Nucleic Acids Research, 2005, Vol. 33, No. 18 ***e157***
doi:10.1093/nar/gni159

Transformation of expression intensities across generations of Affymetrix microarrays using sequence matching and regression modeling

Soumyaroop Bhattacharya[1,2,*] and Thomas J. Mariani[1,2]

[1]Division of Pulmonary Medicine, Pulmonary Medicine, Thorn 908, Brigham and Women's Hospital and
[2]Pulmonary Bioinformatics, The Lung Biology Center, Harvard Medical School, Boston, MA 02115, USA

Received April 20, 2005; Revised July 11, 2005; Accepted September 25, 2005

ABSTRACT

The utility of previously generated microarray data is severely limited owing to small study size, leading to under-powered analysis, and failure of replication. Multiplicity of platforms and various sources of systematic noise limit the ability to compile existing data from similar studies. We present a model for transformation of data across different generations of Affymetrix arrays, developed using previously published datasets describing technical replicates performed with two generations of arrays. The transformation is based upon a probe set-specific regression model, generated from replicate measurements across platforms, performed using correlation coefficients. The model, when applied to the expression intensities of 5069 shared, sequence-matched probe sets in three different generations of Affymetrix Human oligonucleotide arrays, showed significant improvement in inter generation correlations between sample-wide means and individual probe set pairs. The approach was further validated by an observed reduction in Euclidean distance between signal intensities across generations for the predicted values. Finally, application of the model to independent, but related datasets resulted in improved clustering of samples based upon their biological, as opposed to technical, attributes. Our results suggest that this transformation method is a valuable tool for integrating microarray datasets from different generations of arrays.

INTRODUCTION

Microarrays have been widely used as exploratory tools for genome-wide analysis of gene expression and allow for simultaneous comparison of thousands of transcripts for detectable expression profile changes. The application of microarrays has not been limited only to differential gene expression studies. They have been used for gene detection, disease diagnosis, pharmaco- and toxicogenomics. Among the different formats of microarrays, those containing short oligonucleotides synthesized *in situ* using photolithography, developed by Affymetrix, have become the most common method of analysis. Affymetrix GeneChip arrays use sets of 25mer oligos as probes, with each set of probes (probe set) representing a gene/transcript. Expression measurements for sets of probes from individual probe sets are summarized, giving an estimate of the expression of the gene represented by the probe set. With continuing updates of mammalian genome sequence, oligonucleotide probe sequences are periodically changed in order to account for sequence accuracy and uniqueness (1,2). This requires the development of new generations of microarrays.

Gene Expression Omnibus (GEO) (3), a database curated and maintained by the National Institutes of Health (NIH), currently has an abundance of data for two versions of Affymetrix Human Genome (HG) chips; HG-U95Av2 (66 datasets) and HG-U133A (55 datasets). The HG-U95Av2 array represents ~10 000 full-length genes based on UniGene database (Build 95) and associated annotations. The HG-U133A 2.0 array represents 14 500 well-characterized human genes. Sequences used in the design of the array were selected from GenBank®, dbEST and RefSeq. The sequence clusters were created from the UniGene database (Build 133, April 20, 2001) and then were refined by analysis and comparison with a number of other publicly available databases including the Washington University EST trace repository and the University of California, Santa Cruz Golden-Path human genome database (April 2001 release) (Affymetrix Product Support, 2004). Additionally, a substantial number of microarray studies were performed using the Affymetrix HuGeneFL (aka HG-U6800) chip. This array, initially released by Affymetrix in November 1998, enables the relative monitoring of mRNA

*To whom correspondence should be addressed. Tel: +1 617 732 6265; Fax: +1 617 232 4623; Email: sbhattacharya@rics.bwh.harvard.edu

transcripts of ~5600 full-length human genes selected from UniGene Build 18 supplemented with additional genes from GenBank and TIGR.

Even with the latest advancements in microarray technology, the calculated gene expression value contains a substantial amount of noise and heteroscedascity, in part owing to the large number of observations and the wide range of gene expression values (4). Different empirical strategies devised for noise reduction include establishing a variable threshold for fold-changes (5), noise-filtering look up tables (6), non-parametric bootstrap for identification and correction for potential confounding effects (7) and other normalization techniques (8), as well as using technical and biological replicates to estimate the variability in gene expression (9). Application of fold-change thresholding has been by far the most commonly applied method of noise reduction. Technical variability, being smaller in comparison with biological variability, can be overcome by using adequate sample size, with large numbers of replicates (10,11).

Combining results from different studies and/or arrays has been difficult owing to an absence of standardized protocols for background correction, normalization and calculation of expression values. Meta-analysis approaches have been applied to microarray data in order to combine results from different labs, in the absence raw data from which the results have been derived. In certain cases, these approaches have proven very successful (12–14). However, when applied to microarray data, meta-analysis is complicated by additional statistical (e.g. multiple comparisons) and non-statistical issues (e.g. probe annotation). Some attempts have combined *P*-values across platforms (14–16), but such approaches can limit the capture of response magnitude (14) or biological significance (17). Unfortunately, most microarray studies suffer from lack of sufficient sample size. Datasets with sufficient numbers of controls and replicates are rarely generated under a proper experimental design, in order to make appropriate comparisons. Insufficient sample size limits biological insight and contributes to poor reproducibility across sample populations. There may even be substantial variation in the measured intensity levels for the same gene within the same generation of chips and for replicates of single tissue samples (6). Furthermore, unprocessed data files (e.g. CEL and .DAT files) are largely unavailable. Finally, the existence of differences among the many different technologies also makes it hard to combine the data. For instance, some studies have reported cross-platform comparisons and found varying degrees of agreement, ranging from the G–C content, sequence overlap and average signal intensity (2,18,19). Compiling previously published microarray datasets, from studies with similar experimental conditions and study design, can be an efficient way of improving the reliability of results and provide appropriate statistical power. Moreover, as the number of publicly available datasets in public data depositories [e.g. GEO (3); Stanford Microarray Database (20); and ArrayExpress at EBI (21)] grows, it is clear that these datasets should be combined to generate a more comprehensive understanding of underlying biology.

Two previous studies have performed comparisons across generations of Affymetrix arrays. A comparison of HuGeneFL (aka HG-U6800) and HG-U95Av2 arrays concluded that the reproducibility is high when the two probe sets share many exact probes and that it is low when they do not (22). Another comparison of HG-U95A and HG-U133A arrays concluded that data from different generations of microarrays can be combined by filtering the probes based on their sequence overlaps (23). The benefits of probe sequence matching upon measurement precision extend across multiple microarray platforms and technologies (1,2).

In the current study, we propose a method for utilizing available datasets by combining signals derived from different generations of Affymetrix arrays. We developed the methodology using the two previously described intergeneration comparisons of Affymetrix Human arrays (http://www.chip.org/~ashish/Reproducibility and http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE2044) (22,23). First, we identified probe sets that have significant sequence overlap between three generations of Affymetrix human arrays (HuGeneFL, HG-U95Av2 and HG-U133A). Next, we derived regression models for each of the sequence-matched probe sets based on the pairwise correlations from replicate measurements. Finally, we developed an algorithm for transformation of expression intensities obtained from one generation of arrays to another. These predicted values resulted in a significant improvement in correlations and reduction in Euclidian distance in replicate measurements across generations. We suggest this regression-based approach be taken for intergeneration conversion of signal intensities and recommend that this strategy be limited to genes represented by sequence-matched probe sets.

METHODS

Microarray data

Data files were obtained from four previous studies: Hwang *et al*. (23) compared expression in 14 human inflammatory myopathy samples. Total RNA was extracted and a portion was hybridized to HG-U95Av2 arrays while the remaining RNA was frozen and then later hybridized to HG-U133A arrays at the same core facility. Nimgaonkar *et al*. (22) compared expression in seven human muscle samples concurrently hybridized to HuGeneFL and HG-U95Av2 chips. Beer *et al*. (24) analyzed gene expression in 104 human lung tissue samples, including 87 adenocarcinomas and 17 normal lungs, hybridized to HuGeneFL arrays. Bhattacharjee *et al*. (25) analyzed gene expression in 203 human lung tissue samples including adenocarcinomas ($n = 127$), squamous cell carcinomas ($n = 21$), pulmonary carcinoids ($n = 20$), small cell lung cancer ($n = 6$) and normal tissue ($n = 17$) specimens using the HG-U95Av2. We limited our analysis of the Bhattacharjee *et al*. (25) dataset to the adenocarcinomas and normal lung samples which are comparable in sample with those used by Beer *et al*. (24). The raw image files were obtained and processed using MAS 5.0 and the non-normalized, background subtracted signal intensities were extracted.

Probe set matching

The annotation information for each probe set in HuGeneFL, HG-U95Av2 and HG-U133A arrays was retrieved from NetAffx Analysis Center (26). According to the annotation information, HuGeneFL has 7129 probe sets annotated from

5600 well-documented genes, and U95Av2 has 12 625 probe sets, annotated by 9091 UniGene and 8672 LocusLink identifiers. The U133A array is composed of 22 283 probe sets annotated by 13 624 UniGene and 12 769 LocusLink identifiers.

We limited our analysis to those probe sets with the greatest likelihood of representing the same gene across generations by implementing the Affymetrix 'Best Match' algorithm. For sequence comparisons, the best match from the HG-U95Av2 to the HuGeneFL, represents instances where the probability of the two sequences match randomly (P-value < e^{-40}) or percent overlap is >90% over at least 70 bases. Additionally, probe sets were matched if they were identical across the platforms (Affymetrix provides a list of the numbers of probe pairs common for the two generations) (http://www.affymetrix.com/analysis/index.affx).

For comparison of HG-U95Av2 and HG-U133A, Affymetrix provides the probe set matches for comparative analysis. These matching tables were constructed based on the sequence information of probe sets as follows: to begin with, all possible probe set pairs between two generations were checked by their similarity in the representative sequence for selection. Then, among these matching probe set pairs, some were selected as 'Good Match' pairs based on three criteria: (i) if the percent identity between the representative sequences was >90%, (ii) the length of the representative sequence was >100 bp and (iii) at least one perfect match probe of one array generation was perfectly aligned to the probe selection region of the other array generation. Further, a 'Best Match' subset was selected by more stringent criteria on the similarity of probe set pairs. Finally, an overlap of the best match probe sets of both comparisons was obtained. Subsequent use of the term 'Best Match' represents this subset of matching probe sets.

Correlation coefficients

The correlation coefficient was used as a metric of congruency of measurement between matched probe sets across the two generations of Affymetrix arrays. First, the signal intensity data were normalized to a mean of 0 and SD of 1. To assess the cross-generation correlation of gene-expression measurements we computed both the Pearson linear correlation coefficients and Spearman rank-order correlation coefficients for best match probe sets for each array pair. In addition, the expression measure of a probe set for a given platform was computed as the mean of the all the samples for a given array.

Regression model

In order to develop a model for transformation of expression intensities, a linear regression model was developed. For each of the probe sets, as well as for the sample wide means, the expression intensities of replicate (same sample) hybridizations on different generations of arrays was modeled. A representative equation for the linear regression model for a particular probe set can be written as

$$\alpha_i X_{ij} + \beta_i = Y_{ij} \quad \mathbf{1}$$

where X_{ij} = expression intensity of probe set i for sample j in array 1, Y_{ij} = expression intensity of the same probe set in the same sample in array 2, α = slope and β = intercept (for one dataset). The regression models developed were used to generate a modified regression model that would predict the expression intensities of array 2 using the intensities in array 1 based on the expression intensities from one dataset. This process was repeated for the second dataset. Further we also developed two more regression models, one for each of the two datasets. In this case X represented the sample-wide mean expression intensity of all probe sets in array 1 and Y is the sample-wide mean expression intensity of the all probe sets in array 2, α = slope and β = intercept.

We have developed a model based on the linear regression of sample-wide means of expression intensities from different generations of arrays along with the regression model for the expression intensities of all the probe sets across two generations. First we generated a model for sample-wide means for all best match probe sets for HG-U95Av2 and HG-U133A arrays from the Hwang *et al*. (23) dataset

$$Y = a_0 X + b_0 \quad \mathbf{2}$$

where Y represents the outcome variable, which in this case is the sample-wide mean for all 5069 best match probe sets in HG-U95Av2 array while X represents the predictor variable which is the sample-wide mean for the same probe sets in HG-U133A array. a_0 and b_0 represent the slope and intercept, respectively, for the model.

We also generated a set of models, one representing each of the 5069 best match probe sets from the Hwang *et al*. (23) dataset

$$Y_{nm} = a_n X_{nm} + b_n \quad (n = 1, 2, \ldots, 5069;\ m = 1, 14) \quad \mathbf{3}$$

where Y_{nm} represents the expression intensity for each of individual probe sets in samples 1–14 hybridized to HG-U95Av2 arrays while X_{nm} is the expression intensity for the same probe sets for samples (1–14) hybridized to HG-U133A array. a_n and b_n represent the slopes and intercepts, respectively, for each model. This procedure provided us with 5069 individual regression models one for each of the best match probe sets.

Next we developed a model for sample-wide means for the 5069 best match probe sets in HG-U95Av2 and HuGeneFL arrays from the Nimgaonkar *et al*. (22) dataset

$$Y = c_0 X + d_0 \quad \mathbf{4}$$

where Y represents the outcome variable, which in this case is the sample-wide mean for all 5069 best match probe sets in HuGeneFL array while X represents the predictor variable which is the sample-wide mean for the same probe sets in HG-U95Av2 array. c_0 and d_0 represent the slope and intercept, respectively, for the model.

We also generated a set of models, one representing each of the 5069 best match probe sets from Nimgaonkar *et al*. (22) dataset

$$Y_{nm} = c_n X_{nm} + d_n \quad (n = 1, 2, \ldots, 5069;\ m = 1, \ldots, 7) \quad \mathbf{5}$$

where Y_{nm} represents the expression intensity for each of individual probe sets in samples 1–7 hybridized to HuGeneFL arrays while X_{nm} is expression intensity for the same probe sets for samples (1–7) hybridized to HG-U95Av2 array. c_n and d_n represent the slopes and intercepts, respectively, for each model. This procedure provided us with 5069 individual regression models one each of the best match probe sets.

We used the slopes and intercepts from the sample-wide means (overall) model as additive correctional factors in developing our inter-generational transformation model for the individual probe sets. The model for converting HG-U95Av2 values to HG-U133A values [based on Hwang *et al.* (23) data] for each probe set can be represented as

$$Y'_{nm} = (a_n + a_0)X_{nm} + (b_n + b_0) \quad (n = 1, 2, \ldots, 5069;\ m = 1, 14) \qquad \mathbf{6}$$

where Y'_{nm} is the predicted value for HG-U133A expression intensities for each of the 5069 best match probe sets, while X_{nm} represents the expression intensity for the same probe sets for samples HG-U95Av2 array. The slope in the model is represented by sum of a_0 and a_n, where a_0 serves as a correction factor while intercept is the sum of b_0 and b_n with b_0 serving as the correction factor in this case.

In the same fashion, we derived the model for transformation from HuGeneFL to HG-U95Av2 [based on Nimgaonkar *et al.* (22)]

$$Y'_{nm} = (c_n + c_0)X_{nm} + (d_n + d_0) \quad (n = 1, 2, \ldots, 5069;\ m = 1, 7) \qquad \mathbf{7}$$

where Y'_{nm} is the predicted value for HG-U95Av2 expression intensities of individual probe sets, while X_{nm} represents the expression intensity for the probe sets for samples HuGeneFL array. The slope in the model is represented by sum of c_0 and c_n, while intercept is the sum of d_0 and d_n. In absence of a dataset, we could not develop a model for direct transformation of HuGeneFL signal intensities to HG-U133A values.

For the sake of maintaining uniformity we present a universal equation for all inter-generational transformation among Affymetrix arrays that can be represented as

$$(\alpha_i + \alpha_o)X_{ij} + (\beta_i + \beta_o) = Y_{ij} \qquad \mathbf{8}$$

where X_{ij} is the expression intensity of probe set i for sample j in array available while Y_{ij} is the expression intensity of the same probe set in the same sample in array to which the values have to be converted to, α_i is probe set slope and α_o is the correction factor, while β_i is the intercept for the probe sets and β_0 is the correction factor.

Hierarchical clustering

Average linkage agglomerative hierarchical clustering based upon Euclidean distance was implemented using the Gene Expression Data Analyzer (27).

RESULTS

We identified 6375 probe sets among the matching probe set pairs between HuGeneFL and HG-U95Av2, which met all the criteria for best match between the two arrays. Similarly, we identified 10 508 probe sets (representing 9000 unique genes) among the overlapping probe set pairs between HG-U95Av2 and HG-U133A, which met all the criteria for best match between the two arrays. We further identified overlaps among the above two best match sets and focused our analysis on only these 5069 probe sets (Figure 1).

To assess the similarity among different generations of arrays, we examined the correlation of standardized expression intensities of replicate measurements, obtained using the Affymetrix MAS 5.0 algorithm. The correlation coefficient

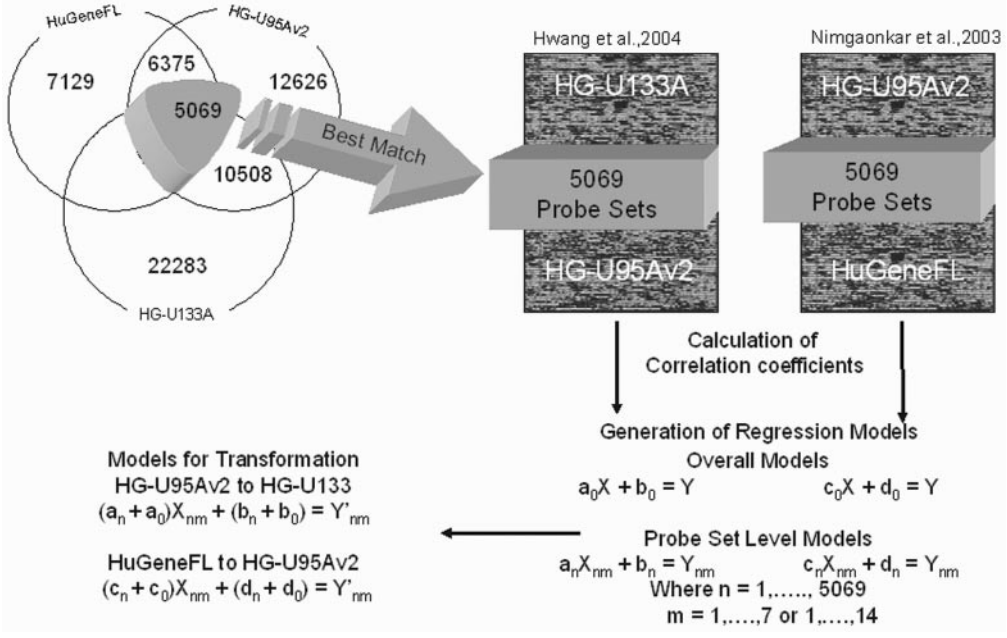

Figure 1. A schematic outline of the procedure. Sequence-matching identified 5069 'best match' probe sets shared across three generations of human microarrays. The expression intensities of those 5069 probe sets from the two datasets were used to compute correlation coefficients. Regression models were generated using expression values of paired probe sets and the sample-wide mean expression intensities. Final models for transformation were developed by applying correction factors to the above regression models. 'Transformed' expression measures were calculated for the 5069 probe sets. In the above regression equations, a and c represent the slopes while b and d represent the intercepts for the two models.

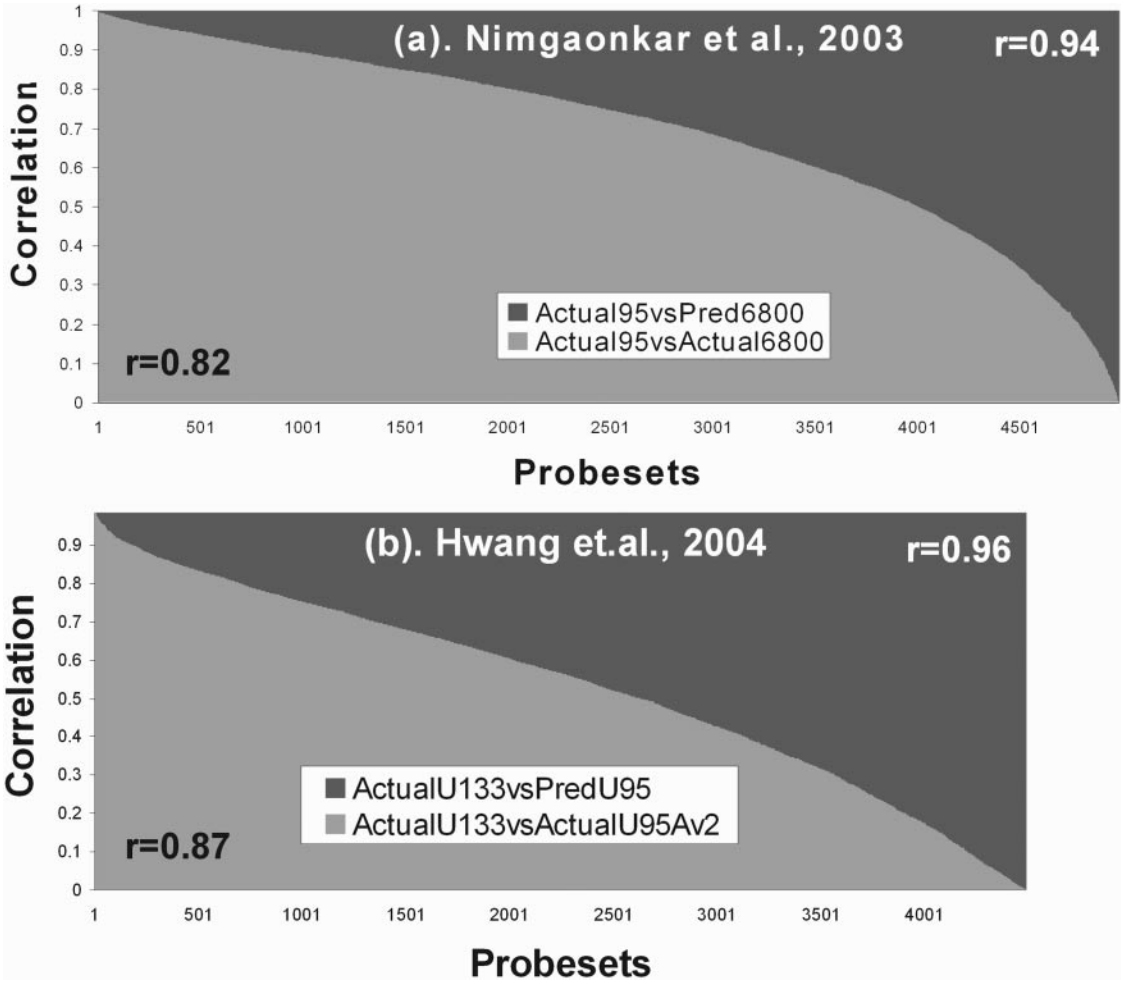

Figure 2. Transformation improves probe set level correlations. A histogram displaying the correlation in expression intensity across two platforms for each sequence matched probe set ranked in descending order. The correlation of individual probe sets for both datasets (**a**) Nimgaonkar *et al*. (22) and (**b**) Hwang *et al*. (23) show significant increase after modeling and transformation of the expression intensities.

between the sample-wide means for all probe sets between HuGeneFL and HG-U95Av2 was 0.82 with a range from 0.9946 to −0.1357 for each of the individual probe sets. Between HG-U95Av2 and HG-U133 the correlation coefficient was 0.87 with a range of 0.9921 to −0.115 for individual probe sets. After modeling and transformation, correlation values increased significantly when calculated using the predicted expression intensity measurements (Figure 2). We observed a correlation of 0.942 between the sample-wide means of the actual HG-U95Av2 and the transformed HuGeneFL intensity values predicted using our regression model. When compared with the sample-wide means of actual HuGeneFL signal intensities, the predicted HuGeneFL values had a correlation of 0.9457. The pairwise correlations for individual probe sets between HG-U95Av2 and the predicted HuGeneFL in most of the cases increased to 1. Similarly, the correlation between sample-wide means of actual HG-U133A and HG-U95Av2 values predicted using our regression model improved to 0.96. The pairwise correlations for individual probe sets between HG-U133A and the predicted values of HG-U95Av2 also increased to 1 in most cases. For both datasets, in some cases the correlation for individual probe sets in the transformed data was −1, owing to a negative correlation prior to transformation. This was limited to a small number of probe sets [∼3.6% in Nimgaonkar *et al*. (22), and ∼12% in Hwang *et al*. (23)]. The sample-wide means correlation between actual and predicted HuGeneFL was 0.962 and between the actual and predicted values of HG-U95Av2 was 0.973. The overall sample-wide mean correlation was improved after transformation, in the HuGeneFL to U95Av2 and the U95Av2 to U133A comparisons. Spearman rank correlations calculated for individual probe sets were also increased (data not shown), but the data were of limited power owing to low sample sizes of 7 and 14 in the two comparisons.

To further assess the improvements in the transformed data, we calculated the distances in Euclidean space among the probe set pairs between actual and predicted values. The Euclidean distance between actual expression intensity values for probe sets between HuGeneFL and HG-U95Av2 had a mean of 0.82 U (range from 0.0915 to 22 U) prior to transformation, and a mean of 0.79 (range from 0.077 to 22 U) after transformation (Figure 3a). Similarly, the distance between actual signal values between HG-U95Av2 and HG-U133A had a mean of 1.49 (range from 0.0095 to 42) before transformation, and a mean of 1.06 (range from 0.0047 to 29) after

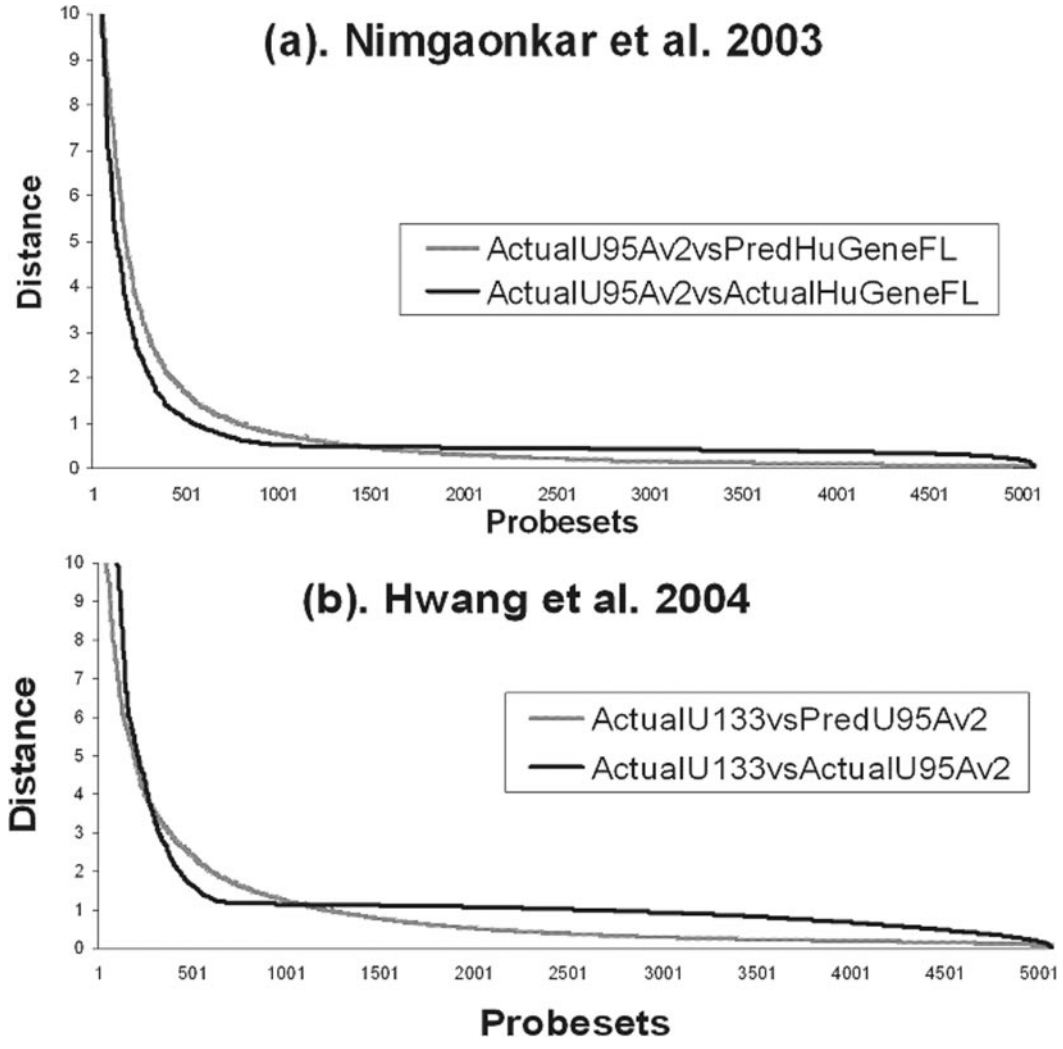

Figure 3. Transformation increases similarity in expression measure for sequence-matched probe sets. A line diagram displaying the Euclidean distance between signal intensity measurements for matched probe sets across the two platforms, ranked in descending order. The Euclidean distances of individual probe sets for both datasets (**a**) Nimgaonkar *et al*. (22) and (**b**) Hwang *et al*. (23) show a significant decrease after modeling and transformation of expression intensities.

transformation (Figure 3b). The overall sample-wide mean Euclidean distance was significantly improved after transformation, with a P-value < 0.001 in both the HuGeneFL to U95Av2 and the U95Av2 to U133A comparison.

Next, in order to further define the success of our model to transform data from one array platform to another, we implemented average linkage hierarchical clustering of the datasets using the 5069 best match probe sets (Figure 4). Clustering of pre-transformation datasets resulted in complete separation of the samples based entirely upon the platform on which the data was generated (Figure 4a and c). We applied our transformation procedure to the data and repeated the clustering, resulting in partial, but incomplete, mixing of samples from different generation of arrays. The absence of complete clustering based upon sample type in the transformed data was possibly owing to residual noise in the datasets. Therefore, we took a standard approach for noise reduction of trimming (thresholding) the transformed dataset and repeated the clustering procedure. For the HuGeneFL to HG-U95Av2 dataset, clustering of 2020 probe sets with a minimum correlation coefficient of $r = 0.80$, resulted in complete separation of samples dependent upon sample type, and independent of platform (Figure 4b). Similar results were found for the HG-U95Av2 and HG-U133A dataset, when clustering 1263 probe sets with a minimum $r = 0.74$ (Figure 4d). For hierarchical clustering, we are using Euclidean distance, and not the correlation, as post scaling squared Euclidean distance is equivalent to correlation distance (28). It has been shown in the past that Euclidean distance performed better than Pearson's correlation coefficient when it comes to the clustering of samples (29).

Clearly, our method of sequence-matching and linear regression modeling of microarray data can transform replicate samples run on multiple platforms. In an effort to determine how our model would function on archived datasets describing related, but independent biological samples, we applied the transformation procedure to two independent datasets (24,25) containing data from human lung specimens. We separated normal tissue samples from tumor samples in order to better test the ability of the transformation procedure to relate biologically similar samples. We merged the MAS 5.0 expression intensities generated *de novo* from the raw data (.CEL) files from both datasets, and performed

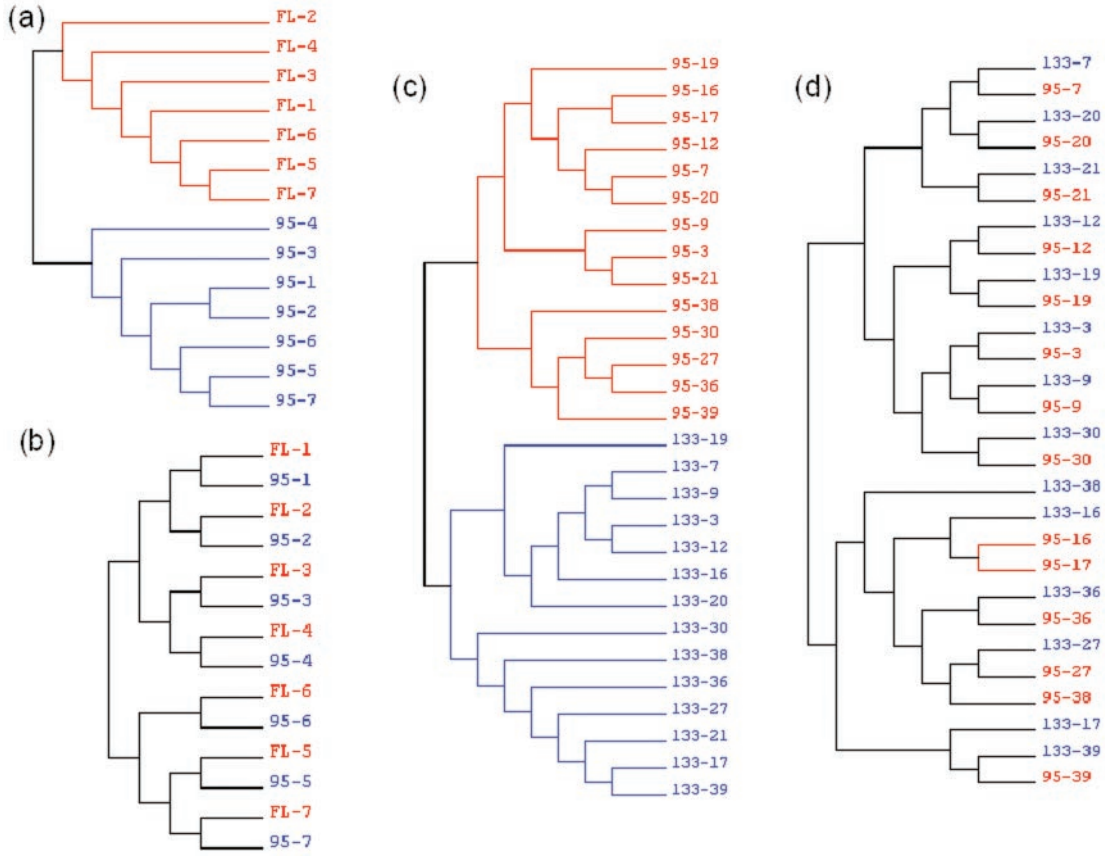

Figure 4. Transformation results in sample-related, instead of platform-related clustering. Average linkage hierarchical clustering was performed on the data from Nimgaonkar *et al*. (22) (**a** and **b**) and Hwang *et al*. (23) (**c** and **d**) before (a and c) and after (b and d) data transformation. The datasets cluster dependent upon the platform used prior to transformation, but cluster dependent upon the biological sample after transformation.

hierarchical clustering (Figure 5a and c). Before transformation, the signal intensities from the 5069 genes showed a sample-wide mean correlation of 0.78 for normal lung tissue samples and 0.82 for adenocarcinoma samples. As expected, the samples formed clusters based entirely upon the array generation. Next, we applied our transformation procedure to convert the signal values from the HuGeneFL dataset (24) to their corresponding HG-U95Av2 values (25). Post-transformation signal intensities showed a sample-wide mean correlation of 0.82 for normal lung tissue samples and 0.94 for adenocarcinoma samples. For the normal samples, clustering of 745 probe sets with minimum correlation coefficient of $r = 0.40$, resulted in platform-independent clustering of samples (Figure 5b). Similar platform-independent sample mixing was observed for adenocarcinoma samples when clustering 584 probe sets with a minimum $r = 0.11$ (Figure 5d).

DISCUSSION

With an increasing number of microarray datasets being deposited in the public domain, real opportunity exists for more reliable information to be generated through the integration of multiple, independently generated datasets focusing on the same biological paradigm. This possibility is hindered by the fact that there is large amount of technical variability associated with microarray data. This is due to a number of factors including, but not limited to, differences in technical platforms, differences in data processing methods and probes sequence errors. Methods that can overcome these problems and allow compilation of diverse datasets will have a profound impact upon the future of this technology.

In this study we aimed to develop a model for transforming expression intensity values obtained from one generation of Affymetrix oligonucleotide array to another. We used three generations of human arrays HuGeneFL, HG-U95Av2 and HG-U133A. Even though all three arrays belonged to the same technological platform type (short oligonucleotide array), there were significant differences among the three generations; such as different sources of probe selection regions (UniGene Builds 18, 95 and 133 for HuGeneFL, U95Av2 and U133A, respectively), different numbers of probe pairs per probe set (20, 16 and 11 in HuGeneFL, U95Av2 and U133A, respectively), and different probe selection methods (30).

Recently, many studies have highlighted the benefit of probe sequence matching in improving the consistency of data from the same biological paradigm generated from multiple microarray platforms. For cross-platform comparison studies, matching of UniGene IDs is the most common method used. However, with each update of the UniGene database, few old tags are removed and new tags are added, and these are hard to track correctly unless the same build of UniGene was used to annotate each platform. Using LocusLink IDs for

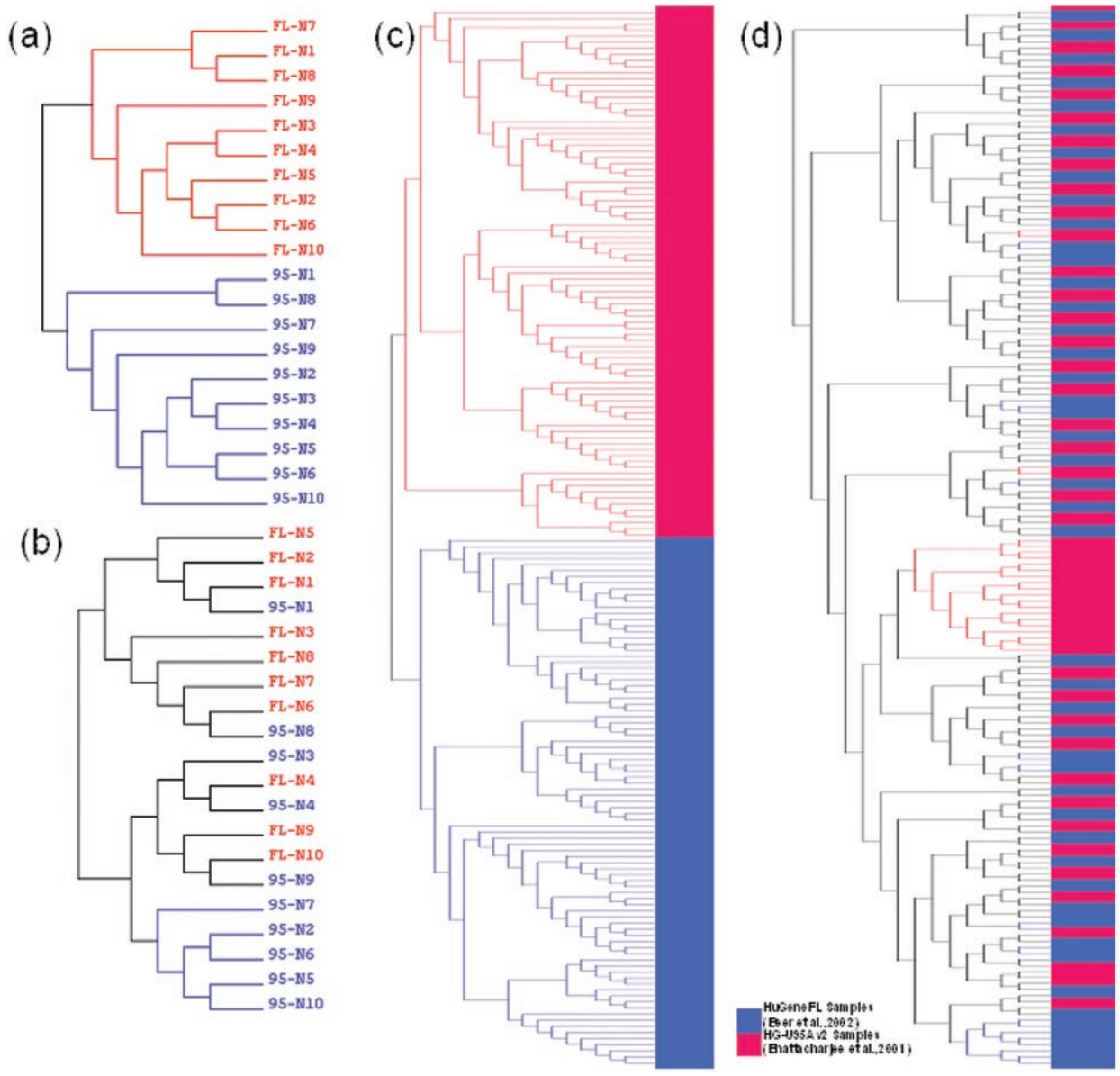

Figure 5. Transformation results in platform-independent clustering of archived datasets describing human lung cancer. Average linkage hierarchical clustering was performed on the human lung tissue data from Beer *et al*. (24) and Bhattacharjee *et al*. (25). Samples from normal lung tissue (**a** and **b**) and adenocarcinoma tissue (**c** and **d**) were analyzed before (a and c) and after (b and d) data transformation. The datasets cluster dependent upon the platform used prior to transformation, but cluster dependent upon the biological sample type after transformation.

matching the genes has been used as an alternative. Recently, we showed a large statistical benefit in measurement precision within and across platforms by limiting analysis to data generated from sequence validated or sequence-matched probes (1,2). Another previous study (23) compared three methods for matching probe sets between two generations of arrays; UniGene IDs and LocusLink IDs, and best match probe sets provided by Affymetrix (Methods). This study concluded that UniGene and LocusLink matching gave similar results, but best match provided higher reproducibility than other matching methods. We therefore used the 'Best Match' algorithm, as outlined in Methods section above, as the criteria for the selection of matching probe sets. These criteria yielded a set of 5069 probe sets that were common in HuGeneFL, HG-U95Av2 and HG-U133A (Figure 1).

Nimgaonkar *et al*. (22) compared the correlations of expression intensities across two generations of the Affymetrix arrays, HuGeneFL and HG-U95Av2. They reported 2200 (27%) of 8044 matched probe sets had negative correlations, i.e. the gene expression patterns changed in opposite directions between the two generations (22). We found the number of probe sets showing negative correlations was considerably reduced using the Best Match criteria (181 probe sets in 5070 common probe sets, ~3.6%). Therefore, sequence-matching using the Best Match criteria (or other sequence matching method) during data preprocessing is an important step for combinatorial analysis in order to reduce the technical variance between platforms. Obviously, limiting the datasets by sequence matching results in the loss of some potentially significant information. For example, only ~50% of the probe

sets in HG-U95Av2 and ~23% of HG-U133A arrays were used in the analysis. This limitation is unavoidable in order to retain the highest level of confidence in the data derived from combining datasets from different arrays. However, the limitation must be contrasted with the advantage of being able to both combine datasets with different generation of arrays and retain the highest level of confidence in the results derived.

Affymetrix microarray technology consistently shows high technical reproducibility, typically in the range of >0.90 (31). The mean correlation in expression for the 5069 best match probe sets were 0.82 and 0.87 when comparing data generated using different generations of arrays, much lower than that reported in past studies for technical replicates using one platform. This difference in correlation coefficients is a measure of the loss of precision when integrating datasets across generations of the same platform. In a previous study involving two different generations of Affymetrix Arabidopsis arrays, the average inter-generation correlation was reported to be 0.81 (32), very similar to our observations. However, this value can be considerably improved by the modeling procedure we describe. Our sequence matching and data modeling methods transformed a vast majority [96% in Nimgaonkar *et al*. (22) and 88% in Hwang *et al*. (23)] of the correlations for sequence matched probe sets to 1, and also produced a significant decrease in Euclidean distance between signal intensities for corresponding probe sets, indicating the success of the procedure.

The transformation procedure we describe here clearly does not remove all components of variation or noise from the dataset. This is observed in the inability to generate ideal clustering (evidenced by mixing of samples from different platforms) using data from all 5069 probe sets, even after transformation. Improved mixing was observed after trimming the datasets in an effort to remove residual noise. One rationale for the remaining noise, even following transformation, is that our model affects the scale, but not the distribution of the datasets. It is commonly observed that expression intensities derived from different arrays of the same platform are scaled and distributed differently (22,23,33). Jiang *et al*. (33) further showed that even after normalization, the scales and distributions of the two datasets remained different. This is attributed to the difference in the hybridization signal intensities between two platforms. In our analysis HuGeneFL had highest overall intensity and lowest variance in comparison with HG-U95Av2 and HG-U133A. These differences may be due to either variations in photo-multiplier tube settings used for data acquisition associated with different arrays, or to differences in probe numbers and the sequences for each gene. As a result, data distributions varied greatly between the pre- and post-transformation versions of the dataset.

Pre-transformation clustering of the replicate datasets resulted in separation of samples according to the platform, while post-transformation clustering resulted in separation according to sample type. Of greatest importance, we observed similar results with independent human lung datasets, where samples were not replicates, but biologically related. This sample type-dependent clustering was improved after removing (thresholding) data from probe sets providing low-quality signals, as determined by the low correlations in expression across the datasets. The success achieved in clustering non-replicate samples from two independent sources validates this model as a unique and rational method for integration of microarray data across platforms.

One of the primary concerns of using linear regression models to generate expression intensities has been that although these models scale the values, they provide no added benefit to analyses of differential expression. Although microarray data can be used for many purposes, such as patient classification (class discovery and class prediction), a major application of the data is to define differential expression. As previously stated, the method described here is most suited to datasets of limited size and statistical power, which describes a vast majority of those available. In such cases, meta-analysis is of limited value owing to the low level of reliability of the individual datasets combined with the concerns of translating gene expression measurements and probe annotation across platforms. Data compilation improves the statistical power of small datasets by increasing the number of observations. The sequence matching and regression modeling approach described here overcomes these problems and allows for the reliable compilation of data from independent platforms. Furthermore, initial results (S. Bhattacharya and T. J. Mariani, unpublished data) indicate that this method leads to increased sensitivity, specificity and reliability when defining differential expression.

CONCLUSION

We have developed an approach for inter-generational transformation of gene expression microarray signal intensity values. This method relies upon; (i) sequence matching of probes across platforms to ensure measurement of the same biological variable, and (ii) linear regression modeling of data derived from technical replicates performed on more than one platform. Our results suggest that success in modeling gene expression microarray data across platforms is achievable. Continuing efforts should lead to dramatic improvements in inter-generation transformation of microarray data.

ACKNOWLEDGEMENTS

The authors would like acknowledge Ashish Nimgaonkar, Peter Park and Sek Won Kong for making the data available. We would also like to thank Dawn Simon, Sorachai Srisuma and Temana Andalcio for reviewing the manuscript. This work was supported by NIH grants HL071885 and HL072303. Funding to pay the Open Access publication charges for this article was provided by NIH Grant HLO71885.

Conflict of interest statement. None declared.

REFERENCES

1. Mecham,B.H., Wetmore,D.Z., Szallasi,Z., Sadovsky,Y., Kohane,I. and Mariani,T.J. (2004) Increased measurement accuracy for sequence-verified microarray probes. *Physiol. Genomics*, **18**, 308–315.
2. Mecham,B.H., Klus,G.T., Strovel,J., Augustus,M., Byrne,D., Bozso,P., Wetmore,D.Z., Mariani,T.J., Kohane,I.S. and Szallasi,Z. (2004) Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acids Res*., **32**, e74.

3. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*., **30**, 207–210.
4. Lee,J.K. (2001) Analysis issues for gene expression array data. *Clin. Chem*., **47**, 1350.
5. Mariani,T.J., Budhraja,V., Mecham,B.H., Gu,C.C., Watson,M.A. and Sadovsky,Y. (2003) A variable fold change threshold determines significance for expression microarrays. *FASEB J*., **17**, 321–323.
6. Mills,J.C. and Gordon,J.I. (2001) A new approach for filtering noise from high-density oligonucleotide microarray datasets. *Nucleic Acids Res*., **29**, E72-2.
7. Bhattacharya,S., Long,D. and Lyons-Weiler,J. (2003) Overcoming confounded controls in the analysis of gene expression data from microarray experiments. *Appl. Bioinformatics*, **2**, 197–208.
8. Kerr,M.K., Martin,M. and Churchill,G.A. (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol*., **7**, 819–837.
9. Lee,M.L., Kuo,F.C., Whitmore,G.A. and Sklar,J. (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl Acad. Sci. USA*, **97**, 9834–9839.
10. Zien,A., Fluck,J., Zimmer,R. and Lengauer,T. (2003) Microarrays: how many do you need? *J. Comput. Biol*., **10**, 653–667.
11. Kendziorski,C., Irizarry,R.A., Chen,K.S., Haag,J.D. and Gould,M.N. (2005) On the utility of pooling biological samples in microarray experiments. *Proc. Natl Acad. Sci. USA*, **102**, 4252–4257.
12. Grutzmann,R., Boriss,H., Ammerpohl,O., Luttges,J., Kalthoff,H., Schackert,H.K., Kloppel,G., Saeger,H.D. and Pilarsky,C. (2005) Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. *Oncogene*, **24**, 5079–5088.
13. Irizarry,R.A., Warren,D., Spencer,F., Kim,I.F., Biswal,S., Frank,B.C., Gabrielson,E., Garcia,J.G., Geoghegan,J., Germino,G. *et al*. (2005) Multiple-laboratory comparison of microarray platforms. *Nature Methods*, **2**, 345–350.
14. Rhodes,D.R. and Chinnaiyan,A.M. (2004) Bioinformatics strategies for translating genome-wide expression analyses into clinically useful cancer markers. *Ann. N. Y. Acad. Sci*., **1020**, 32–40.
15. Ghosh,D., Barette,T.R., Rhodes,D. and Chinnaiyan,A.M. (2003) Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer. *Funct. Integr. Genomics*, **3**, 180–188.
16. Rhodes,D.R., Barrette,T.R., Rubin,M.A., Ghosh,D. and Chinnaiyan,A.M. (2002) Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res*., **62**, 4427–4433.
17. Stevens,J.R. and Doerge,R.W. (2005) Combining Affymetrix microarray results. *BMC Bioinformatics*, **6**, 57.
18. Kuo,W.P., Jenssen,T.K., Butte,A.J., Ohno-Machado,L. and Kohane,I.S. (2002) Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics*, **18**, 405–412.
19. Mitchell,S.A., Brown,K.M., Henry,M.M., Mintz,M., Catchpoole,D., LaFleur,B. and Stephan,D.A. (2004) Inter-platform comparability of microarrays in acute lymphoblastic leukemia. *BMC Genomics*, **5**, 71.
20. Gollub,J., Ball,C.A., Binkley,G., Demeter,J., Finkelstein,D.B., Hebert,J.M., Hernandez-Boussard,T., Jin,H., Kaloper,M., Matese,J.C. *et al*. (2003) The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res*., **31**, 94–96.
21. Brazma,A., Parkinson,H., Sarkans,U., Shojatalab,M., Vilo,J., Abeygunawardena,N., Holloway,E., Kapushesky,M., Kemmeren,P., Lara,G.G. *et al*. (2003) ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res*., **31**, 68–71.
22. Nimgaonkar,A., Sanoudou,D., Butte,A.J., Haslett,J.N., Kunkel,L.M., Beggs,A.H. and Kohane,I.S. (2003) Reproducibility of gene expression across generations of Affymetrix microarrays. *BMC Bioinformatics*, **4**, 27.
23. Hwang,K.B., Kong,S.W., Greenberg,S.A. and Park,P.J. (2004) Combining gene expression data from different generations of oligonucleotide arrays. *BMC Bioinformatics*, **5**, 159.
24. Beer,D.G., Kardia,S.L., Huang,C.C., Giordano,T.J., Levin,A.M., Misek,D.E., Lin,L., Chen,G., Gharib,T.G., Thomas,D.G. *et al*. (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med*., **8**, 816–824.
25. Bhattacharjee,A., Richards,W.G., Staunton,J., Li,C., Monti,S., Vasa,P., Ladd,C., Beheshti,J., Bueno,R., Gillette,M. *et al*. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl Acad. Sci. USA*, **98**, 13790–13795.
26. Liu,G., Loraine,A.E., Shigeta,R., Cline,M., Cheng,J., Valmeekam,V., Sun,S., Kulp,D. and Siani-Rose,M.A. (2003) NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res*., **31**, 82–86.
27. Lyons-Weiler,J., Patel,S. and Bhattacharya,S. (2003) A classification-based machire learning approach for the analysis of genome-wide expression data. *Genome Res*., **13**, 503–512.
28. Chipman,H., Hastie,T. and Tibshirani,R. (2003) *Clustering Microarray Data. Statistical Analysis of Gene Expression Microarray Data*. Chapman and Hall, CRC Press, Florida, USA.
29. Knudsen,S. (2004) *Guide to Analysis of DNA Microarray Data*. 2nd edn. Wiley-VCH Verlag GmbH & Co. KGaA.
30. Mei,R., Hubbell,E., Bekiranov,S., Mittmann,M., Christians,F.C., Shen,M.M., Lu,G., Fang,J., Liu,W.M., Ryder,T. *et al*. (2003) Probe selection for high-density oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **100**, 11237–11242.
31. Baugh,L.R., Hill,A.A., Brown,E.L. and C.P.,Hunter (2001) Quantitative analysis of mRNA amplification by *in vitro* transcription. *Nucleic Acids Res*., **29**, E29.
32. Hennig,L., Menges,M., Murray,J.A. and Gruissem,W. (2003) *Arabidopsis* transcript profiling on Affymetrix GeneChip arrays. *Plant Mol. Biol*., **53**, 457–65.
33. Jiang,H., Deng,Y., Chen,H.S., Tao,L., Sha,Q., Chen,J., Tsai,C.J. and Zhang,S. (2004) Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *BMC Bioinformatics*, **5**, 81.