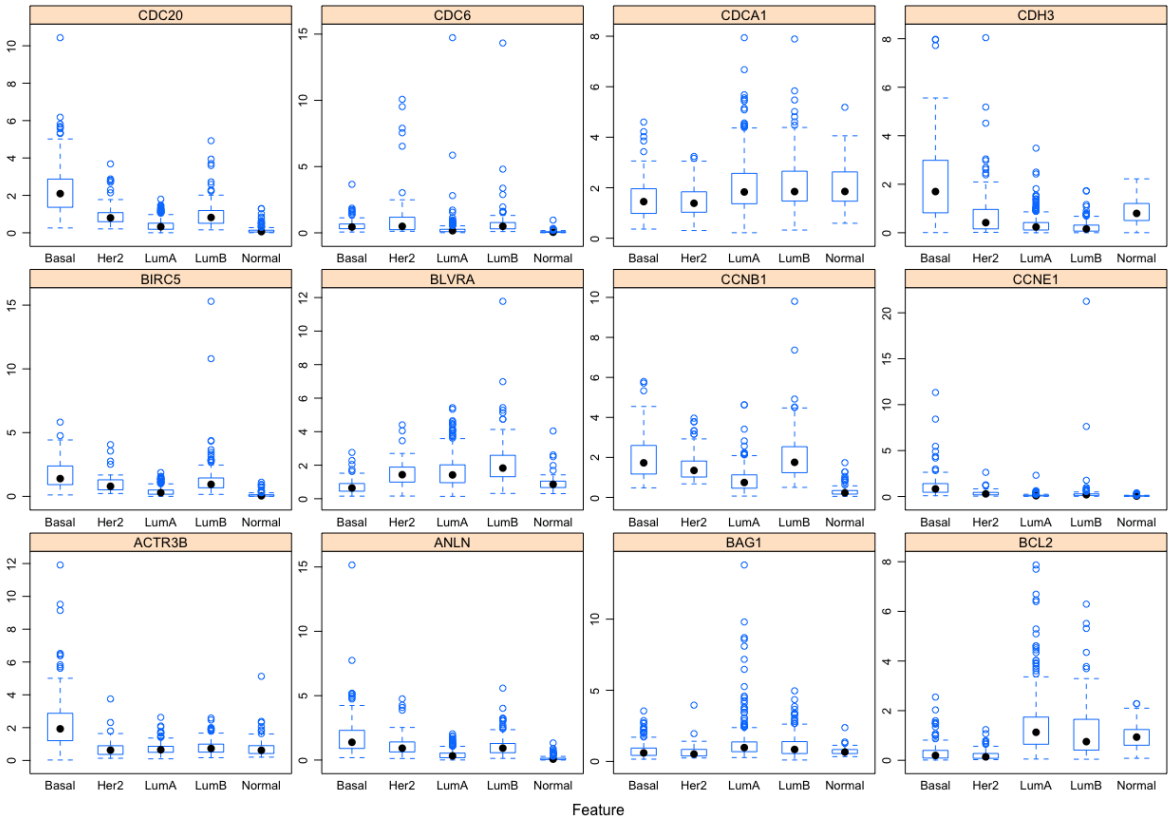


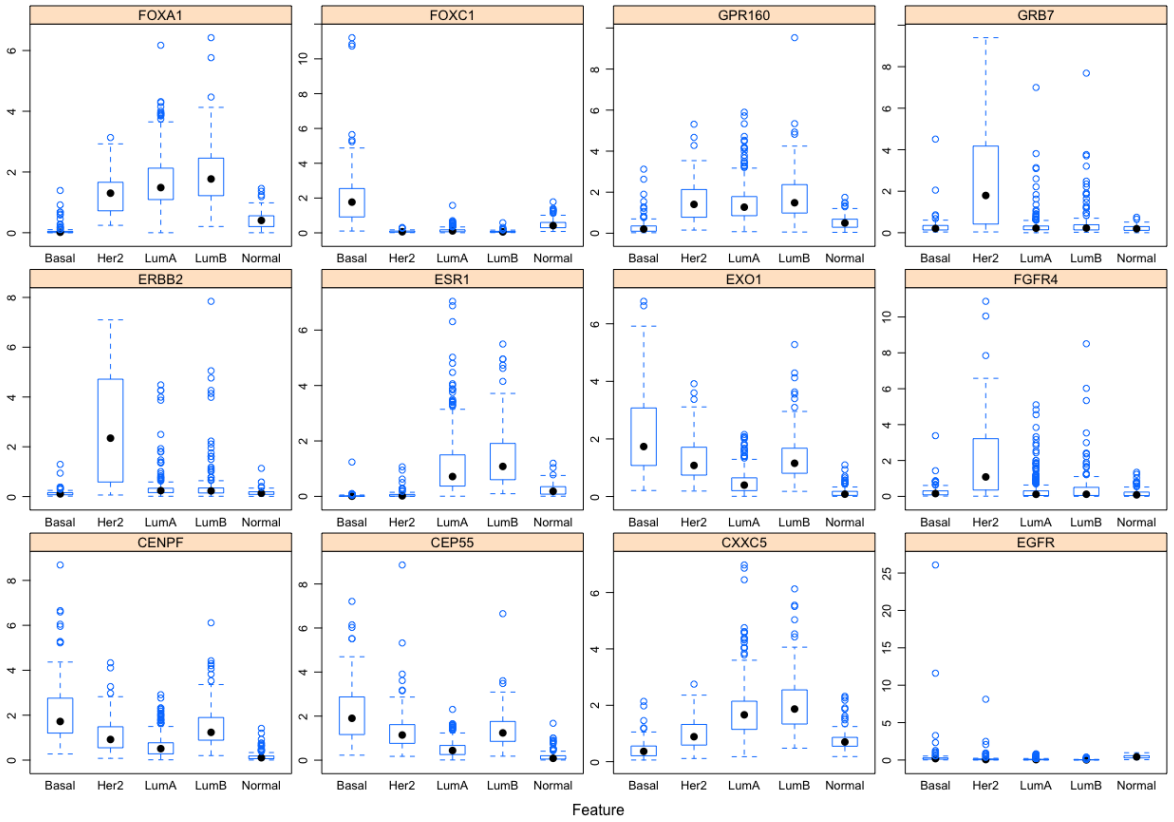
Appendix Figure 3. Distribution of PAM50 genes in TCGA.

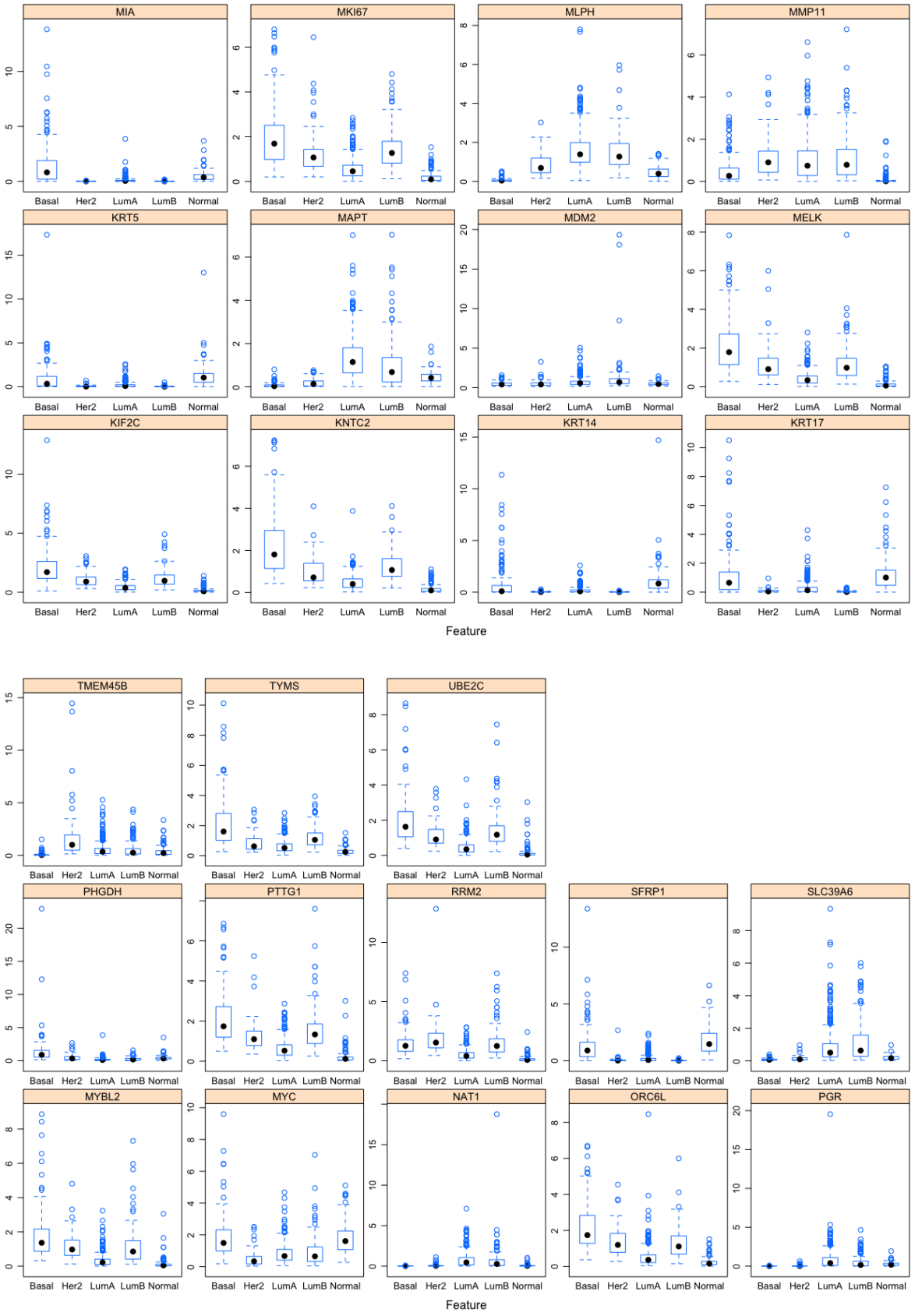

Frequency
ACTR3B
n:90 m:0
ANLN
n:90 m:0
BAG1
n:90 m:0
BCL2
n:90 m:0
BIRC5
n:90 m:0
BLVRA
n:90 m:0
CCNB1
n:90 m:0
CCNE1
n:90 m:0
CDC20
n:90 m:0
CDC6
n:90 m:0
CDCA1
n:90 m:0
CDH3
n:90 m:0
CENPF
n:90 m:0
CEP55
n:90 m:0
CXXC5
n:90 m:0
EGFR
n:90 m:0
ERBB2
n:90 m:0
ESR1
n:90 m:0
EXO1
n:90 m:0
FGFR4
n:90 m:0



Appendix Figure 4. Distribution of PAM50 genes in ABCC.


CDC20
CDC6
CDCA1
CDH3
BIRC5
BLVRA
CCNB1
CCNE1
ACTR3B
ANLN
BAG1
BCL2
Basal
Her2
LumA
LumB
Normal
Feature


FOXA1
FOXC1
GPR160
GRB7
ERBB2
ESR1
EXO1
FGFR4
CENPF
CEP55
CXXC5
EGFR
Basal
Her2
LumA
LumB
Normal
Feature



Appendix Figure 5. Box plot of each feature *(PAM50 genes)* in train data.