Challenges in Projecting Clustering Results Across Gene Expression–Profiling Datasets

Lara Lusa, Lisa M. McShane, James F. Reid, Loris De Cecco, Federico Ambrogi, Elia Biganzoli, Manuela Gariboldi, Marco A. Pierotti

Background Gene expression microarray studies for several types of cancer have been reported to identify previously unknown subtypes of tumors. For breast cancer, a molecular classification consisting of five subtypes based on gene expression microarray data has been proposed. These subtypes have been reported to exist across several breast cancer microarray studies, and they have demonstrated some association with clinical outcome. A classification rule based on the method of centroids has been proposed for identifying the subtypes in new collections of breast cancer samples; the method is based on the similarity of the new profiles to the mean expression profile of the previously identified subtypes.

Methods Previously identified centroids of five breast cancer subtypes were used to assign 99 breast cancer samples, including a subset of 65 estrogen receptor–positive (ER+) samples, to five breast cancer subtypes based on microarray data for the samples. The effect of mean centering the genes (i.e., transforming the expression of each gene so that its mean expression is equal to 0) on subtype assignment by method of centroids was assessed. Further studies of the effect of mean centering and of class prevalence in the test set on the accuracy of method of centroids classifications of ER status were carried out using training and test sets for which ER status had been independently determined by ligand-binding assay and for which the proportion of ER+ and ER– samples were systematically varied.

Results When all 99 samples were considered, mean centering before application of the method of centroids appeared to be helpful for correctly assigning samples to subtypes, as evidenced by the expression of genes that had previously been used as markers to identify the subtypes. However, when only the 65 ER+ samples were considered for classification, many samples appeared to be misclassified, as evidenced by an unexpected distribution of ER+ samples among the resultant subtypes. When genes were mean centered before classification of samples for ER status, the accuracy of the ER subgroup assignments was highly dependent on the proportion of ER+ samples in the test set; this effect of subtype prevalence was not seen when gene expression data were not mean centered.

Conclusions Simple corrections such as mean centering of genes aimed at microarray platform or batch effect correction can have undesirable consequences because patient population effects can easily be confused with these assay-related effects. Careful thought should be given to the comparability of the patient populations before attempting to force data comparability for purposes of assigning subtypes to independent subjects.

J Natl Cancer Inst 2007;99:1715–23

A frequent objective of molecular oncology studies using gene expression microarrays is to identify previously unknown cancer subtypes for which gene expression profiles are homogeneous within a subtype but different between subtypes (1–3). This class discovery objective (4) can be particularly appealing in cancer research where there is often much heterogeneity in patients' clinical outcomes that cannot be explained with standard clinical/pathologic features or biologic markers (5). Discovering new subtypes of a disease might be of great help in the decision-making process related to the choice of existing treatments as well as in the development of new target-specific therapeutics.

To transfer class discovery results from one gene expression microarray study to another in order to independently confirm the results and, most important, to assign new patients to subtypes

Affiliations of authors: Department of Experimental Oncology (LL, JFR, LDC, MG, MAP) and Unit of Medical Statistics and Biometry (EB), Fondazione IRCCS (Istituti di ricovero e cura a carattere scientifico) Istituto Nazionale dei Tumori, Milano, Italy; Molecular Genetics of Cancer Group, IFOM Fondazione Istituto FIRC (Fondazione Italiana per la Ricerca sul Cancro) di Oncologia Molecolare, Milano, Italy (LL, JFR, LDC, MG, MAP); Biometric Research Branch, National Cancer Institute, Bethesda, MD (LMM); Institute of Medical Statistics and Biometry, Università degli Studi di Milano, Milano, Italy (FA).

Correspondence to: Lara Lusa, PhD, IFOM Fondazione Istituto FIRC di Oncologia Molecolare, Via Adamello, 16 I-20139 Milano, Italy (e-mail: lara.lusa@ifom-ieo-campus.it).

See "Funding" and "Notes" following "References."

DOI: 10.1093/jnci/djm216

© The Author 2007. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oxfordjournals.org.

CONTEXT AND CAVEATS

Prior knowledge

Microarray data on the expression of multiple genes in a given sample have been used to classify breast and other cancers into subtypes that are associated with different clinical outcomes. A method had been proposed (the method of centroids) for assigning new samples to these subtypes based on the similarity of their expression profile to the mean expression profile of the previously identified subtypes.

Study design

New samples for which there was prior information on estrogen receptor status were assigned to previously identified breast cancer subtypes using the method of centroids, and the effect of subtype prevalence and systematic differences across datasets on assignment was assessed.

Contribution

This study identified a number of factors that can influence the accuracy of assignment of patient samples to previously identified cancer subtypes.

Implications

Careful consideration must be given to the comparability of patient populations and datasets in assigning samples to previously identified subtypes.

Limitations

A robust classification rule for assigning new samples that are not part of the original dataset from which the clusters were derived remains elusive.

with clinical relevance, one needs to develop an effective and reliable classification rule. Such rules are seldom provided in class discovery studies because class discovery methods such as clustering are suited for exploring the characteristics of a dataset for identifying new subtypes but not for deriving classification rules. Another challenge in projecting clustering results from one dataset to another is the existence of systematic study effects that affect the comparability of data across studies.

Grouping of breast tumors into subtypes (6–8) derived from gene expression microarray class discovery studies has received much attention (9,10). Perou et al. (7) adopted a class discovery approach using hierarchical clustering with the aim of discovering subtypes of breast cancer distinguished by different gene expression profiles. The analysis was later refined (6,8) by increasing the number of patient samples analyzed, redefining the subtypes, and showing an association of the five identified subtypes with clinical outcome. Sotiriou et al. (11) adopted a similar approach, whereas others used a supervised data analysis approach to identify new markers and molecular signatures to predict outcome or response to treatment (12–18).

Sørlie et al. (6) went beyond the mere discovery of the subtypes of breast cancer and defined a classification rule for projecting the clustering results from the original dataset (training set) to independent datasets (test sets). This classification rule, i.e., the method of centroids, was based on the similarity between the gene expression profile of samples from the test set and the centroids of the subtypes, with the centroid of a subtype defined as the vector containing the mean expression profile of all samples assigned to that subtype in the training set; the similarity between the new samples and centroids was measured by Pearson correlation. Subsequently, many investigators have used the method of centroids (19–26) or other ad hoc methods (27–36) to identify the five subtypes of breast cancer in their gene expression microarray studies. The possible implications of the putative subtypes for changes in clinical practice and the feasibility of performing the microarray assays in a robust and reproducible fashion in routine clinical settings have been less thoroughly addressed. Recently, concerns have been raised about the robustness (37) and the reproducibility of the subtypes (38) and about the lack of an operational definition of what constitutes each of them (39).

In this study, we explored the performance, as measured by classification accuracy rates, and robustness of the method of centroids, as proposed by Sørlie et al. (6), for projecting clustering results from one microarray dataset to another using real data examples and selected simulations. We examined how factors such as normalizations applied to microarray datasets and subtype prevalence influenced the ability to reliably project across datasets. The properties of some other classification methods are briefly described for purposes of discussing the generalizability of our results.

Methods

Subtypes of Breast Cancer and Their Centroids

The five subtypes of breast cancer examined in this study and defined by Sørlie et al. (6) are luminal A [based on 28 samples from the dataset of Sørlie et al. (6), 89% of which were estrogen receptor positive (ER+)], luminal B (11 samples, ER+: 82%), ERBB2+ (11 samples, ER+: 64%), basal (19 samples, ER+: 22%), and normal breast-like (10 samples, ER+: two out of three for which the data on ER status were available). Details on how the subtypes were identified by Sørlie et al. (6) in the training set can be found in the original publication.

We used the centroids of the five subtypes from the dataset of Sørlie et al. (6) that were obtained by averaging the expression levels of the intrinsic genes (genes whose expression varied the least in successive samples from the same patient's tumor but which showed the most variation among tumors of different patients) for the samples assigned to each subtype. As a part of the preprocessing and normalization steps that were applied by Sørlie et al. (6), the expression of each gene was transformed, setting the mean (and eventually the median) expression for each gene equal to zero (genes were mean centered and median centered). In practice, to mean (or median) center a gene, the mean (or median) gene expression of the gene across all the arrays is subtracted from the expression of that gene in each array. Mean centering the genes was justified (6) as a necessary step for adjusting for array batch differences within the dataset of Sørlie et al. (6), and it was also used in one of the test sets (40) (see Supplementary Information for discussion of other justifications for mean centering genes and for discussion of other types of normalization [available online]).

Microarray Data

In addition to the centroids of the five subtypes (6), the examples presented in this paper use two previously published two-channel microarray gene expression datasets. One is the dataset of 99 samples from the population-based study of Sotiriou et al. (11)

that included 65 samples that were reported to be ER+ according to the ligand-binding assay. Three hundred thirty-six out of the 552 clones included in the intrinsic gene set of Sørlie et al. (6) were present on the microarrays of Sotiriou et al. (11). The other dataset consists of 117 samples from the study by van't Veer et al. (12) and includes 71 patients that were reported to be ER+ (12). Details on data availability are reported in Supplementary Information (available online).

Subtype Membership Assignment of Breast Cancer Samples

The method of centroids, as proposed by Sørlie et al. (6), was used as the classification rule to predict breast cancer subtype membership for samples included in the dataset of Sotiriou et al. (11). We used the centroids defined by Sørlie et al. (6) and recorded which new samples would be considered nonclassifiable (6) because they had Pearson correlation less than .1 with all the centroids. We compared the classification results obtained with and without mean centering the genes in the dataset of Sotiriou et al. (11), both for the complete 99-sample dataset and for the subset of 65 ER+ samples. We evaluated whether the clusters found in the original dataset were also reliably identified in the new dataset by looking at the expression of genes that were previously used as markers to identify subtypes, by looking at the proportion of ER+ samples classified in each subtype, and by using the in-group proportion measure (38) (see Supplementary Information, available online).

Exploration of the Properties of Class Assignments Methods Using ER Status Prediction

We used the dataset of Sotiriou et al. (11) and the method of centroids to predict ER status and evaluate the effect of various factors on subtype assignments. These factors included the normalization step (i.e., mean centering the genes in training set and/or in the test set), the proportion of samples of each subtype in training and test datasets, the presence of systematic differences across datasets, and the use of an arbitrary cutoff point for the magnitude of the correlation for the purpose of identifying nonclassifiable samples.

ER status was known for all samples, assuming no measurement error. Therefore, we skipped the clustering step of class discovery on the training set and we were able to use the "true" ER status to identify the samples that were correctly classified in the test set.

Assessment of the Effect of Subtype Prevalence. To explore the effect of subtype prevalence in the test set on the accuracy of ER status assignment, we used the 99 samples from the dataset of Sotiriou et al. (11) with expression measurements for 751 clones obtained after elimination of genes showing minimal variation as previously described (11). We obtained a training set by randomly selecting a subset of 10 ER+ and 10 ER− samples from the ER+ and ER− subsets and derived the centroids. We kept the proportion of ER+ samples in the training set fixed because the proportion of samples from each class in the training set does not systematically affect the classification rule when using the method of centroids with Pearson correlation, regardless of whether the genes are mean centered.

Five test sets obtained from the collection of samples not in the training set were as follows: 1) all the samples not included in the training set (55 ER+ and 24 ER−); 2) the same number of ER+ and ER− samples (24 ER+ randomly selected from the 55 ER+ samples and 24 ER− samples); 3) 12 ER+ and 24 ER− samples randomly selected from the 55 ER+ samples and 24 ER− samples; 4) all the ER+ samples not included in the training set (55 ER+); 5) all the ER− samples not included in the training set (24 ER−). For each test set, the method of centroids was applied to predict ER status of the samples in the test set after centering or not centering the genes in the training and test sets. (See Supplementary Information for more details [available online].)

Exploration of the Effect of Systematic Differences Across Datasets Using Real and Simulated Data. We evaluated the effects of systematic assay-related differences between training and test sets with two different approaches. In the first approach, we used real data from Sotiriou et al. (11) and van't Veer et al. (12) to predict ER class membership for the van't Veer et al. (12) samples using the data of Sotiriou et al. (11) as a training set, both with and without mean centering genes in both datasets (see Supplementary Information for the details on which genes were used [available online]). We also conducted simulations in which we artificially introduced systematic study effects. (See Supplementary Information for details [available online].)

Statistical Analyses

To evaluate the accuracy of the method of centroids for predicting ER status under the various situations considered, a class-specific accuracy estimate for a given class was computed as the number of samples in the class that were correctly classified as belonging to it divided by the true number of samples in that class.

All survival curves presented were estimated using the Kaplan–Meier method as implemented in the freely available statistical software package R (41). Relapse-free survival for the patients included in the dataset of Sotiriou et al. (11) was obtained from the original publication and was defined as the interval elapsed between the date of breast surgery and the date of diagnosis of recurrent or second primary breast cancer (11).

The 95% confidence interval (CI) estimates for survival proportion at specific time points were calculated using the survfit() function of the survival library of R and were based on the log survival. Ninety-five percent confidence intervals for hazard ratio (HR) were calculated as in Simon (42), and the standard error was estimated using the survdiff() function of the survival library of R. To compare gene expression microarray–based measurements of estrogen receptor (ESR1) between identified subtypes, a $\log_2$-transformed ratio of the expression of the gene in each sample relative to common reference was calculated, and the means of these $\log_2$-transformed ratios were compared between subtypes using a two-sided t test with unequal variances (Welch test). Ninety-five percent confidence intervals for the ratio of the geometric means of the expression ratios between the two groups were calculated by forming a confidence interval for the difference in the mean $\log_2$-transformed expression ratios and back-transforming to the original scale.

Table 1. Subtype prediction of samples from the dataset of Sotiriou et al. (11)*

	With mean centering of genes†			Without mean centering of genes†		
Class	Predicted (ρ<.1)‡	ER+	Mean correlation (min–max)	Predicted (ρ<.1)	ER+	Mean correlation (min–max)
Luminal A	43 (5)	41	.22 (.07–.42)	59 (1)	55	.24 (.09–.40)
Luminal B	13 (2)	11	.17 (.04–.29)	1 (1)	1	.11 (.11–.11)
ERBB2+	13 (2)	6	.26 (.01–.41)	10 (0)	2	.13 (.04–.21)
Basal	21 (0)	0	.40 (.14–.54)	5 (0)	0	.19 (.14–.28)
Normal	9 (0)	7	.25 (.11–.46)	24 (2)	7	.21 (.05–.42)

* The prediction method is based on the method of centroids in which the centroids were defined using the data and intrinsic gene set of Sørlie et al. (6) (training set). Predicted = number of samples from the test set that were classified in the subtype; ER = estrogen receptor; ER+ = ER– positive status. Correlation = centered Pearson correlation.

† In the test set [data of Sotiriou et al. (11)].

‡ In parentheses, the number of samples for which the Pearson correlation (ρ) with the centroid of the predicted class was less than .1.

Results

Subtype Membership Assignment on Sotiriou et al. Dataset

The method of centroids was used to assign breast cancer subtypes to the 99 samples represented in the microarray dataset of Sotiriou et al. (11), both with and without mean centering the genes (Table 1, centered and noncentered analyses). The subtype assignments were strongly influenced by mean centering the genes; more than one-third of the samples were assigned to a different subtype when comparing centered and noncentered analyses (Fig. 1; Supplementary Table S1, available online).

When genes were not mean centered, luminal B (colored light blue in Fig. 1) and basal (red) subtypes were scarce, and about a quarter of the samples were classified in the normal breast-like subtype (green). Mean centering the genes caused samples to be reallocated, mostly from luminal A (dark blue) to luminal B

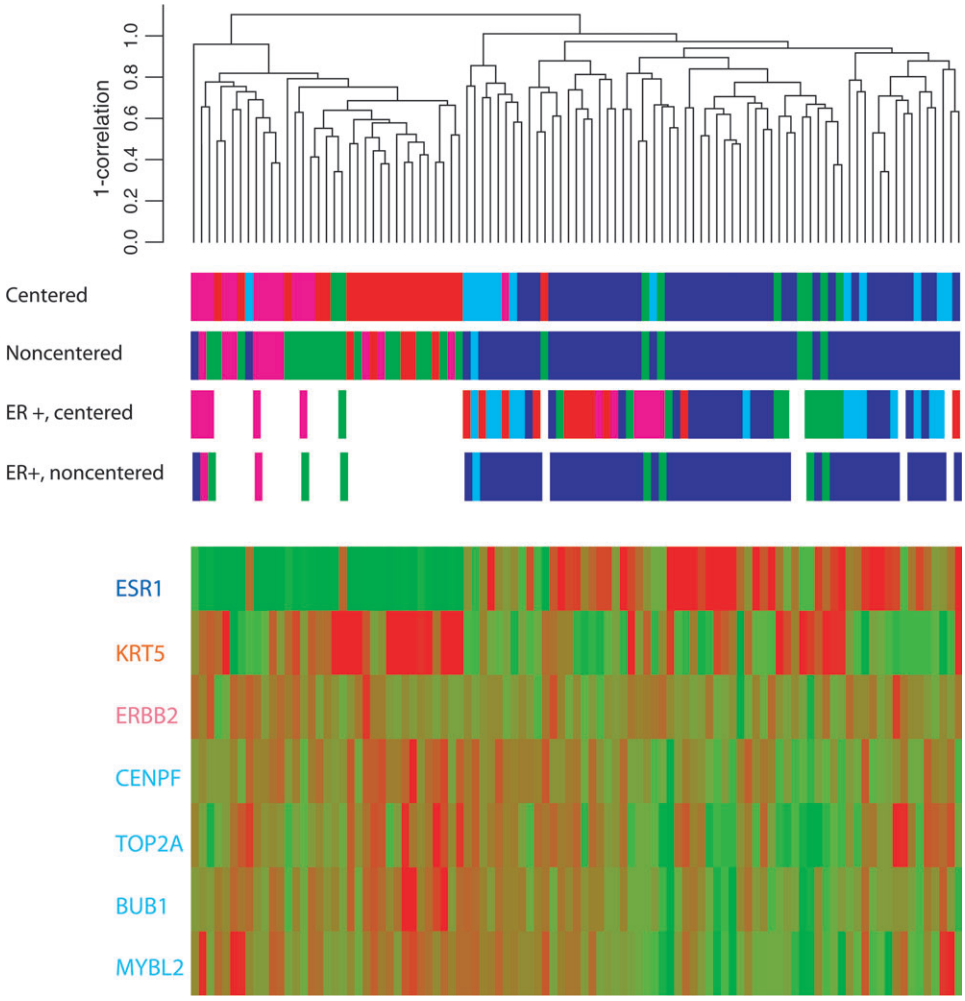

Fig. 1. Subtype prediction of data from the dataset of Sotiriou et al. (11) using method of centroids and gene expression of some genes representative of the subtypes. The dendrogram displays the results of hierarchical clustering of the complete dataset of Sotiriou et al. (11) (genes were median centered for the purpose of this clustering display). The distance metric used in the hierarchical clustering was one minus centered Pearson correlation, and linkage method was average linkage. The **colored bars** below the dendrogram represent the predicted subtype results obtained applying method of centroids mean centering the genes (Centered) and without mean centering the genes (Noncentered) in the Sotiriou et al. (11) dataset; the full dataset and a dataset restricted to estrogen receptor–positive (ER+) cases only were considered. The colors used to represent the subtypes are **dark blue** for luminal A, **light blue** for luminal B, **pink** for ERBB2+, **red** for basal and **green** for normal subtype. The expression of the genes was color coded using colors ranging from **green** (for low relative expression) to **red** (high relative expression). The first three genes shown in the **bottom panel** should be more expressed in the subtype that they represent (ERBB2 for ERBB2+ subtype, ESR1 for luminal A subtype, KRT5 for basal subtype). The set of proliferation genes (MYBL2, BUB1, TOP2A, and CENPF) should be highly expressed in basal and luminal B subtypes but not in the normal subtype. The gene expression of all the genes included in the intrinsic gene set is reported in Supplementary Fig. S1 (available online).

Table 2. Subtype prediction of samples from the dataset of Sotiriou et al. (11) limited to samples with ER+*

	With mean centering of genes†		Without mean centering of genes†	
Class	Predicted (ρ<.1)‡	Mean correlation (min–max)	Predicted (ρ<.1)	Mean correlation (min–max)
Luminal A	19 (6)	.18 (.05–.34)	55 (1)	.25 (.09–.40)
Luminal B	13 (3)	.18 (.04–.29)	1 (0)	.11 (.11–.11)
ERBB2+	11 (1)	.21 (.10–.34)	2 (0)	.11 (.10–.12)
Basal	11 (5)	.12 (.00–.27)	0 (0)	ND
Normal	11 (1)	.23 (.07–.45)	7 (0)	.21 (.10–.42)

* The prediction method is based on the method of centroids in which the centroids were defined using the data and intrinsic gene set of Sørlie et al. (6) (training set). Predicted = number of samples from the test set that were classified in the subtype; ER+ = estrogen receptor positive status; ND = nondefinable estimate.

† In the test set [samples of Sotiriou et al. (11) with ER+].

‡ In parentheses, the number of samples for which the Pearson correlation (ρ) with the centroid of the predicted class was less than .1

subtypes and from normal breast-like subtype to basal and ERBB2 (pink) subtypes.

Although the true distribution of the samples from the test set according to the five previously identified subtypes is unknown, the results obtained with the centered analysis appeared more consistent with the subtype characteristics described by Sørlie et al. (6) than did results obtained without mean centering the genes. For instance, based on the results of Sørlie et al. (6), we would expect that most of the ER− samples would be classified in the basal subtype instead of the normal and we would expect some of the ER+ samples to be classified in the luminal B subtype. The results obtained with the centered analysis on the complete dataset were also more consistent with the patterns of gene expression that distinguish the subtypes as they were described by Sørlie et al. (6), and this conclusion was supported by visual inspection of the hierarchical clustering of the samples as depicted in the dendrogram in Fig. 1 together with their predicted subtypes and gene expression of a selected subset of genes (Supplementary Fig. S1 shows the expression of the complete intrinsic gene set, available online). The first three genes in Fig. 1 should be more expressed in the subtype that they represent (c-erb B2/neu [ERBB2] for ERBB2+ subtype, ESR1 for luminal A subtype, cytokeratin 5 [KRT5] for basal subtype). These genes are conventionally used as markers to subclassify luminal, ERBB2+, and basal subtypes by immunohistochemical staining (10,43). The set of proliferation genes (MYBL2, BUB1, TOP2A, and CENPF) should be highly expressed in basal and luminal B subtypes but not in the normal subtype (44,45).

The gene expression of most of the samples that were reallocated from normal to basal subtype by the centered analysis was consistent with the known characteristics of basal samples: they had low expression of ESR1 and high expression of basal cytokeratin KRT5 and of other genes assigned to the basal cluster of Sørlie et al. (6). Unlike the normal samples, they had high expression of proliferation genes. In addition, the samples reallocated from luminal A to luminal B subtype showed characteristics that were consistent with the definition of the luminal B subtype, i.e., on average they had moderate expression of ESR1 and high expression of proliferation genes.

Next, we restricted our attention to the subset of 65 ER+ samples from the dataset of Sotiriou et al. (11). Subtype assignment of new samples is independent of the characteristics of the other samples in the test set if genes in the test set are not mean centered; therefore, classification results for noncentered data were the same as those discussed for the complete data. The classification did not seem to be reliable on this subset of samples when genes were mean centered. Although the luminal A subtype was somewhat more abundant than the other assigned subtypes, samples were roughly uniformly distributed across all the subtypes (Table 2; Supplementary Table S2, available online), contrary to the expectation based on the results of Sørlie et al. (6) that most of the ER+ samples would be assigned to luminal A or luminal B subtype. About half of the samples were assigned to nonluminal subtypes, which should contain the majority of ER− samples (Fig. 1; Supplementary Fig. S1, available online). The ER+ samples assigned to the basal subtype had on average a lower expression of ESR1, but at the same time their expression of KRT5 and of other genes from the basal cluster was not high, as should be expected for basal samples (Supplementary Fig. S1, available online). Therefore, even though the in-group proportion measure indicated support for the existence of all the subtypes ($P<.10$ for all subtypes; Supplementary Table S3, available online), mean centering the genes did not seem to provide a reliable classification when only ER+ samples were considered. The explanation for the fact that many of the samples were classified to a different subtype than the one assigned in the complete data analysis (28 out of 65) is that gene centering disrupts the correlations between the profiles in the test set and the training set centroids, and these correlations are the basis for the predictions. Moreover, the way in which the profiles in the test set are modified depends on the class distribution in the test set.

Effect of Subtype Prevalence in Test Sets

To explore further the properties of the method of centroids, we switched from breast cancer subtype projection to the simpler problem of predicting known ER status. As previously reported (40), it was possible to predict ER status with high predictive accuracy also on the dataset of Sotiriou et al. (11).

We considered five test sets that included different proportions of ER+ samples (Table 3) using the data of Sotiriou et al. (11). When genes were mean centered both in the training and the test set (Table 3; Supplementary Fig. S2, available online), there was a strong dependence of both overall predictive accuracy and

Table 3. ER class prediction results for five test sets with different prevalence of ER+ samples*

Test set†		With mean centering of genes‡			Without mean centering of genes‡		
True ER+	True ER–	Predicted ER+ (correct/ incorrect)§	Predicted ER– (correct/ incorrect)	Predictive accuracy§ ER+/ER–	Predicted ER+ (correct/ incorrect)	Predicted ER– (correct/ incorrect)	Predictive accuracy ER+/ER–
55	24	46 (43/3)	33 (21/12)	78%/88%	57 (52/5)	22 (19/3)	95%/79%
24	24	25 (21/4)	23 (20/3)	88%/83%	27 (22/5)	21 (19/2)	92%/79%
12	24	16 (11/5)	20 (19/1)	92%/79%	16 (11/5)	20 (19/1)	92%/79%
55	0	29 (29/0)	26 (0/26)	53%/ND	52 (52/0)	3 (0/3)	95%/ND
0	24	9 (0/9)	15 (15/0)	ND/62%	5 (0/5)	19 (19/0)	ND/79%

* The prediction method is based on the method of centroids in which the centroids were defined using the data of the most variable genes of Sotiriou et al. (11). ER = estrogen receptor; ER+ = ER– positive status; ER– = ER– negative status; ND = nondefinable estimate.

† The training set was the same for all the examples and was selected randomly sampling 10 ER+ and 10 ER– samples from the dataset of Sotiriou et al. (11).

‡ In both the training and the test sets.

§ Predicted ER+ is the number of samples from test set that were classified in the ER+ class; correct/incorrect indicates the number of correctly and incorrectly classified samples.

class-specific predictive accuracy on the proportion of ER+ samples in the test set. This was not the case when genes were not mean centered (noncentered analysis, Table 3; Supplementary Fig. S2, available online).

The predictive accuracy of the noncentered analysis was consistently better or equivalent to that obtained with the centered analysis. The drop in the predictive accuracy in the centered analysis was particularly striking when the test set included only ER+ samples (47% of samples misclassified) or only ER– samples (38% of samples misclassified). The same samples were classified with high predictive accuracy when included in the test set with 55 ER+/24 ER– samples, where overall less than 20% of the samples were misclassified (Table 3; Supplementary Fig. S2, available online).

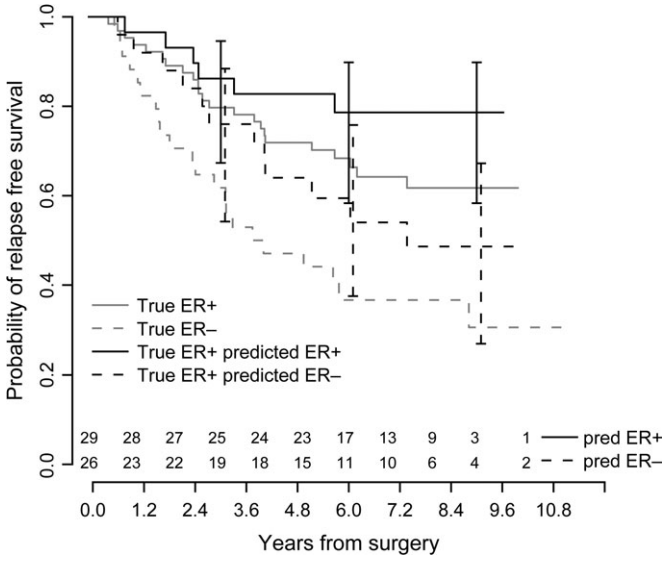

Fig. 2. Kaplan–Meier estimates of disease-free survival curves of true and predicted estrogen receptor–positive (ER+) and ER– samples from the dataset of Sotiriou et al. (11). **Gray lines**: disease-free survival curves estimated with the method of Kaplan–Meier for ER+ (**solid line**) and ER– (**dashed line**) patients included in the dataset of Sotiriou et al. (11). **Black lines**: estimated disease-free survival curves for a test set that included 55 ER+ samples from the dataset of Sotiriou et al. (11), separately reported for the groups of predicted ER+ (**solid line**) and predicted ER– (**dashed line**) samples, together with the 95% confidence intervals estimated at three different time points. Predictions were obtained with the method of centroids, mean centering the genes in both training and test sets. The training set was based on 10 ER+/10 ER– samples. The number of patients at risk in the two predicted groups is reported at the bottom of the figure.

Misclassification of the subtypes could have important implications for clinical interpretation. When considering the test set that included only ER+ samples (test set 4), we observed a statistically significant difference in the relapse-free survival curves between the predicted ER+ and predicted (and misclassified) ER– samples (log-rank test $P = .045$, HR = 2.62 [ER–/ER+], 95% CI = 1.02 to 6.59), and therefore the ER+ patients that were correctly classified had better survival than the full set of ER+ patients. However, the patients whose samples were incorrectly classified as ER– had a statistically significantly better survival than true ER– patients ($P = .044$, HR = 2.08 [ER–/ER+], 95% CI = 1.02 to 4.37, see Fig. 2). The two predicted classes had statistically significantly different mean levels of ER as measured by microarray (mean $\log_2$ ESR1 values 2.26 for predicted ER+ and 1.46 for predicted ER–, 95% CI for the ratio of the geometric means of the expression ratios [ER+/ER–] = 1.22 to 2.51, $P = .003$).

In the centered analysis, the overall predictive accuracy decreased as the proportion of ER+ samples deviated from 50% in the test set and the class-specific predictive accuracy decreased as the proportion of samples in that class in the test set increased (see Table 3 for some examples). This result is not surprising because it can be demonstrated analytically (not shown) that the probability of assigning a sample to a class decreases as the proportion of samples from the test set in that class increases when using the method of centroids and mean centering the genes. Furthermore, the problems we observed in the centered analysis are not overcome when the prevalence of the subtypes in the training set is matched with the unknown prevalences in the test set but depend only on the prevalence of the subtypes in the test set.

We also performed a set of simulations based on multiple resamplings of the dataset of Sotiriou et al. (11) to assess whether the results that were observed using the five different test sets were independent of the specific choices of the training and test sets. Simulation results confirmed that when genes were mean centered there was strong dependence of the predictive accuracy on the

Table 4. ER class prediction results for the samples from the dataset of van't Veer et al. (12) (test set) based on a predictor developed using data from the dataset of Sotiriou et al. (11) (training set)*

	With mean centering of genes†			Without mean centering of genes†		
Class	True ER+‡	True ER–	Mean correlation (min–max)	True ER+	True ER–	Mean correlation (min–max)
Predicted ER+§ (ρ<.1)‖	67 (4)	7 (4)	.42 (.03–.62)	63 (53)	3 (3)	.02 (–.24–.13)
Predicted ER– (ρ<.1)	4 (2)	39 (1)	.26 (.01–.55)	8 (7)	43 (43)	–.03 (–.23–.16)

* The prediction method is based on the method of centroids in which the centroids were defined using the data of the most variable genes of Sotiriou et al. (11). ER = estrogen receptor; ER+ = ER– positive status; ER– = ER– negative status.

† In both the training and the test sets.

‡ True ER+ is the number of samples that belong to the ER+ class in the test set.

§ Predicted ER+ is the number of samples from test set that were classified in the ER+ class.

‖ In parentheses, the number of samples for which the Pearson correlation (ρ) with the centroid of the predicted class was less than .1.

proportion of ER+ samples included in the test set (Supplementary Fig. S3, available online). In terms of overall predictive accuracy, centered and noncentered analyses gave comparable results when the proportion of ER+ samples in the test set was between 30% and 70% (see Supplementary Fig. S3, available online). When the proportion of ER+ and ER− samples was more unbalanced, the class-specific predictive accuracy of the less represented subtype was greatly reduced by mean centering the genes. The Supplementary Information (available online) reports a more thorough presentation of the simulation results, with additional details about the effect of mean centering on the correlations and on the proportion of nonclassifiable samples.

Effect of Systematic Differences Across Datasets

To evaluate the effects of potential systematic differences across datasets, we predicted ER class membership for samples from the dataset of van't Veer et al. (12) using the data of Sotiriou et al. (11) as a training set. Overall predictive accuracy (91%) was the same whether genes were mean centered or not, although there were some minor differences in class-specific predictive accuracies. However, the low correlations observed with the noncentered analysis (always less than .16; Table 4) suggested that when genes were not mean centered, it was not possible to project ER status membership from the dataset of Sotiriou et al. (11) to that of van't Veer et al. (12) for most of the samples (i.e., 91% of the samples had Pearson correlation less than .1 with all the centroids and would have been considered nonclassifiable; Table 4). This result was confirmed using simulated data (Supplementary Information, available online) in which systematic effects were artificially introduced into the data and was in contrast to what was observed in the previous examples, where centering the genes reduced the correlations when training and test sets came from the same dataset. However, in both centered and noncentered analyses, designating samples as nonclassifiable on the basis of low correlation did not prove to be a reliable method of identifying incorrectly classified samples (Supplementary Information, available online).

Discussion

Gene expression microarray experiments and class discovery methods have been used to identify previously unknown subtypes of diseases. We focused on the molecular classification of breast cancer into subtypes (6–8) and addressed some recently raised concerns about the subtypes (37–39) through analysis of several real datasets and through resampling-based simulations.

We showed that many difficulties remain in validating and extending class discovery results to new samples and that projection of clusters from one dataset to another must be done with care. Centering of genes, proportion of samples from each of the subtypes present in the test set, and systematic study effects were identified as factors that play a role in how accurately subtypes that had been discovered in a previous dataset can be identified in an independent dataset. We found that the appropriateness of gene centering depends on the particular situation. If there is a clear additive and strong study effect but the two datasets arise from roughly similar populations, then centering may be helpful, even though we showed that matching the prevalence in the training and test set itself does not guarantee good performance of the classifier. If there are substantial study effects and differences in subtype prevalence, centering will not solve the problems posed by differences in training and test sets, so we recommend that careful consideration should be given to the comparability of patient populations. Unfortunately, in practice, patient population effect can easily be confused with microarray platform or batch effects. Most of the problems that we identified in our study that prevent projection of clusters from one dataset to another persisted when projection methods other than the method of centroids were used (see Supplementary Discussion section for more detailed discussion of this finding [available online]).

Other studies have attempted to reproduce subtypes in a new dataset without using a method of centroids-type projection but simply by clustering the new data using the set of genes defined in a previous dataset and assuming a similar number of clusters (27,28,30–32). This type of class assignment does not suffer from exactly the same problems that we pointed out for the method of centroids, but it still has many limitations. It is always possible to find a given number of subtypes in a new dataset using clustering techniques. However, clusters are not automatically associated to the subtypes and it can be problematic to assess if the subtypes found in new data correspond to those that were previously observed in a different dataset. These kinds of studies generally do not provide additional insight on how to classify new samples because it is still unclear how to classify a sample that is not part of the dataset that has been clustered. A robust classification rule for new samples remains elusive.

Some of the studies that claimed to have "validated" the breast cancer subtypes have focused on comparing the clinical outcome differences between subtypes assigned in an independent cohort with the differences previously reported for the subtypes (8,26,30). Although this approach can provide supporting evidence for an association between the gene expression profiles and clinical outcome, it does not provide a direct measure of the robustness of the specific clinical classification at an individual level, which is essential before assigning patients to subtypes in clinical practice for purposes of risk stratification or therapy selection (26,30). As we showed with the example of the two groups predicted as ER+ and ER− within the group of ER+ samples, even though a statistically significant difference in survival was observed between the two predicted groups, many patients were misclassified. If these incorrect groups had been used to determine which patients received endocrine therapy, it is possible that patients with moderate to weak ER expression (who were misclassified as ER−) could have been denied endocrine therapy from which they might have received clinical benefit. This situation might share some similarities with what has been observed for luminal A and luminal B subtypes of breast cancer (6), with patients with luminal B breast cancer having worse prognosis than patients with the luminal A subtype. Even though most luminal B samples are ER+, this subtype is characterized by lower expression of ER compared with the luminal A subtype (6). It is possible that the luminal A and luminal B subtypes lie on a biological continuum and that no clear delineation of the two groups really exists.

A clinically useful classifier for breast cancer subtypes must satisfy a number of conditions. It must unambiguously classify a new sample into a specific subtype independently of any other samples being considered for classification at the same time. The clinical meaning of the subtype assignment (e.g., survival probability or probability of response to a particular drug) must be stable across populations to which the classifier might be applied. The technology platform(s) that produce(s) the profiles must be stable enough so that when the same sample is assayed on different occasions it will with very high likelihood be classified to the same subtype. We have demonstrated in this paper that the currently claimed breast cancer subtypes fall substantially short of meeting all of these requirements.

Our study had some limitations. We focused solely on breast cancer gene expression microarray datasets, and we considered a limited number of cluster projection methods. It is possible that the strong effect of ER status on gene expression profiles and its major role in defining the putative breast cancer subtypes produced a particularly dramatic effect when ER distribution was perturbed and gene centering was used. However, given the wide attention that the breast cancer subtypes have received, we feel that the breast cancer examples are highly relevant. Although we did not provide a comprehensive evaluation of the performance of a large number of cluster projection methods, the fundamental difficulty in disentangling assay-related study effects from true biological difference in populations leads us to believe that identification of a universally robust cluster projection method with the ability to cross microarray platforms will be difficult. More reproducible approaches to profiling through the standardization of microarray methods or other technologies such as reverse transcription–polymerase chain reaction or immunohistochemistry will likely be helpful in achieving the necessary robustness of results to transform these promising findings to clinically useful tools.

References

(1) Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 2000;406:536–40.

(2) Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403:503–11.

(3) Simon RM, Korn EL, McShane LM, Radmacher MD, Wright GW, Zhao Y. Design and analysis of DNA microarray investigations. New York, NY: Springer; 2004. Chapter 9.

(4) Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531–7.

(5) Bieche I, Lidereau R. Genetic alterations in breast cancer. Genes Chromosomes Cancer 1995;14:227–51.

(6) Sørlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci USA 2003;100:8418–23.

(7) Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature 2000;406: 747–52.

(8) Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA 2001;98: 10869–74.

(9) Brenton JD, Carey LA, Ahmed AA, Caldas C. Molecular classification and molecular forecasting of breast cancer: ready for clinical application? J Clin Oncol 2005;23:7350–60.

(10) Charafe-Jauffret E, Ginestier C, Monville F, Fekairi S, Jacquemier J, Birnbaum D, et al. How to best classify breast cancer: conventional and novel classifications (review). Int J Oncol 2005;27:1307–13.

(11) Sotiriou C, Neo SY, McShane LM, Korn EL, Long PM, Jazaeri A, et al. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. Proc Natl Acad Sci USA 2003; 100:10393–8.

(12) van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:484–5.

(13) van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 2002;347:1999–2009.

(14) Huang E, Cheng SH, Dressman H, Pittman J, Tsou MH, Horng CF, et al. Gene expression predictors of breast cancer outcomes. Lancet 2003;361:1590–6.

(15) Ayers M, Symmans WF, Stec J, Damokosh AI, Clark E, Hess K, et al. Gene expression profiles predict complete pathologic response to neoadjuvant paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide chemotherapy in breast cancer. J Clin Oncol 2004;22:2284–93.

(16) Ma XJ, Wang Z, Ryan PD, Isakoff SJ, Barmettler A, Fuller A, et al. A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. Cancer Cell 2004;5:607–16.

(17) Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 2005;365:671–9.

(18) Foekens JA, Atkins D, Zhang Y, Sweep FC, Harbeck N, Paradiso A, et al. Multicenter validation of a gene expression-based prognostic signature in lymph node-negative primary breast cancer. J Clin Oncol 2006;24: 1665–71.

(19) Zhao H, Langerod A, Ji Y, Nowels KW, Nesland JM, Tibshirani R, et al. Different gene expression patterns in invasive lobular and ductal carcinomas of the breast. Mol Biol Cell 2004;15:2523–36.

(20) Mecham BH, Klus GT, Strovel J, Augustus M, Byrne D, Bozso P, et al. Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. Nucleic Acids Res 2004;32:e74.
(21) Chang HY, Nuyten DS, Sneddon JB, Hastie T, Tibshirani R, Sørlie T, et al. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. Proc Natl Acad Sci USA 2005;102:3738–43.
(22) Farmer P, Bonnefoi H, Becette V, Tubiana-Hulin M, Fumoleau P, Larsimont D, et al. Identification of molecular apocrine breast tumours by microarray analysis. Oncogene 2005;24:4660–71.
(23) Bertucci F, Finetti P, Cervera N, Charafe-Jauffret E, Mamessier E, Adelaide J, et al. Gene expression profiling shows medullary breast cancer is a subgroup of basal breast cancers. Cancer Res 2006;66:4636–44.
(24) Van Laere SJ, Van den Eynden GG, Van der Auwera I, Vandenberghe M, van Dam P, Van Marck EA, et al. Identification of cell-of-origin breast tumor subtypes in inflammatory breast cancer by gene expression profiling. Breast Cancer Res Treat 2006;95:243–55.
(25) Sørlie T, Wang Y, Xiao C, Johnsen H, Naume B, Samaha RR, et al. Distinct molecular mechanisms underlying clinically relevant subtypes of breast cancer: gene expression analyses across three different platforms. BMC Genomics 2006;7:127.
(26) Calza S, Hall P, Auer G, Bjohle J, Klaar S, Kronenwett U, et al. Intrinsic molecular signature of breast cancer in a population-based cohort of 412 patients. Breast Cancer Res 2006;8:R34.
(27) Yu K, Lee CH, Tan PH, Tan P. Conservation of breast cancer molecular subtypes and transcriptional patterns of tumor progression across distinct ethnic populations. Clin Cancer Res 2004;10:5508–17.
(28) Kristensen VN, Sørlie T, Geisler J, Langerod A, Yoshimura N, Karesen R, et al. Gene expression profiling of breast cancer in relation to estrogen receptor status and estrogen-metabolizing enzymes: clinical implications. Clin Cancer Res 2005;11:878s–83s.
(29) Wang ZC, Lin M, Wei LJ, Li C, Miron A, Lodeiro G, et al. Loss of heterozygosity and its correlation with expression profiles in subclasses of invasive breast cancers. Cancer Res 2004;64:64–71.
(30) Rouzier R, Perou CM, Symmans WF, Ibrahim N, Cristofanilli M, Anderson K, et al. Breast cancer molecular subtypes respond differently to preoperative chemotherapy. Clin Cancer Res 2005;11:5678–85.
(31) Weigelt B, Hu Z, He X, Livasy C, Carey LA, Ewend MG, et al. Molecular portraits and 70-gene prognosis signature are preserved throughout the metastatic process of breast cancer. Cancer Res 2005;65:9155–8.
(32) Perreard L, Fan C, Quackenbush JF, Mullins M, Gauthier NP, Nelson E, et al. Classification and risk stratification of invasive breast carcinomas using a real-time quantitative RT-PCR assay. Breast Cancer Res 2006;8:R23.
(33) Muggerud AA, Johnsen H, Barnes DA, Steel A, Lonning PE, Naume B, et al. Evaluation of MetriGenix custom 4D arrays applied for detection of breast cancer subtypes. BMC Cancer 2006;6:59.
(34) Oh DS, Troester MA, Usary J, Hu Z, He X, Fan C, et al. Estrogen-regulated genes predict survival in hormone receptor-positive breast cancers. J Clin Oncol 2006;24:1656–64.
(35) Hu Z, Fan C, Oh DS, Marron JS, He X, Qaqish BF, et al. The molecular portraits of breast tumors are conserved across microarray platforms. BMC Genomics 2006;7:96.
(36) Fan C, Oh DS, Wessels L, Weigelt B, Nuyten DSA, Nobel AB, et al. Concordance among gene-expression–based predictors for breast cancer. N Engl J Med 2006;355:560–9.
(37) Pusztai L, Mazouni C, Anderson K, Wu Y, Symmans WF. Molecular classification of breast cancer: limitations and potential. Oncologist 2006;11:868–77.
(38) Kapp AV, Tibshirani R. Are clusters found in one dataset present in another dataset? Biostatistics 2007;8:9–31.
(39) Loi S, Sotiriou C, Buyse M, Rutgers E, van't Veer L, Piccart M, et al. Molecular forecasting of breast cancer: time to move forward with clinical testing. J Clin Oncol 2006;24:721–2;author reply 722–3.
(40) West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci USA 2001;98:11462–7.
(41) R Development Core Team. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing; 2005.ISBN 3-900051-07-0. Available at: http://www.R-project.org. [Last accessed: October 22, 2007.]
(42) Simon R. Confidence intervals for reporting results of clinical trials. Ann Intern Med 1986;105:429–35.
(43) Nagle RB, Bocker W, Davis JR, Heid HW, Kaufmann M, Lucas DO, et al. Characterization of breast carcinomas by two monoclonal antibodies distinguishing myoepithelial from luminal epithelial cells. J Histochem Cytochem 1986;34:869–81.
(44) Whitfield ML, George LK, Grant GD, Perou CM. Common markers of proliferation. Nat Rev Cancer 2006;6:99–106.
(45) Perreard L, Fan C, Quackenbush JF, Mullins M, Gauthier NP, Nelson E, et al. Classification and risk stratification of invasive breast carcinomas using a real-time quantitative RT-PCR assay. Breast Cancer Res 2006;8:R23.

Funding

Italy–U.S.A. Fellowship of the Istituto Superiore di Sanità on Oncological Pharmacogenomics—Seroproteomics; Associazione Italiana per la Ricerca sul Cancro (to M. A. P. and M. G.); European Community FP6 Project No. 503438 (Transfog) (to M. A. P.).

Notes

We thank an anonymous referee whose thorough and constructive comments helped to substantially improve the manuscript. This study utilized the high-performance computational capabilities of the Biowulf/LoBoS3 cluster at the National Institutes of Health, Bethesda, MD. The authors take full responsibility for the study design, data collection, analysis and interpretation of the data, the decision to submit the manuscript for publication, and the writing of the manuscript.

Manuscript received December 5, 2006; revised September 7, 2007; accepted October 1, 2007.