You might find this additional information useful...

Supplemental material for this article can be found at:
http://physiolgenomics.physiology.org/cgi/content/full/00066.2004/DC1

This article cites 41 articles, 27 of which you can access free at:
http://physiolgenomics.physiology.org/cgi/content/full/18/3/308#BIBL

This article has been cited by 10 other HighWire hosted articles, the first 5 are:

Expression Profiles of the Mouse Lung Identify a Molecular Signature of Time-to-Birth
A. T. Kho, S. Bhattacharya, B. H. Mecham, J. Hong, I. S. Kohane and T. J. Mariani
Am. J. Respir. Cell Mol. Biol., January 1, 2009; 40 (1): 47-57.
[Abstract] [Full Text] [PDF]

Characterization of mismatch and high-signal intensity probes associated with Affymetrix genechips
Y. Wang, Z.-H. Miao, Y. Pommier, E. S. Kawasaki and A. Player
Bioinformatics, August 15, 2007; 23 (16): 2088-2095.
[Abstract] [Full Text] [PDF]

Chips to Bedside: Incorporation of Microarray Data into Clinical Practice
L. Pusztai
Clin. Cancer Res., December 15, 2006; 12 (24): 7209-7214.
[Full Text] [PDF]

An application for assessing quality of RNA hybridized to Affymetrix GeneChips
K. J. Archer and T. Guennel
Bioinformatics, November 1, 2006; 22 (21): 2699-2701.
[Abstract] [Full Text] [PDF]

Assessing quality of hybridized RNA in Affymetrix GeneChip experiments using mixed-effects models
K. J. Archer, C. I. Dumur, S. E. Joel and V. Ramakrishnan
Biostat., April 1, 2006; 7 (2): 198-212.
[Abstract] [Full Text] [PDF]

Updated information and services including high-resolution figures, can be found at:
http://physiolgenomics.physiology.org/cgi/content/full/18/3/308

Additional material and information about *Physiological Genomics* can be found at:
http://www.the-aps.org/publications/pg

This information is current as of September 22, 2009 .

Increased measurement accuracy for sequence-verified microarray probes

**Brigham H. Mecham,[1] Daniel Z. Wetmore,[1] Zoltan Szallasi,[2]
Yoel Sadovsky,[3] Isaac Kohane,[2] and Thomas J. Mariani[1]**
[1]*Division of Pulmonary and Critical Care Medicine, Department of Medicine and Pulmonary Bioinformatics,
the Lung Biology Center, Brigham and Women's Hospital, and* [2]*Children's Hospital Informatics Program,
Harvard Medical School, Boston, Massachusetts 02115; and* [3]*Department of Obstetrics and Gynecology,
Washington University School of Medicine, St. Louis, Missouri 63110*

**Mecham, Brigham H., Daniel Z. Wetmore, Zoltan Szallasi,
Yoel Sadovsky, Isaac Kohane, and Thomas J. Mariani.** Increased
measurement accuracy for sequence-verified microarray probes.
Physiol Genomics 18: 308–315, 2004. First published May 25, 2004;
10.1152/physiolgenomics.00066.2004.—Microarrays have been ex-
tensively used to investigate genome-wide expression patterns. Al-
though this technology has been tremendously successful, it has
suffered from suboptimal individual measurement precision. Signifi-
cant improvements in this respect have been recently made. In an
effort to further explore the underlying variability, we have attempted
to globally assess the accuracy of individual probe sequences used to
query gene expression. For mammalian Affymetrix microarrays, we
identify an unexpectedly large number of probes (greater than 19% of
the probes on each platform) that do not correspond to their appro-
priate mRNA reference sequence (RefSeq). Compared with data
derived from inaccurate probes, we find that data derived from
sequence-verified probes show *1*) increased precision in technical
replicates, *2*) increased accuracy translating data from one generation
microarray to another, *3*) increased accuracy translating data from
oligonucleotide to cDNA microarrays, and *4*) improved capture of
biological information in human clinical specimens. The logical
conclusion of this work is that probes containing the most reliable
sequence information provide the most accurate results. Our data
reveal that the identification and removal of inaccurate probes can
significantly improve this technology.

Affymetrix; Agilent; reference sequence; RefSeq

COMPREHENSIVE GENE EXPRESSION profiling using microarray
technology has facilitated a revolution in the characterization
of cellular regulation and shows great potential for human
disease diagnostics. Microarrays have been successfully used
to identify the targets of transcription factors (7, 41) and
secreted regulatory molecules (25, 37). Transcriptional profil-
ing has also lead to the identification of molecular mechanisms
involved in animal models of disease (13, 42). The technology
has been successfully used for the identification and/or classi-
fication of disease (6, 8, 18, 36) and has also provided insight
into regulatory networks contributing to developmental pro-
cesses (3, 14, 21, 23, 29, 32, 35, 39).

Commercial microarrays have been widely available, and
the Affymetrix oligonucleotide microarrays (GeneChip tech-
nology; Ref. 19) represent a large proportion of previously
published studies. GeneChip technology utilizes multiple in-

dependent probe hybridization events to measure the expres-
sion level for each gene investigated. Additionally, for each
hybridization event, the technology utilizes 25-nt oligomers
(25-mers) and corresponding single nucleotide mismatch 25-
mers to measure specificity. The individual 25-mers are de-
rived from publicly available nucleotide sequence information.

There has been tremendous success applying microarray
technology to disease diagnostic applications. For instance,
multiple groups have shown that microarray data can identify
previously unappreciated molecular subtypes of lung cancer
that differ in their prognoses (1, 2, 5). Unfortunately, poor
reproducibility of results exists across these studies. This
indicates either an underlying distinction in the nature of the
disease investigated or, more likely, a limitation of the tech-
nology in reliably capturing the underlying biology. Microar-
ray technology has been criticized for a lack of individual
measurement accuracy. However, the technology is rapidly
advancing, and improvements in reagents (11) and data anal-
ysis (12, 16) have increased measurement precision. Some
sources of noise, such as those due to hybridization intensity
differences, are systematic and have been successfully defined
(4, 9, 12, 15, 20, 33, 40).

As a tremendous volume of data has been generated (par-
ticularly from human clinical specimens, which cannot be
duplicated) strategies to improve analysis of ("clean up")
existing data sets are of great value. One limitation of the
application of this technology could be due to the failure of
similar studies to measure identical biological parameters. For
instance, Sorlie et al. (31) recently showed that limiting anal-
ysis of data to probes that are verified to query identical
UniGenes improves concordance of results from one microar-
ray data set to another.

Probe sequence inaccuracies are known to exist for both
oligonucleotide and cDNA microarrays. However, there is a
general lack of information regarding the scope of probe
sequence inaccuracies on currently available Affymetrix plat-
forms. In this study we report a global analysis of the probes
used by Affymetrix technology, where we have systematically
attempted to confirm the accuracy of individual probe se-
quences. We find that for a significant number of probe sets, on
both old and current platforms, the probe sequences do not
perfectly correspond with the appropriate mRNA as defined by
the reference sequence (RefSeq). Given the approach Af-
fymetrix uses to determine true mRNA hybridization from
background (e.g., single nucleotide mismatch), any sequence
discrepancy likely renders the probe uninformative. Further-
more, we report that data derived from sequence-verified
probes shows vastly improved precision. Therefore, removing

---

information from inaccurate probes should significantly improve the validity of results from this technology.

EXPERIMENTAL PROCEDURES

Database Information and Probe Verification

All mRNA sequences were retrieved from the NCBI UniGene molecular database. UniGene builds 162 (Sept. 16, 2003), 131 (Dec. 4, 2003), and 125 (Dec. 5, 2003) were used for the human, mouse, and rat genomes, respectively. Affymetrix probe set annotation information was obtained from the Oct. 15, 2003, release of NetAffx (17). Annotation information included individual probe sequences as well as the RefSeq identifier used to relate probe sets to mRNAs. A RefSeq represents a curated, nonredundant transcript sequence (26, 27). The location of all Affymetrix probe sequences was identified in their corresponding mRNAs with the use of ProbeMapper, a Perl script developed specifically for this procedure. "Perfect match" (PM) and "mismatch" (MM) probe sequences that overlap RefSeq information were considered "verified." Probe sets were considered verified if at least one probe sequence was an exact match with the corresponding RefSeq. We chose such a relaxed classification scheme for probe sets in order to account for the incompleteness of 5′ and 3′ untranslated regions (UTR) in the RefSeq database. Agilent probe information, including the GenBank sequence ID, was provided by the manufacturer. All outdated Agilent clones were removed prior to analysis.

Data Processing and Probe Matching

For each Affymetrix microarray experiment, array image files were analyzed with Affymetrix Microarray Suite ver. 5 (MAS5) software, and signal intensities were calculated for MAS5, robust multi-array averaging (RMA) (12), and DNA-Chip Analyzer (dChip) (16) with the aid of Bioconductor software. Expression values for Agilent cDNA data were calculated using the manufacturer's standard software and normalization procedures. Correlations were calculated using MATLAB, and clustering diagrams were generated by the TIGR Multiexperiment Viewer (MeV) (28) software package.

When comparing measurements across generations of Affymetrix platforms, probe sets were matched if they contained at least one probe that corresponds to the same UniGene. When comparing measurements between Agilent and Affymetrix technologies, probe sets were matched if the Affymetrix probe set was for a UniGene containing an Agilent cDNA sequence.

Microarray Data Sets and Analysis Details

Affymetrix technical replicate experiment. The microarray data set used for this analysis has been previously described (20) and is available at the NCBI Gene Expression Omnibus (GEO) (GSE1302). Three independent mRNA samples were studied on the Hu95v2 platform (subarrays A–E). For each sample, a single mRNA target was generated and analyzed using five (sets of U95Av2, U95Bv2...) microarrays. Therefore, for each subarray tested, the three conditions produced three independent sets of five technical replicate samples. Each set of five replicates generated ten comparisons (e.g., *replicate 1* vs. *2*, *1* vs. *3*, . . ., *4* vs. *5*). The probe sets for each platform were classified as verified or unverified as described above. Pearson and Spearman correlations were calculated for each set of ten replicate comparisons, for either the verified or unverified probe sets, using MAS5.

Breast cancer cell line experiments. This microarray data set has been previously described (39) and is available at GEO (GSE1299). For the microarray data set used for this analysis, mRNA from two cancer cell lines and one normal cell line were hybridized to replicate Affymetrix Hu95Av2, Hu133A, and Hu133B and Agilent Human cDNA arrays. Each replicate experiment was compared with a similar experiment on the other platforms (e.g., *cancer cell line 1, replicate 1*

on the Hu95Av2 platform was compared with the *cancer cell line 1, replicate 1* on the Hu133A, Hu133B, and Agilent platforms), resulting in six independent comparisons for each platform. Probe sets were matched across Affymetrix platforms and between Affymetrix and Agilent technologies as described above.

When comparing data across Affymetrix platforms, Pearson and Spearman correlation coefficients were calculated from signal intensity data for the matching verified and unverified probe sets for each of the six comparisons. As Agilent technology reports expression levels as a ratio between two samples, when comparing across technologies the Affymetrix data had to be transformed. Here, the expression level for each Affymetrix probe set was transformed into the log base 2 of the ratio between its signal intensity in a cancer sample with its signal intensity in a normal sample. Pearson and Spearman correlation coefficients were calculated from signal intensity data for both the verified and unverified probe sets.

Breast cancer tissue experiments. Both breast cancer studies were performed on the U95Av2 and have been previously described. The Dana Farber Cancer Institute (DFCI) study included 101 breast cancer tumors and 8 normal breast tissue samples (30) and is publicly available (**http://lungtranscriptome.bwh.harvard.edu/Breast_Cancer_Data.html**). The Duke University study included 89 breast cancer tumors (38) and is publicly available (**http://data.cgt.duke.edu/lancet.php**). A 198 × 12,626 matrix was generated that contained RMA signal intensity values for each probe set in all samples from both the Duke and DFCI studies. hierarchical average linkage clustering (2), with the Pearson correlation as the distance metric, was preformed on the following three sets of data: *1*) The 1,000-probe sets that exhibited the largest standard deviation relative to their mean intensity, *2*) the subset of these 1,000-probe sets that are RefSeq-verified, and *3*) the subset of these 1,000-probe sets that are unverified. Note that the Affymetrix control probe sets (e.g., AFFX-PheX-M_at) were excluded from the clustering so that no bias was introduced by different processing methods that either center might have used.

Significance Testing

Because of hybridization quality variation, the significance of the difference in correlations between sequence-verified and -unverified probe sets was determined using a paired *t*-test.

RESULTS

Probe Sequence Verification

Inaccuracy of probe sequences has been previously recognized in both custom and commercial microarray platforms. However, an investigation of the magnitude and scope of this problem for Affymetrix technology has not been previously reported. In this study, we performed a systematic evaluation of the accuracy of the sequences of the individual probes on 20 of the most commonly used Affymetrix mammalian microarray platforms available as of June 2003. For each platform, we identified all exact probe sequences within each appropriate UniGene database. For every mRNA sequence in a UniGene database, our algorithm identified all 25-nt subsequences that are identical to either a PM, MM, reverse complement PM or reverse complement MM probe sequence. Because of the nature of the technology (e.g., use of single nucleotide mismatch as a control), we only identified probes that exactly match a 25-nt-length subsequence of a transcript. For every identified probe sequence, we defined its inclusion, location, and orientation in the RefSeq database (26, 27). We used the RefSeq, as it is the most highly curated, publicly available definition of mRNA transcript sequences for entire genomes.
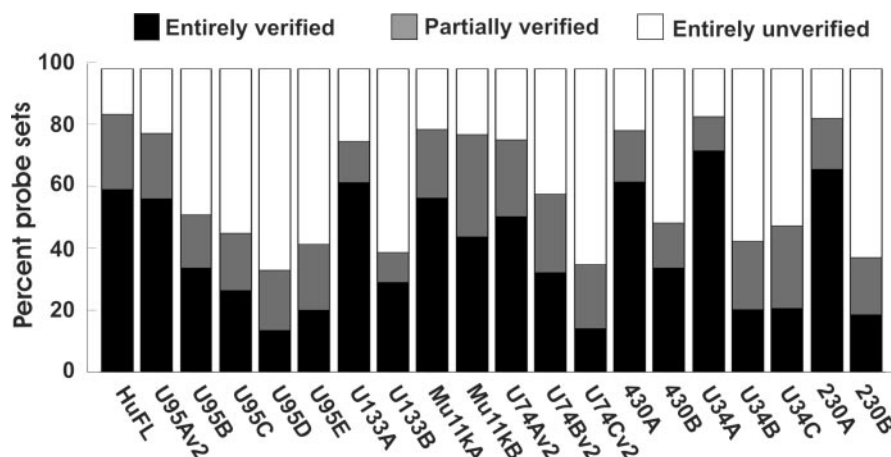
Fig. 1. Proportion of sequence-verified probes sets. The distribution of probes sets on all platforms, classified as *1*) entirely verified when all probes corresponded to the RefSeq mRNA sequence (black bars), *2*) entirely unverified when none of the probes corresponded to the mRNA sequence (open bars), or *3*) partially verified when a subset of probes corresponded to the mRNA (gray bars).

We found that a very high proportion of the individual probes could not be verified as measuring their associated RefSeq (Supplemental Table S1, available at the *Physiological Genomics* web site).[1] Of particular note, the percentages of verified probes on the more recent platforms were no greater than those for the older platforms: U133A, 72% vs. HuFL, 80%; 430A, 75% vs. Mu11kA, 74%; and 230A, 80% vs. U34A, 81%. Moreover, for each platform that contains a set of arrays (e.g., Hu95A–E), there was always a decrease in the percentage of verified probes on the secondary array(s) (e.g., Hu95B–E, max 47%) compared with that of the primary array in the set (e.g., Hu95A, 72%). Interestingly, platforms with higher verification rates generally contained fewer EST sequence-derived probes.

As Affymetrix technology utilizes "sets" of probes to interrogate a single gene, we investigated the distribution of verified probes into probe sets (Fig. 1). A probe set was classified as either *1*) entirely verified if it contained only verified probes, *2*) entirely unverified if none of its probes were verified, or *3*) partially verified if some, but not all, of its probes were verified. For human platforms, we found at best 62.5% (U133A) to at worst 13.8% (U95D) of probe sets were entirely verified. Again, the older and newer platforms showed similar percentages of entirely verified probe sets. Probe sets were generally either entirely verified or entirely unverified, with only 10% (U133B) to 24.7% (HuFL) being partially verified. Additionally, a large number of the partially verified probe sets contained one or two unverified probes. For U133A, more than 50% of the partially verified probe sets have 9 or 10 out of 11 probes verified (Supplemental Fig. S1). This distribution of probes in the partially verified probe sets indicates a preference for sets of probes to identify their intended target.

There were numerous reasons for the lack of probe sequence verification. An obvious source is continuously evolving/improving sequence database quality. For example, probe set 220154_at on the U133A platform was designed to query UniGene Hs.443518 and RefSeq NM_020388. As of UniGene build 133, NM_020388 was a 1,022-nt mRNA sequence, and all 11 probes for 220154_at were accurate. This RefSeq has

subsequently been updated to a 9,348-nt mRNA sequence with no significant similarity to the older sequence (as defined by BLAST). As a result, the probes for 220154_at no longer measure their intended target. Presently, 4 of the 11 probes measure NM_015548, an alternative transcript for Hs.443518. An additional source of inaccuracy is improper annotation of where transcription starts and ends. For example, probe set 202172_at on the U133A platform is intended to monitor NM_007146. However, its probes map to a region ~1,800–2,200 bp 3′ of the transcribed region for NM_007146. Other inaccuracies arise from probes derived from the noncoding strand. Probe set 219533_at on the U133A platform was designed to monitor NM_000076; however, all 11 probes measure the reverse complement of this sequence. As stated above, not all probe sets are composed of entirely verified or unverified probes. One example is probe set 208256_at on the 133A platform, which was designed to measure NM_001405. Currently, four of its probes measure the sense strand, two measure the antisense strand, three are found downstream of the 3′ end of the coding sequence, and the remaining two cannot be identified in the transcript sequence or in the region 10-kb on either side of the transcribed region. This probe set exemplifies the various problems that can occur as a result of database evolution, improperly defined sequence boundaries, and confusion over sense and antisense transcripts. Complete probe sequence mapping information, and verification results concerning individual probe sets on all Affymetrix platforms described here, is available for academic use at **http://lungtranscriptome.bwh.harvard.edu** (Supplemental Fig. S2).

Replicate Precision of Verified Probe Sets

Although we may have decreased confidence in the reliability of unverified probes, these probes may not necessarily show a difference in measurement precision. We sought to test whether precision was related to probe sequence accuracy using data from a previously published replicate microarray data set (20), including three mRNA targets each hybridized to five individual microarrays on each of the U95 (A, B, C, D, and E) platforms. The three conditions produced three sets of five technical replicate data sets, with each set of replicates generating ten comparisons (e.g., *replicate 1* vs. *replicate 2*, *1* vs. *3*, . . . *4* vs. *5*). Verified probe sets were defined as those containing at least one verified probe and unverified probe sets

contained no verified probes. We focused upon a comparison of the behavior between these two groups. There are two major benefits of this approach: *1*) it provides a measure of "quality" for both the verified and unverified probe sets, and *2*) the independence of the verified and unverified sets allow for a statistical comparison to be made.

Pearson and Spearman correlation coefficients were calculated for replicate measurements for both verified and unverified probe sets on each subarray using MAS5 data (Table 1 and Supplemental Table S2). For data generated on U95A, Pearson correlation coefficients ranged from 0.949–0.991 for verified probe sets and 0.883–0.987 for unverified probe sets, with *P* values of 0.014 to 0.000076. Spearman correlation coefficients showed a greater increase in accuracy for verified probe sets. The consistently significant increase in the correlations derived from verified probe sets indicated that probe accuracy affects data reproducibility.

The dramatic difference in Spearman correlation coefficients led us to investigate intensity differences between the two groups of data. Mean signal intensities for nonverified probe sets were less than those of verified probe sets (although not significantly), suggesting that removal of unverified probes might be equivalent to filtering low-intensity signals. We removed all unverified probe sets signals with a maximum intensity value less than 300 and recalculated Pearson correlation coefficients. After filtering low-intensity probe set signals, verified probe sets showed a similar significant increase in correlation (Supplemental Table S3) compared with unverified probe sets. The increased correlations of verified probe sets were not dependent upon signal intensity as shown by the data from the U95B platform, where removal of low signal intensity data from unverified probes resulted in a greater mean signal intensity for this group. Additionally, for U95A replicates (performed on *treatment 1* data), we limited the unverified probe set data to the 500 probe sets with the largest mean intensities and recalculated the replicate correlations. This unverified probe set subgroup showed a greater mean signal intensity (1,919 vs. 1,409), but significantly lower correlations (verified, 0.991–0.947 vs. unverified, 0.986–0.916; *P* < 0.00001) than that for the verified probe sets.

These differences were somewhat surprising, given the measurements were from technical (and not experimental) replicates. However, it is likely that unverified probes capture background hybridization to one or more transcripts. Whether individual unverified probes measure background, an alternate transcript, or multiple transcript variants, it is intuitive that unverified probe sets would have greater variability than probe sets that are verified to measure a single transcript. The difference in correlations clearly indicated a difference in measurement reproducibility and led us to further test the effects of probe sequence accuracy on data reliability.

Accuracy of Verified Probe Sets Across Platforms and Technologies

Having shown that verified probe sets capture information with greater precision in replicate experiments performed on one platform, we tested the effects of sequence verification on measurement accuracy across complementary microarray technologies. We used a data set consisting of experiments on each of two breast cancer cell lines and one normal mammary epithelial cell line (22). For each cell line, replicates were performed on the U95A, U133A, U133B, and Agilent cDNA microarray platforms.

We tested the effect of sequence verification upon measurement accuracy across generations of Affymetrix platforms. Probe sets from the U95 and U133 platforms were matched using UniGene identifiers. Probe sets that were shared across platforms were classified as verified if at least one probe matched its RefSeq. For example, UniGene Hs.20952 is measured by two probe sets on both the Hu95Av2 (37761_at, 33386_at) and Hu133A (209821_s_at, 208886_at) platforms. However, the RefSeq for this UniGene, NM_001682, is only measured by 37761_at and 209821_s_at. Therefore, we classified these probe sets as verified, whereas 33386_at and 208886_at were classified as unverified. Spearman and Pearson correlation coefficients were calculated for both the verified and unverified (matched) probe set signal intensities. To confirm that the observed effects were not due to specific data processing algorithms, we repeated the analyses using MAS5, RMA, and dChip.

For each comparison between two platforms there were six correlations calculated. Consistently significant (*P* < 0.00001) increases in correlations were observed for verified probe sets compared with unverified probe sets (Table 2 and Supplemental Fig. S3*A*) using all analysis methods. For example, *P* values for the Pearson correlation coefficients in the U95A to U133A comparison were all less than 0.000001 for MAS, RMA, and dChip. The differences were unrelated to sample size, as RefSeq-verified probe sets comprised a majority in the U95A to U133A comparison and a minority the U95A to U133B comparisons. An interesting aspect of this analysis is the extremely low correlations for unverified probe sets in the U95A to U133B comparison, likely due to the high amount of EST-derived probe sequences on the U133B platform.

As in the technical replicate experiment, verified probe sets had higher mean signal intensities than unverified probe sets. Again, filtering out low signal intensity data did not remove the improvement in correlations for RefSeq-verified probe sets (Supplemental Table S4). For example, *P* values for the Pearson correlation coefficients in the U95A to U133A comparison were less than 0.000001 for MAS, RMA, and dChip. These

Table 1. *Increased measurement precision for sequence verified probes in technical replicates*

Array	Treatment	Spearman Correlation			Pearson Correlation			No. of Probe Sets Tested	
		Verified	Unverified	*P* value	Verified	Unverified	*P* value	Verified	Unverified
U95A	*1*	0.896–0.935	0.825–0.886	<0.00001	0.949–0.991	0.935–0.987	<0.00001	9,559	3,068
	2	0.838–0.903	0.756–0.835	<0.00001	0.965–0.986	0.912–0.985	<0.015	9,559	3,068
	3	0.849–0.918	0.825–0.886	<0.00001	0.961–0.987	0.883–0.971	<0.0015	9,559	3,068

Table 2. *Increased measurement accuracy for sequence verified probes across multiple versions of Affymetrix platforms and across Affymetrix and Agilent cDNA technologies*

Comparison			Spearman Correlation			Pearson Correlation			No. of Probe Sets Tested	
Platform A	Platform B	Algorithm	Verified	Unverified	P value	Verified	Unverified	P value	Verified	Unverified
U95A	U133A	MAS5	0.68–0.72	0.53–0.57	<0.00001	0.69–0.77	0.43–0.57	<0.00001	8,839	2,239
		RMA	0.60–0.65	0.50–0.52	<0.00001	0.62–0.67	0.49–0.53	<0.00001	8,839	2,239
		dChip	0.54–0.60	0.42–0.45	<0.00001	0.64–0.73	0.34–0.45	<0.00001	8,839	2,239
U95A	U133B	MAS5	0.54–0.63	0.11–0.17	<0.00001	0.62–0.67	0.04–0.11	<0.00001	310	796
		RMA	0.44–0.56	0.13–0.17	<0.00001	0.50–0.60	0.13–0.15	<0.001	310	796
		dChip	0.34–0.47	0.09–0.12	<0.00001	0.58–0.63	0.05–0.06	<0.00001	310	796
U95A	Agilent	MAS5	0.560–0.597	0.395–0.457	<0.05	0.576–0.663	0.386–0.492	<0.001	6,800	621
		RMA	0.634–0.679	0.507–0.562	<0.001	0.627–0.734	0.423–0.659	<0.05	6,800	621
		dChip	0.613–0.628	0.480–0.523	<0.001	0.630–0.729	0.400–0.623	<0.05	6,800	621
U133A	Agilent	MAS5	0.562–0.591	0.347–0.489	<0.05	0.560–0.597	0.395–0.457	<0.05	9,510	990
		RMA	0.625–0.653	0.490–0.560	<0.05	0.634–0.727	0.395–0.616	<0.025	9,510	990
		dChip	0.587–0.637	0.415–0.543	<0.05	0.636–0.743	0.370–0.611	<0.05	9,510	990
U133B	Agilent	MAS5	0.380–0.488	0.390–0.444	0.27	0.421–0.564	0.334–0.497	<0.05	994	427
		RMA	0.505–0.573	0.498–0.533	0.53	0.513–0.634	0.418–0.590	<0.05	994	427
		dChip	0.483–0.537	0.458–0.501	0.61	0.518–0.647	0.390–0.573	<0.05	994	427

MAS5, Affymetrix Microarray Suite v. 5; RMA, robust multi-array averaging; dChip, DNA-Chip Analyzer.

data clearly show that removing unverified probes are not the equivalent of removing low signals.

Another technique used by some researchers to remove low-quality information is to utilize the Affymetrix "detection call" algorithm. We investigated the rate of "absent" and "present" calls for both the verified and unverified probe sets. For each of the U95A, U133A, and U133B platforms the percentage of verified and unverified probe sets called present, absent, or "marginal" were calculated (Supplemental Table S5). Verified probe sets were scored present at a higher rate than unverified probe sets (56% vs. 45%) for U95A (58% vs. 46%), for U133A, and (53% vs. 34%) for U133B. However, the on-average greater than 40% score of present for unverified probe sets clearly indicates using the detection call as a basis for removing low-quality information from microarray experiments does not compensate for probe sequence inaccuracies. Additionally, we recalculated the correlations for the verified and unverified probe sets between the Hu95A and Hu133A platforms after removing all probe sets not consistently scored present. Verified probe sets retained significantly higher correlations than unverified probe sets ($P < 0.00001$, data not shown). These data further support the conclusion that these probes are measuring a nonspecific transcript, albeit at lower reproducibility, and this likely explains the improved measurement precision of verified probe sets in the technical replicate experiment.

Finally, we investigated the effects of sequence verification on measurement accuracy across Affymetrix and Agilent cDNA microarray technologies. There is greater concordance between cDNA and Affymetrix data when the probes used to measure expression contain similar information (22). Therefore, we only included probe sets from the U95 or U133 platforms if they overlap a sequence used on the Agilent array. Probe sets that matched Agilent cDNA probe sequences were defined as verified when at least one Affymetrix probe was RefSeq verified. For example, we identified a match between probe set 207038_at on the U133A platform with Agilent clone U79745. All 11 Affymetrix probes from probe set 207038_at were verified in RefSeq NM_004694 and the GenBank mRNA

for U79745. Probe sets that matched Agilent cDNA probe sequences were defined as unverified when no Affymetrix probe was RefSeq verified. For example, we identified a match between probe set 215264_at on the U133A platform and Agilent clone X68879. None of the probes from probe set 215264_at was contained in a RefSeq. As described in the MATERIALS AND METHODS, log base 2-transformed Affymetrix expression ratios were correlated with sequence-matched cDNA expression ratios. We used MAS5, RMA, and dChip expression values to control for any bias introduced by various data processing algorithms.

Again, significantly increased correlations were consistently observed for sequence-verified probe sets (Table 2 and Supplemental Fig. S3B). For example, P values for the Pearson correlation coefficients in the U95A to Agilent comparison were 0.003 for MAS5, 0.022 for RMA, and 0.019 for dChip. Spearman correlation coefficients for U133A to Agilent were equivalent to Pearson correlation coefficients, with similar P values. However, when comparing data from U133B to Agilent, Spearman correlation coefficients were dramatically lower and not significantly increased for verified probe sets. This suggests poorer overall quality measurements on the U133B subarray, with improvements due to sequence verification being insufficient to reach significance.

Diagnostic Accuracy of Verified Probe Sets

In addition to the value of expression profiling in understanding biological mechanism, this technology has been used as a predictive measure for defining disease states. One limitation of this application has been the inability to translate predictive value between complementary experiments (31). The goal of any cancer classification study is to uncover shared biology that can be used to identify additional cases of cancer tumorigenesis or metastasis. Moreover, the most basic test of any discrimination method should involve the detection of a difference between diseased and normal samples. We used these assumptions to test the effects of probe sequence accuracy in data from two independent breast cancer expression

profiling studies preformed on the U95A platform (30, 38) (10). Note we did not address the results from these individual studies, but investigated the effects of probe sequence verification upon the ability to compare multiple patient-related data sets.

To minimize artifacts arising from the experiments being performed at separate facilities (handling, hybridization conditions, scanner settings, etc.), we generated a single expression matrix using RMA. We identified the 1,000 genes that exhibited the highest mean value relative to their standard deviation for unsupervised hierarchical clustering. This filtering strategy was used to minimize the effects of noise or artifacts. Cluster analysis was then performed (Supplemental Fig. S4) for *1*) all 1,000 probe sets, *2*) the subset of these 1,000 probe sets that were unverified, and *3*) the subset of these 1,000 probe sets that were sequence verified. When all 1,000 probe sets were used for clustering, the samples separated into two major nodes each predominantly composed of samples from either study, alone. Moreover, the normal samples were split into three separate groups. This suggested the noise in the system was much greater than the captured underlying biology. Clustering of the samples using only the unverified probe sets produced similar results. However, clustering of the samples using only the verified probe sets produced a striking improvement. First, all normal samples clustered tightly, separated from a majority of tumors. Additionally, although samples still separated into two major nodes, there was significant mixing of the two nodes with samples from both studies. The observed increase in diseased sample overlap (shared biology) and grouping of normal samples as highly similar indicates that restricting data to sequence-verified probes can improve the diagnostic power of microarray technology. This result does not address a particular classification scheme but indicates that removing unverified probe sets allows for the major component of change to be related to the underlying biology of breast cancer as opposed to the source of the experiments.

DISCUSSION

Although inaccuracies in microarray probe sequences are an appreciated source of experimental noise, little information is available regarding their magnitude and contribution to microarray data variability. In the studies presented here, we have systematically assessed probe sequences on the most commonly used microarray technology. We find that numerous probes fail to correspond with current high-quality definitions of transcribed sequences as defined by the RefSeq database (Fig. 1 and Supplemental Table S1). We chose the RefSeq database, as it is the most highly curated, publicly available genome-wide definition of transcript sequence information. Although it is imperfect, it is likely to be the genomics "gold standard" for sequences of those regions of the genome that are transcribed. As described in RESULTS, there are many causes for these probe sequence inaccuracies, but most notably there has been constant improvement in sequence information databases over time. At the time of probe design, the sequences provided in publicly available databases might have been inconclusive (and further complicated due to alternate transcripts). Subsequently, the target sequence may have become outdated as a result of the consolidation of sequence information into RefSeq mRNAs. Regardless of the nature of probe sequence inaccu-

racies, we clearly show that sequence-verified probes perform more consistently, and with higher accuracy, within replicates and across different versions of the technology. Apparently, Affymetrix has come to the same conclusion and has recently released a platform containing RefSeq-verified probes. We conclude that verification of probe sequences against the best available transcript information is warranted.

As probe sequence inaccuracies would seem to be a likely source for measurement error, we directly tested this possibility. Within large technical replicate experiments (Table 1, and Supplemental Tables S2 and S3), across multiple generations of Affymetrix platforms (Table 2 and Supplemental Table S4) and across Affymetrix and (Agilent) cDNA technologies (Table 2), we consistently found significantly increased measurement accuracy for sequence-verified probe sets. We classified probe sets as completely verified, partially verified, or completely unverified using the individual probe verification information (Fig. 1, Supplemental Fig. S1, and Supplemental Table S1). For analysis of measurement accuracy, we considered probe sets verified if they included a single verified probe, as it is likely that the RefSeq database is not exhaustive and some transcript sequences are truncated. Further limiting verified probe sets to completely verified probes would likely increase the benefit of sequence verification but at the cost of excluding another ~10–20% of the data set. We tested measurement accuracy for verified probe sets and compared them to the accuracy of unverified probe sets; this allowed for a quantification of "quality" and statistical comparison of these independent sets. Compared with the entire data set (all probe sets), sequence-verified probe sets consistently showed higher correlations. Therefore, the improvement in data accuracy, when filtering for verified probe sets, is directly related to the number of unverified probe sets. Although the benefit of sequence verification can be either modest or large, dependent upon the platform used, the rationale for including data derived from inaccurate probes is unclear. The logical conclusion of these studies is that interrogation of the most accurate sequences generates the most accurate data.

The utility of using multiple probes (probe sets) for monitoring individual gene expression has been refined by the implementation of probe-level normalization methods such as those utilized by dChip and RMA (12, 16). In theory, inaccuracy of individual probe measurements could be compensated for by these algorithms. However, our data clearly show that these methods do not completely compensate for probe sequence inaccuracies and that sequence verification adds additional benefit to microarray data analysis. Another simple explanation for the benefit of sequence verification would be that unverified probe sets showed lower signal intensities, due to failure to accurately measure any transcript, and that removal of unverified probe sets was equivalent to removing low signal intensity measurements. Others have shown a dependency of measurement accuracy upon signal intensity (24). Indeed, signal intensities for unverified probe sets were less than those for verified probe sets. However, verified probe sets showed greater measurement precision than unverified probe sets, even after removing low signal intensity probe sets (Supplemental Tables S3 and S4). Furthermore, although verified probe sets were more often scored "present" by the MAS5 detection call, nearly half of all unverified probe sets were also present (Supplemental Table S5). Although there is a relation-

ship between verified probe sequences and signal intensity, these data reveal one limitation of simply removing probe sets with low signal intensities from microarray data and further show the benefit of sequence verification independent of thresholding for signal intensity or detection call.

Importantly, we have shown that the benefit of probe sequence verification extends beyond controlled in vitro experimental samples to potential diagnostic and predictive applications of microarrays (Supplemental Fig. S4). Unsupervised clustering of two independent human breast cancer data sets resulted in nearly complete separation of all patients from each study. A possible explanation could be that the breast cancer entities are distinct and related to different, regional biological mechanisms. A more plausible explanation for this result is that the signal-to-noise ratio of the data is insufficient to reliably capture the consistent underlying biology. Our data clearly show that sequence-verified probe sets improve the capture of the underlying biology as evidenced by the improved grouping of normal samples and improved mixing of samples from the two data sets. Unfortunately, since no clear subclassification for breast cancer exists, it is difficult to prove the clustering is "better." However, the increased clustering of normal samples, which represent a distinct group, using verified probes supports this conclusion.

We are not the first to use probe sequence-based information to assess microarray data accuracy. For instance, Tan et. al. (34) reported increased consistency of replicate measurements across Affymetrix, Agilent, and Amersham technologies. Sorlie et al. (31) used UniGene-matched probes to combine information from Affymetrix and Stanford cDNA microarrays in an effort to improve breast cancer classification. In all cases, postexperiment filtering using sequence information improves data quality. As combining data from multiple microarray platforms/technologies is certain to prove a common method, our results showing increased accuracy of sequence-verified probes across platforms (oligo vs. oligo and oligo vs. cDNA) substantiate the importance of using the most reliable information to verify equivalence of measurement across technologies. This can be facilitated by using the probe mapping files available at **http://lungtranscriptome.bwh.harvard.edu**, which includes lists of verified and unverified probe sets for each Affymetrix platform described in this study, as well as additional information (Supplemental Fig. S2) regarding the location of individual probes within RefSeqs. Alternatively, we encourage end-user verification with the most recent, publicly available sequence information.

REFERENCES

1. **Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JM, Iannettoni MD, Orringer MB, and Hanash S.** Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med* 8: 816–824, 2002.

2. **Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, and Meyerson M.** Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci USA* 98: 13790–13795, 2001.

3. **Diaz E, Yang YH, Ferreira T, Loh KC, Okazaki Y, Hayashizaki Y, Tessier-Lavigne M, Speed TP, and Ngai J.** Analysis of gene expression in the developing mouse retina. *Proc Natl Acad Sci USA* 100: 5491–5496, 2003.

4. **Drobyshev AL, Machka C, Horsch M, Seltmann M, Liebscher V, Hrabe de Angelis M, and Beckers J.** Specificity assessment from fractionation experiments (SAFE): a novel method to evaluate microarray probe specificity based on hybridisation stringencies. *Nucleic Acids Res* 31: E1–1, 2003.

5. **Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, van de Rijn M, Rosen GD, Perou CM, Whyte RI, Altman RB, Brown PO, Botstein D, and Petersen I.** Diversity of gene expression in adenocarcinoma of the lung. *Proc Natl Acad Sci USA* 98: 13784–13789, 2001.

6. **Golub T, Slonim D, Tamayo P, Huard C, Gaasenbeek M, Mesirov J, Coller H, Loh M, Downing J, Caligiuri M, Bloomfield C, and Lander E.** Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286: 531–537, 1999.

7. **Gupta RA, Brockman JA, Sarraf P, Willson TM, and DuBois RN.** Target genes of peroxisome proliferator-activated receptor gamma in colorectal cancer cells. *J Biol Chem* 276: 29681–29687, 2001.

8. **Haslett JN, Sanoudou D, Kho AT, Han M, Bennett RR, Kohane IS, Beggs AH, and Kunkel LM.** Gene expression profiling of Duchenne muscular dystrophy skeletal muscle. *Neurogenetics* 4: 163–171, 2003.

9. **Hekstra D, Taussig AR, Magnasco M, and Naef F.** Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays. *Nucleic Acids Res* 31: 1962–1968, 2003.

10. **Huang E, Cheng SH, Dressman H, Pittman J, Tsou MH, Horng CF, Bild A, Iversen ES, Liao M, Chen CM, West M, Nevins JR, and Huang AT.** Gene expression predictors of breast cancer outcomes. *Lancet* 361: 1590–1596, 2003.

11. **Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR, Kobayashi S, Davis C, Dai H, He YD, Stephaniants SB, Cavet G, Walker WL, West A, Coffey E, Shoemaker DD, Stoughton R, Blanchard AP, Friend SH, and Linsley PS.** Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol* 19: 342–347, 2001.

12. **Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, and Speed TP.** Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 31: e15, 2003.

13. **Karp CL, Grupe A, Schadt E, Ewart SL, Keane-Moore M, Cuomo PJ, Kohl J, Wahl L, Kuperman D, Germer S, Aud D, Peltz G, and Wills-Karp M.** Identification of complement factor 5 as a susceptibility locus for experimental allergic asthma. *Nat Immun* 1: 221–226, 2000.

14. **Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, Eizinger A, Wylie BN, and Davidson GS.** A gene expression map for *Caenorhabditis elegans*. *Science* 293: 2087–2092, 2001.

15. **Li C and Hung Wong W.** Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol* 2: RESEARCH0032, 2001.

16. **Li C and Wong WH.** Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci USA* 98: 31–36, 2001.

17. **Liu G, Loraine AE, Shigeta R, Cline M, Cheng J, Valmeekam V, Sun S, Kulp D, and Siani-Rose MA.** NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res* 31: 82–86, 2003.

18. **Lock C, Hermans G, Pedotti R, Brendolan A, Schadt E, Garren H, Langer-Gould A, Strober S, Cannella B, Allard J, Klonowski P, Austin A, Lad N, Kaminski N, Galli SJ, Oksenberg JR, Raine CS, Heller R, and Steinman L.** Gene-microarray analysis of multiple sclerosis lesions yields new targets validated in autoimmune encephalomyelitis. *Nat Med* 8: 500–508, 2002.

19. **Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, and Brown EL.** Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 14: 1675–1680, 1996.

20. **Mariani TJ, Budhraja V, Mecham BH, Gu CC, Watson MA, and Sadovsky Y.** A variable fold change threshold determines significance for expression microarrays. *FASEB J* 17: 321–323, 2003.
21. **Mariani TJ, Reed JJ, and Shapiro SD.** Expression profiling of the developing mouse lung: insights into the establishment of the extracellular matrix. *Am J Respir Cell Mol Biol* 26: 541–548, 2002.
22. **Mecham BH, Klus GT, Strovel J, Augustus M, Byrne D, Bozso P, Wetmore DZ, Mariani TJ, Kohane IS, and Szallasi Z.** Sequence-matched probe sets produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acids Res* 32: e74, 2004.
23. **Mody M, Cao Y, Cui Z, Tay KY, Shyong A, Shimizu E, Pham K, Schultz P, Welsh D, and Tsien JZ.** Genome-wide gene expression profiles of the developing mouse hippocampus. *Proc Natl Acad Sci USA* 98: 8862–8867, 2001.
24. **Nimgaonkar A, Sanoudou D, Butte AJ, Haslett JN, Kunkel LM, Beggs AH, and Kohane IS.** Reproducibility of gene expression across generations of Affymetrix microarrays. *BMC Bioinformatics* 4: 27, 2003.
25. **Oliver TG, Grasfeder LL, Carroll AL, Kaiser C, Gillingham CL, Lin SM, Wickramasinghe R, Scott MP, and Wechsler-Reya RJ.** Transcriptional profiling of the Sonic hedgehog response: a critical role for N-myc in proliferation of neuronal precursors. *Proc Natl Acad Sci USA* 100: 7331–7336, 2003.
26. **Pruitt KD and Maglott DR.** RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res* 29: 137–140, 2001.
27. **Pruitt KD, Tatusova T, and Maglott DR.** NCBI Reference Sequence project: update and current status. *Nucleic Acids Res* 31: 34–37, 2003.
28. **Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, Sturn A, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovich E, Borisovsky I, Liu Z, Vinsavich A, Trush V, and Quackenbush J.** TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* 34: 374–378, 2003.
29. **Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, and Friedman N.** Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* 34: 166–176, 2003.
30. **Signoretti S, Di Marcotullio L, Richardson A, Ramaswamy S, Isaac B, Rue M, Monti F, Loda M, and Pagano M.** Oncogenic role of the ubiquitin ligase subunit Skp2 in human breast cancer. *J Clin Invest* 110: 633–641, 2002.
31. **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, and Botstein D.** Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci USA* 100: 8418–8423, 2003.
32. **Stathopoulos A and Levine M.** Whole-genome expression profiles identify gene batteries in *Drosophila*. *Dev Cell* 3: 464–465, 2002.
33. **Strand AD, Olson JM, and Kooperberg C.** Estimating the statistical significance of gene expression changes observed with oligonucleotide arrays. *Hum Mol Genet* 11: 2207–2221, 2002.
34. **Tan PK, Downey TJ, Spitznagel EL Jr, Xu P, Fu D, Dimitrov DS, Lempicki RA, Raaka BM, and Cam MC.** Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res* 31: 5676–5684, 2003.
35. **Tanaka TS, Jaradat SA, Lim MK, Kargul GJ, Wang X, Grahovac MJ, Pantano S, Sano Y, Piao Y, Nagaraja R, Doi H, Wood WH, 3rd Becker KG, and Ko MS.** Genome-wide expression profiling of midgestation placenta and embryo using a 15,000 mouse developmental cDNA microarray. *Proc Natl Acad Sci USA* 97: 9127–9132, 2000.
36. **van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, and Friend SH.** Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415: 530–536, 2002.
37. **Verrecchia F, Chu ML, and Mauviel A.** Identification of novel TGF-beta/Smad gene targets in dermal fibroblasts using a combined cDNA microarray/promoter transactivation approach. *J Biol Chem* 276: 17058–17062, 2001.
38. **West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JA Jr, Marks JR, and Nevins JR.** Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci USA* 98: 11462–11467, 2001.
39. **White KP, Rifkin SA, Hurban P, and Hogness DS.** Microarray analysis of *Drosophila* development during metamorphosis. *Science* 286: 2179–2184, 1999.
40. **Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, and Speed TP.** Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* 30: e15, 2002.
41. **Zhao R, Gish K, Murphy M, Yin Y, Notterman D, Hoffman WH, Tom E, Mack DH, and Levine AJ.** The transcriptional program following p53 activation. *Cold Spring Harb Symp Quant Biol* 65: 475–482, 2000.
42. **Zuo F, Kaminski N, Eugui E, Allard J, Yakhini Z, Ben-Dor A, Lollini L, Morris D, Kim Y, DeLustro B, Sheppard D, Pardo A, Selman M, and Heller RA.** Gene expression analysis reveals matrilysin as a key regulator of pulmonary fibrosis in mice and humans. *Proc Natl Acad Sci USA* 99: 6292–6297, 2002.