Computational and Structural Biotechnology Journal 21 (2023) 2940–2949

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

Development and validation of a prognostic 15-gene signature for stratifying HER2+/ER+ breast cancer

Qian Liu [a,b,c,d], Shujun Huang [a,b], Danielle Desautels [e,f], Kirk J. McManus [b,e], Leigh Murphy [b,e], Pingzhao Hu [a,b,c,e,*]

[a] Department of Biochemistry, Western University, London, Ontario, Canada
[b] Department of Biochemistry and Medical Genetics, University of Manitoba, Winnipeg, Manitoba, Canada
[c] Department of Computer Science, University of Manitoba, Winnipeg, Manitoba, Canada
[d] Department of Statistics, University of Manitoba, Winnipeg, Manitoba, Canada
[e] CancerCare Manitoba Research Institute, Winnipeg, Manitoba, Canada
[f] Department of Internal Medicine, Winnipeg, Manitoba, Canada

ARTICLE INFO

Article history:
Received 8 November 2022
Received in revised form 2 May 2023
Accepted 3 May 2023
Available online 4 May 2023

Keywords:
Breast cancer
Subtyping
Consensus clustering
Gene signature
Machine learning

ABSTRACT

Background: Human epidermal growth receptor 2-positive (HER2+) breast cancer (BC) is a heterogeneous subgroup. Estrogen receptor (ER) status is emerging as a predictive marker within HER2+ BCs, with the HER2+/ER+ cases usually having better survival in the first 5 years after diagnosis but have higher recurrence risk after 5 years compared to HER2+/ER-. This is possibly because sustained ER signaling in HER2+ BCs helps escape the HER2 blockade. Currently HER2+/ER+ BC is understudied and lacks biomarkers. Thus, a better understanding of the underlying molecular diversity is important to find new therapy targets for HER2+/ER+ BCs.

Methods: In this study, we performed unsupervised consensus clustering together with genome-wide Cox regression analyses on the gene expression data of 123 HER2+/ER+ BC from The Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA) cohort to identify distinct HER2+/ER+ subgroups. A supervised eXtreme Gradient Boosting (XGBoost) classifier was then built in TCGA using the identified subgroups and validated in another two independent datasets (Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) and Gene Expression Omnibus (GEO) (accession number GSE149283)). Computational characterization analyses were also performed on the predicted subgroups in different HER2+/ER+ BC cohorts.

Results: We identified two distinct HER2+/ER+ subgroups with different survival outcomes using the expression profiles of 549 survival-associated genes from the Cox regression analyses. Genome-wide gene expression differential analyses found 197 differentially expressed genes between the two identified subgroups, with 15 genes overlapping the 549 survival-associated genes.

XGBoost classifier, using the expression values of the 15 genes, achieved a strong cross-validated performance (Area under the curve (AUC) = 0.85, Sensitivity = 0.76, specificity = 0.77) in predicting the subgroup labels. Further investigation partially confirmed the differences in survival, drug response, tumor-infiltrating lymphocytes, published gene signatures, and CRISPR-Cas9 knockout screened gene dependency scores between the two identified subgroups.

Conclusion: This is the first study to stratify HER2+/ER+ tumors. Overall, the initial results from different cohorts showed there exist two distinct subgroups in HER2+/ER+ tumors, which can be distinguished by a

Abbreviations: AUC, Area under the ROC curve; BC, Breast cancer; CC, Consensus clustering; DepMap, The Cancer Dependency Map; DepOI, Dependency of interest; ER, Estrogen receptor; FDR, False discovery rate; GENIUS, Gene progNostic Index Using Subtypes; GEO, Gene Expression Omnibus; GGI, Gene expression Grade Index; GSEA, Gene set enrichment analysis; HER2, Human epidermal growth receptor 2; HR, Hormone receptor; IHC, Immunohistochemistry; KM, Kaplan-Meier; LogCPM, Log transformed count per million; METABRIC, Molecular Taxonomy of Breast Cancer International Consortium; NMF, Non-negative matrix factorization; PCA, Principal component analysis; PCR, Pathological complete response; PPR, Pathological partial response; PR, Progesterone receptor; ROC, Receiver operating characteristic curve; SHAP, Shapley Additive Explanation; TCGA-BRCA, The Cancer Genome Atlas Breast Invasive Carcinoma; TILs, tumor-infiltrating lymphocytes; TIMER, Tumor immune estimation resource; TMM, Trimmed Mean of M values; XGBoost, eXtreme Gradient Boosting

* Correspondence to: Department of Biochemistry, Western University, Medical Sciences Building Rm. 342, London N6A 5C1 Ontario, Canada.
E-mail address: phu49@uwo.ca (P. Hu).

https://doi.org/10.1016/j.csbj.2023.05.002

15-gene signature. Our findings could potentially guide the development of future precision therapies targeted on HER2+/ER+ BC.

1. Background

Breast cancer (BC) is the most commonly diagnosed cancer and the second leading cause of cancer mortality in U.S. women, accounting for nearly 31% of all new cancer cases and 15% of cancer-related deaths [1]. Though often referred to as a single disease, BC is heterogeneous in histology, progression, therapeutic response, and clinical outcome [2]. In the clinic, the expression levels of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth receptor 2 (HER2) are routinely tested using immunohistochemistry (IHC) to determine BC subtypes, which is fundamental for treatment decision and prognosis of BC [3]. Around 15–20% of BCs express HER2 and thus are referred to as HER2+ tumors. In the past, patients with HER2+ tumors had the worst prognosis among all subtypes of invasive BC. But with the development of anti-HER2 therapy, such as trastuzumab and lapatinib, survival rates for both early- and late-stage HER2+ disease have increased [3] with varying degrees.

Though HER2+ BC is a heterogeneous disease showing different relapse, survival outcomes, and treatment responses, there are currently no prognostic and/or predictive biomarkers[4]. Recently, ER status is emerging as a robust predictive marker within HER2+ BCs. Approximately half of all HER2+ BCs co-express hormone receptors (HRs), namely ER+ and/or PR+ [5]. Of HER2+/HR+ BC patients, over 95% are HER2+/ER+ , who usually have better survival in the first 5 years following a diagnosis but have higher recurrence risk after 5 years compared to those with HER2+/ER- tumors [6,7], possibly due to sustained HR (mainly ER) signaling helps tumor escape from HER2 blockade [5,8]. Multiple studies found that HER2+/ER+ patients treated with anti-HER2 therapy showed lower pathological complete response (PCR) rates than patients with other types of BCs [9–11]. Given the predictive value of ER status, it would be reasonable for the next wave of clinical trials to target HER2+/ER+ and HER2+/ER- patients separately. As well, the development of new therapeutic strategies is of utmost importance to overcome the limitations to targeted therapies and improve treatment for HER2+/ER+ breast cancer. Currently, there is no comprehensive study focused specifically on HER2+/ER+ BC. Thus, a better understanding of the molecular diversity of the ER+/HER2+ BC could pave the way to breakthroughs in HER2+/ER+ treatments.

Gene expression has been used in cancer stratification and gene signature identification since 1999 [12]. Gene expression signatures can help improve patient care by classifying tumors into distinct groups, providing guidance for personalized clinical decisions. Molecular classification of BC based on gene expression profiles has been extensively explored, with the most established subtyping scheme as the intrinsic classification (also known as PAM50) [13]. Using the PAM50 classification, BCs can be divided into 5 subtypes (luminal A, luminal B, HER2-enriched, basal-like, and normal-like), but not all HER2+ tumors fall into the HER2-enriched subtype [14]. Based on the PAM50 genes, Prosigna (rorS) was developed as a gene expression signature estimating distant recurrence risk of ER+ , PR+ , hormone-treated, postmenopausal women with BC [15]. Besides rorS, there are several other commercialized BC gene expression signatures which can be used to estimate different risks for different BC subgroups. Oncotype DX and EndoPredict are gene signatures that estimate the distant recurrence in ER+/HER2- and hormone-treated BC from the expression of 21 genes and 11 genes expression, respectively [16,17]. PIK3CA-GS is derived from exon 20 (the kinase domain) mutations and is able to predict PIK3CA mutation status and tamoxifen sensitivity of ER+/HER2- BC [18]. MammaPrint (also called GENE70) is a 70-gene expression signature that could predict the benefit of adjuvant therapy for BC patients under the age of 61 [15,19]. Gene Prognostic Index Using Subtypes (GENIUS) is a prognostic gene expression signature applicable for any subtype of BC [20]. Gene expression Grade Index (GGI) is a 97-gene signature generated from differentially expressed genes between different histological grades of BC and can estimate the prognostic and recurrence risks of ER+ BC patients [21]. However, currently there is no such gene signature for HER2+/ER+ BC patients. Thus, exploring the gene expression profiles of HER2+/ER+ BC and generating prognostic and predictive gene signatures is critical for better HER2+/ER+ BC clinical guidance.

To better understand the heterogeneity of HER2+/ER+ breast tumors and improve the current HER2+/ER+ BC stratification, we sought to establish prognostic gene expression signatures for identifying reproducible HER2+/ER+ BC subgroups. To this end, we applied unsupervised clustering using Cox regression filtered genes on HER2+/ER+ BCs and subsequently performed genome-wide expression differential analysis between the identified HER2+/ER+ BC subgroups. A gene expression signature was generated based on the significant genes in both Cox regression analysis and gene expression differential analysis. A supervised classifier was then trained based on the expression of the proposed gene signature and validated in two independent HER2+/ER+ BC cohorts.

2. Methods

2.1. Data sources

An overall workflow of this study is shown in Fig. 1. The raw read counts of RNA-sequencing data and clinical information of The Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA) cohort [22] were downloaded through R package “TCGAbiolinks” [23]. TCGA-BRCA has 1102 patients, with each having more than 55,368 gene expression values. After filtering for the patients showing ER+ and HER2+ by immunohistochemistry (IHC), there are a total of 123 HER2+/ER+ BC patients. TCGA also provides the gene expression data of 113 normal adjacent breast tissue samples. The microarray-based gene expression data of Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) [24] and Gene Expression Omnibus (GEO) (accession number GSE149283) BC cohorts [25] were also collected. METABRIC provides more than 24,368 gene expression values for all 1904 BC patients, with 104 patients being HER2+/ER+. GSE149283 provides neoadjuvant trastuzumab therapy response information and 24,352 gene expressions for 18 BC samples. Among these 18 samples, 14 are HER2+/ER+ BCs.

TCGA-BRCA was used as the discovery data in this study, while METABRIC and GSE149283 were used for validation. A total of 15,850 common genes present in all three cohorts were kept. Batch effect removal was performed among TCGA-BRCA, METABRIC and GSE149283 data using “ComBat_seq” function in the “sva” R package [26]. The gene expression data of TCGA-BRCA and normal samples before and after batch effect removal were visualized using principal component analysis (PCA) plots (Supplementary Fig. 1A). The PCA

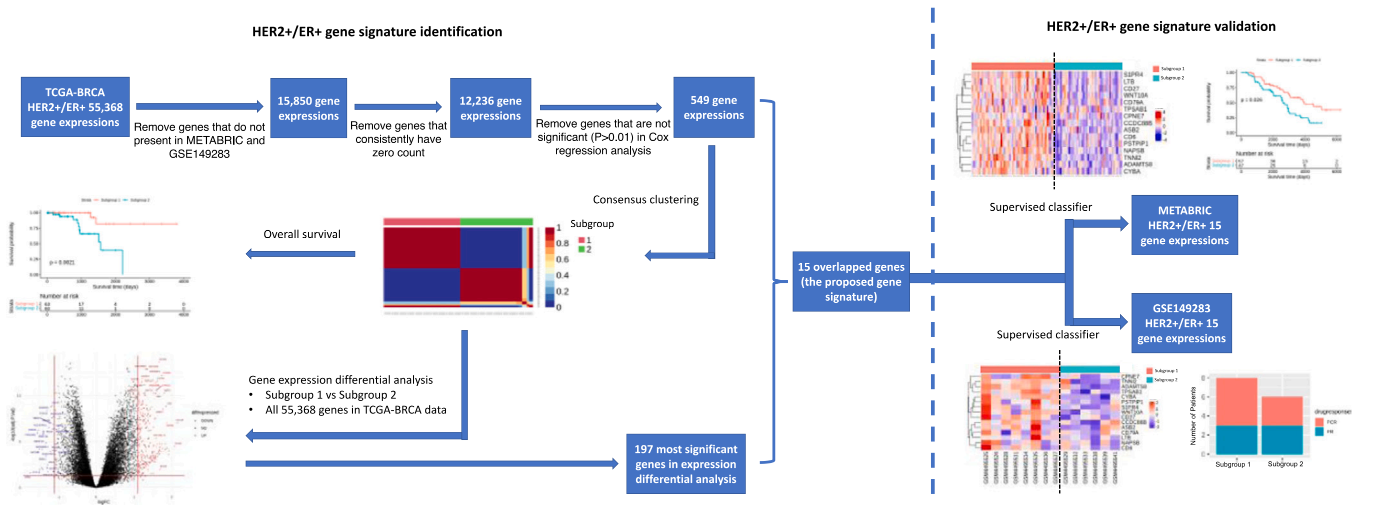
HER2+/ER+ gene signature identification
TCGA-BRCA HER2+/ER+ 55,368 gene expressions
Remove genes that do not present in METABRIC and GSE149283
15,850 gene expressions
Remove genes that consistently have zero count
12,236 gene expressions
Remove genes that are not significant (P>0.01) in Cox regression analysis
549 gene expressions
Consensus clustering
Subgroup
1
2
Overall survival
Gene expression differential analysis
Subgroup 1 vs Subgroup 2
All 55,368 genes in TCGA-BRCA data
197 most significant genes in expression differential analysis
15 overlapped genes (the proposed gene signature)
HER2+/ER+ gene signature validation
Supervised classifier
METABRIC HER2+/ER+ 15 gene expressions
Supervised classifier
GSE149283 HER2+/ER+ 15 gene expressions
Subgroup 1
Subgroup 2

Fig. 1. Overall workflow of this study. 15,850 genes are in common among TCGA-BRCA, METABRIC, and GSE149283 HER2+/ER+ patients. Of them 12,236 genes with at least one count in one sample are kept and input into a Cox regression-based feature selection step, which results in 549 significant genes based on the criteria of *p-value* < 0.01. Consensus clustering are then performed to stratify TCGA-BRCA HER2+/ER+ patients based on gene expression profile of these 549 significant genes. Gene differential analysis is done among the identified subtypes to identify most differentially expressed genes. Genes that are significant in both Cox regression analysis and gene expression differential analysis are selected to form the proposed gene signature. Validation of this gene signature is performed on METABRIC and GSE149283 HER2+/ER+ cohorts. A XGBoost classifier is trained using the proposed gene signature on TCGA-BRCA data, and then applied to assign METABRIC and GSE149283 BCs into two subgroups. For METABRIC, survival difference of the predicted subgroups is tested. For GSE149283, the drug response difference between the predicted subgroups is tested.

plots of the expression data from the three data sets before and after batch effect removal are shown in Supplementary Fig. 1B.

For the TCGA-BRCA data, we removed the genes that consistently have zero counts, which resulted 12,236 genes left for the following analysis. We first normalized the raw count data based on Trimmed Mean of M values (TMM) [27] and then calculated the log transformed count per million (LogCPM) value using R package “edgR”. We subsequently performed feature selection using Cox regression model with a cut-off *p-value* of 0.01. Genes significant in univariate Cox regression analysis were selected, resulting in 549 survival-associated genes.

The expression data of the TCGA normal samples was used as reference for the later gene expression differential analysis. We also used the Bayesian tensor factorization (BTF) integrated and encoded 17 multi-omics features from the copy number variation (CNV), DNA methylation, and gene expression data of 68 out of the 123 TCGA-BRCA HER2+/ER+ patients with the three types of data available to compare the single-omics and multi-omics-based subtyping for HER2+/ER+ BCs. The details of the BTF method and the integrated multi-omics features could be found in [28].

2.2. Unsupervised clustering on discovery data

We used the R package “CancerSubtypes” [29] to run the consensus clustering (CC) [30] on the TCGA-BRCA HER2+/ER+ data matrix (549 genes × 123 patients). Traditional subtyping methods such as K-mean [31] and hierarchical clustering have some limitations. For K-mean clustering, a pre-defined K, which is the number of clusters, is needed. However, for most of the unsupervised clustering problems, the number of clusters is unknown. Although hierarchical clustering could provide us a tree-based results, a pre-defined cut-off point is still needed to decide the number of clusters. However, CC is a resampling-based clustering algorithm which could estimate the number of clusters and obtain robust clustering result according to the consensus among several clustering runs [30,32].

In our case, CC first subsampled the gene expression data matrix for 30 times. Then, non-negative matrix factorization (NMF) was applied on these 30 sample sets to obtain 30 clustering results. The maximum number of subtypes (k) was set as 10, which means we could expect up to 10 subtypes in each of the 30 clustering runs. The 30 clustering results for each subtype number (from 1 to k, which is 10 in our case) were then used to calculate the pairwise consensus value, which is defined as the probability of two items being clustered together [30]. These consensus values formed a consensus matrix for each cluster number. Which means, we obtained 10 consensus matrices, each one corresponding to a specific number of clusters. In the end, an agglomerative hierarchical consensus clustering was applied to each of the 10 consensus matrices, to obtain the final 10 clustering results.

Next, we calculated the silhouette value for each patient, which measures how similar a sample is to its own cluster compared to other clusters. The silhouette value ranges from −1 to 1 with a high value indicating that the sample is well matched to its own cluster and poorly matched to other clusters. If most samples have a high positive value, then the clustering configuration is appropriate. We then did survival analysis to test whether there are survival differences between the identified subtypes. The result was visualized using the Kaplan-Meier (KM) curve [33], which is widely used in clinical and healthcare fundamental research. It shows what the probability of an event (survival) is at a certain time interval. To compare the subtyping results based on single-omics and multi-omics, we applied the same CC and NMF clustering method on the TCGA-BRCA HER2+/ER+ multi-omics data matrix (17 BTF features × 68 patients). The silhouette value for the multi-omics based HER2+/ER+ subtypes was also calculated, and the survival difference was visualized by KM plot as well.

2.3. Gene expression differential analysis for identified subgroups

We performed the differential analyses of each of the identified subgroups relative to the normal tissue samples in TCGA-BRCA cohort using R package “limma” [34]. We also performed differential analyses between the identified HER2+/ER+ subgroups. The differential expressed genes of each HER2+/ER+ subgroup versus the normal samples were then used to perform gene set enrichment analysis (GSEA; detailed later) to explore the enriched biological pathways. The differentially expressed genes between the identified subgroups were further used to filter for the genes for gene signature construction.

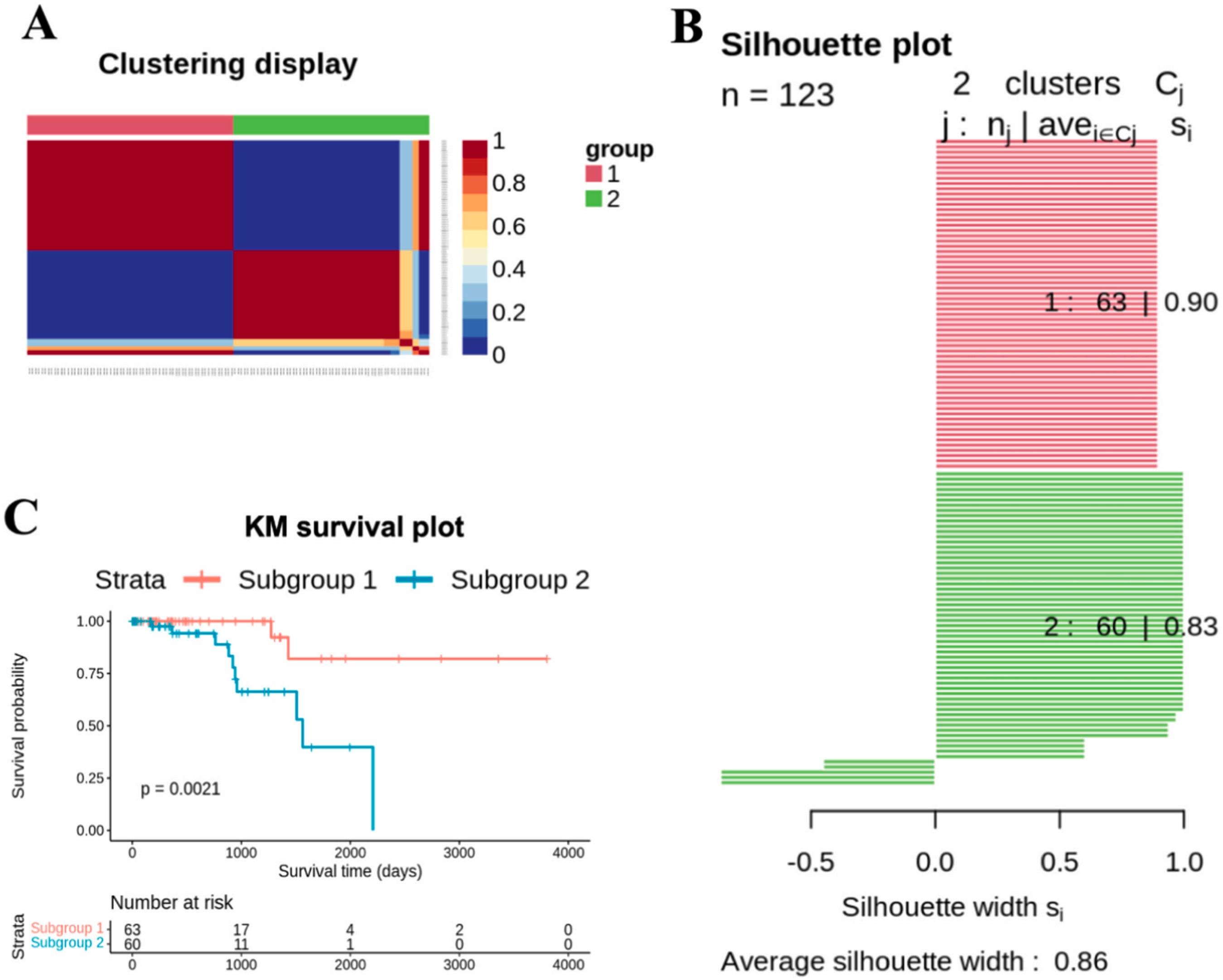

Fig. 2. Results of consensus clustering on TCGA-BRCA data. A: Symmetric consensus matrix hierarchical clustering heatmap for TCGA-BRCA data. Columns and rows are patients. The color represents the probability that two patients were clustered together. B: Silhouette plot for the TCGA-BRCA data. Each horizontal line represents a sample, and the length of the line is the silhouette value for the sample. The color represents different subtypes: red ones are in Subgroup 1, while green ones are in Subgroup 2. A high value indicates that the sample is well matched to its own cluster and poorly matched to other clusters. If most samples have a high positive value, then the clustering configuration is appropriate. The overall silhouette value is 0.91, which means the clustering is appropriate. C: KM plot of two subgroups identified by CC.

2.4. Gene set enrichment analysis for identified subgroups

GSEA was applied on the differentially expressed genes identified in the previous step, and only included genes with adjusted *p-value* less than 0.05 and |logFC| (absolute value of the log2 transformed folder change) large than 0.5. We pre-ranked the genes based on their adjusted *p-values* in the differential analysis, then input them into the GSEA software [35].

2.5. Gene signature creation and validation

Using eXtreme Gradient Boosting (XGBoost)-based supervised classification, we built a gene signature from the genes that are significant in both Cox regression analysis and the expression differential analysis between the identified HER2+/ER+ subgroups. The TCGA-BRCA data was then used to train and evaluate the XGBoost classifier in a 5-fold cross validation way using the R package “caret” [36]. The model’s performance was measured by the area under of the curve (AUC) of receiver operating characteristic (ROC), accuracy, sensitivity, and specificity. Shapley Additive Explanation (SHAP) values were calculated to increase the interpretability of the XGBoost model [37]. A higher SHAP value of a given feature in the model implies stronger influence on the model’s decision.

We applied the well-trained XGBoost model to classify the METABRIC and GSE149283 cohorts using the expression profile of the proposed gene signature. Survival analysis and hierarchical heatmap were also performed to visualize the predicted subgroup on METABRIC. For GSE149283, we visualized the difference of the neoadjuvant trastuzumab therapy response between each subgroups using bar plot. To evaluate the extendibility of the identified gene signature, we also applied the well-trained XGBoost model to classify all the TCGA-BRCA patients (not limited to HER2+/ER+ BCs). Survival analysis and hierarchical heatmap were also performed to visualize the predicted subgroup on the entire TCGA-BRCA patients.

We further checked the uniquely mutated genes in each subgroup for both TCGA-BRCA and METABRIC cohorts using the CBioPortal OncoPrint function [38]. The tumor immune estimation resource (TIMER) [39] quantified abundance of the tumor-infiltrating lymphocytes (TILs) (B cells, CD4 + T cells, CD8 + T cells, neutrophils, macrophages, and dendritic cells), the PAM50 intrinsic subtypes, rorS, GENIUS, GENE70, and GGI scores of each sample in both TCGA-BRCA and METABRIC cohorts were also tested.

The Cancer Dependency Map (DepMap) project provides systematically identifies genetic and pharmacologic dependencies that were measured on CRISPR-Cas9 knockout cancer cell lines [40]. To extend the cell line DepMap to tumors, Chiu et al. developed a deep learning model named DeepDep to predict the effect scores of the dependency of interest (DepOI) from genomics data [41]. We applied DeepDep on our HER2+/ER+ BC gene expression data to get the predicted effect score of the DepOIs for TCGA-BRCA and METABRIC cohorts, and then checked the difference between two HER2+/ER+ BC subtypes.

3. Results

3.1. Two distinct subgroups within HER2+/ER+ identified

To stratify TCGA-BRCA HER2+/ER+ patients, we performed CC on survival significant genes. We first obtained 549 genes associated with survival outcome according to the genome-wide univariate Cox regression analysis. When the unsupervised CC cluster number equals to 2, we could get the most significant survival difference (*p-value* = 0.0021). The consensus matrix heatmap, silhouette plot, gene expression heatmap, and KM plot are shown in Fig. 2. We observed a clear two-cluster pattern on both consensus matrix heatmap (Fig. 2A) and gene expression heatmap (Supplemental Fig. 2). The average silhouette value is 0.91, indicating two robust different subgroups existing in the TCGA-BRCA HER2+/ER+ cohort (Fig. 2B). Sixty-three patients were assigned to Subgroup 1, while the other 60 patients were assigned to Subgroup 2. The survival difference between these two subgroups is significant (*p-value* = 0.0021). Patients in Subgroup 2 suffered poor prognosis (Fig. 2C). In addition, from the differential analyses between Subgroup 1 versus normal and Subgroup 2 versus normal (Supplemental Fig. 3), Subgroup 2 has more differentially expressed genes than Subgroup 1 relative to normal. GSEA results (Supplemental Fig. 3) based on the two pre-ranked lists of differentially expressed genes (Subgroup 1 versus normal and Subgroup 2 versus normal) found that only the Martens Bound by Promyelocytic leukemia (PML) retinoic acid receptor alpha (RARA) Fusion pathway is significantly downregulated in Subgroup 2 relative to normal (false positive rate (FDR) = 0.006).

The demographic and clinical information of the two identified subgroups for TCGA-BRCA HER2+/ER+ cohort is shown in Table 1. Except for age (*p-value* = 0.0258), no other significant demographic or clinical differences exist between these two subgroups.

The differential analyses of Subgroup 1 versus Subgroup 2 identified 197 genes differentially expressed (|logFC| > 1 and adjusted *p-value* < 0.05) (Supplementary Fig. 3). Among these 197 differentially expressed genes, 15 overlapped with the 549 survival-associated genes from Cox regression analyses: *TNNI2*, *CCDC88B*, *CYBA*, *ASB2*, *LTB*, *S1PR4*, *PSTPIP1*, *CD6*, *CD27*, *WNT10A*, *NAPSB*, *CD79A*, *ADAMTS8*, *CPNE7*, and *TPSAB1*. These 15 genes have both survival significance and subgroup distinguishing significance and constitute the proposed gene signature. In this way, we decreased the gene number from 549 to 15 for easier application to other datasets Fig. 3.

The consensus matrix heatmap, silhouette plot, gene expression heatmap, and KM plot of the multi-omics-based subtyping are shown in Supplementary Fig. 4. Multi-omics method also resulted a clear two-cluster pattern on consensus matrix heatmap (Supplementary Fig. 4A). However, the overall silhouette value of multi-omics-based subtyping (0.62) is smaller than that of the single-omics-based subtyping (0.91), which means the single-omics-based subtyping methods might be more appropriate than multi-omics-based subtyping. As can be noted in Supplementary Fig. 4B, the survival difference between the two multi-omics-based HER2+/ER+ subtypes (*p-value* = 0.027) is less significant than the single-omics-based HER2+/ER+ subtypes (*p-value* = 0.0021). The difference of the subtype assignment between the single-omics and multi-omics were compared in Supplementary Fig. 5, most of the patients (45 of 68) were assigned to the same subtypes based on the two methods (single-omics and multi-omics). The 23 differently assigned patients were mainly Luminal subtypes with good or moderate prognosis (Supplementary Fig. 5B and 5C).

Table 1
Demographic and clinical information of the identified HER2+/ER+ subgroups in TCGA-BRCA cohort.

		Subgroup 1	Subgroup 2	p-value
No. patients		63	60	
Age	min	34	29	0.02581*
	max	88	90	
	mean	57.13	62.65	
	Standard deviation	13.67	13.46	
T	T1, T1b, T1c	14	10	0.2133
	T2	41	39	
	T3	7	7	
	T4 (T4, T4b)	1	4	
N	N0, N0 (i-), N0 (i +)	31	23	0.2202
	N1, N1a, N1b	23	19	
	N2, N2a	6	11	
	N3, N3a	3	6	
	NX	0	1	
M	cM0 (i +), M0	50	50	0.1991
	M1	1	1	
	MX	12	9	
Stage	I, IA	7	9	0.2414
	II, IIA, IIB	43	28	
	IIIA, IIIB, IIIC	12	21	
	IV	1	1	
	X	0	1	
Surgery	Lumpectomy	8	4	0.2650
	Modified Radical Mastectomy	17	20	
	Simple Mastectomy	12	4	
	Other	19	28	
	Not Available	7	4	

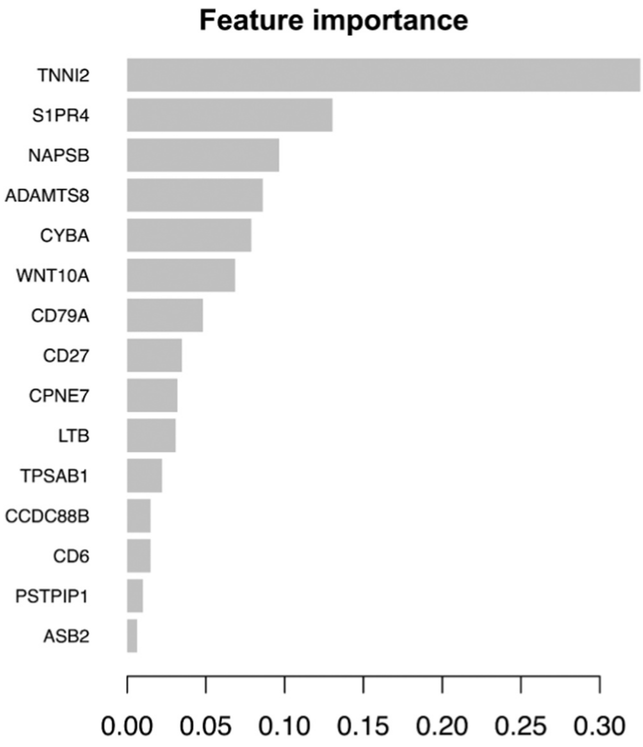

Fig. 3. The SHAP importance score of each gene in the XGBoost classifier.

Table 2
Hyperparameter finetuning process and 5-fold cross-validated model performance of the XGBoost classifier.

Learning rate	Max depth	AUC	Sensitivity	Specificity
0.005	6	0.84	0.78	0.73
	8	0.82	0.78	0.75
	10	0.84	0.78	0.75
	12	0.82	0.73	0.75
0.01	6	0.81	0.76	0.73
	8	0.83	0.81	0.75
	10	**0.85**	**0.76**	**0.77**
	12	0.83	0.79	0.77
0.05	6	0.81	0.79	0.77
	8	0.81	0.79	0.73
	10	0.82	0.78	0.72
	12	0.79	0.79	0.73
0.07	6	0.80	0.78	0.72
	8	0.82	0.76	0.75
	10	0.80	0.76	0.75
	12	0.80	0.79	0.72
0.1	6	0.80	0.78	0.72
	8	0.82	0.79	0.72
	10	0.80	0.81	0.75
	12	0.79	0.81	0.73

3.2. HER2+/ER2 + BC gene signature development

A supervised classification model (XGBoost) was trained to classify the TCGA-BRCA HER2+/ER+ patients into the CC identified two subgroups. The fine-tuning process and 5-fold cross-validated performance of the XGBoost classifier are shown in Table 2. The 5-fold cross-validated AUC, sensitivity, and specificity on TCGA-BRCA data are 0.85, 0.76, and 0.77, respectively. The importance score of each gene in the XGBoost model is shown in Fig. 3. TNNI2 is the most important feature for the model to make the prediction decision.

HER2+/ER2 + BC gene signature validation.

The well-trained XGBoost classifier was applied to two external datasets (METABRIC HER2+/ER+ cohort and GSE149283 HER2+/ER+ cohort) and one extended dataset (entire TCGA-BRCA cohort, which is not restricted to the 123 HER2+/ER+ BC cases) for validation. The expression profiles of the 15-gene signature on METABRIC HER2+/ER+ cohort, GSE149283 HER2+/ER+ cohort, and entire TCGA-BRCA cohort are presented in Fig. 4A, B, and Supplementary Fig. 6A, respectively. METABRIC HER2+/ER+ patients, GSE149283 HER2+/ER+ patients, and entire TCGA-BRCA patients that are assigned into Subgroup 2 showed overall lower

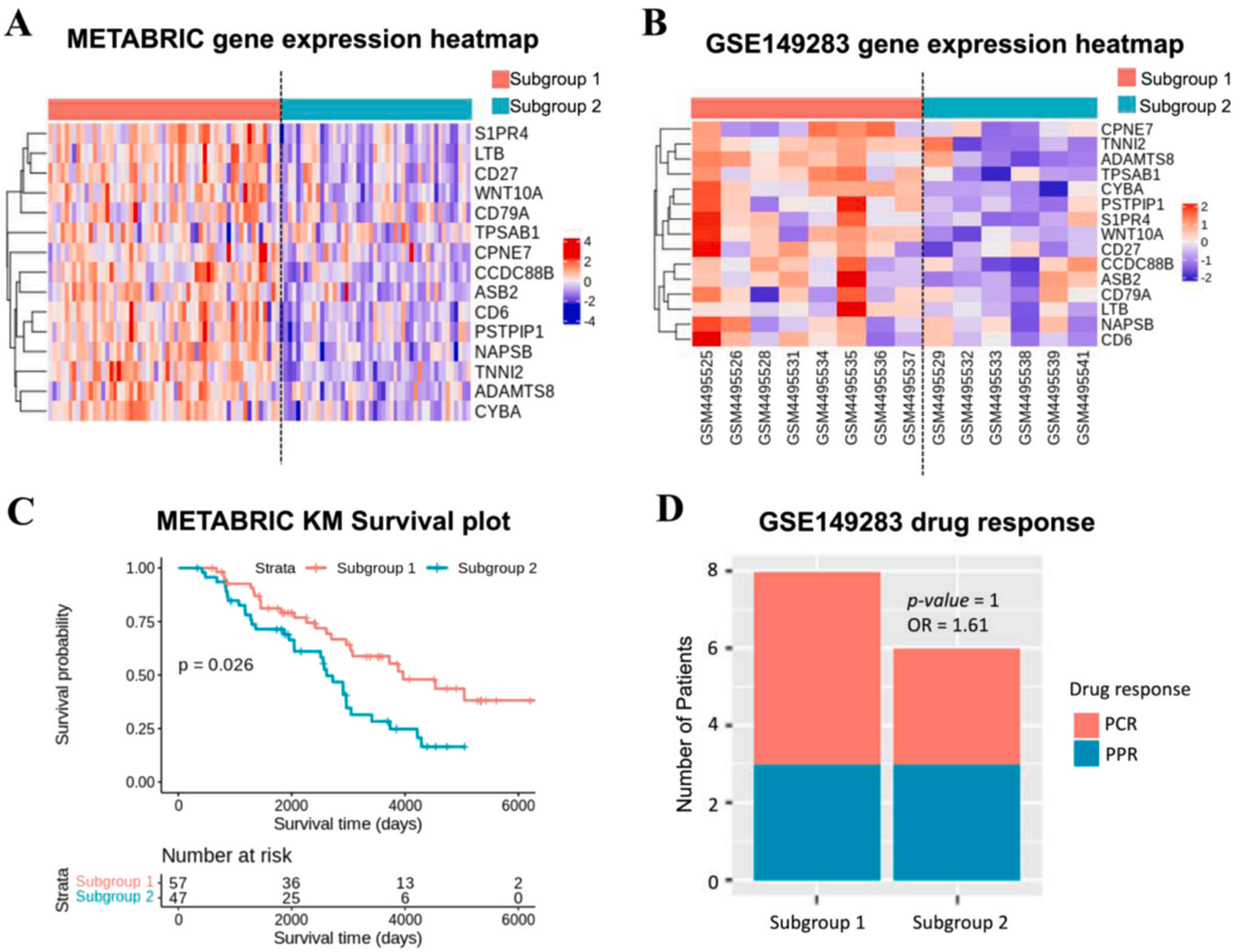

Fig. 4. Predicted subgroups of external validation HER2+/ER+ BC cohorts. A: The expression profile of the proposed 15-gene signature on METABRIC HER2+/ER+ BC cohort. Columns are 104 patients, while rows are 15 genes. The XGBoost predicted subgroup labels are shown in the top side bar. B: The expression profile of the proposed 15-gene signature on GSE149283 HER2+/ER+ BC cohort. Columns are 14 patients, and rows are 15 genes. The XGBoost predicted subgroup labels are shown in the top side bar. C: KM plot of the two subgroups of METABRIC cohort predicted by XGboost. D: The stacked histogram of the trastuzumab therapy response for the XGBoost predicted subgroups. PCR, pathological complete response; PPR, pathological partial response; OR, odds ratio.

Table 3
METABRIC demographic information.

	Subgroup 1	Subgroup 2	*p-value*	
No. patients	57	47		
Age	Min	29	39	0.1035
	Max	87	87	
	Mean	60	64	
	Standard deviation	13.72	12.20	
Tumor size	Min	5	10	0.1074
	Max	70	50	
	Mean	27	23.17	
	Standard deviation	13.60	7.60	
Lymph nodes positive	Min	0	0	0.0894
	Max	25	25	
	Mean	1.83	3.68	
	Standard deviation	4.61	6.11	
Grade	1	2	1	0.2318
	2	18	14	
	3	35	32	
	Null	2	0	
Stage	0	15	9	0.2424
	1	14	8	
	2	13	19	
	3	2	2	
	4	1	0	
	Null	12	9	
Treatment	Chemotherapy (CT)	2	1	0.2578
	Radiotherapy (RT)	4	2	
	Hormonotherapy (HT)	12	9	
	CT/RT	3	2	
	CT/HT	0	3	
	CT/HT/RT	5	6	
	HT/RT	23	21	
	None	8	3	

expression values of the 15 gene signature than Subgroup 1, which is similar to what is observed in the TCGA-BRCA HER2+/ER+ cohort.

For the 104 HER2+/ER+ patients from the METABRIC cohort, 57 were assigned to Subgroup 1, while 47 were assigned to Subgroup 2 by the XGBoost model. The survival difference of these two subgroups was significant with Subgroup 2 showing worse survival than Subgroup 1 (Fig. 4C), which is similar to the two subgroups inthe TCGA-BRCA HER2+/ER+ cohort (Fig. 2C). However, unlike the TCGA-BRCA HER2+/ER+ cohort findings, there were no significant differences in demographic or clinical characteristics between these two subgroups (Table 3). For the 14 HER2+/ER+ patients in GSE149283, eight patients were in Subgroup 1 and six patients were assigned to Subgroup 2. According to Fig. 4D, there is a higher proportion of patients in Subgroup 2 (three out of six) showed partial response to trastuzumab than Subgroup 1 (three out of eight patients). However, the significance of the difference was not significant according to the Fisher's exact test (*p-value* = 1, odds ratio = 1.61). For the entire 1102 TCGA-BRCA patients, 807 were assigned to Subgroup 1, while 295 were assigned to Subgroup 2 by the XGBoost model. The survival difference of these two subgroups was significant with Subgroup 2 showing worse survival than Subgroup 1 (Supplementary Fig. 6B), which is consistent to that observed in the TCGA-BRCA HER2+/ER + and METABRIC HER2+/ER+ cohorts.

3.3. *Computational characterization of the two subgroups (external validation)*

Uniquely mutated genes, TILs, PAM50 subtypes, some other published gene signatures such as rorS, GENIUS, GENE70, GGI scores, and DepMap dependency were calculated to characterize the identified two HER2+/ER+ subgroups. We found that there are 1914 mutated genes in the genome of TCGA-BRCA Subgroup 1 patients that were not observed within Subgroup 2 patients' genome, while TCGA-BRCA Subgroup 2 patients' genome have 3293 mutated genes that are absent in Subgroup 1′s genome. There are six genes (*CDKN1B, PRKCE, ACVRL1, UBR5, AGMO, SMARCC2*) commonly mutated in both TCGA-BRCA Subgroup 1 and METABRIC Subgroup 1 (Fig. 5A top left). While another six common genes (*PALLD, DCAF4L2, MAP3K13, RPGR, SHANK2, FANCA*) are altered in both TCGA-BRCA Subgroup 2 and METABRIC Subgroup 2 (Fig. 5A bottom left).

The quantified abundance of six immune cell types were estimated using TIMER on both TCGA-BRCA and METABRIC cohorts to check the TILs difference in two HER2+/ER+ subgroups (Fig. 5B). The infiltration of dendritic cells, neutrophils, and CD4 + T cells are significantly lower in TCGA-BRCA Subgroup 2 than in Subgroup 1 (*p-values* are 0.0009, 0.0500, and 0.0028, while absolute fold changes are 0.7171, 0.7812, and 0.7132). The combined Subgroup 2 patients (TCGA-BRCA Subgroup 2 plus METABRIC Subgroup 2) also show fewer dendritic cell (*p-value* = 0.0026, absolute fold change = 0.8286) and CD4 + T cell infiltrations (*p-value* = 0.0016, absolute fold change = 0.7883) than the combined Subgroup 1 patients (TCGA-BRCA Subgroup 1 plus METABRIC Subgroup 1) also showed lower dendritic cell (*p-value* = 0.0026, absolute fold change = 0.8286) and CD4 + T cell infiltrations (*p-value* = 0.0016, absolute fold change = 0.7883).

Fig. 5C and Fig. 5D are showing the PAM50 intrinsic subtypes and the published gene signatures (rorS and GENIUS) of different subgroups in both TCGA-BRCA and METABRIC cohorts. There is a lower proportion of Normal PAM50 subtype in TCGA-BRCA HER2+/ER+ Subgroup 2 compared with TCGA-BRCA HER2+/ER+ Subgroup 1. For METABRIC, there is a lower proportion of LumA type in the HER2+/ER+ Subgroup 2 than Subgroup 1. TCGA-BRCA HER2+/ER+ Subgroup 2 showed significantly higher intrinsic rorS score and GENIUS score than Subgroup 1. However, other published gene signatures didn't show difference between the two HER2+/ER+ subgroups in both cohorts. The predicted effect scores of DepOIs are visualized in Supplementary Fig. 7. Please be noted that only top 15 DepOIs with most significant differences between HER2+/ER+ Subgroup 1 and Subgroup 2 are shown in the heatmaps. As can be seen, there are visible differences in TCGA-BRCA cohort, but not in METABRIC cohort. TCGA-BRCA HER2+/ER+ Subgroup 1 shows lower dependency effect scores of the top 15 DepOIs.

4. Discussion

We have identified a 15-gene expression signature which could stratify HER2+/ER+ BC patients into two prognostically different subgroups in both unsupervised and supervised manners. This 15-gene expression signature could be extended to predict prognosis in all BCs, not just HER2+/ER+ BCs. The prognostic difference between the two HER2+/ER+ subgroups was observed in both TCGA-BRCA and METABRIC cohorts, not confounded by other clinical characteristics, including tumor size, grade, or stage. The two subgroups also tend to exhibit difference in terms of their response to trastuzumab in GSE149283 with 14 samples, suggesting the predictive potential of the proposed 15-gene signature. However, no statistical significance was observed in the GSE149283 cohort, which might be due to the small sample size (n = 14). Thus, further validation is required once a large drug response dataset becomes available. According to the GSEA results of the differentially expressed genes, Martens Bound by PML RARA Fusion pathway was significantly enriched in Subgroup 2 but not in Subgroup 1. This pathway is a diagnostic marker of the acute promyelocytic leukemia [42] and maybe used as a diagnostic marker for HER2+/ER+ Subgroup 2 in the near future.

Six genetic alterations were found in Subgroup 1 that were not seen in Subgroup 2. Among them, *UBR5* amplification was observed in 17% TCGA-BRCA Subgroup 1 patients' genome and 37% METABRIC Subgroup 1 patients' genome. *UBR5* encodes a HECT-domain containing E3 ubiquitin ligase that is involved in regulating DNA damage response, cell cycle, metabolism, transcription, and apoptosis [43]. Multiple studies have demonstrated that elevated expression of *UBR5* is implicated in different cancers, including breast and ovarian cancers, and is closely associated with advanced clinical stage,

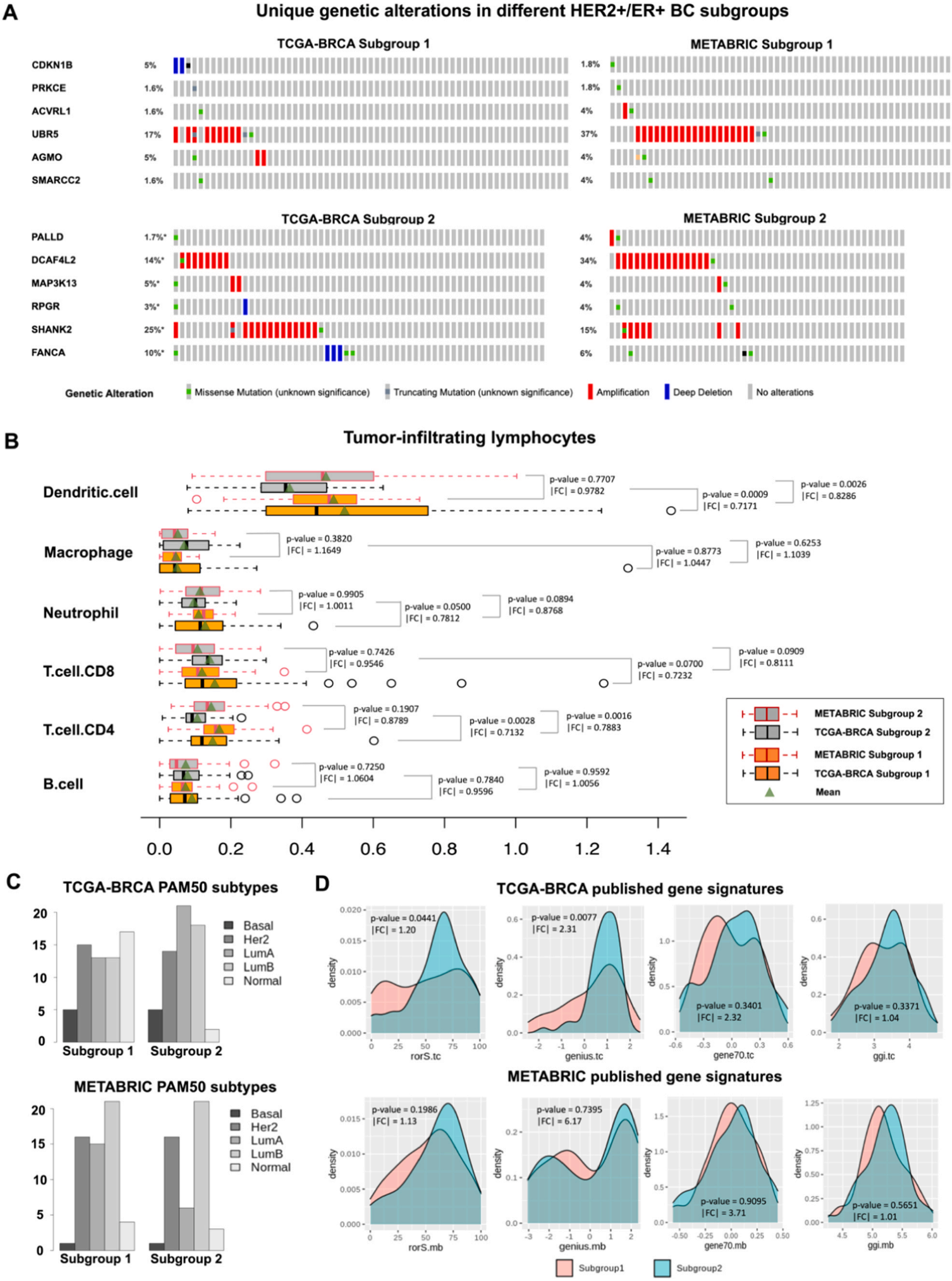

Fig. 5. Computational characterization of the HER2+/ER+ subgroups for both TCGA-BRCA cohort and METABRIC cohort. A top panel: The common genes that are mutated in both TCGA-BRCA Subgroup 1 and METABRIC Subgroup 1. A bottom panel: The common genes that are altered in both TCGA-BRCA Subgroup 2 and METABRIC Subgroup 2. B: The TIMER quantified abundances of tumor-infiltrating lymphocytes for both TCGA-BRCA and METABRIC cohorts. T-test were used to test the significance of the differences. C: Histograms of the PAM50 intrinsic subtypes distributions for two subgroups. D**:** Density plots of the published gene signatures (rorS, GENIUS, GENE70, GGI) of different subgroups in both TCGA-BRCA and METABRIC cohorts.

distant metastasis, and shorter overall survival in patients [43]. *UBR5* exhibits oncogene-like characteristics as it is proposed to promote breast and ovarian cancer growth and metastasis, which makes it an attractive therapeutic target for aggressive BC [44]. Our study further suggests that UBR5 might offer a potential way to target HER2+/ER+ Subgroup 1 as it is amplificated in Subgroup 1 (17% TCGA-BRCA HER2+/ER+ BC Subgroup 1 patients; 37% METABRIC HER2+/ER+ BC Subgroup 1 patients) but not in Subgroup 2. Similarly, several other genetic alterations were unique to Subgroup 2, including *DCAF4L2* amplification (14% TCGA-BRCA HER2+/ER+ BC Subgroup 2 patients; 34% METABRIC HER2+/ER+ BC Subgroup 2 patients) and *SHANK2* amplification (25% TCGA-BRCA HER2+/ER+ BC Subgroup 2 patients; 15% METABRIC HER2+/ER+ BC Subgroup 2 patients). DCAF4L2 belongs to the WD-repeat domain (WDR) protein family, which commonly functions mediating protein-protein interactions [45]. *DCAF4L2* overexpression in human colorectal cancer is associated with a more advanced clinical stage as in lymphatic and distant metastasis. Moreover, overexpression was also found to promote cell migration, invasion, and epithelial-mesenchymal-transition (EMT) through activating NFκB signal pathway [46]. High expression of *DCAF4L2* may be positively associated with poor overall survival of BC [45]. However, it remains to be determined how *DCAF4L2* is implicated HER2+/ER+ BC pathology and whether it could be used as a novel candidate target for HER2+/ER+ treatment. *SHANK2* is one of the *SHANK* family of master scaffolding proteins and plays important roles in regulating synapse plasticity. The *SHANK* family proteins were recently found to be involved in cancer cell invasion [47]. Methylation of *SHANK2* could promote BC cell migration through activating endosome focal adhesion kinase FAK signaling [47]. FAK is a cytoplasmic protein-tyrosine kinase which is important in cell adhesion, survival and migration [48]. *SHANK2* methylation was also identified as a potential biomarker of BC metastasis [47]. However, the precise roles of *SHANK2* in HER2+/ER+ BC have yet to be determined.

The TIMER estimated tumor infiltrations of two types of immune cells (dendritic cells and CD4 + T cells) were significantly lower in HER2+/ER+ BC Subgroup 2 than in Subgroup 1. According to Jin et al., lower tumor infiltration of these two types of immune cells were associated with worse prognosis [49], which is consistent with our findings that Subgroup 2 patients have lower TILs and worse prognosis. PAM50 intrinsic subtypes, the related rorS score, and the Genius score also showed different distributions between two HER2+/ER+ BC subgroups in TCGA-BRCA cohort, indicating a higher survival risk of Subgroup 2 than Subgroup 1. However, these differences were not observed in METABRIC HER2+/ER+ BC cohort. Other published gene signatures such as GENE70 and GGI were not significantly different between two subgroups in both TCGA-BRCA HER2+/ER+ and METABRIC cohorts, suggesting their unsuitable for HER2+/ER+ BC. DeepDep predicted DepMap dependency scores showed visible different pattern between two subgroups in TCGA-BRCA HER2+/ER+ cohort according to our results, which further confirmed the difference between these two subgroups. However, this pattern cannot be reproduced in METABRIC HER2+/ER+ cohort.

In summary, we found that some of the differences between the proposed two HER2+/ER+ subgroups were observed in TCGA-BRCA cohort but not in METABRIC cohort, such as the tumor infiltrations of dendritic and CD4 + T cells, two published gene signatures (rorS and GENIUS), DeepDep predicted gene dependency scores, etc., possibly due to the different acquisition technologies used to obtain the raw gene expression data of these two cohorts [50]. TCGA-BRCA gene expression data were obtained through RNA sequencing, while METABRIC used microarray technology. Another possible reason is that TCGA-BRCA biospecimens were collected from newly diagnosed patients who had received no prior treatment, while lymph node-positive METABRIC patients received chemotherapy before the biopsy [22] [24].

5. Conclusion

In conclusion, our study is the first to explore the heterogeneity within HER2+/ER+ BCs. We identified and validated the potential subgroups of HER2+/ER+ breast tumors with reproducible prognostic and other properties. We tried both single-omics and multi-omics methods, but we decided to focus on single-omics due to the limited sample size of the multi-omics data and the better performance of the single-omics method. We provided a well-trained HER2+/ER + subgroup classifier to assign new patients to a specific subgroup. We also discussed potential biological explanations of the identified subgroups and linked it with existing knowledge of BC. Most important, our findings may provide guidance for future new target therapies of the HER2+/ER+ BC patients.

Ethics approval and consent to participate

Not applicable.

Funding

This work was supported in part by CancerCare Manitoba Foundation, the Canada Research Chairs Tier II Program and the Canadian Institutes of Health Research. P.H. is the holder of Manitoba Medical Services Foundation (MMSF) Allen Rouse Basic Science Career Development Research Award.

Consent for publication

Not applicable.

CRediT authorship contribution statement

QL: Conceptualization, Data curation, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. **SH:** Conceptualization, Data curation, Methodology, Writing – review & editing. **DD:** Conceptualization, Funding, Writing – review & editing. **KM:** Conceptualization, Funding, Writing – review & editing. **LM:** Conceptualization, Supervision, Funding, Writing – review & editing. **PH:** Conceptualization, Data curation, Methodology, Supervision, Funding, Writing – review & editing.

Declaration of Competing Interest

There is no conflict of interest.

Acknowledgements

We thank for the authors of the TCGA-BRCA, METABRIC and DepMap to make the data sets publicly available.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.05.002.

References

[1] American Cancer Society. Cancer Facts & Figures 2022. Atlanta: American Cancer Society. 2022.

[2] Dai X, Xiang L, Li T, Bai Z. Cancer hallmarks biomarkers and breast cancer molecular subtypes. J Cancer 2016;7:1281–94.

[3] Rye IH, Trinh A, Sætersdal AB, Nebdal D, Lingjærde OC, Almendro V, et al. Intratumor heterogeneity defines treatment-resistant HER2+ breast tumors. Mol Oncol 2018;12:1838–55.

[4] Brandão M, Caparica R, Malorni L, Prat A, Carey LA, Piccart M. What Is the real impact of estrogen receptor status on the prognosis and treatment of HER2-positive early breast cancer? Clin Cancer Res 2020;26:2783–8.

[5] Gingras I, Gebhart G, De Azambuja E, Piccart-Gebhart M. HER2-positive breast cancer is lost in translation: time for patient-centered research. Nat Rev Clin Oncol 2017;14:669–81.

[6] Hwang KT, Kim J, Jung J, Chang JH, Chai YJ, Oh SW, et al. Impact of breast cancer subtypes on prognosis of women with operable invasive breast cancer: a population-based study using SEER database. Clin Cancer Res 2019;25:1970–9.

[7] Cameron D, Piccart-Gebhart MJ, Gelber RD, Procter M, Goldhirsch A, de Azambuja E, et al. 11 years' follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive early breast cancer: final analysis of the HERceptin Adjuvant (HERA) trial. Lancet 2017;389:1195–205.

[8] Bender LM, Nahta R. Her2 cross talk and therapeutic resistance in breast cancer. Front Biosci 2008;13:3906–12.

[9] Gianni L, Pienkowski T, Im YH, Roman L, Tseng LM, Liu MC, et al. Efficacy and safety of neoadjuvant pertuzumab and trastuzumab in women with locally advanced inflammatory or early HER2-positive breast cancer (NeoSphere): a randomised multicentre open-label phase 2 trial. Lancet Oncol 2012;13:25–32.

[10] Baselga J, Bradbury I, Eidtmann H, Di Cosimo S, De Azambuja E, Aura C, et al. Lapatinib with trastuzumab for HER2-positive early breast cancer (NeoALTTO): a randomised open-label multicentre phase 3 trial. Lancet 2012;379:633–40.

[11] Carey LA, Berry DA, Cirrincione CT, Barry WT, Pitcher BN, Harris LN, et al. Molecular heterogeneity and response to neoadjuvant human epidermal growth factor receptor 2 targeting in CALGB 40601, a randomized phase III trial of paclitaxel plus trastuzumab with or without lapatinib. J Clin Oncol 2016;34:542–9.

[12] Qian Y, Daza J, Itzel T, Betge J, Zhan T, Marmé F, et al. Prognostic cancer gene expression signatures: current status and challenges. Cells 2021;10:1–17.

[13] Bernard PS, Parker JS, Mullins M, Cheung MCUU, Leung S, Voduc D, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 2009;27:1160–7.

[14] Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, et al. Comprehensive molecular portraits of human breast tumours. Nature 2012;490:61–70.

[15] Joel S, Parker Mullins, Cheang M, Leung MCU, Voduc S, Vickery T D, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 2009;27:1160–7.

[16] Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of Tamoxifen-treated node-negative breast cancer. New Engl J Med 2004;351:2817–26. https://doi.org/10.1056/NEJMoa041588

[17] Filipits M, Rudas M, Jakesz R, Dubsky P, Fitzal F, Singer CF, et al. A new molecular predictor of distant recurrence in ER-Positive HER2-negative breast cancer adds independent information to conventional clinical risk factors. Clin Cancer Res 2011;17:6012–20. https://doi.org/10.1158/1078-0432.CCR-11-0926

[18] Loi S, Haibe-Kains B, Majjaj S, Lallemand F, Durbecq V, Larsimont D, et al. PIK3CA mutations associated with gene signature of low mTORC1 signaling and better outcomes in estrogen receptor - positive breast cancer. Proc Natl Acad Sci 2010;107:10208–13.

[19] van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530–6. https://doi.org/10.1038/415530a

[20] Bontempi G, Sotiriou C, Haibe-Kains B, Rothé F, Piccart M, Desmedt C. A fuzzy gene expression-based computational approach improves breast cancer prognostication. Genome Biol 2010;11:R18.

[21] Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst 2006;98:262–72.

[22] Liu J, Lichtenberg TM, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell 2018;173(400–416):e11.

[23] Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. e71–e71 Nucleic Acids Res 2016;44. https://doi.org/10.1093/NAR/GKV1507

[24] Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 2012;486:346–52.

[25] Cecco L.De. Gene expression profiling of primary HER2-positive breast cancers treated with neoadjuvant trastuzumab. 2021. ⟨https://www.omicsdi.org/dataset/geo/GSE149283⟩. Accessed 2 May 2022.

[26] Leek J.T., Johnson W.E., Parker H.S., Fertig E.J., Jaffe A.E., Zhang Y., Storey JD TL. sva: Surrogate Variable Analysis. 2022. ⟨https://bioconductor.org/packages/release/bioc/html/sva.html⟩.

[27] Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 2010;113(11):1–9. https://doi.org/10.1186/GB-2010-11-3-R25

[28] Liu Q, Cheng B, Jin Y, Hu P. Bayesian tensor factorization-drive breast cancer subtyping by integrating multi-omics data. J Biomed Inf 2022;125:103958.

[29] Xu T, Le TD, Liu L, Su N, Wang R, Sun B, et al. CancerSubtypes: an R/Bioconductor package for molecular cancer subtype identification validation and visualization. Bioinformatics 2017;33:3131–3.

[30] Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering a resampling-based method for class discovery and Vi - monti - mach learn. Mach Learn 2003;52:91–118.

[31] Lloyd SP. Least squares quantization in PCM. IEEE Trans Inf Theory 1982;28:129–37.

[32] Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics 2010;26:1572–3.

[33] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958;53:457–81. https://doi.org/10.1080/01621459.1958.10501452

[34] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015.

[35] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci 2005;102:15545–50. https://doi.org/10.1073/pnas.0506580102

[36] Kuhn M. Building predictive models in R using the caret package. J Stat Softw 2008;28:1–26.

[37] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017:4766–75. 2017-Decem Section 2.

[38] Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 2013:6. https://doi.org/10.1126/SCISIGNAL.2004088

[39] Li B, Severson E, Pignon JC, Zhao H, Li T, Novak J, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biol 2016:17.

[40] Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, et al. Defining a cancer dependency map. Cell 2017;170:564–76.

[41] Chiu YC, Zheng S, Wang LJ, Iskra BS, Rao MK, Houghton PJ, et al. Predicting and characterizing a cancer dependency map of tumors with deep learning. Sci Adv 2021:7.

[42] Martens JHA, Brinkman AB, Simmer F, Francoijs KJ, Nebbioso A, Ferrara F, et al. PML-RARα/RXR alters the epigenetic landscape in acute promyelocytic leukemia. Cancer Cell 2010;17:173–85. https://doi.org/10.1016/j.ccr.2009.12.042

[43] Xiang G, Wang S, Chen L, Song M, Song X, Wang H, et al. UBR5 targets tumor suppressor CDC73 proteolytically to promote aggressive breast cancer. Cell Death Dis 2022 2022;135(13):1–14. https://doi.org/10.1038/s41419-022-04914-6

[44] Song M, Wang C, Wang H, Zhang T, Li J, Benezra R, et al. Targeting ubiquitin protein ligase E3 component N-recognin 5 in cancer cells induces a CD8+ T cell mediated immune response. Oncoimmunology 2020:9. https://doi.org/10.1080/2162402X.2020.1746148/SUPPL_FILE/KONI_A_1746148_SM1333.ZIP

[45] Hu DJ, Shi WJ, Yu M, Zhang LI. High WDR34 mRNA expression as a potential prognostic biomarker in patients with breast cancer as determined by integrated bioinformatics analysis. Oncol Lett 2019;18:3177–87.

[46] Wang H, Chen Y, Han J, Meng Q, Xi Q, Wu G, et al. DCAF4L2 promotes colorectal cancer invasion and metastasis via mediating degradation of NFκb negative regulator PPM1B. Am J Transl Res 2016;8:405 /pmc/articles/PMC4846892/. Accessed 20 Sep 2022.

[47] Liu Y, Li L, Liu X, Wang Y, Liu L, Peng L, et al. Retraction: arginine methylation of SHANK2 by PRMT7 promotes human breast cancer metastasis through activating endosomal FAK signalling. Elife 2021;10:1–27.

[48] Alanko J, Ivaska J. Endosomes: emerging platforms for integrin-mediated FAK signalling. Trends Cell Biol 2016;26:391–8. https://doi.org/10.1016/j.tcb.2016.02.001

[49] Jin YW, Hu P. Tumor-infiltrating cd8 t cells predict clinical breast cancer outcomes in young women. Cancers 2020:12.

[50] Bismeijer T, Canisius S, Wessels LFA. Molecular characterization of breast and lung tumors by integration of multiple data types with functional sparse-factor analysis. PLoS Comput Biol 2018;14:1–28.