Breast Cancer Res Treat (2009) 118:433–441
DOI 10.1007/s10549-008-0242-8

PRECLINICAL STUDY

Meta-analysis of gene expression profiles related to relapse-free survival in 1,079 breast cancer patients

Balazs Győrffy · Reinhold Schäfer

Received: 8 August 2008 / Accepted: 28 October 2008 / Published online: 5 December 2008

Abstract The transcriptome of breast cancers have been extensively screened with microarrays and large sets of genes associated with clinical features have been established. The aim of this study was to validate original gene sets on a large cohort of raw breast cancer microarray data with known clinical follow-up. We recovered 20 publications and matched them to Affymetrix HGU133A annotations. Raw Affymetrix HGU133A microarray data were extracted from GEO and MAS5 normalized. For classifying patients using the selected gene sets, we applied prediction analysis of microarrays and constructed Kaplan–Meier plots. A new classification including all patients was generated using supervised principal components analysis. Seven studies including 1,470 patients were downloaded from GEO. Notably, we uncovered 641 microarrays representing 251 individual tumor specimens among them, which were repeatedly described under independent GEO identifiers. We excluded all redundant data and used the remaining 1,079 samples. Eight of the 20 gene sets were able to predict response at a significance of $P < 0.05$. The discrimination of good and poor prognosis groups exclusively relying on gene expression data resulted in high significance ($P = 1.8\text{E}{-12}$). A model including genes fitted by both gene expression and clinical covariates (lymph node status and grade) contains 44 genes and can predict response at $P = 9.5\text{E}{-7}$. The outcome provides a ranking of the gene lists regarding applicability on an independent dataset. We established a consensus predictor combining the available clinical and gene expression data. The database comprising expression profiles of 1,079 breast cancers can be used to classify individual patients.

Keywords Microarray · Gene expression signature · Breast cancer prognosis · Bioinformatics

Introduction

Although molecular markers like expression of estrogen and growth factor receptors, pS2, metallothionein, CD24, cathepsin D, ERBB2, and mutations in the TP53 gene all have been correlated to breast cancer prognosis, the use of single marker provides limited information for the prognosis of an individual patient [1, 2]. In view of the molecular heterogeneity of breast tumors and the large number of marker genes involved, studying multiple genetic alterations simultaneously is of utmost importance. With the arrival of microarray technologies, searches for tumor markers can be performed in a discovery-driven manner in high through-put.

The first microarray-based breast cancer studies have revealed distinct clinical phenotypes. Two major types, basal and luminal, have been identified, each with the

Electronic supplementary material The online version of this article (doi:10.1007/s10549-008-0242-8) contains supplementary material, which is available to authorized users.

B. Győrffy
Research Group for Pediatrics and Nephrology, Hungarian Academy of Sciences and Semmelweis University, Budapest, Hungary

B. Győrffy (✉)
Children's Hospital Boston Informatics Program, Harvard-MIT Health Sciences and Technology, 300 Longwood Avenue, Enders 150.6, Boston, MA, USA
e-mail: zsalab2@yahoo.com

R. Schäfer
Laboratory of Molecular Tumor Pathology and Laboratory of Functional Genomics, Charité, Universitätsmedizin Berlin, Berlin, Germany

potential to be subdivided into additional subtypes [2–5]. Although histological grade can provide clinically important prognostic information, as many as 30–60% of tumors are classified as grade 2. This grade is associated with an intermediate risk of recurrence and is thus not informative for clinical decision making. Gene expression signatures capable of discerning tumors of grade 1 (G1) and grade 3 (G3) histology might provide a more objective measure of grade with prognostic benefit for patients with G2 disease [5].

Estrogen receptor (ER) status is the only globally accepted treatment predictive factor for hormonal therapy in primary breast cancer. As only a small proportion (7%) of cells in the normal mammary epithelium express ER [6], receptor status is the main discriminator in the high proportion of ER+ tumors. The ER status of breast tumors has been suggested to either reflect tumor progression with ER− tumors evolving from ER+ precursors, or to indicate a distinct origin from different types of epithelial cells in the mammary gland. Metastases from ER+ tumors may be ER− [7] supporting the hypothesis that ER-expressing and ER-negative breast cancers represent different disease entities [8]. In contrast, a large proportion of the patients with ER+ breast cancer do not respond to tamoxifen. These unsolved issues led to a significant number of studies investigating ER status and prognosis in breast cancer [8–12].

While many of these studies presented promising results, most proposed markers were not reproduced in consecutive studies. The proposed best discriminatory genes rarely match in different studies. A major criticism has been that in 90% of early reports the validation set of patients overlaps with the training set [13]. Additional critical issues regarding the use of microarray data for prognostic classification include gene selection bias, error estimation, fragility of gene signatures, and overoptimistic performance estimation due to model over-fitting [14]. Over-fitting means finding a discriminatory gene pattern by chance. This can happen when large numbers of variables (genes) are assessed for a small number of samples [13].

Michiels et al. re-analyzed data from seven large published studies that have attempted to predict prognosis of cancer patients using data from DNA microarray analysis. They expanded the standard strategy based on unique training and validation sets by using multiple random sets. The list of genes identified as predictors of prognosis was highly unstable; molecular signatures strongly depended on the selection of patients in the training sets, the proportion of misclassified patients decreased as the number of patients in the training set increased. Five of the seven studies did not classify patients better than chance [15].

Thus, information based on microarray analysis requires independent validation in distinct data sets [16]. While the use of microarray technology as a diagnostic tool can potentially revolutionize current breast cancer management, critical scientists advocate further studies with profiling of larger cohorts using single microarray platforms before prospective clinical use of molecular classifiers can be contemplated [17].

In present study we aimed to perform a large-scale meta-analysis to compare several different gene sets for predicting relapse in a set of raw microarray data accessible through Gene expression omnibus (GEO). A second aim of the study was to establish a consensus predictor combining all suggested genes and available patient samples.

Methods

Included raw microarray studies

We systematically searched GEO (http://www.ncbi.nlm.nih.gov/geo/) using the keywords "breast cancer" and "gpl96" (platform accession for Affymetrix HGU133A microarrays). Only studies publishing data for more than 20 patients with available clinical information were considered. Seven studies published such raw data in GEO, which were downloaded.

Included gene lists

We have searched Pubmed using the keywords "breast cancer" and "microarray". The search was then limited to studies in English with Pubmed accessible text. Only genome-wide association studies were selected. Studies investigating <20 patients or publishing <5 genes were excluded from the study.

Annotation and matching to the Affymetrix IDs was performed using the Affymetrix Netaffx analysis centre (http://www.affymetrix.com/analysis/index.affx). For the construction of the matched gene lists the available gene identifiers (Unigene ID, Genbank Accession, gene symbol, Affymetrix ID) were used. The datasets were combined using Microsoft Access 2007.

Statistical analyses

The downloaded data was MAS 5.0 normalized in the *R* statistical environment (http://www.R-project.org). MAS5.0 applies normalization on an individual chip; it has excellent specificity and good sensitivity. As MAS 5.0 it is the factory-default normalization method, in the future even single microarrays can be added to our table.

In order to apply the gene sets we used the updated version of the "Prediction Analysis for Microarrays" (PAM) [18] which uses a semi-supervised method to

predict patient survival [19]. PAM was performed in a leave-one-out cross validation and the threshold was set to include all genes in the prediction. When investigating datasets with GEO available microarrays, the original dataset was excluded from the analysis. PAM is a modification of the nearest-centroid method and was applied as previously described.

For establishing a new, consensus predictor, the BRB Arraytools 3.6.0-beta_3 package was used (developed by Dr. Richard Simon and Amy Peng Lam, available at http://linus.nci.nih.gov/BRB-ArrayTools.html). Instead of using a separate test-set and training-set we performed a leave-one-out cross validation to assess the performance of the new predictor. When investigating the gene sets of the seven studies delivering the GEO data, the samples resulting from the corresponding study were excluded from the analysis to avoid inclusion of the training set (the original samples the gene list was derived from) in the test set. An overview of the applied analytical pathway is depicted on Fig. 1.

Using a new approach implemented in BRBArrayTools we evaluated whether the expression data provides more accurate predictions than that provided by the two other significant clinical covariates, lymph-node status and grade. Here, an additional model is developed for a combination of the covariates and the expression data. For each cross-validated training set, genes are selected which add to predicting survival over the predictive value provided by the covariates. The principal components of those genes are computed and a model fitted containing the covariates and the supervised principal components. The survival risk group for the patient omitted from that training set is predicted using that composite model. Finally a *P* value is determined which measures whether the expression data adds significantly to risk prediction compared to the covariates.

Descriptive statistics, significance for clinical variables and survival plots were constructed using the Winstat for Excel software.

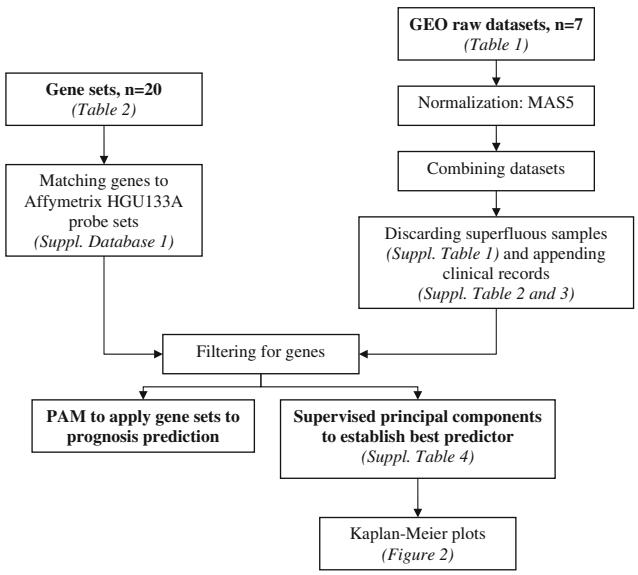

Fig. 1 Overview of the applied analytical pathway

Results

Creating a non-redundant database for breast cancer specimens subjected to expression profiling

Altogether, GEO listed 1,470 raw GPL96 microarray samples from published studies. When surveying the clinical data, we observed a high similarity between some studies. Therefore, we compared the gene expression data of all microarrays and identified 641 redundant samples related to 251 individual raw microarray files first published in GEO under the series accession number GSE3494. The remaining datasets (GSE2990, GSE4922 and GSE6532) include 389 microarrays identical to GSE3494 but supplemented with additional clinical information (some microarrays were deposited more than twice). We have listed the redundant GSE3494 samples in Supplemental Table 1 for future reference. The database includes the GEO series accession numbers, the GEO sample accession numbers and the average of normalized expression values of all transcripts for a given sample. We found identical average expression values only on identical microarrays. As none of the publications showed the complete clinical information, we merged the individual clinical features into one database (Table 1). The complete dataset containing the normalized expression values of the 1,079 chips is shown in Supplemental Table 2. The Supplemental Table 3 containing the detailed clinical information records includes all available clinical data for each patient. We excluded all redundant samples for the statistical analysis totaling 1,079 microarrays.

Meta-analysis of published gene sets

Twenty-four published studies representing 20 gene sets with discriminatory potential for clinical relapse were included in the study. If several gene lists were available in these studies, we selected the most extensive one exhibiting statistical significance. The gene lists are summarized in Table 2.

For the meta-analysis of the previously published breast cancer associated genes we used Prediction Analysis of Microarrays to predict the risk of relapse using the non-redundant set of microarrays. The analysis was performed independently for all 20 published discriminatory gene

Table 1 Descriptive statistics of the raw datasets used in the study

GEO ID	Grade: 1/2/3	Proportion of ER+	Proportion of lymph node+	Relapse event	Average relapse free survival	Age (year)	Size (mm)	Number of published CEL files	Number of included GSE3494 CEL files	Number of individual CEL files	Reference
GSE1456	28/58/61	NA	NA	40 (25%)	6.2 ± 2.3	NA	NA	159	0	159	[26]
GSE2034[a]	NA	209 (73%)	0	107 (37%)	6.5 ± 3.5	NA	NA	286	0	286	[27]
GSE3494	67/128/54	213 (85%)	84 (33%)	NA	NA	62 ± 14	2.2 ± 1.3	251	(251)	251	[28]
GSE2990	64/48/55	147 (78%)	30 (16%)	67 (35%)	6.6 ± 3.9	56 ± 12	2.2 ± 1.1	189	87	102	[3]
GSE4922	68/126/54	211 (85%)	81 (33%)	89 (36%)	7.1 ± 4.3	62 ± 14	2.2 ± 1.3	249	247[b]	1	[5]
GSE6532	1/94/4	114 (88%)	55 (43%)	44 (34%)	5.6 ± 3.2	64 ± 10	2.6 ± 1.2	138	56	82	[10]
GSE7390	30/83/83	134 (68%)	NA	91 (46%)	9.3 ± 5.6	46 ± 7	2.2 ± 0.8	198	0	198	[29]
TOTAL	258/537/312	1,028 (79%)	250 (30%)	438 (36%)	7.0 ± 4.2	58 ± 14	2.3 ± 1.4	1,470	641		
TOTAL[c]	123/206/151	700 (77%)	121 (13%)	386 (36%)	9.0 ± 3.4	56 ± 13	2.3 ± 1.2		251	1,079	

NA not available

[a] Only MAS5 data; [b] one additional CEL file is from GSE2990; [c] excluding redundant samples, only these arrays were included in the analysis

sets. Eight of them were able to predict relapse at a P value <0.05 (Table 2). Using PAM we also constructed the training error plots to estimate the performance of top discriminatory genes defined in these studies. Notably, even gene sets exhibiting low overall predictive ability contained some top genes capable of discriminating patient tumors with or without relapse (data not shown).

Establishing a new predictor for relapse-free survival

We have computed a new predictor based on all genes associated with breast cancer and all available microarrays. In this setting only the gene expression data was used. In the supervised principal component analysis the threshold was set to 0.001 to fit Cox proportional hazard model. The best discriminatory gene signature contains 376 genes (Supplemental Table 4) and the result of a leave-one-out discrimination has a significance of 1.8E−12. The Kaplan–Meier survival plot is presented in Fig. 2.

We have calculated the predictive power of the available clinical variables: lymph node status ($P = 0.01$) and grade ($P = 0.0005$) were significant, while ER status ($P = 0.11$) was not predictive for relapse-free survival (Fig. 2). Therefore, the predictive calculation was extended to evaluate whether the expression data provides more accurate predictions than that provided by the two other significant clinical covariates, lymph-node status and grade. Forty-four genes performed over the adjusted clinical covariates. The prediction based on these genes and the clinical covariates was highly significant ($P = 9.5\text{E}{-7}$) and is presented in Fig. 2 and the gene list in Supplemental Table 3.

Establishing a database for future predictions

All available MAS 5.0 normalized microarray expression data were merged into a large training set to permit inclusion of new patient data as test samples and to allow classification of those in a straight forward manner. The complete data set of all samples with complete clinical information (and the necessary experiment descriptor file) is available as Supplemental Tables 5 and 6 in a BRB-ArrayTools compatible format for independent application.

Discussion

We have critically assessed the potential of microarray data for predicting relapse-free survival in breast cancer patients based on a large cohort of tumor samples ($n = 1{,}079$) previously collected in and documented by different clinical centers. The gene lists which were developed to predict prognosis generally outperform those

Table 2 List of gene sets included in the study

Summary of the study						Analysis results	
Number of patients	Platform	Number of markers in published set	Classification problem	Reference	Remark	Number of matched for HGU133A (=markers used for analysis)	Relapse *P* value: Kaplan–Meier survival
Main focus on prognosis							
117	Hu25K microarray	231 genbank accession numbers	Clinical outcome	[30]	Used Hclust and multivariate analysis. Validated in [31] and [14]	144 transcripts	0.39
99	cDNA microarray, 7,650 spots	485 genes	Survival gene list	[32]	Used parametric *t* tests	564 transcripts	0.47
100	cDNA array, 1,200 genes	38 genes	Progression-associated signature	[33]	Used Hclust and linear discriminant analysis	38 transcripts	0.59
279	Affymetrix Hu6800 and Hu35KsubA	17 genes	Signature of metastasis correlates with clinical outcome	[34]	Used and own algorithm and Hclust	39 transcripts	0.13
89	Affymetrix HGU95Av2	56 metagenes	Predict nodal metastatic states and relapse	[35]	Used Bayesian classification tree analysis	74 transcripts	**0.017**
24	Affymetrix HGU95Av2	92 genes	Predict docetaxel response	[36]	Used different statistical methods	128 transcripts	0.88
159	Affymetrix HGU133A	64 genes	Separate patients with good and bad prognosis	[26]	Used diagonal linear discriminant analysis	72 transcripts	**0.00081 (gse1456 excluded)**
251	Affymetrix HGU133A and B chips	32 transcripts	A p53 downstream signature predicting prognosis	[28]	Original list of 32 transcripts contained both HGU133A and B chips. Used diagonal linear discriminant analysis	18 transcripts	0.96
286	Affymetrix HGU133A	76 transcripts	Predict distant metastasis in lymph-node negative ca.	[27]	Only MAS5 data. Used complex analysis including different techniques. Validated in [29] and [37]	76 transcripts	**0.00001 (gse2034 excluded)**
180	As of [27]		Validation of predictor	[37]		–	–
198	As of [27]		Validation of predictor	[29]	Unavailable online software used for analysis	–	–
651	As of [9]		Prediction of chemotherapy benefit	[38]	Used Cox proportional hazards model	–	–
244	In silico	7 genes	Prognostic signature for ER− tumors	[39]	Used PACK	19 transcripts	**0.033**[a]
162	cDNA array with 10,368 spots	21 genes	Signature to predict breast cancer outcome	[40]	Used PAM, SAM, and correlation	30 transcripts	**0.018**
135	Agilent human 1A oligo microarray	70 genes	Prognostic signature	[17]	Used Cox-clustering	91 transcripts	**0.00071**

Table 2 continued

Summary of the study						Analysis results	
Number of patients	Platform	Number of markers in published set	Classification problem	Reference	Remark	Number of matched for HGU133A (=markers used for analysis)	Relapse *P* value: Kaplan–Meier survival
Main focus on estrogen receptor							
58	cDNA microarray, 6,728 genes	50 genes	List of genes which discriminate according to ER status	[8]	Used Artificial neural network and Hclust	75 transcripts	0.99
49	Affymetrix HuGeneFL	100 genes	Discrimination of estrogen receptor status	[12]	Used a Bayesian regression model	149 transcripts	0.88
668	RT-PCR	16 + 5 genes	Recurrence of tamoxifen-treated node-negative breast cancer	[9]	Used a multistep statistical approach	35 transcripts	0.44
65	Agilent whole-genome microarrays	822 genes	Estrogen regulated genes predict survival in ER+ patients	[11]	Used SAM	331 transcripts	**0.044**[b]
335	As of [3]		Clinical outcome in Tamoxifen-treated ER+ breast cancer	[10]	Study includes additional samples from [3]	–	–
Main focus on differentiating subtypes							
78	cDNA array, 8,102 genes	264 cDNA clones	To identify tumor subclasses	[2]	Used SAM and Hclust. Validated in [41]	329 transcripts	0.09
189	Affymetrix HGU133A	242 transcripts	Histological grade	[3]	Used complex analysis to calculate the Gene Grading Index. 128 transcripts in an unpublished list.	242 transcripts	0.64
105	Agilent human oligo arrays	306 genes	Tumor classification into subclasses	[4]	Used SAM and Hclust	256 transcripts	0.87
249	Affymetrix HGU133A and B	18 transcripts	Genetic grade signature to differentiate G1 and G3	[5]	Used PAM and SWS	13 transcripts	**0.00007**

The analysis results correspond to the *P* value of a Kaplan–Meier after classification of patients described in Table 1 with PAM using the matched set of genes published

The values in bold denote significance at $P < 0.05$

[a] $P = 0.08$ if ER+ samples excluded; [b] with batch adjustment for ER status, without adjustment $P = 0.08$

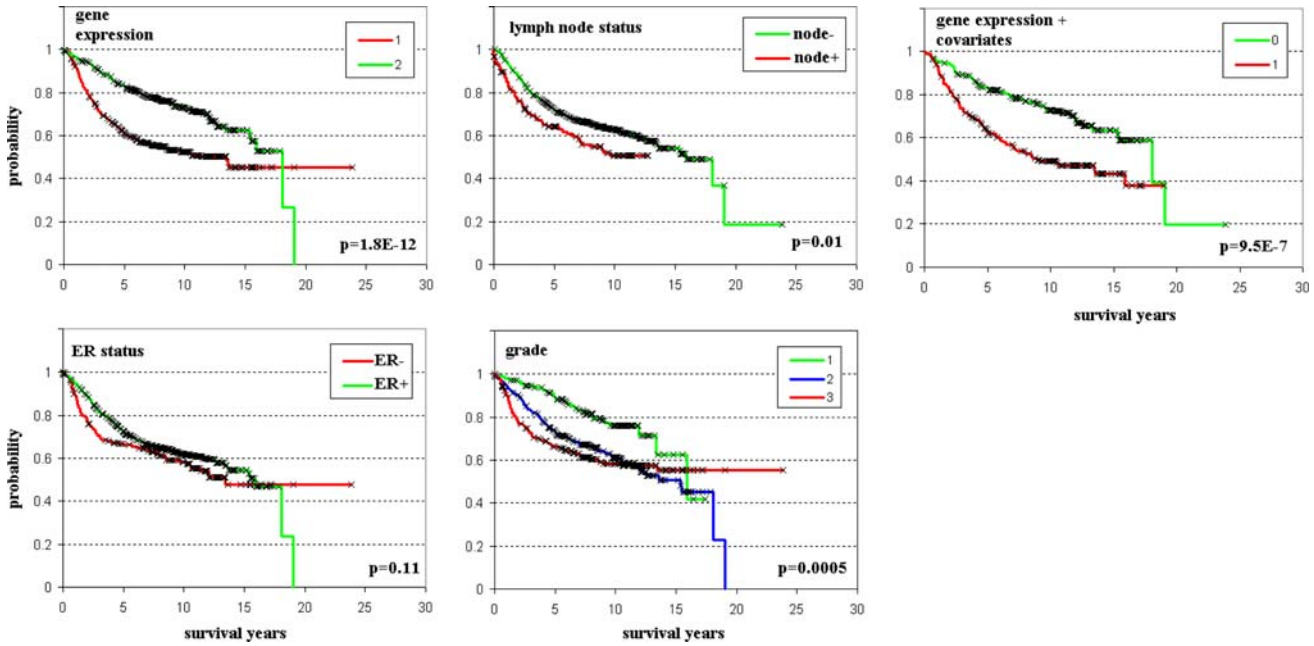

Fig. 2 Kaplan–Meier survival plot for all patients using the best 376 genes, lymph-node status, ER status, grade and the top 44 genes fitted by both gene expression and covariates (lymph node status and grade)

focusing on ER receptor or subtype. Our results provide a ranking of published predictors with regard to their applicability on independent data sets. The best discriminative pattern can be reached by including all available samples. When screening relevant publications and entries in public databases, we detected 389 microarray-based data sets that had been repeatedly entered in GEO with new accession numbers. Since the inclusion of repeated microarray data that were not derived independently could result in over-optimistic classification and lack of reproducibility in validation studies, we emphasize the necessity for curating public databases to eliminate such anomalies.

Another comparative analysis was based on 374 genes extracted from published gene lists relevant for breast cancer prognosis [20], but did not validated the original gene sets like our study. The authors included redundant samples from GSE2990 and GSE4990. Although their repeatedly validated gene-set predicted clinical response better than tumor size, lymph node status, ER status and grade, the presence of redundant data might influence their results.

Predictions of clinical outcome based on gene expression patterns were met with some skepticism, because multiple, non-overlapping gene sets were able to predict molecular phenotypes correctly. For example, a recent study on microarray data related to breast cancer, renal tumors and lymphoma and including clinical information compared the prediction errors using different training sets. The results suggested that expression profiles established in this way showed little overlap [21]. We achieved significant prediction success using different gene sets established on different microarray platforms, and therefore we provide additional support to this finding.

To increase the significance of the prediction, we have used all available samples to build the consensus predictor instead of splitting the microarray data into a training and test set. Ntzani et al. found by investigating 84 diverse microarray studies that significant associations were 3.5 times more likely when the sample size was doubled and 9.7 times more likely when the number of microarray probes were increased tenfold [22]. These authors also advocated the use of complete cross validation in order not to inflate the predictive power [22]. Accordingly, we included all samples in the initial training set and the Kaplan–Meier plot was based on the results of a leave-one-out cross validation (LOOCV). The LOOCV provides a nearly unbiased estimate of the true error rate of the classification procedure [23]. At the end of the LOOCV process, we have constructed different models for each sample only to estimate the prediction error. The model that is suggested for future predictions is the one constructed at the beginning using all 1,079 samples.

We used the supervised principal component analysis and the prediction analysis of microarrays for classification. More sophisticated algorithms do not perform better than the simple ones as shown by Dudoit and colleagues who have evaluated simple—diagonal linear discriminant analysis and nearest-neighbour classification—and complex—classification trees and machine-learning techniques such as bagging and boosting—classification methods [24].

The analysis was performed after mapping the gene sets to a single platform. This mapping relies on the proven fundamental assumption that different microarrays are capable to reproducibly measure gene expression [25].

In our study we demonstrate that different microarray datasets can be used to predict relapse in an independent dataset established using single channel microarrays. In this context, our study is a validation for the original studies using a much larger patient cohort. Finally, we established a database incorporating the genes from 20 microarray studies and gene expression data for 1,079 patients. This BRB Arraytools compatible database in an easily extendable format and can be used to validate future studies or to classify individual patients.

Acknowledgment This study was supported by a Bolyai fellowship to BG.

References

1. Surowiak P, Materna V, Gyorffy B et al (2006) Multivariate analysis of oestrogen receptor alpha, pS2, metallothionein and CD24 expression in invasive breast cancers. Br J Cancer 95:339–346. doi:10.1038/sj.bjc.6603254
2. Sorlie T, Perou CM, Tibshirani R et al (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA 98:10869–10874. doi:10.1073/pnas.191367098
3. Sotiriou C, Wirapati P, Loi S et al (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst 98:262–272
4. Hu Z, Fan C, Oh DS et al (2006) The molecular portraits of breast tumors are conserved across microarray platforms. BMC Genomics 7:96. doi:10.1186/1471-2164-7-96
5. Ivshina AV, George J, Senko O et al (2006) Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. Cancer Res 66:10292–10301. doi:10.1158/0008-5472.CAN-05-4414
6. Petersen OW, Hoyer PE, van DB (1987) Frequency and distribution of estrogen receptor-positive cells in normal, nonlactating human breast tissue. Cancer Res 47:5748–5751
7. Kuukasjarvi T, Kononen J, Helin H et al (1996) Loss of estrogen receptor in recurrent breast cancer is associated with poor response to endocrine therapy. J Clin Oncol 14:2584–2589
8. Gruvberger S, Ringner M, Chen Y et al (2001) Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. Cancer Res 61:5979–5984
9. Paik S, Shak S, Tang G et al (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 351:2817–2826. doi:10.1056/NEJMoa041588
10. Loi S, Haibe-Kains B, Desmedt C et al (2007) Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. J Clin Oncol 25:1239–1246. doi:10.1200/JCO.2006.07.1522
11. Oh DS, Troester MA, Usary J et al (2006) Estrogen-regulated genes predict survival in hormone receptor-positive breast cancers. J Clin Oncol 24:1656–1664. doi:10.1200/JCO.2005.03.2755
12. West M, Blanchette C, Dressman H et al (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci USA 98:11462–11467. doi:10.1073/pnas.201162998
13. Ransohoff DF (2004) Rules of evidence for cancer molecular-marker discovery and validation. Nat Rev Cancer 4:309–314. doi:10.1038/nrc1322
14. Buyse M, Loi S, van't Veer L et al (2006) Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Natl Cancer Inst 98:1183–1192
15. Michiels S, Koscielny S, Hill C (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet 365:488–492. doi:10.1016/S0140-6736(05)17866-0
16. Ioannidis JP (2005) Microarrays and molecular research: noise discovery? Lancet 365:454–455
17. Naderi A, Teschendorff AE, Barbosa-Morais NL et al (2007) A gene-expression signature to predict survival in breast cancer across independent data sets. Oncogene 26:1507–1516. doi:10.1038/sj.onc.1209920
18. Tibshirani R, Hastie T, Narasimhan B et al (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci USA 99:6567–6572. doi:10.1073/pnas.082099299
19. Bair E, Tibshirani R (2004) Semi-supervised methods to predict patient survival from gene expression data. PLoS Biol 2:E108. doi:10.1371/journal.pbio.0020108
20. Lauss M, Kriegner A, Vierlinger K et al (2008) Consensus genes of the literature to predict breast cancer recurrence. Breast Cancer Res Treat 110:235–244. doi:10.1007/s10549-007-9716-3
21. Gormley M, Dampier W, Ertel A et al (2007) Prediction potential of candidate biomarker sets identified and validated on gene expression data from multiple datasets. BMC Bioinformatics 8:415. doi:10.1186/1471-2105-8-415
22. Ntzani EE, Ioannidis JP (2003) Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. Lancet 362:1439–1444. doi:10.1016/S0140-6736(03)14686-7
23. Simon R, Radmacher MD, Dobbin K et al (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 95:14–18
24. Dudoit S, Fridlyand J, Speed T (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc 97:77–87. doi:10.1198/016214502753479248
25. Shi L, Reid LH, Jones WD et al (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol 24:1151–1161. doi:10.1038/nbt1239
26. Pawitan Y, Bjohle J, Amler L et al (2005) Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. Breast Cancer Res 7:R953–R964. doi:10.1186/bcr1325
27. Wang YX, Klijn JGM, Zhang Y et al (2005) Gene-expression pro-files to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 365:671–679
28. Miller LD, Smeds J, George J et al (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proc Natl Acad Sci USA 102:13550–13555. doi:10.1073/pnas.0506230102
29. Desmedt C, Piette F, Loi S et al (2007) Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. Clin Cancer Res 13:3207–3214. doi:10.1158/1078-0432.CCR-06-2765
30. van 't Veer LJ, Dai H, van de Vijver MJ et al (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415:530–536. doi:10.1038/415530a

31. van de Vijver MJ, He YD, van 't Veer LJ et al (2002) A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347:1999–2009. doi:10.1056/NEJMoa021967
32. Sotiriou C, Neo SY, McShane LM et al (2003) Breast cancer classification and prognosis based on gene expression profiles from a population-based study. Proc Natl Acad Sci USA 100:10393–10398. doi:10.1073/pnas.1732912100
33. Ma XJ, Salunga R, Tuggle JT et al (2003) Gene expression profiles of human breast cancer progression. Proc Natl Acad Sci USA 100:5974–5979. doi:10.1073/pnas.0931261100
34. Ramaswamy S, Ross KN, Lander ES et al (2003) A molecular signature of metastasis in primary solid tumors. Nat Genet 33:49–54. doi:10.1038/ng1060
35. Huang E, Cheng SH, Dressman H et al (2003) Gene expression predictors of breast cancer outcomes. Lancet 361:1590–1596. doi:10.1016/S0140-6736(03)13308-9
36. Chang JC, Wooten EC, Tsimelzon A et al (2003) Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. Lancet 362:362–369. doi:10.1016/S0140-6736(03)14023-8
37. Foekens JA, Atkins D, Zhang Y et al (2006) Multicenter validation of a gene expression-based prognostic signature in lymph node-negative primary breast cancer. J Clin Oncol 24:1665–1671. doi:10.1200/JCO.2005.03.9115
38. Paik S, Tang G, Shak S et al (2006) Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. J Clin Oncol 24:3726–3734. doi:10.1200/JCO.2005.04.7985
39. Teschendorff AE, Miremadi A, Pinder SE et al (2007) An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. Genome Biol 8:R157. doi:10.1186/gb-2007-8-8-r157
40. Korkola JE, Blaveri E, DeVries S et al (2007) Identification of a robust gene signature that predicts breast cancer outcome in independent data sets. BMC Cancer 7:61. doi:10.1186/1471-2407-7-61
41. Sorlie T, Tibshirani R, Parker J et al (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci USA 100:8418–8423. doi:10.1073/pnas.0932692100