A Simple Method for Robust and Accurate Intrinsic Subtyping of Breast Cancer

Cancer Informatics
Volume 22: 1–14
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/11769351231159893

Mehdi Hamaneh and Yi-Kuo Yu

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA.

ABSTRACT

MOTIVATION: The PAM50 signature/method is widely used for intrinsic subtyping of breast cancer samples. However, depending on the number and composition of the samples included in a cohort, the method may assign different subtypes to the same sample. This lack of robustness is mainly due to the fact that PAM50 subtracts a reference profile, which is computed using all samples in the cohort, from each sample before classification. In this paper we propose modifications to PAM50 to develop a simple and robust single-sample classifier, called MPAM50, for intrinsic subtyping of breast cancer. Like PAM50, the modified method uses a nearest centroid approach for classification, but the centroids are computed differently, and the distances to the centroids are determined using an alternative method. Additionally, MPAM50 uses unnormalized expression values for classification and does not subtract a reference profile from the samples. In other words, MPAM50 classifies each sample independently, and so avoids the previously mentioned robustness issue.

RESULTS: A training set was employed to find the new MPAM50 centroids. MPAM50 was then tested on 19 independent datasets (obtained using various expression profiling technologies) containing 9637 samples. Overall good agreement was observed between the PAM50- and MPAM50-assigned subtypes with a median accuracy of 0.792, which (we show) is comparable with the median concordance between various implementations of PAM50. Additionally, MPAM50- and PAM50-assigned intrinsic subtypes were found to agree comparably with the reported clinical subtypes. Also, survival analyses indicated that MPAM50 preserves the prognostic value of the intrinsic subtypes. These observations demonstrate that MPAM50 can replace PAM50 without loss of performance. On the other hand, MPAM50 was compared with 2 previously published single-sample classifiers, and with 3 alternative modified PAM50 approaches. The results indicated a superior performance by MPAM50.

CONCLUSIONS: MPAM50 is a robust, simple, and accurate single-sample classifier of intrinsic subtypes of breast cancer.

KEYWORDS: Breast cancer, intrinsic subtyping, gene expression

RECEIVED: November 2, 2022. **ACCEPTED:** February 7, 2023.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Intramural Research Program of the National Library of Medicine at the National Institutes of Health [grant number: ZIA LM092404].

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Yi-Kuo Yu, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA. Email: yyu@ncbi.nlm.nih.gov

Introduction

The intrinsic subtypes of breast cancer, namely Luminal A (LumA), Luminal B (LumB), HER2-enriched (Her2), Basal-like (Basal), and Normal-like (Normal), have distinct molecular characteristics and prognostic attributes.[1] Thus, it is of great importance to have a robust breast cancer intrinsic subtype classifier. Unlike the clinical subtypes of breast cancer, which are identified using immunohistochemical biomarkers such as estrogen receptor (ER), the intrinsic subtypes are determined based on expression profiling of a gene signature. The most widely used signature/method for intrinsic subtyping of breast cancer, PAM50 uses the expression of a set of 50 genes in conjunction with a nearest centroid approach. Specifically, PAM50 computes the rank correlations between the preprocessed samples and 5 centroids, found using the PAM algorithm,[2] each corresponding to one of the subtypes. Each sample is then assigned the subtype corresponding to its nearest centroid (the one with highest rank correlation).

Despite its popularity and widespread use, PAM50 subtyping has been shown to suffer from lack of robustness: that is, the subtype assigned to a sample from a cohort may change depending on the other samples included in the cohort.[3-7] For example, Patil et al[3] showed that both the number of samples and the percentage of ER-positive samples in a dataset significantly affect the subtyping results. This lack of robustness is due to sample preprocessing, which includes subtraction of a reference profile from each sample. The reference is computed using all samples included in the analysis, and so removing or adding samples to the dataset generally changes the reference and consequently may change the assigned subtypes. Even for the same set of samples, the assigned subtypes may depend on the applied gene expression normalization method.[8] Additionally, there are different PAM50 implementations that differ in the way the reference is computed. The results of these different implementations have also been shown to have less than optimal concordance with each other.[6,9]

To resolve the robustness issue, Patil et al[3] suggested using PAM50 with no preprocessing. On the other hand, 2 recent publications proposed new methods to subtype a breast cancer sample in an "absolute" way, namely in a manner that is independent of all other samples. Raquett and Hallett[6] proposed a set of 151 genes and a set of 100 rules comparing the pair-wise unnormalized expressions of these genes to classify breast cancer samples. Note that, this method, called AIMS, uses unnormalized (no between-sample normalization) gene expressions, thus eliminating the effect of other samples. The authors showed good agreement between the PAM50 subtyping and their results when the method was applied to an independent dataset. To improve on this work, Seo et al developed MiniABS,[7] a method that utilizes an 11-gene signature in conjunction with a Random Forest model (again using unnormalized data) for intrinsic subtype classification. MiniABS was tested on multiple datasets demonstrating improved performance.

This paper aims to classify breast cancer samples in a way that is robust and more accurate than previously published approaches. To this end, we propose a modified version of PAM50, called MPAM50, which uses unnormalized expression values for subtyping and thus avoids the issue of lack of robustness. We show that: (1) MPAM50- and PAM50-assigned subtypes are in overall good agreement, (2) MPAM50 preserves the prognostic value of the intrinsic subtypes, and (3) MPAM50 and PAM50 perform comparably in terms of agreement between the assigned intrinsic subtypes and the reported clinical subtypes. Additionally, we compare our results to those of AIMS, MiniABS, and 3 alternative modified PAM50 approaches (including the one suggested by Patil et al[3]), and show a superior performance by MPAM50. These findings suggest MPAM50 is a robust and accurate method for intrinsic subtyping of breast cancer.

Methods

Overview

MPAM50, like PAM50, takes a nearest centroid approach for intrinsic subtyping of breast cancer. In other words, each intrinsic subtype is represented by a centroid vector, calculated using expression profiles form a training set, and a patient is assigned the subtype whose centroid is nearest to the patient's expression profile. Additionally, MPAM50 utilizes the PAM50 genes, namely each centroid is a 50-dimensional vector whose elements correspond to these 50 genes. However, MPAM50 and PAM50 differ in 3 ways: in MPAM50 (1) each centroid is simply calculated as the average of the weighted unnormalized log-transformed expression profiles in the training set (see the next subsection for details), (2) each sample is classified independently using unnormalized log-transformed expression values (no reference subtraction), and thus the issue of lack of robustness is resolved, and (3) Pearson (instead of rank) correlations are used to measure the distances from the centroids.

We used publicly available data (see the "Data" subsection) to find the MPAM50 centroids and to assess the performance of MPAM50. We tested MPAM50 in 3 ways: (1) computing the prediction accuracy of MPAM50 (and comparing the accuracy to those achieved by previously published methods), (2) comparing the survival probability curves predicted by MPAM50 and PAM50, and (3) evaluating the agreement between the predicted intrinsic subtypes and the corresponding clinical subtypes. Detailed descriptions of these tests are given in the following subsections. To give a visual overview of the study, a flowchart describing the step by step procedure is shown in Figure 1.

The MPAM50 centroids

In this subsection, the method we used to find the MPAM50 centroids is explained. Given a training set with known PAM50-assigned subtypes, the method uses the average of the weighted (log-transformed) unnormalized samples in each intrinsic subtype as the centroid corresponding to that subtype. We chose averages as the centroids because such centroids have been reported to lead to excellent results for subtyping of lung cancer across tens of independent datasets obtained using different technologies.[10,11] To achieve a robust classification each sample should be classified independently, and so classification is performed on the unnormalized samples, namely samples that have not been between-sample normalized or preprocessed in a way that is dependent on other samples (although the samples may be within-sample normalized/preprocessed). Hence, the unnormalized samples are used for calculating the centroids. Mathematically, our approach can be formulated as follows. Let m and n be respectively the number of genes in the signature and the number of subtypes present in the data (here $m = 50$ and $n = 5$). Also, let E_j denote the $m \times 1$ matrix containing the unnormalized but log-transformed expression levels of the m genes in the j th sample of the training set. For the k th subtype, the i th element of the $m \times 1$ centroid vector is given by

$$X_{ik} = \frac{1}{M_k} \sum\nolimits_{j \in C_k} \frac{E_{ij}}{\sigma\left(E_j\right)}, \tag{1}$$

where C_k is the set of all samples in the training set that are in the k th subtype, M_k is the number of such samples, and $\sigma(*)$ denotes the standard deviation of the elements of the vector $*$. Since we are using unnormalized data, expression levels in different samples are generally not comparable to each other. Thus, to make the samples somewhat comparable, we scale them to have the same standard deviation. Note that there is no need for centering (making the means of the samples equal), because centering would only add a constant to the average (centroid) and would not affect the correlations.

Given the centroids obtained above, the j th sample is assigned to the k th subtype if X_k is the closest centroid to the sample's log-transformed, unnormalized expression profile $\boldsymbol{E_j}$ (a nearest neighbor approach). The measure of closeness between $\boldsymbol{E_j}$ and $\boldsymbol{X_k}$ is given by

$$R_{jk} = r\left(E_j, X_k\right), \tag{2}$$

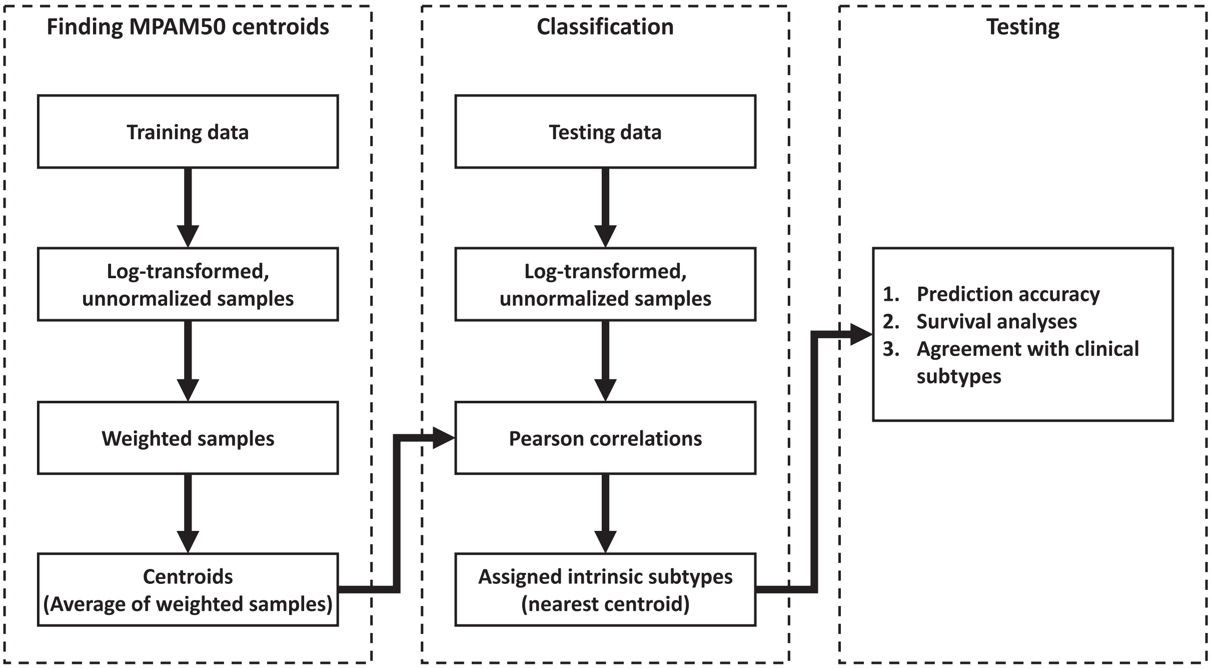

Finding MPAM50 centroids
Training data
Log-transformed, unnormalized samples
Weighted samples
Centroids (Average of weighted samples)
Classification
Testing data
Log-transformed, unnormalized samples
Pearson correlations
Assigned intrinsic subtypes (nearest centroid)
Testing
1. Prediction accuracy
2. Survival analyses
3. Agreement with clinical subtypes

Figure 1. Overview of the study. In this study, we first found the MPAM50 centroid corresponding to each intrinsic subtype. For a given subtype, the centroid was calculated as the weighted average of the log-transformed unnormalized samples (in the training set) that had been assigned that subtype by PAM50 (see equations (1) and (3)). The centroids were then used to classify all samples in the testing set (See the Data subsection and Table 1). Each sample was assigned the subtype whose centroid had the largest Pearson correlation with the sample. The performance of MPAM50 was then assessed in 3 ways: (1) comparing the prediction accuracy of MPAM50 with (a) concordance between 2 implementations of PMA50, and (b) accuracy of the previously published robust classifiers, (2) comparing the survival curves predicted by MPAM50 and PAM50, and (3) comparing MPAM50 and PAM50 in terms of the agreement between the assigned intrinsic subtypes and the reported clinical subtypes.

where $r\left(\boldsymbol{E}_j, \boldsymbol{X}_k\right)$ is the Pearson correlation between these 2 vectors. Henceforth, we refer to R_{jk} as the score of the j th sample relative to the k th centroid.

Generalization to more than one training dataset. To increase the number of samples, and to include expression data obtained using different technologies (eg, RNA-Seq and microarray), one may want to have multiple datasets in the training set.[6,7,12] In such a case, we define the generalized score of the j th sample relative to the k th subtype as $R_{jk} = \left(1/L\right)\sum_{l=1}^{L} r\left(E_j, X_k^l\right)$ where L is the number of datasets included in the training set, and $\boldsymbol{X}_k^l$ denotes the centroid (defined above) of the l th dataset corresponding to the k th subtype. Note that R_{jk} is proportional to $r\left(\boldsymbol{E}_j, \boldsymbol{X}_k\right)$, where

$$X_k = \sum_{l=1}^{L} \frac{X_k^l}{\sigma\left(X_k^l\right)}. \tag{3}$$

In other words, when multiple datasets are present in the training set, the centroids for each dataset are first calculated independently and then the overall centroids are found using equation (3).

Data

Training set. To include both RNA-Seq and microarray data, and to increase the number of samples, we included $L = 2$ datasets in our training set as described below. We used RNA-Seq data from The Cancer Genomic Atlas (TCGA) as a part of the training set. The within-sample normalized, level-3 TCGA expression data (RPKM values) were downloaded from the Broad Institute's Firehose (https://gdac.broadinstitute.org/), and the assigned PAM50 subtypes were obtained from Ciriello et al.[13] Patients for which either PAM50 subtyping or gene expression levels were not available were excluded, leaving 108 Basal, 56 Her2, 317 LumA, 158 LumB, and 18 Normal samples.

Also added to the training set was the largest microarray dataset (GSE115577) that we could find in the Gene Expression Omnibus (GEO) database.[14] For GSE115577,[9] Affymetrix CEL files (tumor samples only) and PAM50 subtypes were downloaded from GEO. Each sample of GSE115577 has been assigned 2 potentially different PAM50 subtypes, using 2 different reference calculation methods. Any sample that had been assigned different subtypes using the 2 methods was excluded. The remaining data included 110 Basal, 119 Her2, 346 LumA, 153 LumB, and 32 Normal microarray samples. Before using the data from these 2 datasets in the training set, the expression levels were preprocessed as explained in the "Preprocessing" subsection below.

Testing set. To construct a testing set, a search was conducted in the GEO database to find independent breast cancer

Table 1. Datasets used for testing MPAM50.

DATASET ID	BASAL	HER2	LUMA	LUMB	NORMAL	TOTAL	TECHNOLOGY
GSE112063	24	35	108	61	10	238	qRT-PCR
GSE126870	5	14	113	40	6	178	microarray (Illumina)
GSE148426	764	279	995	335	124	2497	NanoString
GSE18229	80	40	96	61	28	305	microarray (Agilent; two-color)
GSE22226	48	22	43	28	8	149	microarray (Agilent; two-color)
GSE22358	45	22	47	25	15	154	microarray (Agilent; two-color)
GSE25066	189	37	160	78	44	508	microarray (Affymetrix)
GSE26304	20	16	24	30	19	109	microarray (Agilent; two-color)
GSE41119	53	21	48	33	8	163	microarray (Agilent; two-color)
GSE41998	110	23	91	33	22	279	microarray (Affymetrix)
GSE53031	50	22	53	33	9	167	microarray (Affymetrix)
GSE54275	40	22	90	86	5	243	microarray (Agilent; two-color)
GSE56493	30	38	12	34	6	120	microarray (Affymetrix)
GSE59246	16	17	32	21	16	102	microarray (Agilent; one-color)
GSE80999	45	42	158	89	44	378	microarray (Agilent; one-color)
GSE81538	57	65	156	105	22	405	RNA-Seq
GSE86374	14	18	50	27	14	123	microarray (Affymetrix)
GSE92977	27	41	80	76	22	246	NanoString
GSE96058	339	327	1657	729	221	3273	RNA-Seq
Total	1956	1101	4013	1924	643	9637	

datasets that: (1) had publicly available PAM50-assigned intrinsic subtypes, (2) included raw data or within-sample (but not between-sample) normalized values, and (3) contained at least 100 samples and included all 5 intrinsic subtypes. The last criterion was adopted to ensure a reliable assessment of the performance of MPAM50. Of note, there are different implementations of PAM50 and the way the samples are preprocessed before calculating their rank correlations with the PAM50 centroids is different in various implementations. We did not limit our search to a specific implementation or a particular way of calculating the reference (as long as such a reference had been subtracted from the data). We found 19 datasets satisfying our criteria comprising 9637 samples including 1956 Basal, 1101 Her2, 4013 LumA, 1924 LumB, and 643 Normal samples (Table 1). The raw (or, in the case of RNASeq, within-sample normalized) expression levels and, if available, the clinical data for these 19 datasets were downloaded from GEO. The downloaded expression levels were preprocessed as explained in the subsection below. For each dataset, excluding GSE41998, the PAM50 subtype assignments were obtained from the GEO record or the corresponding publication. The assigned PAM50 subtypes of samples in GSE41998 were collected from Prat et al.[15] In GSE54275 expression levels have been profiled using 2 different microarray technologies. For this study we used the ones obtained using the Agilent platform.

Preprocessing. For data obtained using different technologies, the downloaded data were preprocessed as follows:

RNA-seq: A Pseudocount of 1 was added to the downloaded RPKM values of the TCGA (training) data before log-transformation. For the 2 RNA-Seq datasets included in the testing set (Table 1) the log-transformed, within-sample normalized expression levels were downloaded and no preprocessing was performed.

microarray: For datasets obtained using the Affymetrix platforms, including GSE115577 used for training, the oligo[16] (or, for some of the older platforms, affy[17]) package of Bioconductor was employed to apply the robust multi-array average (RMA) algorithm to each raw CEL file independently. In other words, each sample was separately background-adjusted and expression values were summarized to the probe-set level, but no between-sample normalization was performed. Each sample in one-color Agilent datasets was independently background-corrected using the limma[18] package of Bioconductor. In the 2-color Agilent datasets, only the channel corresponding to the tumor was considered and background-corrected.[6,7] The limma package was also used to separately background-correct each sample of the sole included microarray dataset (GSE126870) acquired using an Illumina platform. For each microarray dataset (Affymetrix, Agilent, or Illumina), the probe level background-adjusted expression values were log-transformed. After log-transformation, replicate probes in Agilent samples (1- or 2-color) were collapsed by averaging their expression values. Probe IDs were mapped to gene IDs using the corresponding annotation file in GEO. If multiple probes mapped to a gene, the expression level of the gene was computed by averaging those of the corresponding probes.

NanoString: Each sample from GSE148426 was separately background-corrected by subtracting the median count of the negative probes (in the sample) from the counts of all genes. Any corrected count smaller than 1 was set to 1 and the data were subsequently log-transformed. In the case of GSE92977 the negative probes' counts were not available, and so we used the log-transformed raw counts.

RT-PCR: Except for log-transformation, no preprocessing was performed on the raw data in GSE112063 (the only dataset obtained using RT-PCR that we found).

Prediction accuracy

To assess the performance of MPAM50, the samples in each of the 19 datasets in the testing set were classified and the resulting subtype assignments were compared with PAM50-assigned subtypes. Specifically, we computed the prediction accuracy (ACC), that is the number of correct predictions divided by the total number of predictions, and the subtype-specific prediction accuracy for subtype k defined, as $ACC_k = N_k / M_k$ ($k = 1,2,\ldots,5$). Here, N_k is the number of correctly-classified samples in subtype k, and M_k is the total number of samples in that subtype. The balanced accuracy ACC_b, that is the unweighted average of the subtype-specific accuracies, was also calculated ($ACC_b = \left(1/n\right)\sum_{k=1}^{n} ACC_k$). As the name suggests ACC_b measures how balanced the performance of a method is in terms of predicting different subtypes. Since, in the case of breast cancer, LumA is the most prevalent subtype (see Table 1), a method with a high ACC_{LumA} may achieve a high ACC without being that effective in identifying the other subtypes. Thus, it is important to look at ACC_b when comparing different classifiers. Specifically, if 2 methods have comparable ACC s we regard the one with larger ACC_b a better classifier.

To see if the ACC achieved by MPAM50 was acceptable, for each dataset in the testing set we compared the ACC with the concordance between different implementations of PAM50. Concordance, denoted by *C*, is defined as the number of samples that have been assigned the same subtype using 2 implementations of PAM50, divided by the total number of samples. We used the genefu package[19] of Bioconductor to subtype the samples in each of the aforementioned 19 datasets. As a result, corresponding to each dataset there were 2 sets of PAM50-assigned subtypes: one reported in the literature, and one determined by us using genefu. (In the rest of the paper, we refer to these sets as the "reported" and "genefu" subtypes respectively.) For each dataset, we then compared the 2 sets of PAM50-assigned subtypes and computed the concordance. Of note, even within genefu there are multiple approaches to preprocess the gene expression before calculating the rank correlations with the PAM50 centroids. We used "pam50.robust" as the classification model. As input to genefu, except for the RNA-Seq datasets, we used the between-sample normalized and log-transformed expression levels (downloaded from GEO). For the 2 RNA-Seq datasets (Table 1) only the log-transformed within-sample normalized expression values were available, and thus we used these values for PAM50 subtyping using genefu.

One way to assess performance across multiple datasets is to pool the scores and calculate the overall performance measures using the pooled scores. However, Table 1 indicates a large imbalance between the numbers of samples in the included datasets (GSE148426 and GSE96058, contain more than half of the samples). Given that our goal was to evaluate the performance of MPAM50 on data obtained employing different platforms and subtyped using various implementations of PAM50, and due to the large differences in the numbers of samples, pooling the scores was not the best approach for overall performance assessment. Hence, we opted to use the mean, denoted by $\langle * \rangle$, and the median, denoted by $Med\left(*\right)$. Here $*$ denotes any of the performance measures defined above.

Survival analyses

As a second way of testing MPAM50, the MPAM50-predicted survival probabilities for different subtypes were compared with those predicted by PAM50. For datasets reporting survival data (see Table 2), survival analyses were performed using the MPAM50- and PAM50-assigned subtypes. GSE59246 was excluded from this analysis because survival data were available for few samples in this dataset (eg, only 1 Basal sample). Kaplan-Meier survival analyses were performed using the survival package of Bioconductor and the statistical significance of the differences between survival probabilities were assessed using the log-rank test. The differences between survival curves were regarded as significant if the *p*-value was smaller than 0.05. The

Table 2. Datasets containing survival data.

DATASET	BASAL	HER2	LUMA	LUMB	NORMAL	TOTAL	SURVIVAL DATA TYPE
GSE18229	57	29	83	49	23	241	OS
GSE18229	57	29	84	49	23	242	RFS
GSE22226	48	22	43	28	8	149	OS
GSE22226	48	22	43	28	8	149	RFS
GSE25066	189	37	160	78	44	508	RFS
GSE26304	20	16	24	30	19	109	OS
GSE41119	52	21	43	33	7	156	OS
GSE53031	50	22	53	33	9	167	RFS
GSE96058	339	327	1657	729	221	3273	OS

Note that survival information for some of the samples in these datasets are not available, and so the numbers mentioned in this table may be different from those given in Table 1.

figures depicting the survival probabilities (see Results) were generated using the survminer package. Some datasets mentioned in Table 1 contain both overall survival (OS) and relapse-free survival (RFS) data. For these datasets, survival analyses were performed for both OS and RFS data.

Agreement between intrinsic and clinical subtypes

As another way of testing MPAM50, we investigated the degree of overlap between the intrinsic subtypes assigned by MPAM50 and their corresponding clinical subtypes. The clinical subtypes are assigned based on the status of the biomarkers ER, PR, HER2, and sometimes KI67, determined using immunohistochemistry. If the status of all 4 markers are known, 4 clinical subtypes can be distinguished that are luminal A-like (LumA-like), luminal B-like (LumB-like), HER2 + (non-luminal), and triple negative (TN). Based on the status of the 4 markers, the clinical subtypes are defined as follows[1,20]:

- LumA-like: ER+, PR+, HER2–, and KI67 low
- LumB-like: (1) ER+, HER2–, and (KI67 high or PR–), or (2) ER+ and HER2+
- HER2 + (non-luminal): ER–, PR– and HER2+
- TN: ER–, PR–, and HER2-.

Here +/– after a marker means the status of the marker is positive/negative. These 4 clinical subtypes correspond respectively to LumA/Normal, LumB, Her2, and Basal intrinsic subtypes. (Note that the Normal class does not have its own corresponding clinical subtype.) Based on these definitions, for each clinical subtype, we calculated the prediction ACC, that is the fraction of the samples in the clinical subtype that have been assigned the corresponding intrinsic subtype. We denote the ACC s for the LumA-like, LumB-like, HER2 + (non-luminal), and TN by ACC_{LAL}^{cl}, ACC_{LBL}^{cl}, ACC_{HER2+}^{cl}, and ACC_{TN}^{cl} respectively. Here, the superscript ^{cl} indicates that the clinical subtypes are chosen as the gold standard.

Table 3. Numbers of LumA-like and LumB-like samples in the 3 datasets that have reported the status of all 4 markers.

DATASET	LUMA-LIKE	LUMB-LIKE
GSE26304	19	65
GSE81538	182	141
GSE96058	519	742
Total	720	948

To distinguish LumA-like from LumB-like, the status of KI67 must be known. However, only 3 of the 19 datasets have included information regarding the status of all 4 markers (Table 3). We used these datasets to calculate ACC_{LAL}^{cl} and ACC_{LBL}^{cl}. Seven additional datasets have reported the status of only ER, PR, and HER2 markers (Table 4). To find ACC_{HER2+}^{cl} and ACC_{TN}^{cl}, samples from all of these 10 datasets were included. For comparison, the ACC^{cl} s were also computed for the reported PAM50 subtypes.

Results

This section is organized as follows. First, in the "Finding the centroids" subsection, we present the MPAM50 centroids computed using the approach described in Methods. In the subsequent subsections the results of testing MPAM50 are presented. As mentioned in Methods, and shown in Figure 1, testing was performed in 3 areas: (1) prediction accuracy, (2) agreement between the predicted intrinsic subtypes and the corresponding reported clinical subtypes, and (3) comparing the survival curves predicted by MPAM50 and those predicted by PAM50. Finally, we compare the performance of MPAM50

Table 4. The number of triple negative and HER2 + (non-luminal) samples in datasets containing the status of the 3 markers ER, PR, and HER2.

DATASET	TRIPLE NEGATIVE	HER2 + (NON-LUMINAL)
GSE18229	58	18
GSE22358	50	19
GSE25066	178	3
GSE26304	3	13
GSE41998	140	16
GSE53031	33	8
GSE54275	41	12
GSE59246	10	11
GSE81538	63	17
GSE96058	143	56
Total	719	173

(in terms of prediction accuracy) with AIMS, MiniABS, and 3 alternative modified PAM50 methods.

Finding the centroids

We used our training set (consisting of data from TCGA and GSE115577) in conjunction with equations (1) and (3) to find a set of centroids for robust subtyping of breast cancer samples (using unnormalized samples and employing a nearest neighbor approach; see Methods for details). The expression values for 2 genes (KRT5 and KRT17) were missing in most of the GSE115577 samples. When calculating the standard deviations of samples with missing expression values, these genes were excluded (see equation (1)). Also, samples with missing values were excluded when computing the averages for these 2 genes. The resulting centroids are given in Table 5.

Prediction accuracy

MPAM50 was tested using 19 independent datasets containing 9637 samples (Table 1). Since the datasets have been obtained using various technologies covering different genes, expression levels for some of the PAM50 genes were not available in some of the datasets. In such cases the missing genes were ignored, and the sample scores were calculated using only the available expression values. All samples in each dataset were classified using MPAM50 and the resulting subtype assignments were compared with the reported subtypes and the performance measures were calculated. The ACC s for each dataset as well as the overall performance measures that are the median ($Med(ACC)$) and average ($\langle ACC \rangle$) are shown in Figure 2 (and Supplemental Table S1). The figure indicates overall good agreement between the predictions of MPAM50 and those of PAM50 ($Med(ACC) = 0.792$ and $\langle ACC \rangle = 0.773$), although a significant level of variability is observed in the individual ACC s. We address this issue at the end of this subsection.

For comparison, the concordance between the reported and genefu subtypes for each dataset and the overall concordance measures are also given in Figure 2, showing overall comparable ACC and C (with $Med(ACC) = 0.792$ barely higher than $Med(C) = 0.791$ and $\langle ACC \rangle = 0.773$ slightly lower than $\langle C \rangle = 0.803$). We also used the genefu subtypes as the gold standard and calculated the ACC^gs, $\langle ACC^g \rangle$, and $Med(ACC^g)$, where the superscript g indicates that the genefu subtypes were used as the gold standard. (Note that the concordance C remains the same.) The results, given in Supplemental Figure S1, show even higher performance measures: $Med(ACC^g) = 0.812$ and $\langle ACC^g \rangle = 0.784$, with no statistically significant difference between $Med(ACC^g)$ and $Med(C)$ ($p = 0.084$; Wilcoxon signed-rank test). We thus conclude that the overall ACC of MPAM50 is comparable to the concordance between the 2 implementations of PAM50. Because of the high concordance between the reported and genefu subtypes and the good agreement between MPAM50 subtypes with both, in the rest of the paper all comparisons are made with the reported subtypes.

Note that in 2 of the 19 datasets (GSE54275 and GSE59246) the reported subtypes were obtained using the genefu package. In other words, for these 2 datasets the reported and genefu subtypes should be essentially the same. However, Figure 2 shows smaller than unity, although high, concordance for each of these 2 datasets, indicating that even using the same package for PAM50 subtyping may lead to slightly different results. These differences are presumably due to slightly different preprocessing steps (eg, when multiple probes map to the same genes, using maximum expression level instead of average) and/or various input parameters for genefu (eg, choosing "pam50.scale" rather than "pam50.robust" as the subtyping model).

To show how samples misclassified by MPAM50 are distributed among different subtypes, the average row-normalized confusion matrix is shown in Figure 3. (The individual confusion matrices were first row-normalized and then averaged over the 19 datasets.) The diagonal elements of the matrix are the average subtype-specific ACCs. The subtype-specific ACCs and ACC_b for each individual dataset are given in Supplemental Tables S2 to S7. The off diagonal elements show, on average, what fraction of the samples in each subtype have been assigned to the other subtypes. The figure indicates rather large (> 0.2) confusion between LumA and LumB/Normal. However, in the following subsection we show that there is no statistically significant difference between the LumA survival probabilities predicted by MPAM50 and PAM50. We also show that the same statement is true for LumB. Additionally,

Table 5. The centroids.

GENE	BASAL	HER2	LUMA	LUMB	NORMAL
NAT1	−1.8815	−1.1336	0.6876	0.1344	−0.8979
BIRC5	1.0642	0.9932	0.1667	0.7708	0.2418
BAG1	−0.0440	0.0139	0.4936	0.2386	0.2916
BCL2	−1.0482	−1.1864	0.4463	0.0777	−0.1383
BLVRA	0.4190	1.1671	0.9563	1.0159	0.8235
CCNB1	−0.1642	−0.2964	−0.6547	−0.2424	−0.7080
CCNE1	−0.1984	−0.5136	−1.0553	−0.9446	−0.8388
CDC6	−0.8473	−0.6287	−1.1969	−0.7754	−1.2769
CDC20	0.5239	0.3183	−0.2775	−0.0333	−0.1311
CDH3	0.7597	0.2234	−0.1977	−0.4803	0.5443
CENPF	0.1401	−0.6188	−0.8698	−0.3119	−1.0058
EGFR	−0.2164	−0.6461	−0.7872	−1.4150	0.2559
ERBB2	0.6577	2.9088	1.3654	1.4447	1.0610
ESR1	−1.5101	−0.8863	1.5317	1.5446	0.0423
FGFR4	−0.7874	0.4590	−0.5658	−0.4579	−0.4078
FOXC1	0.6055	−0.7535	−0.7232	−1.2302	−0.0559
GRB7	−0.2234	1.2796	0.0310	0.0820	−0.0707
FOXA1	−1.0925	1.4685	1.8434	1.8405	0.6735
KRT5	2.3604	0.0954	1.1495	−0.3498	2.7535
KRT14	1.0643	−0.1618	0.5366	−0.6446	1.9523
KRT17	2.0644	0.6693	1.1621	0.0708	2.2951
MAPT	−1.2827	−0.6363	0.6428	0.1810	−0.3186
MDM2	−0.1837	−0.3420	0.4540	0.4513	0.2107
MKI67	0.1118	−0.3320	−0.8012	−0.2035	−0.7981
MMP11	1.1185	2.1429	1.5898	1.6648	0.9654
MYBL2	0.7435	0.7709	−0.1983	0.5395	−0.0843
MYC	1.7727	1.1208	1.0994	1.2417	1.6922
PGR	−2.3740	−2.1231	−0.2876	−0.8880	−1.2773
RRM2	0.1274	0.3175	−0.4639	−0.0338	−0.3440
SFRP1	1.0482	−1.2237	−0.3756	−1.6827	1.4060
TYMS	1.0059	0.4207	0.2914	0.7431	0.3003
MIA	0.5431	−0.4382	−0.2900	−0.5077	0.3275
EXO1	−1.0473	−1.2751	−1.6132	−1.3873	−1.5378
PTTG1	0.2019	0.0010	−0.4422	0.0131	−0.2941
MELK	−0.3493	−0.7280	−1.2107	−0.8469	−1.1098
NDC80	−0.7847	−1.2764	−1.5320	−1.2364	−1.3839

(Continued)

Table 5. (Continued)

GENE	BASAL	HER2	LUMA	LUMB	NORMAL
KIF2C	−0.3243	−0.5926	−0.9626	−0.7322	−0.8662
UBE2C	1.6995	1.7526	0.6396	1.5067	0.6018
ORC6	−0.1906	−0.3984	−0.7401	−0.6807	−0.6072
SLC39A6	0.4881	0.7485	2.5871	2.6863	1.4932
PHGDH	0.6422	0.2431	−0.3315	−0.4334	0.3016
GPR160	−1.0812	0.1392	0.3226	0.2460	−0.1291
UBE2T	0.2997	0.0650	−0.3758	0.1827	−0.5809
CXXC5	0.1591	1.3373	1.1573	1.5406	0.7239
ANLN	−0.1825	−0.6269	−1.1305	−0.6945	−1.0938
CEP55	−0.6040	−0.9049	−1.3520	−1.0114	−1.1786
ACTR3B	−0.1884	−0.7469	−0.4118	−0.5548	−0.4771
MLPH	−0.7425	1.2217	1.7598	1.6378	0.7796
NUF2	−0.7247	−1.1265	−1.3708	−1.0677	−1.4124
TMEM45B	−1.5474	−0.2815	−0.6960	−1.0080	−0.7125

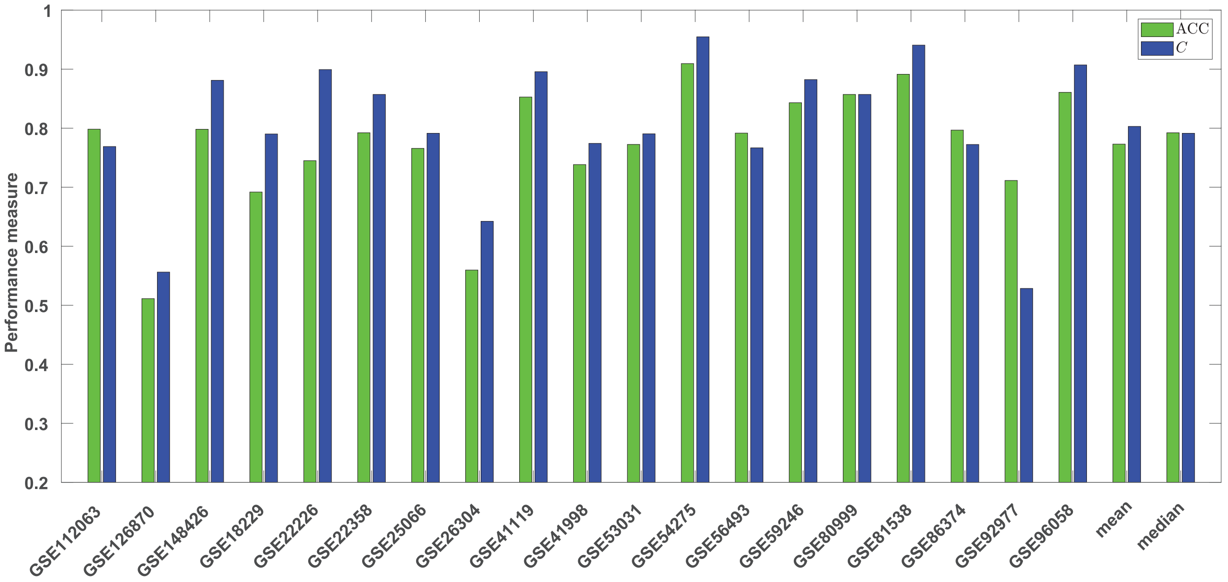


Figure 2. Prediction accuracy. The performance measures ACC of MPAM50 and the concordance C between the reported and genefu subtypes are plotted for each of the 19 datasets included in the testing set. Also shown are the average and median of the 2 measures.

we demonstrate that MPAM50 and PAM50 subtypes have overall comparable concordance with the clinical subtypes defined in the Methods. These observations suggest that MPAM50 can result in robust subtyping without loss of clinical relevance.

We now address the issue of variability in ACC s (Figure 2). We note that:

- The same level of variability is seen in the concordance between the 2 implementations of PAM50. In fact, $\sigma(C) = 0.119$ is slightly larger than $\sigma(ACC) = 0.102$.
- The variability does not appear to be due to differences in gene expression quantification technologies used in various datasets. For example, the data in both GSE26304 and GSE54275 have been obtained using Agilent microarray technology. However, the ACC for the former (0.560) is significantly smaller than that for the latter (0.909). Interestingly, although the data in the training set was obtained using microarray/RNASeq, the ACCs for 2 out of the 3 datasets not using microarray or RNA-seq technology (GSE148426 and GSE112063) are higher than average. These observations suggest that

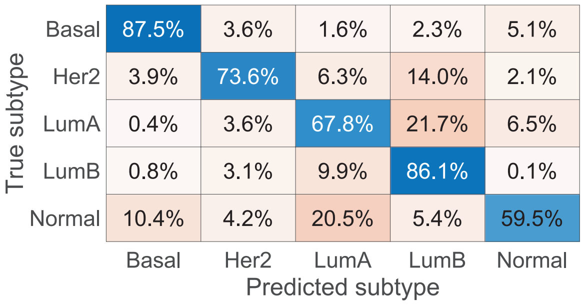
MPAM50 performs well regardless of the platform used for expression profiling.

- A low ACC does not necessarily mean bad performance in terms of survival prediction or overlap between the intrinsic and clinical subtypes. For example, in the following subsections we show that, in the case of GSE26304 ($ACC = 0.560$), there is no statistically significant difference between the results of PAM50 and MPAM50 in terms of survival probabilities or overlap with clinical subtypes.

Based on these observations we conclude that the variability in ACCs does not diminish the usefulness of MPAM50.

Agreement between intrinsic and clinical subtypes

We first considered the samples that, based on the status of ER, PR, and HER2 markers, were assigned to HER2 + (non-luminal) or triple negative clinical subtypes (892 samples from 10 datasets; see Table 4). Using the clinical subtypes as the gold standard and MPAM50 subtypes as predicted ones, for each of the 10 included datasets ACC^{cl}_{HER2+} and ACC^{cl}_{TN} were calculated (see Methods for details). Similarly, the ACCs for the 2 remaining clinical subtypes, LumA-like and LumB-like (ACC^{cl}_{LAL} and ACC^{cl}_{LBL}), were computed for the samples in the 3 datasets that included the status of KI67 in addition to those of ER, PR, and HER2 (1668 samples; see Table 3). The medians of ACC^{cl}_{HER2+}, ACC^{cl}_{TN}, ACC^{cl}_{LAL}, and ACC^{cl}_{LBL} are shown in Figure 4. The corresponding values for individual datasets and also the mean values are given in Supplemental Tables S8 and S9. For comparison, the median subtype-specific ACC^{cl}s achieved by the reported PAM50 subtypes are also shown in the figure. The figure shows similar performances, with MPAM50 performing slightly better although the differences are not statistically significant ($p > 0.1$ in all cases; Wilcoxon signed-rank test). Figure 5 shows the ACC^{cl} s achieved by MPAM50 and PAM50 for the 3 datasets for which all 4 clinical subtypes were defined. The figure indicates comparable performances by the 2 methods in the cases of GSE81538 and GSE96058, but shows a higher PAM50 ACC^{cl} for GSE26304. However, in this case the difference seen in the figure is not statistically significant ($p = 0.21$ McNemar's test). Thus, we conclude that MPAM50 and PAM50 perform comparably in predicting clinical subtypes.



	Basal	Her2	LumA	LumB	Normal
Basal	87.5%	3.6%	1.6%	2.3%	5.1%
Her2	3.9%	73.6%	6.3%	14.0%	2.1%
LumA	0.4%	3.6%	67.8%	21.7%	6.5%
LumB	0.8%	3.1%	9.9%	86.1%	0.1%
Normal	10.4%	4.2%	20.5%	5.4%	59.5%

Figure 3. The confusion matrix. The mean row-normalized confusion matrix, averaged over the 19 datasets, is shown.

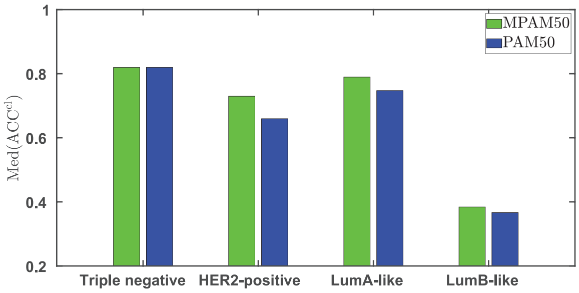


Figure 4. The median class-specific accuracy of predicting the clinical subtypes. The median subtype-specific accuracies achieved by PAM50 and MPAM50 have been plotted for each of the 4 clinical subtypes. For a given clinical subtype the median was computed over all datasets for which that clinical subtype was defined (see text).

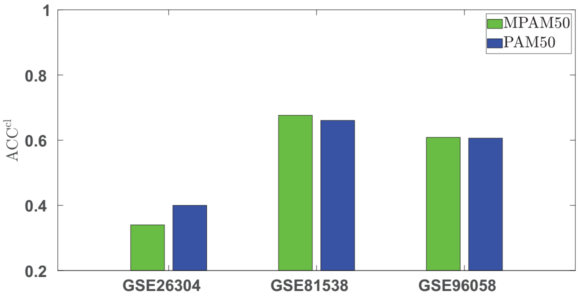


Figure 5. The accuracy of predicting the clinical subtypes. For the 3 datasets with available information regarding the status of all 4 markers, the ACC^{cl} s achieved by PAM50 and MPAM50 are plotted.

Survival analysis

Perhaps the most useful aspect of the intrinsic subtypes is their prognostic value, with LumA, LumB, Normal, and Her2/Basal having respectively good, intermediate, intermediate, and poor prognosis.[1] To investigate if subtypes assigned using MPAM50 preserve the distinction between survival probabilities of different subtypes, we performed survival analyses for the 7 datasets that have reported survival data (Table 2). Survival analysis was performed for each of the 7 datasets separately; that is, because of the large differences in median follow-up times (ranging from 29 to 87 months) and the significant imbalance in the numbers of samples, we did not pool the data. These 7 datasets contain overall survival (OS) or relapse-free survival (RFS) data, with 2 (GSE182229 and GSE22226) having both types of data available (Table 2). For these 2 datasets survival analyses were performed for both OS and RFS data. For comparison, survival analyses were also performed using the reported intrinsic subtypes assigned by PAM50.

The ability of MPAM50 to preserve the prognostic value of the subtypes was assessed in 2 different ways. First, for each subtype and each dataset, we investigated if there was any statistically significant difference between the survival curves obtained using the 2 methods (MPAM50 and PAM50). We

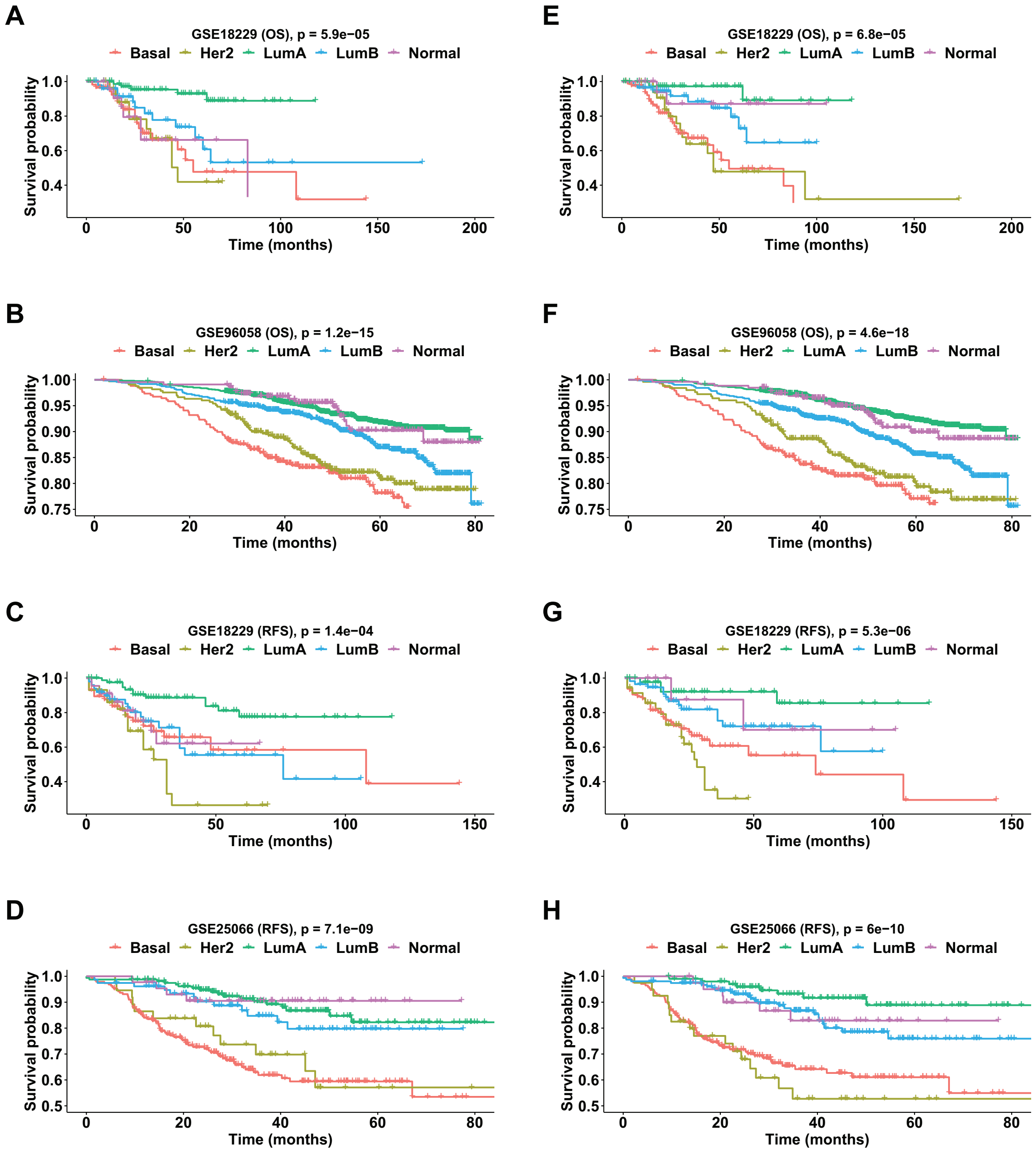


Figure 6. Survival analyses. For different datasets survival probabilities are plotted as functions of time for subtypes obtained using PAM50 (A–D) and using MPAM50 (E–H).

found $p>0.09$ in all cases but one, indicating no significant difference between the survival probabilities (Supplemental Table S10). The exception, with barely significant $p = 0.049$, occurred in the case of the Normal subtype and for the OS data from GSE18229.

As a second way of verifying the prognostic capacity of MPAM50 subtypes, we compared the survival probabilities of the 5 MPAM50-assigned subtypes for each dataset and found the corresponding p-value. These comparisons were also made for PAM50-assigned subtypes. For datasets with more than 200 samples (Table 2), the survival curves (and the corresponding p-values) are shown in Figure 6A to D (PAM50), and E-H (MPAM50). The remaining survival curves are depicted in Supplemental Figures S2 to S6. Figure 6A to H demonstrate significant differences ($p < 0.001$) between survival probabilities of the subtypes regardless of the subtyping method used (PAM50 or MPAM50). These figures show either comparable p-values for both methods or indicate a smaller p for MPAM50. Given the previously mentioned subtypes' prognoses, one expects to see the survival curve of LumA above those of LumB and Normal, which in turn are expected to be above the survival curves of Basal and Her2. These expected trends are seen in the survival curves obtained by MPAM50 shown Figure 6E to H. These trends are also mostly observed in the

PAM50 survival curves (Figure 6A–D), but there are some exceptions. For example, in Figure 6C the Basal curve is not clearly below that of LumB. In other words, in this case MPAM50 appears to perform better than PAM50, but we should emphasize that we did not find any significant difference between the LumB/Basal curves in Figure 6C and the corresponding LumB/Basal curves in Figure 6G (see Supplemental Table S10).

For datasets with fewer than 200 samples, Supplemental Figures S2 to S6 indicate comparable p-values (that have the same order of magnitude) for both methods, except for GSE53031 with $p(PAM50) = 0.047$ and $p(MPAM50) = 0.167$. However, in this case $p(PAM50) = 0.047$ is barely statistically significant. Interestingly, in the case of GSE26304, Supplemental Figure S4 indicates $p > 0.3$ for both methods. This is presumably because of the small number of samples in the dataset (109; Table 2). The fact that PAM50-assigned subtypes of GSE26304 also do not have significantly different survival probabilities, suggests that the poor agreement between PAM50 and MPAM50 ($ACC = 0.560$; Figure 2) does not necessarily mean worse performance by MPAM50 in terms of prognosis prediction. (In this case both PAM50 and MPAM50 perform equally badly due to the low number of samples.) Based on the results discussed in this subsection, we conclude that MPAM50 performs comparably with PAM50 in terms of prognosis prediction.

Of note, the survival curves plotted in Supplemental Figures S2 to S6 do not clearly show the previously mentioned expected survival probability patterns. For example, in Supplemental Figure S2 (GSE22226, OS data) the survival curve of Her2 (obtained by either method) is not clearly lower than that of LumA or LumB. What is important here is the comparison between the curves produced by the 2 methods. We have already mentioned that for each subtype and each dataset there was no statistically significant difference between the survival probabilities (except for the Normal subtype in the case of GSE18229 (OS) where the difference was barely significant $p = 0.049$; see Supplemental Table S3). In other words, in some cases both MPAM50 and PAM50 fail to produce the expected results presumably due to the small number of samples included in the dataset.

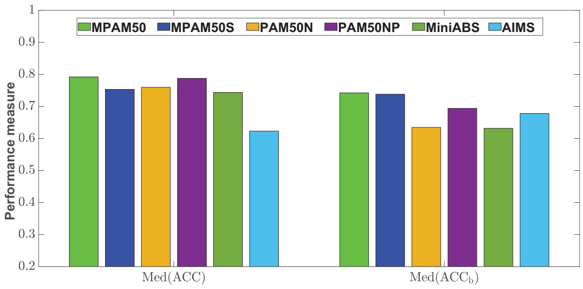


Figure 7. Performance comparison. For MPAM50, MPAM50S, PAM50N, PAM50NP, MiniABS, and AIMS the $Med(ACC)$ and $Med(ACC_b)$ are compared. PAM50N, and PAM50NP employ the same centroids as PAM50, but use unnormalized data with no reference subtraction. PAM50NP scores the samples by calculating Pearson correlation instead of rank correlation used by PAM50 and PAM50N. The only difference between MPAM50S and MPAM50 is that the former uses the Spearman rank correlation.

Performance comparison

In this section the performance of MPAM50 is compared with those of AIMS[6] and MiniABS,[7] which were specifically developed for robust classification of the intrinsic subtypes using unnormalized data. Additionally, performance comparisons are made between MPAM50 and 3 alternative modified PAM50 approaches, including the one suggested by Patil et al[3] that is available as a part of the genefu package. In this approach the samples are classified the same way as they are in PAM50, but no reference subtraction is performed. We henceforth call this method PAM50N. The second alternative modified PAM50 method, referred to as PAM50NP, works just like PAM50N but uses Pearson correlation for scoring the samples (same centroids as PAM50, no reference subtraction, and Pearson instead of rank correlation). The third modified version, called MPAM50S, works similarly to MPAM50 but scores the samples using the Spearman rank correlation. Note that, to the best of our knowledge, PAM50NP and MPAM50S have not been used by other investigators. We included these 2 alternative methods in our comparisons only to see how each of the 2 main modifications to PAM50 (using a different set of centroids and using Pearson instead of Spearman correlations) affects the results.

For subtyping using AIMS and MiniABS we employed the R packages provided by the authors of these 2 studies. In the case of AIMS, as suggested by the authors, the unnormalized data were not log-transformed. In all other cases log-transformed unnormalized data were used. The PAM50 centroids (needed for subtyping with PAM50N and PAM50NP) were obtained from the genefu package. Each of the 19 datasets were subtyped using each of the methods mentioned in the previous paragraph, and the ACCs were computed by comparing the resulting subtypes with the reported ones. The ACCs, subtype-specific ACCs, and balanced ACCs for each of the datasets and their mean and median values are given in Supplemental Tables S1 to S7. The performance measures $Med(ACC)$ and $Med(ACC_b)$ for all of the methods are plotted in Figure 7.

Figure 7 demonstrates that MPAM50 outperforms the other methods in terms of ACC and/or ACC_b. Especially, the differences in $Med(ACC_b)$ are larger, indicating a more balanced performance for MPAM50. Interestingly, the figure suggests that the 2 alternative methods proposed here (MPAM50S and PAM50NP), which respectively share the centroids and the scoring method with MPAM50, have the closest performance measures to those of MPAM50. On the other hand, the figure indicates that MPAM50 improves $Med(ACC)$ ($Med(ACC_b)$) by 27%, 7%, and 4% (10%, 18%,

Table 6. The *P*-values assessing the statistical significance of the differences shown in Figure 7.

	MPAM50S	PAM50N	PAM50NP	MINIABS	AIMS
$Med(ACC)$	0.0003	0.0702	0.2273	0.0196	0.0001
$Med(ACC_b)$	0.0126	0.0070	0.0218	0.0001	0.0005

and 17%) in comparison with AIMS, MiniABS, and PAM50N respectively. The statistical significance of the differences shown in Figure 7 were assessed by Wilcoxon signed-rank tests. The results, given in Table 6, suggest that all differences in $Med(ACC_b)$ shown in the figure are significant (at the 0.05 level) although some of improvements in $Med(ACC)$ are not. Based on these observations we conclude that MPAM50 is superior to the other methods, because it significantly increases $Med(ACC)$ and/or $Med(ACC_b)$.

Of note, MiniABS has the benefit of using only 11 genes to classify the samples. However, we believe that the primary goal of a classification method should be achieving a higher accuracy and that the number of genes used for classification should be regarded as a secondary factor in choosing the best approach. In other words, if 2 methods have comparable accuracies, then the method using the lower number of genes can be regarded as superior. But when one classification approach clearly outperforms the other (as is the case here) the number of genes should not be a factor in deciding which approach is better.

Discussion

PAM50 is the most widely-used signature/method for intrinsic subtyping of breast cancer. However, several publications[3-5] have reported robustness issues with this method, showing significant dependence of the results on the number and composition of the samples included in the dataset to be classified. This is due to the fact that PAM50 uses all samples in a dataset to calculate the reference to be subtracted from each sample before the samples are scored. A robust single-sample intrinsic subtype classifier is thus desirable. In this paper we introduce MPAM50, a modified version of PAM50 that classifies each sample independently, thereby avoiding the robustness issue.

MPAM50 is able to classify each sample independently because it does not subtract a reference from the samples and uses unnormalized gene expression. Like PAM50, MPAM50 uses the nearest centroid approach for classification, but the centroids are obtained by simply averaging the (weighted) unnormalized samples in the training data and the similarity to each centroid is measured by Pearson correlation rather than Spearman rank correlation. Both of these modifications (using different centroids and different similarity measure) contribute to the success of MPAM50 as a singles-sample classifier. This can be deduced from the fact that MPAM50 outperforms the 3 alternative modified PAM50 methods (PAM50N, MPAM50S, and PAM50NP) that lack one or both of these modifications. The fact that a simple nearest centroid method, which combines averaged centroids and Pearson correlation, can perform well across many datasets/platforms has been already demonstrated in other contexts[10,11] (although in these cases the averaging is done over unweighted normalized expression values). In this paper we confirm that such a simple method also performs well in the context of robust intrinsic subtyping of breast cancer. Specifically, we show that MPAM50 outperforms AIMS[6] and MiniABS,[7] 2 previously published platform-independent robust classifiers. (We did not compare MPAM50 with the robust subtyping method of Cascianelli et al[4] as it is platform-specific and is meant for subtyping only RNA-Seq samples.)

We have also shown that MPAM50 and PAM50 predict the clinical subtypes with similar accuracy: that is the intrinsic subtypes assigned by the 2 methods agree with the clinical subtypes at a comparable level. As also reported by other studies (see Kim et al[21] and references therein), the agreement is suboptimal, especially in the case of LumB-like subtype. However, the point here is that PAM50 and MPAM50 perform comparably in this regard. Note that in some studies (including Kim et al[21]) LumA-like and LumB-like clinical subtypes have been defined differently (in terms of the status of the relevant biomarkers), and so for these 2 subtypes quantitative comparisons between such studies and ours should be avoided. For each of the remaining clinical subtypes (HER2+ and triple negative), the prediction accuracy of MPAM50 reported here is comparable with that of PAM50 reported by Kim et al[21] (MPAM50: $Med\left(ACC^{cl}_{HER2+}\right)=0.73$, $Med\left(ACC^{cl}_{TN}\right)=0.82$ and PAM50: $ACC^{cl}_{HER2+}=0.74$, $ACC^{cl}_{TN}=0.81$), again confirming that MPAM50 and PAM50 perform comparably. It is also worth mentioning that improving the agreement between the intrinsic and clinical subtypes is not the goal of this study, and thus we have not compared MPAM50 to methods like PCA-PAM50[22] that have been specifically developed for this purpose.

Another important feature of MPAM50 is preserving the prognostic value of the intrinsic subtypes. We have demonstrated this by performing survival analyses and showing that MPAM50 and PAM50 predict survival probability curves that do not significantly differ. Of note, we have not compared MPAM50 with methods that aim to improve prognostication (see, eg, Pu et al[23]), because our goal is to propose a method that, while performing comparably to PAM50 in terms of prognostication, results in robust classification.

Our results indicate that, by independently subtyping individual samples, MPAM50 achieves our stated goal of robust subtyping. However, in some cases lack of robustness may be due to reasons other than preprocessing (subtracting a reference profile from) the samples, and thus may not be eliminated by single-sample subtyping. Specifically, a patient may be assigned 2 different subtypes even by a single-sample classifier depending on the sample preparation approach[24] or expression profiling method[25] used to obtain the expression profiles (because these factors affect the unnormalized expression values). Hence, one may wonder how MPAM50 compares to PAM50 in terms of the concordance between 2 sets of subtypes assigned using 2 sets of samples obtained from the same group of patients. Unfortunately, due to insufficient data, we were unable to directly test the performance of MPAM50 in this regard. For the only dataset (GSE54275) in our testing set that includes samples from the same patients obtained using 2 platforms, we found comparable concordances for the 2 methods (0.81 and 0.85 for MPAM50 and PAM50 respectively). This result and the overall good agreement between MPAM50- and PAM50-assigned subtypes suggest comparable performance by both methods, although more datasets are needed for drawing a definitive conclusion.

A limitation of MPAM50 is that it does not propose a gene selection algorithm (it uses the PAM50 genes). This may limit its applicability to other subtyping problems. However, the goal of this paper is not to propose a general-purpose classification algorithm. The aim is rather to propose a simple and easy-to-use method specifically for robust and accurate intrinsic subtyping of breast cancer. Our results suggest that MPAM50 is successful in achieving this goal, and thus can be helpful to breast cancer researchers.

Acknowledgements

This study utilized the computational resources of the Biowulf system at the National Institutes of Health, Bethesda, MD (http://biowulf.nih.gov). The results published here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Author contributions

MH and YKY designed research; MH performed research; MH and YKY analyzed the data and wrote the paper.

Supplemental material

Supplemental material for this article is available online.

REFERENCES

1. Dai X, Li T, Bai Z, et al. Breast cancer intrinsic subtype classification, clinical use and future trends. *Am J Cancer Res*. 2015;5:2929-2943.
2. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci*. 2002;99: 6567-6572.
3. Patil P, Bachant-Winner PO, Haibe-Kains B, Leek JT. Test set bias affects reproducibility of gene signatures. *Bioinformatics*. 2015;31:2318-2323.
4. Cascianelli S, Molineris I, Isella C, et al. Machine learning for RNA sequencing-based intrinsic subtyping of breast cancer. *Sci Rep*. 2020;10:1-13.
5. Weigelt B, Mackay A, A'hern R, et al. Breast cancer molecular profiling with single sample predictors: a retrospective analysis. *Lancet Oncol*. 2010;11:339-349.
6. Paquet ER, Hallett MT. Absolute assignment of breast cancer intrinsic molecular subtype. *J Natl Cancer Inst*. 2015;107:357.
7. Seo MK, Paik S, Kim S. An improved, assay platform agnostic, absolute single sample breast cancer subtype classifier. *Cancers*. 2020;12:3506.
8. Lusa L, McShane LM, Reid JF, et al. Challenges in projecting clustering results across gene expression-profiling datasets. *J Natl Cancer Inst*. 2007;99:1715-1723.
9. Kensler KH, Sankar VN, Wang J, et al. Pam50 molecular intrinsic subtypes in the nurses' health study cohorts. *Cancer Epidemiol Biomarkers Prev*. 2019; 28:798-806.
10. Hamaneh M, Yu YK. An 8-gene signature for classifying major subtypes of non-small-cell lung cancer. *Cancer Inform*. 2022;21:11769351221100718.
11. Girard L, Rodriguez-Canales J, Behrens C, et al. An expression signature as an aid to the histologic classification of non–small cell lung cancer. *Clin Cancer Res*. 2016;22:4880-4889.
12. Li X, Shi G, Chu Q, et al. A qualitative transcriptional signature for the histological reclassification of lung squamous cell carcinomas and adenocarcinomas. *BMC Genomics*. 2019;20:881.
13. Ciriello G, Gatza ML, Beck AH, et al. Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*. 2015;163:506-519.
14. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002; 30:207-210.
15. Prat A, Fan C, Fernández A, et al. Response and survival of breast cancer intrinsic subtypes following multi-agent neoadjuvant chemotherapy. *BMC Med*. 2015;13:1-11.
16. Carvalho BS, Irizarry RA. A framework for oligonucleotide microarray preprocessing. *Bioinformatics*. 2010;26:2363-2367.
17. Gautier L, Cope L, Bolstad BM, Irizarry RA. Affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*. 2004;20:307-315.
18. Ritchie ME, Phipson B, Wu D, et al. Limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43:e47-e47.
19. Gendoo DMA, Ratanasirigulchai N, Schröder MS, et al. Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. *Bioinformatics*. 2016;32:1097-1099.
20. Goldhirsch A, Winer EP, Coates AS, et al. Personalizing the treatment of women with early breast cancer: highlights of the St Gallen international expert consensus on the primary therapy of early breast cancer 2013. *Ann Oncol*. 2013;24:2206-2223.
21. Kim HK, Park KH, Kim Y, et al. Discordance of the PAM50 intrinsic subtypes compared with immunohistochemistry-based surrogate in breast cancer patients: potential implication of genomic alterations of discordance. *Cancer Res Treat*. 2019;51:737-747.
22. Raj-Kumar PK, Liu J, Hooke JA, et al. PCA-PAM50 improves consistency between breast cancer intrinsic and clinical subtyping reclassifying a subset of luminal A tumors as luminal B. *Sci Rep*. 2019;9:1-3.
23. Pu M, Messer K, Davies SR, et al. Research-based PAM50 signature and long-term breast cancer survival. *Breast Cancer Res Treat*. 2020;179:197-206.
24. Lien TG, Ohnstad HO, Lingjærde OC, et al. Sample preparation approach influences pam50 risk of recurrence score in early breast cancer. *Cancers*. 2021;13:6118.
25. Larsen MJ, Thomassen M, Tan Q, Sørensen KP, Kruse TA. Microarray-based RNA profiling of breast cancer: batch effect removal improves cross-platform consistency. *Biomed Res Int*. 2014;2014:651751.