Comparison of the Predictive Accuracy of DNA Array-Based Multigene Classifiers across cDNA Arrays and Affymetrix GeneChips

James Stec,^{*,†} Jing Wang,[‡] Kevin Coombes,[‡]
Mark Ayers,^{*} Sebastian Hoersch,^{*} David L. Gold,[‡]
Jeffrey S. Ross,^{*} Kenneth R. Hess,[‡]
Stephen Tirrell,^{*} Gerald Linette,^{*}
Gabriel N. Hortobagyi,[§] W. Fraser Symmans,[¶] and
Lajos Pusztai[§]

From the Departments of Biostatistics,[‡] Breast Medical Oncology,[§]
and Pathology,[¶] University of Texas M.D. Anderson Cancer
Center, Houston, Texas; Millennium Pharmaceuticals, Inc.,^{*}
Cambridge, Massachusetts; and Praecis Pharmaceuticals,[†]
Waltham, Massachusetts

We examined how well differentially expressed genes and multigene outcome classifiers retain their class-discriminating values when tested on data generated by different transcriptional profiling platforms. RNA from 33 stage I-III breast cancers was hybridized to both Affymetrix GeneChip and Millennium Pharmaceuticals cDNA arrays. Only 30% of all corresponding gene expression measurements on the two platforms had Pearson correlation coefficient $r \geq 0.7$ when UniGene was used to match probes. There was substantial variation in correlation between different Affymetrix probe sets matched to the same cDNA probe. When cDNA and Affymetrix probes were matched by basic local alignment tool (BLAST) sequence identity, the correlation increased substantially. We identified 182 genes in the Affymetrix and 45 in the cDNA data (including 17 common genes) that accurately separated 91% of cases in supervised hierarchical clustering in each data set. Cross-platform testing of these informative genes resulted in lower clustering accuracy of 45 and 79%, respectively. Several sets of accurate five-gene classifiers were developed on each platform using linear discriminant analysis. The best 100 classifiers showed average misclassification error rate of 2% on the original data that rose to 19.5% when tested on data from the other platform. Random five-gene classifiers showed misclassification error rate of 33%. We conclude that multigene predictors optimized for one platform lose accuracy when applied to data from another platform due to missing genes and sequence differences in probes that result in differing measurements for the same gene. (*J Mol Diagn* 2005, 7:357-367)

Transcriptional profiling is increasingly explored as a potential diagnostic tool in an attempt to classify cancer into clinically relevant subgroups or predict disease outcome.¹ In the gene expression profiling literature, prediction of clinical outcome is usually made based on the constellation of a relatively large numbers of genes. Comparison of prediction results across several laboratories is critically important to determine the practical value and technical limitations of a proposed new test. Cross-laboratory validation of gene expression-based predictors is challenging. Different laboratories frequently use different microarray platforms that usually contain different sets of genes. Different platforms may also measure the expression of the same gene with different precision, on a different scale and with different dynamic range.²⁻⁴ When a gene signature-based predictor discovered in one laboratory is tested on cases at a different laboratory, poorer predictive performance may result from at least two major factors. The first is biological; a predictor developed from a particular set of clinical specimens may not be generalizable to a different set of cases. This is more likely to occur if the sample size used for marker discovery is too small. The second factor is technical; predictors generated from data obtained with one platform may become compromised when applied to data generated by another platform. This could be due to missing informative genes or to measurement differences. Different gene expression profiling platforms measure the expression of the same gene with different relative accuracy on a different numerical scale.

The goal of the current research was to compare gene expression data generated by Affymetrix UB3A (Santa Clara, CA) oligonucleotide arrays with those generated by cDNA microarrays from the same 33 RNA samples

Supported in part by the Nellie B. Connally Breast Cancer Research Fund, by grants from Millennium Pharmaceuticals, The Dee Simmons Fund, and the University of Texas M.D. Anderson Cancer Center Aventis Drug Development Award and R01 CA106290-01 (to L.P.), and by grant LF2002-044HM from The Susan G. Komen Breast Cancer Foundation (to W.F.S.).

J.S. and J.W. contributed equally to the development of this manuscript.

Accepted for publication February 7, 2005.

Address reprint requests to Lajos Pusztai, Department of Breast Medical Oncology, University of Texas M.D. Anderson Cancer Center, Unit 424, 1515 Holcombe Boulevard, Houston TX 77030-4009. E-mail: pusztai@mdanderson.org.

extracted from fine-needle aspirates of breast cancer. We examined three questions. 1) How closely do normalized gene expression measurements for the same gene correlate across platforms? 2) Do genes that discriminate between responders and nonresponders to chemotherapy in hierarchical cluster analysis on one platform retain discriminating value when tested on data generated by another platform? 3) How well do multigene predictors generated on one platform hold up on data generated by the other platform? Because the same RNA was profiled on both platforms, this design allowed us to evaluate the impact of the profiling method without the confounding effect of sampling variation. This analysis was not designed to compare the accuracy of the two platforms. This would have required a third, external, gold standard measurement of mRNA expression. Similarly, our goal was not to develop the best possible predictor of pathological CR from this data. This would have required a somewhat different statistical approach to optimize the predictor for validation on independent samples.

Materials and Methods

Patients and Samples

Fine-needle aspiration (FNA) specimens from 33 patients with newly diagnosed stage I-III breast cancer were included in the cross-platform analysis. These cases were selected from an ongoing clinical trial at the University of Texas M.D. Anderson Cancer Center, and patient characteristics are presented in Table 1. The study was approved by the Institutional Review Board (IRB) of the University of Texas M.D. Anderson Cancer Center, and all patients signed an informed consent. The goal of the larger trial is to develop a multigene predictor of pathological complete response to preoperative paclitaxel and 5-fluorouracil, doxorubicin, cyclophosphamide chemotherapy. Early results of this response marker discovery project were reported separately.⁵ The cases included in the current analysis were selected because they had gene expression profiling performed on two different platforms. This was done as part of a transition from limited-supply, proprietary, cDNA nylon membrane arrays (Millennium Pharmaceuticals, Inc., Cambridge, MA) to the commercially available Affymetrix oligonucleotide DNA-chip platform. Twenty-seven additional specimens that were collected consecutively were included in a small independent validation of the performance of the classifier that performed well on both platforms.

All FNAs were obtained before any systemic therapy and were collected into vials containing RNAlater solution (Ambion, Austin, TX) and stored at -80°C until profiling. The RNA yield and cellular composition of these samples have previously been reported.⁶ Estrogen receptor (ER) and HER-2 receptor status of each cancer was assessed on a separate diagnostic core biopsy as part of routine care. ER was assessed by immunohistochemistry using 1D5 antibody from Zymed (San Francisco, CA).

Table 1. Clinical Information and Demographics of Patients Included in the Cross-Platform Comparison

Female	33 (100%)
Median age	50 years (range, 29 to 75)
Race	
Caucasian	22 (67%)
Asian	3 (9%)
Hispanic	4 (12%)
African American	4 (12%)
Histology	
Invasive ductal	28 (85%)
Mixed ductal/lobular	3 (9%)
Invasive lobular	1 (3%)
Invasive mucinous	1 (3%)
TNM stage	
T1	3 (9%)
T2	19 (59%)
T3	4 (12%)
T4	6 (18%)
N0	17 (52%)
N1	12 (36%)
N2	4 (12%)
Black's modified nuclear grade	
1	2 (6%)
2	12 (36%)
3	19 (58%)
ER positive*	19 (58%)
ER negative	14 (42%)
HER-2 positive†	9 (27%)
HER-2 negative	24 (73%)
Neoadjuvant therapy‡	
Weekly T (80 mg/m ²) × 12 + FAC × 4	23 (70%)
3-weekly T (225 mg/m Cl) × 4 + FAC × 4	10 (30%)
pCR§	10 (30%)
RD	23 (05%)

*ER, estrogen receptor and cases where $\geq 10\%$ of tumor cells stained positive for ER with IHC were considered positive.

†HER-2 status was determined by IHC or fluorescent *in situ* hybridization. Cases that showed either 3+ IHC staining or had gene copy number >2.0 were considered HER-2 positive.

‡T, paclitaxel; FAC, 5-fluorouracil, doxorubicin, and cyclophosphamide.

§Pathological complete response was defined as complete disappearance of all invasive cancer from the breast and lymph nodes at the time of surgery.

Microarray Hybridization

RNA was extracted from FNA samples using the RNA-easy kit (Qiagen, Valencia, CA). The amount and quality of RNA was assessed with a DU-640 U.V. Spectrophotometer (Beckman Coulter, Fullerton, CA) and by an Agilent 2100 Bioanalyzer RNA 6000 LabChip kit (Agilent Technologies, Palo Alto, CA). Two microarray platforms were tested in this study: the Affymetrix Human Genome U133 A and B gene chip sets and cDNA nylon membranes proprietary to and printed by Millennium Pharmaceuticals. The cDNA arrays contained 21,594 independent human sequence-verified clones. For the cDNA array hybridization, first-strand cDNA synthesis was performed with Superscript II (Invitrogen, Carlsbad, CA) in the presence of [³²P]dCTP (100 mCi/ml; Amersham, Little Chalfont, UK) from 1 to 2 μg of total RNA. The isotope-labeled cDNA probes were hybridized without further amplification to high-density nylon arrays. Affymetrix profiling was also performed without second round amplification following the standard protocol using 1 μg of total

RNA from the same pool used for the cDNA array experiments. Briefly, double-stranded cDNA was synthesized, followed by *in vivo* transcription reaction to generate biotinylated cRNA. Biotin-labeled and fragmented cRNA was hybridized to the Affymetrix U133A and B gene chips overnight at 42°C. Procedures followed standard operating practice outlined in the Affymetrix technical manual. The Affymetrix GeneChip system was used for hybridization and scanning of the probe arrays, and Microarray Analysis Suite (MAS) 5.0 was used for data acquisition.

Data Processing

Results of the cDNA membrane array experiments and data acquisition were previously reported.⁷ Gene expression values of the cDNA array experiments were normalized to the median expression value of all genes on each membrane array. The median values were set to 1 by the normalization process; the postnormalization mean expression values ranged from 2.535 to 4.787. For the Affymetrix data, several standard metrics were examined to assess the quality of each hybridization result. We assessed T7 amplification and labeling efficiency by calculating the 3' to 5' ratios of β -actin and glyceraldehyde-3-phosphate dehydrogenase. The median ratio for β actin was 1.45 (range, 1.21 to 1.60); for glyceraldehyde-3-phosphate dehydrogenase, 1.13 (range, 0.86 to 1.67). To assess brightness, we used dCHIP V1.3 to calculate the percentages of array outliers and of single outliers for each chip.⁸ MAS 5.0 was used to produce *P* values for signal detection. Chips with more than 5% array or single outliers or with less than 15% detection *P* values <0.01 were flagged. All 33 Affymetrix profiles included in this analysis have passed each QC step. Affymetrix data were quantified and normalized with dCHIP V1.3 (<http://biosun1.harvard.edu/complab/dchip>).⁸

Statistical Analysis

The primary goal of this research was cross-platform testing of informative genes and multigene outcome predictors in an experimental setting in which the only variable is the gene expression profiling technique. Normalized expression data from both platforms was transformed by computing the base-two logarithm before further analysis. Pearson correlation coefficients were calculated on the normalized, log-transformed data for each gene represented on both platforms, with matching based on UniGene build 160. Further analysis was performed only on the subset of genes shared on both the Affymetrix and cDNA platforms. When multiple cDNA probes or Affymetrix probe sets targeted the same gene, separate Pearson and Spearman correlation coefficients were calculated for each distinct probe set and matching cDNA probe. The distribution of all gene expression measurements on both platforms was near normal. To identify differentially expressed genes between the two response outcome groups, pathological complete response versus residual cancer, we explored two methods: the two-sample *t*-test and an unequal variance *t*-test on the ranks. The

resulting *P* values were analyzed as a β -uniform mixture, and the relationship between *P* values and false discovery rate (FDR) was assessed.^{9,10} Hierarchical clustering was performed using the informative genes identified in the univariate analysis. The distance metric was computed on log-transformed gene expression values. Data analyses were performed using the statistical software package S-Plus 2000 (Insightful Corp., Seattle, WA).

Multigene classifiers were built by combining a genetic algorithm (GA) with linear discriminant analysis (LDA).^{11,12} This classification strategy consists of two main components. The GA is used as a gene selector to identify the best combination of discriminating genes, and the class prediction is performed by LDA. The working and optimization principle of GA is analogous to an evolutionary process and does not rely on rank-based gene selection as a *t*-test does.¹¹ This method has previously been applied successfully to high-dimensional gene expression and proteomic profile data.^{13–16} A potential advantage of this approach is that the GA can identify combination of genes that are predictive together even if individual genes that contribute to the classifier have limited predictive value and would be missed by two-sample *t* statistics. However, to reduce computation time, we performed GA/LDA on a subset of the data that was enriched in informative genes. The subsets in each platform were selected based on the FDR computed from the two-sample *t*-tests. By setting FDR < 0.6, we selected 1032 clones on the cDNA platform and 2010 probe sets on the Affymetrix platform. This preselection of genes was performed to reduce computation time by using genes with somewhat increased class-discriminating value. Twenty-seven cases were used as an independent validation set to test the predictive accuracy of the predictor that had the highest cross-validation accuracy on both platforms.

Results

Identification of Shared Genes Present on Both the Affymetrix U133 A/B Chips and the cDNA Membrane Arrays

We used GenBank accession numbers to identify matching genes on both platforms (UniGene build 160, Spring 2003). The U133A chip contained 22,215 probe sets representing 13,736 unique UniGene clusters, and the U133B chip contained 22,577 probe sets representing 17,132 unique UniGene clusters. The cDNA nylon membranes contained 30,720 clones corresponding to 21,594 unique UniGene clusters. We found 9402 unique UniGene clusters that overlapped between the U133A chip and the nylon cDNA array (corresponding to 14,897 probe sets on the U133A chip and to 14,104 clones on the cDNA array). We also found 8430 clusters that overlapped between the U133B chip and the cDNA arrays (corresponding to 11,472 probe sets on the U133B chip and to 12,079 clones on the cDNA array). Figure 1 illustrates these results.

Overlapped Unique UniGene

Figure 1. Venn diagram of the genes common to both microarray platforms. There are 9402 unique UniGene clusters common to Affymetrix GeneChip U133A chip and the Millennium cDNA array, and there are 8430 unique UniGene clusters common to the U133B and cDNA array.

Correlation of Gene Expression Measurements between Affymetrix and cDNA Data

We calculated the Pearson correlation coefficients of corresponding gene expression measurements for the genes represented on both platforms using normalized log-transformed data. Of the 24,100 individual matching probe set measurements, 3232 had a Pearson correlation coefficient $r \geq 0.7$, another 7590 measurements had $r \geq$

0.5, and 12,466 measurements had an $r \geq 0$ but ≤ 0.5 . There were 4044 measurements that had negative correlation coefficient. The distribution of all Pearson coefficients indicated a modest overall correlation of individual gene expression measurements across the two platforms (Figure 2A). Similar results were seen when Spearman correlation coefficient was used (data not shown). Concordance correlation of t scores for the matching probes on the two platforms was also modest, $r = 0.458$. Closer examination of the two data sets revealed that the U133A often contains multiple probe sets corresponding to a particular gene on the cDNA platform. Of the 9402 shared genes on U133A chip and on the cDNA array, only 4309 had a 1:1 ratio of probe sets to cDNA clones. When multiple probe sets with distinct sequences targeted the same gene, substantial differences in correlation with the corresponding cDNA array measurement were seen. Some probe sets displayed very high correlation, whereas others showed minimal or no correlation with the cDNA result. To further address this problem, we performed BLAST sequence match for each paired cDNA clone and Affymetrix probe set. cDNA clone sequences and the Affymetrix 25-mer probes were aligned against the National Center for Biotechnology Information Reference Sequences (RefSeq) of each target gene. If both the cDNA clone and the Affymetrix probe had ≥ 25 nucleotides overlapping with the RefSeq sequences, then the pair was considered "RefSeq-matched" using software *bl2seq*. Three categories of probe pairs were created: 1) probes pairs that targeted the same gene but were misaligned and shared no overlap in RefSeq target sequence ($n = 2108$), 2) pairs that matched to the same RefSeq sequence as described above ($n = 3169$), and 3) pairs where cDNA and Affymetrix probe sequences directly overlapped ($n = 4008$). As expected, closer se-

Figure 2. Distribution of Pearson correlation coefficients of corresponding gene expression measurements on Affymetrix and cDNA platforms. The bell curve represents the expected distribution of coefficients if the null hypothesis of no correlation between individual measurements is true. The actual distribution is shown by the bar graphs along with a curve fitted to the data. **A:** All corresponding measurements based on UniGene ID ($n = 9402$) included. **B:** Only probe pairs with direct sequence overlap ($n = 4008$) are included.

Figure 3. Human ER α gene structure and mapping of 4 Affymetrix oligonucleotide probe sets to the sequence of the ER gene. The reverse transcription reaction starts at the 3' end of the cDNA, and the efficacy of the reaction drops off as it moves toward the 5' end, therefore, probes close to the 3' end give stronger signal and show the best correlation with the overlapping cDNA probe.

quence matches between the cDNA and Affymetrix probes yielded greater correlation. Figure 2B illustrates the distribution of Pearson correlation coefficient for the Affymetrix and cDNA probe pairs that had direct sequence overlap.

These results suggest that DNA microarray-based gene expression measurements are in fact robust and reproducible across platforms if the probe sets have overlapping sequences. However, this includes only a relatively small percentage of all probes present on these two particular array platforms, 29% (Affymetrix U133A) and 19% (Millennium Pharmaceuticals cDNA arrays), respectively.

Sequence-Based Explanation for the Discrepant Expression Measurements on the Two Platforms

To illustrate the impact of sequence matching and location of Affymetrix probe target sequence in relation to the 3' end of the transcript, we examined the expression results for the ER obtained with the two platforms. The expression of estrogen ER is routinely measured in breast cancer specimens by immunohistochemistry. ER protein expression correlates closely with mRNA expression of ER.^{7,17} These routine clinical results, which were available for all cases, provided an external "gold standard" reference for ER expression. We observed very good correlation between ER expression determined by immunohistochemistry and by cDNA microarray (estrogen receptor 1 α , IMAGE clone 725321, Hs.1657; NM_000125) or by Affymetrix gene chip when the probe set "205225_at" was used for comparison. The expression values of ER measured by cDNA arrays and by the Affymetrix probe set "205225_at" also correlated with each other very highly ($r = 0.976$). However, there are multiple probe sets that target the human ER α gene on the Affymetrix U133A chip. The correlation coefficients of these probe sets with the ER cDNA result (and with the gold standard immunohistochemistry (IHC) result) varied considerably, ranging from 0.146 to 0.976 (Figure 3). The normalized mean expression intensities of these ER

probe sets were also highly variable, ranging from 62 to 1291 (arbitrary units).

We examined the sequence and structure of the human ER α gene in relation to the IMAGE clone ER sequence represented on the cDNA array and the various Affymetrix oligonucleotide ER probe sequences. The closer the region of an Affymetrix probe set was to the region of the IMAGE clone, the higher the correlation was between the two measurements (Figure 3). Probe set "205225_at" had direct sequence overlap with the cDNA probe sequence and showed the highest correlation and greatest signal intensity. For probe sets "211233_x_at" and "215552_s_at", the correlation with the cDNA measurement was still strong, but the signal intensities were less than one-tenth of the signal produced by probe set "205225_at". This may be explained by the distance of these sequences from the 3' end of the ER transcript. The target sequence for "205225_at" is located at the 3' end of the transcript in a long untranslated region; the other two probes are located closer to the 5' end, and there are 3500 nucleotides between probe set 205225_at and the other two probes with lesser signal intensity. Considering that the reverse transcription reaction starts at the 3' end, a bias favoring the 3' end of the cDNA among the labeled probes is expected. However, additional mechanisms cannot be excluded. Alternative poly-A addition is a possibility, especially for very long untranslated regions. Also, the presence of ER transcript variants in human breast cancers could contribute to lower intensity values for these probes. GenBank sequences S80316 or AF258449/AF258450 are possible ER variants that lack exons corresponding to at least parts of probe sets "215552_s_at" and "211233_x_at".

The lack of correlation of probe set "215551_at" with ER gene expression may also be explained by the peculiarities of that sequence. This sequence maps to an exon that is only reported in one ER-related expressed sequence tag sequence. This raises the possibility of an alternative 3' end that gives rise to a shorter ER transcript variant. The lack of correlation with the cDNA result and with the clinical ER status coupled with the low signal intensity for this particular probe suggest a low-abundance ER transcript variant that could be expressed at low levels uniformly in all samples. These observations suggest that the relatively poor overall correlation between the cDNA array and Affymetrix results is largely due to sequence differences between the probes present in the two platforms.

Do Informative Genes Identified on One Platform Retain Their Discriminating Function on Another Platform in Hierarchical Cluster Analysis?

We identified genes that were differentially expressed between the 10 cases with pathological complete response (pCR) and the remaining 23 cases that had residual cancer (RD) after completion of preoperative chemotherapy (Supplementary Table S1, A and B at

A. cDNA data n=45 informative genes

B. Affymetrix U 133 data the same 45 informative genes (62 probe sets) were used for clustering

Figure 4. Supervised hierarchical clustering of cDNA and Affymetrix gene expression data using 45 informative genes identified on the cDNA platform. Cases with pathological complete response are underlined.

jmd.amjpathol.org/). In the cDNA array data, we used unequal variance *t*-test on the ranks to identify 45 genes as differentially expressed each with $P \leq 0.00236$ that corresponded to $FDR \leq 0.40$. This high false discovery rate suggested that the most informative genes on the cDNA platform were removed when analysis was restricted to probes that had a matched pair on the U133A chip. However, high FDR rate was acceptable for our purpose because we wanted to identify a large number of discriminating genes in this particular data and to test how well their discriminating value holds up when tested on expression data generated from the same RNA with a different profiling platform. One approach to test the discriminating value of these genes is to perform supervised hierarchical clustering with these informative genes. Average linkage supervised hierarchical clustering of the cDNA array data using these 45 genes visually demonstrated that these genes had high discriminating value on

the original data set, with 10 pCR/3 RD in one main cluster and 0 pCR/20 RD in the other that represents a 91% overall clustering accuracy (Figure 4). When the same 45 genes, corresponding to 62 Affymetrix probe sets, were used to cluster the Affymetrix data, the genes lost a substantial amount of discriminating value. Six cases of pCR and three RD clustered together in one main arm of the dendrogram and 4 pCR/20 RD were in the other main cluster, representing a 79% overall clustering accuracy. This decrease in discriminating value was predicted based on the modest overall correlation of individual gene expression values on the two platforms.

Applying the same selection criterion ($FDR \leq 0.40$) to the Affymetrix data yielded 182 probe sets corresponding to 166 distinct genes each with $P \leq 0.00607$. When these genes were used for supervised hierarchical clustering, they also performed well on the original data. Eight cases of pCR and 1 case of RD were clustered in one

Table 2. 17 Genes Associated with Pathological CR on Both Platforms

Probe sets	Correlations	Symbol	Description
206401_s_at	0.94	MAPT	Microtubule-associated protein tau
203929_s_at	0.95	MAPT	Microtubule-associated protein tau
203928_x_at	0.94	MAPT	Microtubule-associated protein tau
203930_s_at	0.90	MAPT	Microtubule-associated protein tau
203930_s_at	0.84	MAPT	Microtubule-associated protein tau
206401_s_at	0.89	MAPT	Microtubule-associated protein tau
203929_s_at	0.89	MAPT	Microtubule-associated protein tau
203928_x_at	0.88	MAPT	Microtubule-associated protein tau
218726_at	0.81	DKFZp762E1312	Hypothetical protein DKFZp762E1312
218726_at	0.82	DKFZp762E1312	Hypothetical protein DKFZp762E1312
200934_at	0.92	DEK	DEK oncogene (DNA binding)
200934_at	0.47	DEK	DEK oncogene (DNA binding)
200934_at	0.79	DEK	DEK oncogene (DNA binding)
212841_s_at	0.80	PPFIBP2	PTPRF interacting protein, binding protein 2
212844_at	0.82	KIAA0179	KIAA0179 protein
212846_at	0.79	KIAA0179	KIAA0179 protein
206550_s_at	0.41	NUP155	Nucleoporin 155 kd
206550_s_at	0.75	NUP155	Nucleoporin 155 kd
201798_s_at	0.86	FER1L3	fer-1-like 3, myoferlin (<i>Caenorhabditis elegans</i>)
211864_s_at	0.84	FER1L3	fer-1-like 3, myoferlin (<i>C. elegans</i>)
217895_at	0.70	FLJ20758	Hypothetical protein FLJ20758
217895_at	0.59	FLJ20758	Hypothetical protein FLJ20758
50314_i_at	0.21	C20orf27	Chromosome 20 open reading frame 27
218081_at	0.67	C20orf27	Chromosome 20 open reading frame 27
212689_s_at	0.57	JMJD1	Jumonji domain containing 1
212689_s_at	0.76	JMJD1	Jumonji domain containing 1
212689_s_at	0.75	JMJD1	Jumonji domain containing 1
218034_at	0.77	CGI-135	CGI-135 protein
206682_at	0.35	HML2	Macrophage lectin 2 (calcium dependent)
212494_at	0.80	TENC1	Tensin-like C1 domain-containing phosphatase
201976_s_at	-0.16	MYO10	Myosin X
201976_s_at	0.94	MYO10	Myosin X
201976_s_at	0.94	MYO10	Myosin X
202529_at	0.87	PRPSAP1	Phosphoribosyl pyrophosphate synthetase-associated protein 1
202392_s_at	0.74	PISD	Phosphatidylserine decarboxylase
202392_s_at	0.84	PISD	Phosphatidylserine decarboxylase
202392_s_at	-0.17	PISD	Phosphatidylserine decarboxylase
202361_at	0.69	SEC24C	SEC24- related gene family, member C (<i>Saccharomyces cerevisiae</i>)

main cluster and 2 pCR and 22 RD formed the other main cluster (91% clustering accuracy). However, when the same 166 genes (corresponding to 249 cDNA clones) were used to cluster the cDNA data, a practically complete loss of discriminating value was seen (data not shown).

These results indicate that informative genes identified on one gene expression profiling platform lose some of their class-discriminating value when measured with a different profiling method. This is further illustrated by the observation that only 17 genes were common to the cDNA and Affymetrix differentially expressed gene lists generated from the same cases (Table 2). As expected, the correlation coefficients for these 17 genes across the platforms were high. When these 17 genes common to both lists were used for hierarchical clustering, almost all cases of pathological CR clustered together in both types of expression data. It was, therefore, possible to identify genes and probe sets that retained discriminating value across platforms by using only genes that show high correlation coefficients when measured by the two distinct platforms. Such genes may be identified by sequence matching of the probes. Hierarchical clustering is not a class prediction tool, and assignment of molecular

class or expected outcome to new cases based on dendrogram results is not appropriate. There are several mathematical methods that are better suited to formulate outcome predictions based on the constellation of multiple genes.

Multigene Classifiers of Response to Therapy Generated on One Platform Show Diminished Classification Accuracy When Applied to Data Obtained with Another Platform

We used a genetic algorithm (for gene selection) in combination with linear discriminant analysis (for class prediction) to develop a large number of multigene classifiers that could separate cases with pathological CR from those with residual disease within each data set. As a preliminary step, we used a variant of the "greedy" algorithm method to determine how many genes should be permitted in the classifiers. To begin, we selected the gene with the largest (absolute) *t* score and next computed the Mahalanobis distance between the two groups using every pair of genes. The pair with the largest Ma-

Figure 5. Classification performance of 100 sets of five-gene GA/LDA classifiers identified from the cDNA expression data. The *x* axis represents the number of misclassified cases, and the *y* axis is the number of predictor sets with a given classification performance. **A:** The performance of five-gene predictors identified from the cDNA data; **B:** the same predictors applied to the Affymetrix data. **C** and **D:** five-gene random classifiers applied to both data sets, respectively.

halanobis distance was then included in the LDA classifier, and the classification accuracy was calculated. If the classification accuracy was improved, then this gene was added to the classifier. We repeated this process, adding additional genes until the classification accuracy stopped improving. On both platforms, the algorithm terminated at five genes. Five-gene LDA classifiers could perfectly separate pathological CR from RD in each data set. We ran the GA on data from each platform to identify 100 different sets of five genes that in combination with the LDA class prediction algorithm showed the best predictive accuracy. This analysis yielded 100 different GA/LDA classifiers that could be tested on the original data and on the data from the other platform.

The 100 sets of five-gene classifiers built from the cDNA data misclassified on average 0.69 cases, corresponding to 2% misclassification error rate (MER), on the original platform. Forty-eight of the 100 sets yielded 0 misclassifications. Thirty-seven sets yielded 1 and 13 sets yielded 2 misclassifications (Figure 5). When these 100 predictors developed from the cDNA data were tested on the Affymetrix results, the classification accuracy dropped. The average number of misclassified cases was 6.42 corresponding to 19.5% average MER. No set produced perfect classification when tested on the Affymetrix results, and only four sets yielded ≤ 2 misclassified cases (Figure 5). However, the cross-platform classification performance of these predictors (optimized for the cDNA results) was still better than that observed with random predictors. We tested the performance of LDA with 100 randomly selected five-gene sets

from the cDNA data. The average number of misclassified cases was 10.88 (33% average MER) on the cDNA platform and 9.62 (29% average MER) on the Affymetrix data.

We repeated the same analysis using the Affymetrix results for developing five-gene predictors. The performance of these predictors showed similar characteristics to what was observed with the cDNA-based predictors. The average number of misclassified cases was 1.2 (3.6% average MER) on the Affymetrix platform and 5.13 (15.5% average MER) on the cDNA data. The 100 sets of random Affymetrix predictors yielded an average of 11.43 (34.6% average MER) and 9.65 (29.2% average MER) misclassified cases on each platform, respectively.

When we compared the two sets of 100 five-gene classifiers identified from the cDNA and Affymetrix data, respectively, we could identify only three classifiers that were common and therefore performed well on both platforms. The classifier that performed the best on both platforms included the following five genes: PPFIBP2 (Hs.12953), PCNT1 (Hs.184352), HNRAPA2B1 (Hs.232400), BBS4 (Hs.26471), and SEC24C (Hs.81964). This predictor classified all cases correctly on the Affymetrix platform and misclassified only one case on the cDNA platform. The second best cross-platform classifier included genes RNF111 (Hs.12504), PPFIBP2 (Hs.12953), SCD4 (Hs.247474), SNRPN (Hs.48375), and PRPSAP1 (Hs.77498) and misclassified 1 case on both platforms. The third classifier including PPFIBP2 (Hs.12953), SAMHD1 (Hs.23889), C20orf27 (Hs.274422), ZNF75 (Hs.355015), and PRPSAP1 (Hs.77498) misclas-

sified one case on the cDNA and two cases on the Affymetrix platforms.

Assessment of the Predictive Accuracy of the Best Five-Gene Classifier on Independent Cases

A predictor that consists of a handful of genes that are reliably and comparably measured by both Affymetrix and cDNA platforms could be very useful, particularly if it predicts outcome accurately in new cases. To estimate the true predictive accuracy of the five-gene LDA predictor that performed the best on both platforms, we tested it on an independent set of 27 patients profiled on Affymetrix. The classifier showed a 70% overall prediction accuracy (95% CI = 50%, 86%). This suggests that it may be possible to develop multigene predictors that perform well across platforms and also show reasonably good predictive accuracy in independent cases.

It is important to consider that the five-gene LDA predictor examined in this analysis may not be the best of all possible predictors that can be developed from each particular data set. Many of the most informative genes (in univariate analysis) in the Affymetrix data were not represented on the cDNA arrays and vice versa. Also, there are a large number of supervised classification methods including support vector machines, k-nearest neighbor, and various types LDA algorithms that can be used to generate predictors. The comparison of different classification methods and selection of the optimal predictor for independent validation was not the main goal of this study, and it is the subject of a separate analysis that includes larger number of cases.

Discussion

More and more laboratories explore gene expression profiling with DNA microarrays as a potential diagnostic tool for classification of cancer. There are a large number of different profiling platforms available for research including commercially available arrays like Affymetrix gene chips and various proprietary and custom-made cDNA and oligonucleotide arrays. Some platforms may only be available at the single laboratory that has assembled it. It is integral to the scientific process to evaluate the performance of novel gene expression-based classifiers in different laboratories and on independent data sets. However, this cross-laboratory validation often involves cross-platform testing. The performance of any classifier in this type of cross-platform and cross-sample validation depends on 1) how generalizeable the initial observation is, and 2) how closely the gene expression measurements of the discriminating genes correspond to each other on the different platforms. Even an accurate outcome predictor may show limited cross-platform reproducibility if the two platforms do not measure the informative, discriminating genes similarly.

We examined the concordance between gene expression data generated from the same RNA specimens by

two different DNA microarray platforms. One platform, Millennium Pharmaceuticals cDNA arrays, contained 21,594 unique UniGene clusters; the Affymetrix HU133 A and B chips contained 30,868 clusters. Only 17,832 clusters were present on both platforms. We observed modest overall correlation between paired measurements of the same genes when probes were matched by UniGene. This was apparent regardless of the correlation metric used including Pearson and Spearman correlation coefficients or Concordance correlation of *t* scores. Our findings are consistent with several other reports that indicated modest overall correlation of gene expression measurements across platforms.^{18–20}

Some Affymetrix probe sets displayed very high correlation with matching cDNA array results, but other probe sets targeting the same gene showed minimal or no correlation. We hypothesized that this variable correlation may be partly due to sequence differences between the cDNA probes and the various Affymetrix probe sets and the location of the target sequence within the gene to be measured. In the case of the ER gene, we could demonstrate that both arrays could measure the expression of ER very accurately compared with an external clinical gold standard and in a highly concordant manner. However, each ER probe set present on the U133A chip showed different levels of expression intensities and various degrees of correlation with ER gene expression defined by IHC or cDNA array. These differences in expression intensity could be explained by the location of the oligonucleotide target sequences within the ER cDNA. When we examined the correlation between cDNA and Affymetrix probes that had direct sequence overlap, the correlation was quite high. Most of the discrepant correlations between Affymetrix probe sets and cDNA measurements could be explained by the sequence differences of the probes. Important factors that contribute to the different signal intensities generated by distinct probes that target the same gene at different locations include differing GC-content, sequence length, intraplatform cross-match opportunities, and the location of the probe sequence in relation to the 3' end.¹⁸

Because many microarray studies draw their conclusions from hierarchical clustering analyses, cross-platform preservation of clustering results is important. We examined whether genes that are differentially expressed between tumors that had complete pathological response to preoperative chemotherapy and those with residual cancer retain their class-discriminating value when used in supervised hierarchical clustering across platforms. Dendograms should be similar if intraplatform relationships between measurements are similar on the two different platforms. Because only 68% of the Affymetrix U133A chip and 44% of the cDNA array genes were common to both platforms, to avoid complexities due to missing informative genes, we restricted our analysis to select differentially expressed genes from the subset of genes represented on both platforms (*n* = 9402). There was only limited overlap between the lists of informative genes that distinguished cases with pathological CR from those with residual disease generated from the same samples with two different profiling plat-

forms. Only 17 genes were common to the 45-gene long cDNA and 168-gene long Affymetrix lists. Not surprisingly, informative genes performed well in supervised hierarchical clustering on the original data but showed decreased discriminating value when applied to data generated on the other platform.

We also examined the cross-platform performance of 100 sets of five-gene GA/LDA response predictors. The average misclassification error rate of these five-gene classifiers developed from the cDNA data were 2% on the original data. When the same classifiers were tested on the Affymetrix data, the average misclassification error rate has risen to 19.5%. For comparison, five-gene random classifiers produced 33% average misclassification error rates. Essentially identical results of diminished classification accuracy were observed when classifiers developed from the Affymetrix data were applied to cDNA results.

It is important to recognize that we compared two very different profiling platforms, cDNA nylon arrays hybridized to radioactive-labeled samples versus oligonucleotide Affymetrix GeneChips hybridized to biotin-labeled samples. Platforms that are more similar, for example two different versions of Affymetrix GeneChips, may show greater concordance of results because of greater similarity of the probe sequences. We only examined the performance of supervised hierarchical clustering and class (response outcome) prediction based on linear discriminant analysis, therefore it is possible that other class prediction methods may be more robust for cross-platform application. However, there is no reason to believe that any prediction algorithm would perform well across platforms if the concordance between the expression measures of the informative genes is low. Essentially all published results that attempted cross-platform testing of informative genes regardless of class prediction methodology reported diminished (but not completely lost) classification accuracy on data generated by platforms other than the original platform.^{21,22}

In summary, many genes with class-discriminating value on one profiling platform lose some of their discriminating value when measured with another profiling method. It is possible to select a subset of genes that retain much of their class-discriminating value across platforms based on high degree of sequence overlap between the probes. However such paired probes represent only a small minority of all probes present in any particular platform. Although it is reassuring that multi-gene predictors do hold up to some extent when applied across platforms, cross-platform application of multigene classifiers may have limited clinical value because of substantial and unpredictable loss of classification accuracy due to 1) missing informative genes and 2) often suboptimal measurement of informative genes between platforms. These observations underscore the importance of collaborative efforts to create uniform gene expression databases across various laboratories using standard operating procedures and a common platform to test the true diagnostic potential of this technology.

References

1. Ramaswamy S, Golub TR: DNA microarrays in clinical oncology. *J Clin Oncol* 2002, 20:1932–1941
2. de Bolle X, Bayliss CD: Gene expression technology. *Methods Mol Med* 2003, 71:135–146
3. King HC, Sinha AA: Gene expression profile analysis by DNA microarrays: promise and pitfalls. *J Am Med Assoc* 2001, 286:2280–2288
4. Ali TR, Li MS, Langford PR: monitoring gene expression using DNA arrays. *Methods Mol Med* 2003, 71:119–134
5. Ayers M, Symmans WF, Stec J, Damokosh A, Clark E, Hess K, Lecocke M, Metivier J, Bolt A, Brown J, Booser D, Ibrahim N, Valero V, Royce M, Arun B, Whitman G, Ross J, Sneige N, Hortobagyi GN, Pusztai L: Gene expression profiles predict complete pathologic response to neoadjuvant paclitaxel/FAC chemotherapy in breast cancer. *J Clin Oncol* 2004, 22:2284–2293
6. Symmans WF, Ayers M, Clark EA, Stec J, Hess KR, Sneige N, Buchholz TA, Krishnamurthy S, Ibrahim NK, Buzdar AU, Theriault RL, Rosales MFM, Thomas ES, Gwyn KM, Green MC, Syed AR, Hortobagyi GN, Pusztai L: Fine needle aspiration and core needle biopsy samples of breast cancer provide similar total RNA yield, but different stromal gene expression profiles. *Cancer* 2003, 97:2960–2971
7. Pusztai L, Ayers M, Stec J, Clark E, Hess K, Stivers D, Damokosh A, Sneige N, Buchholz TA, Esteva FJ, Arun B, Booser D, Rosales M, Valero V, Adams C, Hortobagyi GN, Symmans WF: gene expression profiles obtained from single passage fine needle aspirations (FNA) of breast cancer reliably identify prognostic/predictive markers such as estrogen (ER) and HER-2 receptor status and reveal large scale molecular differences between ER-negative and ER-positive tumors. *Clin Cancer Res* 2003, 9:2406–2415
8. Li C, Wong WH: Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol* 2001, 2:1–11
9. Benjamini Y, Hochberg Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B* 1995, 57:289–300
10. Pounds S, Morris SW: Estimating the occurrence of false positive and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics* 2003, 19:1236–1242
11. Goldberg DA: *Genetic Algorithms in Search, Optimization and Machine Learning*. New York, Addison-Wesley, 1989
12. Liu J, Iba H, Ishizuka M: Selecting informative genes with parallel genetic algorithms in tissue classification. *Genome Informatics Series: Proceedings of the Workshop on Genome Informatics* 2001, 12:14–23
13. Ooi CH, Tan P: Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. *Bioinformatics* 2003, 19:37–44
14. Li L, Weinberg CR, Darden TA, Pedersen LG: Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics* 2001, 17:1131–1142
15. Hamadeh HK, Bushel PR, Jayadev S, DiSorbo O, Bennett L, Li L, Tennant R, Stoll R, Barrett JC, Paules RS, Blanchard K, Afshari CA: Prediction of compound signature using high density gene expression profiling. *Toxicol Sci* 2002, 67:232–240
16. Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC, Liotta LA: Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* 2002, 359:572–577
17. Gruvberger S, Ringner M, Chen Y, Panavally S, Saal LH, Borg A, Ferno M, Peterson C, Meltzer PS: Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res* 2001, 61:5979–5984
18. Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado L, Kohane IS: Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* 2002, 18:405–412
19. Yuen T, Wurmbach E, Pfeffer RL, Ebersole BJ, Sealfon SC: Accuracy and calibration of commercial oligonucleotide and custom of cDNA microarrays. *Nucleic Acids Res* 2002, 30:1–9
20. Tan PK, Downey TJ, Spitznagel EL Jr, Xu P, Fu D, Dimitrov DS,

- Lempicki RA, Raaka BM, Cam MC: Evaluation of gene expression measurements from commercial microarray platforms *Nucleic Acids Res* 2003, 31:5676–5684
21. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J, Perou CM, Lonning PE, Brown PO, Borresen-Dale AL, Botstein D: Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci USA* 2003, 100:8418–8423
22. Sotiriou C, Neo SY, McShane LM, Korn EL, Long PM, Jazaeri A, Martiat P, Fox SB, Harris AL, Liu ET: Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci USA* 2003, 100:10393–10398