

Appendix 1: Data Normalization and Clustering Techniques

A1-1: Data normalization. The Wang et al data (1) (Affymetrix U133A, MASS5.0) was log₂ transformed and the median of the transformed data was computed for each array across all genes. Next the median of medians m_m was computed across all arrays. Each entry in each array was normalized by subtracting m_m . Before each clustering experiment, the expression values of each gene was standardized (mean 0, variance 1) across all samples being considered in the clustering.

A1-2: Consensus Ensemble Clustering. Unsupervised clustering algorithms divide data into meaningful groups or clusters such that the intra-cluster similarity is maximized and the inter-cluster similarity is minimized (2). Clustering is an NP-hard problem. However, many heuristic methods exist and they can be categorized into hierarchical, partitioning and grid-based methods. We apply all these methods in an unsupervised way to the data, i.e. without assuming a predefined label on the objects to be classified. Unsupervised clustering is known to produce unstable solutions which are sensitive to various data parameters and/or perturbations and to the clustering techniques used. A relatively recent solution which corrects for this instability is consensus ensemble clustering (3, 4): Given several methods of clustering data, consensus ensemble clustering finds a combination of these methods which improves the quality of the individual methods.

The consensus ensemble approach can be divided up into two parts: a method which generates a collection of clustering solutions, and a consensus function that combines them to produce a single output clustering of the data. There is an implicit assumption in this that combining the results of several clustering techniques will give groupings that are more reliable and less biased to a particular technique. This has been demonstrated in supervised classification schemes where it was shown that multiple solutions may reduce the variance of the error and at the same time, increase the robustness of the result

(5, 6). The ensemble clustering technique was introduced in (3, 4) and its effects were described in several subsequent studies e.g., (7).

In our study the challenge posed to the ensemble consensus clustering approach was to identify meaningful clusters which were stable and robust both to perturbations of the data and the choice of clustering methods used. This goal was approached in multiple ways. (a) If the method was stochastic, we reduced the effect of the stochastic variation by applying the method repeatedly and taking an appropriate average. (b) To reduce the sensitivity to random variation in the data, we applied each clustering method to multiple sample datasets obtained by bootstrapping in both the features (gene probes) used in the clustering as well as in the samples.

A1-2a: Clustering high dimensional data

To correct for the fact that many gene probes have low variability across the samples and merely add noise to the data, we cluster on subspaces of attributes (genes) rather than on the entire space. The subset of gene probes on which data is clustered may have an important influence on the clustering solution. Since data is high dimensional we restricted the clustering procedures only to those gene probes which were determined as “reliable” through an initial principal component analysis (PCA), ie, they occur with high weights (top 25% in absolute value) in the eigenvectors corresponding to those eigenvalues which represent a significant fraction (85% was the value chosen in the present paper) of variation of the data across the samples. The largest k eigenvalues $\{\lambda_1, \dots, \lambda_k\}$ out of N that represent a variation $V\%$ can be

obtained by solving the equation $NV / 100 = \sum_{i=1}^k \lambda_i$.

The details of our clustering method are as follows:

Step 1. For each $k=2, \dots, 50$, we created k clusters on re-sampled and random projected datasets based on individual clustering methods. We generated 150 datasets as follows: 50 datasets were created by bootstrapping the samples, 50 datasets by projecting the data onto subsets of reliable gene probes bootstrapped from the data and 50 additional datasets by first projecting the data on a bootstrapped subset of gene probes and then bootstrapping samples on the resulting dataset. We then applied representative methods for each major class of known clustering techniques. We discuss these briefly below.

Partitioning Relocation Methods. These methods divide data into several subsets and use certain greedy heuristics in the form of iterative optimization to reassign points between the k clusters. The optimization is applied to an objective function defined on unique cluster representatives (e.g., centroid, medoid), which is usually a dissimilarity measure.

We applied the following algorithms:

- (i) *Partition around medoids* (PAM (8)): PAM is an iterative optimization that relocates the points between perspective clusters by re-nominating the points as potential medoids.
- (ii) *Clustering Large Applications* (CLARA (8)): This method uses several samples of the data and subjects each of them to PAM. The dataset is then reassigned to the resulting medoids and the best system of medoids is retained.
- (iii) *K-means* (9): To each cluster, this method associates the mean (centroid) of its points and uses as the objective function the sum of distances between a point and its centroid.
- (iv) *Graph partitioning* (10). In this method the points (samples) are associated with vertices in a graph and each point is connected to the closest neighbor. The resulting graph is then split into k -clusters by applying a min-cut approach.

Clusters produced by centroid methods (*k*-means, PAM, CLARA) work by identifying samples into clusters if they form a spheroid shape. Thus, they are suitable for clustering datasets with uniform and relatively low variation among samples. Graph partitioning methods produce clusters in which samples are added in if they are “close” to at least one sample in the candidate cluster. Thus graph partitioning approaches can successfully identify clusters with unequal variance along the feature coordinates (ie, they can find a “long” shape).

Agglomerative methods. These methods build the clusters gradually by trying to establish a hierarchical order (8). One starts by assigning each sample to its own cluster and then recursively merging two or more most similar clusters until a stopping criterion is fulfilled. The similarity between clusters is usually computed based on a linkage metric which reflects the connectivity and similarity between the clusters. In our study we applied hierarchical clustering techniques based on the following metrics:

Average linkage metric – computes the distance between two clusters as the average of the distances between the pairs of points in these clusters.

Complete linkage metric – computes the distance between two clusters as the maximum distance between the pairs of points in the two clusters.

Single linkage metric - computes the distance between two clusters as the minimum distance between the pairs of points in the two clusters.

Mcquitty metric computes the distance between two clusters as the average distance between the subclusters of the two clusters.

Centroid metric - computes the distance between two clusters as the distance between the centroids of the two clusters

Ward metric: computes the distance between two clusters as the distance between the centroids of the two clusters averaged to the reciprocal mean of the sizes of the two clusters

(vii) In addition we applied a *hybrid biased agglomerative method (bagglo (11))* which combines partitioning clustering with the agglomerative hierarchical approach. For n samples, we start with an initial partition into $n^{1/2}$ clusters and augment the original feature space by adding $n^{1/2}$ dimensions corresponding to the initial clusters. The agglomerative clustering approach is then applied to this augmented dataset.

Methods based on probability: In these methods, data is considered to be a sample independently drawn from a mixture model of several probability distributions and the clusters are associated with the area around the mean of each distribution. It is assumed that each point is assigned to a unique cluster. The probabilistic clustering method optimizes the log-likelihood of the data to be drawn from a given mixture model.

In our approach we applied the *expectation maximization (EM)* method (12-14). EM is a two step procedure which starts with estimating for each point the probability of belonging to a certain cluster. In the second step EM finds an approximation to the mixture model by maximizing the log-likelihood in an iterative way until the convergence to an optimal solution is reached.

Entropy-based-clustering (ENCLUST):

This method (15) starts by dividing the interval associated to each attribute into 1-dimensional bins (cells) and retaining only the cells with a high density. The iterative step consists in creating cells of higher dimensions by joining the cells with low dimension and retaining only those cells which have the entropy below a certain threshold as optimal for clustering.

Self Organizing Maps (SOM):

This method (16) is both a data visualization and a clustering technique which reduces the dimensions of data through the use of self-organizing neural networks. The way SOMs reduces dimensions is by producing a map of usually 1 or 2 dimensions which plot the similarities of the data by grouping similar data points together.

Step 2. Each method was applied 50 times with different parameter initialization on the full dataset, and once on each of the 150 datasets obtained as described in Step 1. Based on the 200 clustering results, we constructed an agreement matrix for each method whose entries m_{ij} represent the fraction of times the pair of samples (i,j) occurred in the same cluster out of the total number of times the pair of samples was selected in the 200 datasets.

Using $d_{ij}=1-m_{ij}$ as the distance between the samples (i,j) , we apply simulated annealing to find k “consensus” clusters which achieve the maximum value for the average internal similarity and the average external dissimilarity. At the end of this step, each method will give us its best clustering into k clusters.

Step 3. For each k , we combine the results from Step 2 by using an agreement matrix to create a consensus of all the individual clustering techniques and once again use simulated annealing to identify the optimal clusters.

Further details of our ensemble clustering method for identifying haplogroups in mRNA gene expression data will be described in a separate paper. There we also discuss why, in comparison with the traditional approach of a single clustering technique, the consensus ensemble clustering approach in combination with principal component analysis has better average performance across datasets, and a lower sensitivity to noise, outliers and sampling variation.

Appendix 2: Protocol to identify HER2+ samples using HER2 amplicon gene levels.

The HER2+ samples were identified using mRNA levels of Her2/neu, GRB7, STARD3 and PPARB which are on the Her2/neu chr17 q12-q21 amplicon (17). The results of this analysis for the present study are shown in **Figure A2-1**. For each of the four genes, the sorted values of each gene across all samples has a “high value” and a “low value” regime. The “high” value regime cutoff value was determined by linear least squares fit to the plots in **Figure A2-1**. The samples identified as HER2+ were those which simultaneously had “high” expression of Her2/neu and at least two other from the set (GRB7, STARD3, PPARB).

This classification model was validated on a publicly available breast cancer U133a Affymetrix data which had 133 samples (18), <http://bioinformatics.mdanderson.org/supplements.html>, for which the Her2/neu scores obtained through IHC and FISH were provided (results not shown).

Caption for Figure A2-1.

Figure A2-1. Identification of HER2+ breast cancer samples based on gene amplification in the Her2/neu amplicon chr17 q12-21. We use consensus up-regulation of the mRNA levels of the genes Her2/neu, GRB7, STARD3 and PPARB as the signature to identify HER2+ samples. As shown in the figure, we see that for each amplicon gene, there is a change in slope from a “low” expression value to a “high” expression value. The cutoff value of up-regulation for each gene is identified by linear least squares fit to these plots and is marked as a red dot in **Figures A2-1 a-e**. At least 3 of the amplicon genes (including Her2/neu) are up-regulated in the 42 samples identified as HER2+ and highlighted red in **Figure A2-1 e**. In 10 other samples (highlighted green in **Figure A2-1 e**) Her2/neu is up-regulated together with at most one other amplicon gene. In the remaining 234 samples (highlighted blue) Her2/neu is not up-regulated.

Appendix 3: Identification of genes which can distinguish the subtype clusters.

Microarray datasets suffer from an overabundance of genes, most of which do not contribute to the signal. Identifying differentially expressed genes for a given set of phenotypes is a difficult problem for which many methods have been proposed. These can be divided into two major groups ((19, 20):

(i) Filtering or Variable Ranking methods: These select features based on quality scores. They include the fold change test (20,21), the t-test (21, 22), the Wilcoxon-Mann-Whitney test (23), the Signal-to-Noise Ratio (SNR) test (24), the J5 test, the D1 test (26) etc. Another set of methods measure the "separability" of data into different phenotype classes. These include simple or weighted separability (25), envelope eccentricity (26), etc. A third class uses information-theoretic methods such as the entropy criterion (28), mutual information (29), etc. Finally, there are the statistical impurity measures (27) which include the two-ing rule, the Gini index, max-minority, sum-minority, sum-of-variances etc.

(ii) Feature Subset Selection Methods: One such method selects those features which are useful for classification for a given machine learning algorithm (SVM (28), ANN (29), kNN (30) etc). More sophisticated approaches are embedded methods which include the selection of features as part of the training process for the classifier. These methods are computational intensive and require efficient search strategies or a preliminary filtering of the non-reliable genes to reduce the dimensionality of the problem.

The existence of such a variety of feature selection methods poses a challenge in microarray data analysis. There have been recent attempts to combine various approaches into a meta selection procedure based on "majority-voting" using ranking by predictive content across many data perturbations and machine learning methods (34,35). Several studies (18,27) have shown that variables

which are only weakly correlated with phenotype are very useful when used in combinations. This principle has led to the development and study of combinatorial markers or patterns (31, 32). In the present study, we used a single feature selection method (namely the SNR test, (24)) which has been shown (33) to have good performance on genomic and proteomic data.

We identified a large pool of uni-gene markers for each core that distinguish it from the others using the signal-to-noise statistic. For gene i , if $\mu_1(i)$ and $\mu_2(i)$ be the average gene expression levels for the core and its complement and $\sigma_1(i)$ and $\sigma_2(i)$ the corresponding standard deviations, the signal-to-noise ratio (SNR) is defined as $SNR = (\mu_0 - \mu_1) / (\sigma_0 + \sigma_1)$. The t-test statistic is the same as the SNR except that the denominator is $(\sigma_0^2 + \sigma_1^2)^{1/2}$. Since $(\sigma_0 + \sigma_1) > (\sigma_0^2 + \sigma_1^2)^{1/2}$. SNR is a more conservative criterion than the t-test.

The SNR statistic is preferred over the t-test in situations when the sample size in a class is small (less than 30) because it does not assume a Gaussian distribution for the underlying variables; an assumption which is implicit in the t-test. When combined with a permutation test for measuring p-values, the SNR statistic is a powerful and widely used technique (25,38,39,40) for feature selection and class discrimination and is implemented in several software packages (eg, GenePattern and Gene Set Enrichment Analysis (GSEA), available at <http://www.broad.mit.edu/tools/software.html>).

The signal-to-noise (SNR) was computed for each gene and for each of the bootstrap sample perturbation experiments for the core samples. The selected genes were those whose p -value for the SNR was below 0.01 and the significance of the SNR for false discovery rate (FDR, (34)) was above 0.95 in each experiment.

Appendix 4: How to stratify unknown samples into our subtypes

Supplementary Table 1 shows a simple (but not necessarily complete or optimal) classifier based on 120 genes which can stratify a sample into our subtypes. This classifier uses 10 top markers for each hierarchical category and 5 levels of hierarchy. Panels 1 through 5 perform the classification along the following hierarchy: $BCA \rightarrow (\text{Luminal}, \text{Basal}, \text{HER2}^+)$; $\text{Luminal} \rightarrow (\text{Luminal A}, \text{Luminal B})$; $\text{Luminal B} \rightarrow (\text{LB}_1, \text{LB}_2, \text{LB}_3)$; $\text{Basal} \rightarrow (\text{BA}_1, \text{BA}_2)$; $\text{HER2}^+ \rightarrow (\text{HER2}^{+I}, \text{HER2}^{+NI})$. To classify a sample, the analysis would begin by using panel 1 and proceed to the other panels depending on the result. Panels 1, 2 and 3 classify a sample into the Luminal subtypes, panels 1 and 4 into the Basal subtypes and panels 1 and 5 into the HER2+ subtypes.

In panel 1 the two classifiers separate the samples into Luminal, Basal, HER2+ using up-regulation signatures for the genes ESR1, GATA3, TFF3, SLC39A6 (Liv-1), SCUBE2, XBP1, FOXA1, CA12, VAV3 and KRT18 for Luminals; FOXC1, DSC2, TTK, KRT6B/17/5, CDH3, CRYAB, CCNE1 and CX3CL1 for Basals and Her2/neu, STARD3, GRB7, PPARBP, CLCA2, GCHFR, CAP1, KMO, S100A9 and TCAP for HER2+. In other panels, the stratification uses further genes discovered by our analysis to further stratify into the 8 subtypes.

We emphasize that **Supplementary Table 1** is presented to show that it is possible to define a simple classifier with a reasonable accuracy which uses genes that make biological sense. It is by no means the most accurate or the most optimum way to classify an unknown sample. To create an optimum classification for a set of unknown samples, the procedure would be to first correct for possible variance bias between the new dataset and the dataset used here (given in **Supplementary Table 2**) and then reanalyze the combined data using the techniques we describe in the paper. The classification of the unknown samples would then follow from their strong association with the samples in the eight subtype clusters identified in the present dataset.

Given a new breast cancer dataset whose samples are to be classified into our subtypes using **Supplementary Table 1**, the relative variance of each gene must first be adjusted so that it matches the variance in the dataset used in the present paper. This can be done using the expression values for all the samples in our study for the genes listed in Table 1. The data necessary to do this is given in **Supplementary Table 1**, which contains the log₂ expression values for all the genes. The procedure for stratifying an unknown sample is described below:

Combine the log₂ transformed new data with **Supplementary Table 1** and normalize the combined data as described in **Appendix 1**. Next, possible relative systematic biases between the data in our study and the new data must be corrected. A simple procedure that accomplishes this is to apply the Distance Weighted Discrimination (DWD) tool (35). DWD shifts the “cloud” of points in one of the datasets in an optimal way such that it best overlaps the “cloud” of points in the other dataset.

Once the systematic biases are corrected and the new data is adjusted to have the same variance as the data used in the present paper, we must renormalize the entries for each gene for all the samples in **Supplementary Table 1** as they may have shifted under the bias correction. This normalization is done by shifting each gene entry in each sample array by the median value of the gene expression for the sample over all genes in **Table 1**. Then, for each gene i , each sample gene entry x_i^s for gene i is standardized by the transformation $x_i^s \rightarrow (x_i^s - \text{mean}_i) / \text{stdev}_i$ where mean_i and stdev_i for gene i are computed across all the samples in the adjusted combined data.

Now, for each level of classifier in Table 1, for each gene used at that level, we determine the average (centroid) value of the gene over all the samples in the corresponding subtype cluster in the data used in the present paper. Thus each level in **Table 1** will have a list of genes and their average (centroid) values for each subtype.

The classification then proceeds using the Spearman rank order correlation coefficient (36). The Spearman rank correlation of two vectors x and y is given by,

$$\rho_S(x, y) = \frac{\sum_i (rx_i - r\bar{x})(ry_i - r\bar{y})}{\sqrt{\sum_i (rx_i - r\bar{x})^2 \sum_i (ry_i - r\bar{y})^2}}$$

where rx_i is the rank of the i th value of x , ry_i is the rank of the i th value of y , and $r\bar{x}$ and $r\bar{y}$ represent the mean of the rank values for x and y .

At each level, for each sample to be classified, we compute the Spearman correlation coefficient with the centroid vector for each subtype in the panel, say $\rho_{S1}, \dots, \rho_{Sp}$, and assign the sample to the phenotype k whose score $\frac{\rho_{Sk}}{\rho_{S1} + \dots + \rho_{Sp}}$ is the largest. If the value is < 0.75 , we consider the

classification to be ambiguous and label the sample as unclassified at the level being considered.

The final identification into one of the eight subtypes is done by computing on all panels and assigning the final subtype if all branches from the top of the hierarchy (BCA) to the given subtype are in agreement. The classification accuracies of this protocol is given in the table as a pair of numbers representing the number of samples correctly classified versus the total number of samples in the subtype.

In supplementary material **Figure A4-1** we show a heatmap for 10 samples chosen randomly from each subtype over the 120 genes in **Supplementary Table 1** to visually illustrate its accuracy. The first three

blocks of rows represent stratification into Luminal, Basal, HER2+ subtypes. The next five blocks of rows show how to further sub-stratify the samples identified as Luminal by the first three blocks. The next two sets of two row blocks each show the sub-stratification of the samples placed in the Basal and HER2+ subtypes respectively.

Supplementary Table 2 shows the projection of the data for all the Wang et al samples including the “potentially HER2+”, “potentially Luminal” and “potentially Basal” onto the 120 genes used in **Supplementary Table 1** for the classification.

Appendix 5. Protocol used to stratify HER2+ samples in the microarray dataset of Harris et al using the markers identified in the Wang et al dataset.

The mRNA gene expression dataset of Harris et al (36) on Affymetrix U133 2plus chips was obtained. Gene expression data and histological information was available for 21 breast cancer samples. Data was normalized in the same way as the Wang et al data (1). PCA applied to the Harris et al data identified 2624 reliable genes and the optimum number of clusters inferred from these was at most three.

Consensus hierarchical clustering of the Harris et al data using the 105 lymphocyte-associated genes in **Supplementary Table 3a** found three clusters of size 6, 7 and 8. The first cluster had a very strong lymphocytic signature, suggesting that the samples in this cluster are HER2+I. The second cluster had a very low lymphocytic signature, suggesting that the samples are Her2+NI. The samples in the third cluster had an intermediate lymphocytic response and they remained unclassified.

Caption for Supplementary Tables 1, 2, 3 and Figure A4-1:

Supplementary Table 1. A simple hierarchical classification system for assignment of a sample to a BCA subtype. The genes listed at panels 1-5 classify a sample as follows: panel 1: (Luminal vs Basal vs HER2+), panel 2: (Luminal A vs Luminal B), panel 3: (LB₁ vs LB₂ vs LB₃), panel 4: (BA₁ vs BA₂),

panel 5: (HER2_{+I} vs HER2_{+NI}). Given an unknown BCA sample array, assuming that there is no variance bias (see Appendix 4 for details), normalize it so that the median expression value across all genes is zero. Next, standardize each gene value using the mean and stdev values in the table by the rule $x \rightarrow (x - \text{mean}) / \text{stdev}$. Going down the panels as appropriate from 1 \rightarrow 5, compute the Spearman rank correlation coefficient between the sample values and the gene values (given in rows for each category) using all the genes in the panel under consideration. Thus, for Panel 1, three coefficients are calculated using the genes in rows for Luminal A, Basal and HER2₊; ie, over the 30 dimensional space of genes listed. The sample is assigned to the category with the largest Spearman rank correlation. If the classification in panel 1 is Luminal and a further stratification is desired, then panel 2 may be used to determine whether the assignment is LA or LB. If the latter, then panel 3 is used to decide whether it is LB₁ or LB₂ or LB₃. If panel 1 assigns the samples to Basal or HER2₊, then panels 4 or 5 respectively are used for further stratification. At any given panel of stratification, all previous panels are computed first. The final assignment is made to the leaf category which matches up to the root of the panel hierarchy. The classification accuracy of this classifier for the core subtype samples as well as the samples set aside (see text) is given as two pairs of numbers in each panel. The first number in each pair is the number of correctly assigned samples in the category and the second is the actual number of samples in the category. We emphasize that this classifier is presented to show that it is easy to get a reasonably accurate classifier using genes that are biologically meaningful. The more accurate way of classification of a set of unknown samples would be to combine them into the dataset after variance bias correction and redo the entire clustering exercise, using all gene expression values. The assignment would then follow from the subtype cluster to which the unknown samples were placed by the full procedure described in this paper.

Supplementary Table 2. Projection of the data of Wang et al. 2005 on the 120 genes used in the classifier of **Supplementary Table 1**. Column 1 shows the sample id as used in Wang et al dataset.

Columns 2 and 3 show respectively the subtype assignment based on our consensus clustering protocol and on the nearest centroid classifier described in **Supplementary Table 1**.

Supplementary Table 3a. The top 105 lymphocyte-associated genes (up-regulated in the HER2_{+I}) which can accurately separate HER2_{+I} from the HER2_{+NI}. These were used in the analysis of the Harris et al data. The first 9 genes, marked with *, are chemokines near the HER2+ amplicon at chr17q12. The remaining genes are sorted by their chromosomal location (from chr 1 to chr X), in decreasing order of their signal-to-noise-ratio (SNR).

Supplementary Table 3b. Complete list of lymphocytic-associated genes for the HER2_{+I} subtype based on the SNR score (p-val ≤ 0.005 and FDR≤0.05). A positive score represents genes up-regulated in HER2_{+I} and a negative score represents genes up-regulated in HER2_{+NI}. The first 9 genes, marked with *, are chemokines near the HER2+ amplicon at chr17q12. The remaining are sorted based by chromosomal location (from chr 1 to chr X) and signal-to-noise ratio (SNR).

Figure A4-1. Heatmap showing the accuracy of the classifier in **Supplementary Table 1** based on 120 genes. For simplicity, only 10 core samples in each subtype are shown. The classification proceeds along the hierarchy: (Luminal vs. Basal vs. HER2₊), (Luminal A vs. Luminal B), (LB₁ vs. LB₂ vs. LB₃), (BA₁ vs. BA₂) and (HER2_{+I} vs. HER2_{+NI}). The first three row-blocks of genes classify a sample into Luminal, Basal or HER2₊. The next two row-blocks distinguish Luminal A from Luminal B and so on. The last two rows show the HER2_{+I} vs. HER2_{+NI} separation. Up-regulated genes are shown in red and downregulated genes in green. The genes used in the assignment rules for each level are given in **Supplementary Table 1** and the stratification procedure is described in **Appendix 4**.

Appendix 6: Kaplan-Meier curves for BA₁ vs BA₂ and HER2₊/ER₊ vs HER2₊/ER₋ :

Figure A6-1 shows the Kaplan-Meier recurrence free survival curves for the two Basal subtypes BA₁ and BA₂. The difference in survival is not significant.

Figure A6-2 shows the Kaplan-Meier recurrence free survival curves for the HER2+ subtype if the samples are split by ER status. The difference in survival is not significant.

Caption for Figures A6-1 and A6-2.

Figure A6-1. Kaplan-Meier distant metastasis-free survival curves for the two Basal subtypes BA₁ and BA₂ identified by our methods. The log-rank p value for difference in survival is 0.6 so this difference is not significant. However, this does not preclude a biological basis for the two subtypes.

Figure A6-2. Kaplan-Meier distant metastasis-free survival curves if the HER2+ subtype is split by ER status. The log-rank p value for the difference in survival between HER2+/ER+ and HER2+/ER- is p = 0.34, which is not statistically significant.

References

1. Wang Y, Klijn JG, Zhang Y, *et al.* Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 2005;365(9460):671-9.
2. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 1987;20:53-65.

3. Monti S, Tamayo P, Mesirov J, Golub T. Consensus Clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning Journal* , 52(1-2):91-118, 2003 2003;52(1-2):91-118.
4. Strehl A, Ghosh J. Cluster ensembles: a knowledge reuse framework for combining partitionings. *Eighteenth national conference on Artificial intelligence*; 2002 July 28-August 01, 2002; Edmonton, Alberta, Canada
2002. p. 93-8.
5. Bhanot G, Alexe G, Venkataraghavan B, Levine AJ. A robust meta-classification strategy for cancer detection from MS data. *Proteomics* 2006;6(2):592-604.
6. Prodromidis AL, Stolfo SJ. A comparative evaluation of meta-learning strategies over large and distributed data sets. *Workshop on Meta-learning, Sixteenth Intl Conf Machine Learning*; August 1999 August 1999; Bled, Slovenia; August 1999. p. 18-27.
7. Topchy A, Jain AK, Punch W. Clustering ensembles: models of consensus and weak partitions. *IEEE transactions on pattern analysis and machine intelligence* 2005;27(12):1866-81.
8. Kaufmann L, Rousseeuw PJ. *Finding Groups in Data: An Introduction to Cluster Analysis*. 1 ed: John Wiley & Sons; 1990.
9. Hartigan JA. Clustering. *Annual review of biophysics and bioengineering* 1973;2:81-101.
10. Karypis G, Kumar V. Multilevel graph partitioning schemes. *24th International Conference on Parallel Processing* 1995; New York CRC Press; 1995. p. 113-22.

11. Karypis G, Kumar V. Multilevel graph partitioning schemes. 24th International Conference on Parallel Processing 1995; New York CRC Press; 1995. p. 113-22.
12. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 1977;39:1-38.
13. Fraley B, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* 2002;97:611-31.
14. Cheeseman P, Kelly J, Self M, Stutz J, Taylor W, Freeman D. AutoClass: A Bayesian Classification System. The Fifth International Conference on Machine Learning; 1988 June 12-14 1988; Ann Arbor, MI: Morgan Kaufmann Publishers, San Francisco; 1988. p. 54-64.
15. Cheng CH, Fu AW, Zhang Y. Entropy-based Subspace Clustering for Mining Numerical Data. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-99); 1999; San Diego; 1999.
16. Kohonen T. *Self-Organizing Maps*: Springer, Berlin, Heidelberg, New York; 2001.
17. Kauraniemi P, Kallioniemi A. Activation of multiple cancer-associated genes at the ERBB2 amplicon in breast cancer. *Endocr Relat Cancer* 2006;13(1):39-49.
18. Hess KR, Anderson K, Symmans WF, *et al.* Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *J Clin Oncol* 2006;24(26):4236-44.

19. Guyon I, Elisseeff A, 3:1157-1182. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research* 2003;3:1157-82.
20. Jeffery IB, Higgins DG, Culhane AC. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC bioinformatics [electronic resource]* 2006;7:359.
21. Gosset WS. The probable error of a mean. *Biometrika* 1908;6(1):1-25.
22. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* 2001;98(9):5116-21.
23. Lehmann E. *Nonparametrics: Statistical Methods Based on Ranks*: San Francisco: Holden-Day, Inc. ; 1975.
24. Golub TR, Slonim DK, Tamayo P, *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science (New York, NY)* 1999;286(5439):531-7.
25. Patel S, Lyons-Weiler J. caGEDA: a web application for the integrated analysis of global gene expression patterns in cancer. *Applied bioinformatics* 2004;3(1):49-62.
26. Alexe G, Alexe S, Vizvari B, Hammer PL. Pattern-Based Feature Selection in Genomics and Proteomics. *Annals of Operations Research* 2006:in press.

27. Su Y, Murali TM, Pavlovic V, Schaffer M, Kasif S. RankGene: identification of diagnostic genes based on expression data. *Bioinformatics (Oxford, England)* 2003;19(12):1578-9.
28. Vapnik V. *Statistical Learning Theory*: Wiley-Interscience; 1998.
29. Bishop C. *Neural Networks for Pattern Recognition*: Oxford, Oxford University Press; 1995.
30. Ripley B. *Pattern Recognition and Neural Networks*: Cambridge 1996.
31. Alexe G, Hammer PL. Spanned patterns for the Logical Analysis of Data Discrete Applied Mathematics 2006;154(7):1039 - 49
32. Crama Y, Hammer P, Ibaraki T. Cause-effect relationships and partially defined Boolean functions. *Annals of Operations Research* 1988;16:299-326.
33. Alexe G, Bhanot G, Venkataraghavan B, *et al.* A Robust Meta-classification Strategy for Cancer Diagnosis from Gene Expression Data. *Proceedings / IEEE Computational Systems Bioinformatics Conference, CSB 2005*:322-5.
34. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* 1995;57:289-300.
35. Benito M, Parker J, Du Q, *et al.* Adjustment of systematic microarray data biases. *Bioinformatics (Oxford, England)* 2004;20(1):105-14.

36. Spearman C. The proof and measurement of association between two rings. *American Journal of Psychology* 1904;15:72-101.