Selecting Reliable mRNA Expression Measurements Across Platforms Improves Downstream Analysis

Pan Tong[1,*], Lixia Diao[1,*], Li Shen[1], Lerong Li[1], John Victor Heymach[2], Luc Girard[3], John D. Minna[4], Kevin R. Coombes[5], Lauren Averett Byers[2] and Jing Wang[1]

[1]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [2]Department of Thoracic and Head and Neck Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [3]Department of Pharmacology, University of Texas Southwestern Medical Center, Dallas, TX, USA. [4]Department of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, TX, USA. [5]Department of Medical Informatics, The Ohio State University, Columbus, OH, USA. *These authors contributed equally to this work.

ABSTRACT: With increasing use of publicly available gene expression data sets, the quality of the expression data is a critical issue for downstream analysis, gene signature development, and cross-validation of data sets. Thus, identifying reliable expression measurements by leveraging multiple mRNA expression platforms is an important analytical task. In this study, we propose a statistical framework for selecting reliable measurements between platforms by modeling the correlations of mRNA expression levels using a beta-mixture model. The model-based selection provides an effective and objective way to separate good probes from probes with low quality, thereby improving the efficiency and accuracy of the analysis. The proposed method can be used to compare two microarray technologies or microarray and RNA sequencing measurements. We tested the approach in two matched profiling data sets, using microarray gene expression measurements from the same samples profiled on both Affymetrix and Illumina platforms. We also applied the algorithm to mRNA expression data to compare Affymetrix microarray data with RNA sequencing measurements. The algorithm successfully identified probes/genes with reliable measurements. Removing the unreliable measurements resulted in significant improvements for gene signature development and functional annotations.

KEYWORDS: beta-mixture model, correlation coefficients, cross-validation, gene expression, probe selection, RNA sequence

CITATION: Tong et al. Selecting Reliable mRNA Expression Measurements Across Platforms Improves Downstream Analysis. *Cancer Informatics* 2016:15 81–89 doi: 10.4137/CIN.S38590.

TYPE: Original Research

RECEIVED: January 08, 2016. **RESUBMITTED:** March 10, 2016. **ACCEPTED FOR PUBLICATION:** March 20, 2016.

ACADEMIC EDITOR: J. T. Efird, Editor in Chief

PEER REVIEW: Six peer reviewers contributed to the peer review report. Reviewers' reports totaled 1685 words, excluding any confidential comments to the academic editor.

FUNDING: Funding was provided in part by The University of Texas SPORE in Lung Cancer, 5P50 CA070907-10, UT MD Anderson Cancer Center Physician Scientist Award, LUNGevity Foundation, United Against Lung Cancer, The Sheikh Khalifa Bin Zayed Al Nahyan Institute for the Personalized Cancer Therapy's (IPCT's) Center for Professional Education and Training, and the National Cancer Institute through the Cancer Center Support Grant, CA016672. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: JVH discloses advisory board memberships with Genentech, AstraZeneca, Novartis, GlaxoSmithKline, Lilly, Boerhinger Ingelheim, Synta and Exelixis, and grants for work outside that presented in this paper, from AstraZeneca, GlaxoSmithKline and Bayer. Other authors disclose no potential conflicts of interest.

CORRESPONDENCE: Jingwang@mdanderson.org

Introduction

Gene expression experiments have been extensively conducted to identify candidate genes and gene signatures in cancer genomic research. Because of many potential technical artifacts, microarray-based mRNA experiments are often associated with a high degree of technical variability, which results in false discoveries among the identified genes. The MicroArray Quality Control project is assessing the quality of microarrays and has found good intraplatform and interplatform reproducibility.[1,2] Nevertheless, concerns have been raised about the quality of microarray data in genomic medicine,[3] such as the possibility of cross-hybridization resulting in microarray probes that actually measure untargeted genes. Therefore, the ability to filter out unreliable measurements and identify the most robust probes on an individual gene expression platform is an important analytical objective that is critical for downstream analysis.

Different methods have been applied to identify probes of good quality, including the use of probes closest to the 3′ end of a nucleic acid,[2] filtering by detection calls,[4] averaging across probes for the same gene,[5] and selecting an optimal probe based on probe specificity, coverage, and robustness.[6] Researchers have also investigated discrepancies in signals from multiple probes that are mapped to the same gene.[4] To improve measurement accuracy, we propose a statistical approach that borrows information across measurement platforms based on correlation coefficients. We hypothesize that unreliable expression measurements for an individual gene (ie, where probes of varying quality are available for a given gene) will lead to poor correlations between profiling platforms, whereas reliable measurements will usually correlate well across platforms. The two groups of correlations constitute a mixed population that can be modeled by a finite mixture model,[7,8] with one component representing the population of high-quality measurements and the other representing the low-quality measurements.

We illustrate our approach through three applications and two public data sets. In the first application, we compared the expression levels between monocytes and macrophages as described in the publication by Maouche et al.[4] We first defined the high-quality probes for two platforms using the beta-mixture model (BMM), which we later used to identify differentially expressed genes. We then applied Gene Ontology (GO) enrichment analysis to define biological relevance. This exercise achieved consistent biological findings when we used only the probes of highest quality. In the second application, we characterized epithelial–mesenchymal transition (EMT), a process associated with the loss of cell adhesion, increased invasion, and cell mobility, migration, and proliferation.[9–12] The EMT gene expression signature developed by Byers et al successfully classified nonsmall-cell lung cancer (NSCLC) cells into epithelial or mesenchymal groups.[13] Applying the BMM, we reduced the dimension of the EMT signature significantly but preserved the performance, which was the same as that of the original signature. In the third application, we compared microarray data to RNAseq data and found that the BMM was able to separate poor and good measurements. Further, we showed that the removal of poor measurements did not affect the performance of the downstream analysis such as functional annotation and signature development based on the reduced feature.

Methods

We applied a BMM on Pearson's correlation coefficients to separate reliable probes from unreliable ones. Originally, a BMM with variable number of components was proposed by Ji et al to model correlations of gene expression levels.[7] Our BMM model was a special case of this approach by modeling correlations between two platforms with a two-component BMM. We calculated Pearson's correlation coefficients for each matched probe pair between platforms, although Spearman's rank correlation would also apply. A linear transformation $y_i = (x_i + 1)/2$ was applied to all correlations as described by Ji et al.[7], so that the values were between 0 and 1. The transformed values can be modeled by a mixture of two beta distributions with a density function $f(y_i|\pi, \alpha_1, \beta_1, \alpha_2, \beta_2)$ as follows:

$$f\left(y_i \mid \pi,\alpha_1,\beta_1,\alpha_2,\beta_2\right)=\pi f_1\left(y_i \mid \alpha_1,\beta_1\right) + (1-\pi) f_2\left(y_i \mid \alpha_2,\beta_2\right),$$

where $f_l(y|\alpha_l, \beta_l)$, $l = (1, 2)$ is the probability density function for a beta distribution with mean $\alpha_l/(\alpha_l + \beta_l)$, variance $\alpha_l\beta_l/((\alpha_l+\beta_l)^2(\alpha_l+\beta_l+1))$, and π is the mixing proportion for the first component (the group with poor correlation). The parameters $(\pi, \alpha_1, \beta_1, \alpha_2, \beta_2)$ can be estimated by maximizing the following likelihood function through the expectation-maximization algorithm (a closed form solution for the E-step and M-step is provided in Ref. 7) or direct optimization procedure:

$$L\left(\pi,\alpha_1,\beta_1,\alpha_2,\beta_2 \mid y_1,y_2,\ldots,y_n\right) = \prod_{i=1}^{i=n}\left\{\pi f_1\left(y_i \mid \alpha_1,\beta_1\right)+(1-\pi) f_2\left(y_i \mid \alpha_2,\beta_2\right)\right\}.$$

Denote the event of y_i coming from the first component as the latent variable z_i. By Bayes' theorem, we have

$$P\left(z_i = 1 \mid y_i;\pi,\alpha_1,\beta_1,\alpha_2,\beta_2\right) = \frac{\pi f_1\left(y_i \mid \alpha_1,\beta_1\right)}{\pi f_1\left(y_i \mid \alpha_1,\beta_1\right)+(1-\pi) f_2\left(y_i \mid \alpha_2,\beta_2\right)},$$

which is the probability of y_i coming from the first component. By solving $P(z_i = 1|y_i; \pi, \alpha_1, \beta_1, \alpha_2, \beta_2) = 0.5$ after plugging in the estimates of $(\pi, \alpha_1, \beta_1, \alpha_2, \beta_2)$, we can obtain the model-based threshold τ. At the original correlation scale, the cutoff $2\tau - 1$ (calculated through the inverse transformation of $y_i = (x_i + 1)/2$) can separate the probe sets into a group with good correlation and a group with poor correlation.

Results

To demonstrate the applicability of the proposed method, we first performed a simulation study with known gold standard. We then applied BMM to three real applications to show the feasibility of separating good probes from probes with low quality, thereby improving the efficiency and accuracy of data analysis.

Simulation. *Simulation setup.* To evaluate the performance of the proposed BMM method, we simulated cross-platform gene expression measurements with both good and poor qualities quantified by correlation strength. In particular, we simulated $G = 5000$ correlation values (ρ) from a two-component BMM with parameters $(\alpha_1, \beta_1, \alpha_2, \beta_2) = (28, 6, 27, 22)$ using different mixture proportions, $\pi = (0.2, 0.4, 0.6, 0.8)$ representing percentages of good-quality measurements. For each pair of gene expression measurements $i(i = 1{:}G)$ in platform $j(j = 1{:}2)$, we simulated $N = (50, 100, 200)$ samples to evaluate the effect of sample size. In total, this led to 12 simulation scenarios. Correlated gene expression data were then simulated from bivariate Gaussian distribution with mean μ_{ij} and covariance matrix

$$\begin{pmatrix} \sigma_{i1}^2 & \rho\sigma_{i1}\sigma_{i2} \\ \rho\sigma_{i1}\sigma_{i2} & \sigma_{i2}^2 \end{pmatrix}.$$

To better mimic the heterogeneity of real data, the mean (μ_{i1}, μ_{i2}) and variance $\left(\sigma_{i1}^2, \sigma_{i2}^2\right)$ for gene i in platform j are randomly sampled from RNAseq data used in application 3. Of note, the parameters specified here were motivated by real data estimates.

BMM successfully recovered good-quality measurements. We fitted BMM model on simulated data for all 12 scenarios. The estimated mixture density (transformed back to correlation scale, solid lines) and true values (dashed lines) are shown in Figure 1A. Model-based thresholds as well as corresponding true-positive rates (TPRs) and false-positive rates (FPRs) were also indicated. Receiver–operator characteristic (ROC) curves evaluating the effect of mixture proportion and sample size across varying decision thresholds are shown in Figure 1B.

In general, the BMM approach successfully recovered the mixture structure. As sample size N and mixture proportion π increased, the fitted densities came closer to their true values. At $N = 50$ and $\pi = 0.2$, there were significant deviation between the true density and estimated density due to inaccurate estimates of the correlation coefficients. However, the threshold estimate $\nu = 0.49$ was not affected severely compared to $\nu = 0.46$ at $N = 200$. At $N = 200$ and $\pi = 0.8$, the best performance of BMM across all simulation scenarios was achieved with a TPR of 0.98 and an FPR of 0.06. The model-based threshold provided an objective way to discern good-quality measurements. As the ROC curves suggest, more stringent or loose cutoffs might be used, depending on requirements of different applications.

Application 1: analysis of microarray gene expression from Affymetrix and Illumina arrays to compare human monocytes and monocyte-derived macrophages. *Data set and probe selection by BMM.* We downloaded the normalized expression values for five monocyte and monocyte-derived macrophage samples from the National Center for Biotechnology Information Gene Expression Omnibus (GEO) repository (http://www.ncbi.nlm.nih.gov/geo/), with GEO

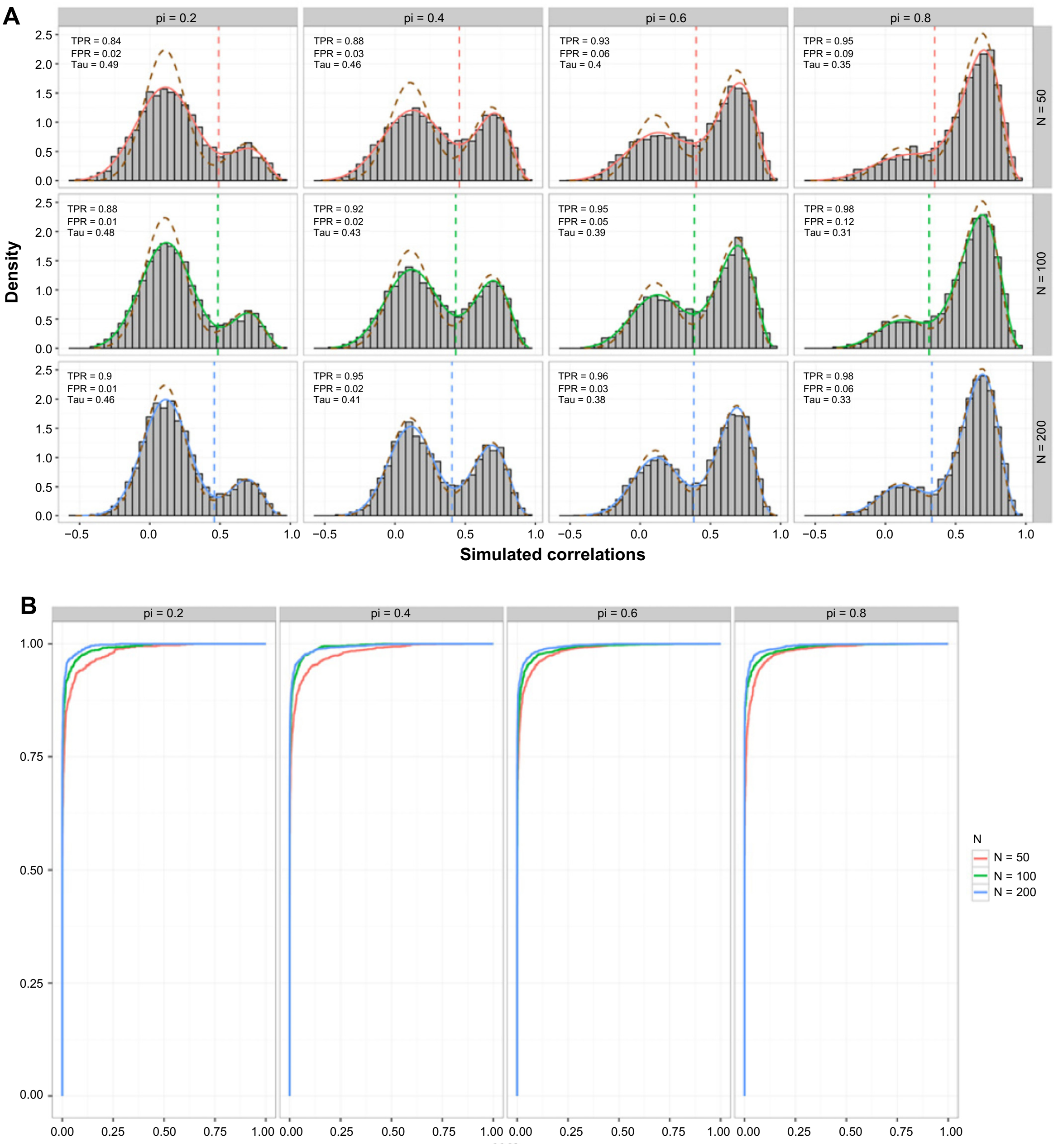

Figure 1. (**A**) Density estimates of the BMM model on simulated data. (**B**) ROC curves for simulated data.
Notes: Solid lines represented estimated density. Dashed lines represented true density.
Abbreviations: TPR, true-positive rate; FPR, false-positive rate; Tau, model-based threshold estimates.

series accession numbers GSE10213 (Illumina Human6-V1) and GSE11430 (Affymetrix U133 plus 2.0). The annotation files were downloaded from GEO (GPL570 Affymetrix and GPL6097 Illumina). Using the 10 samples (five monocyte and five monocyte-derived macrophage samples) profiled on both Affymetrix HGU133 plus 2.0 and Illumina Human-6 v1, we computed Pearson's correlation coefficients for each matched probe set pair based on the gene symbols. After filtering out the unmatched probes, there were 24,670 and 13,951 probes in Affymetrix and Illumina, respectively, and 31,622 pairs of probe sets that correspond to 11,565 unique gene symbols. If the probe set is highly correlated with at least one of the matched paired probe sets on the other platform, we would expect this probe set to be of reasonably good quality.

We applied BMM to the transformed quantities of these correlation coefficients, the density of which is given in Figure 2. The parameter estimates of the BMM model were $(\alpha_1, \beta_1, \alpha_2, \beta_2, \pi) = (3.435, 2.833, 11.359, 1.262, 0.803)$. The means of the fitted beta distributions, respectively, equaled 0.55 and 0.90, which corresponded to 0.10 and 0.80 on the correlation coefficient scale. Therefore, the probe sets with high correlation, which suggested good-quality measurements, corresponded to the component with a mean correlation coefficient of 0.80 (range, 0.711–0.999), while the probe sets with weak or no correlations corresponded to the component with a mean correlation coefficient of 0.10 (range, −0.96–0.71). The model-based threshold ν at the correlation scale was 0.7110. We also considered more stringent cutoff thresholds (0.8 and 0.9) to evaluate the stability of the results based on the hypothesis that a higher correlation provides more reliable measurements (Table 1).

The Affymetrix percent present call is one of the criteria used to define the quality of probe sets. Figure 3A shows the densities of the Affymetrix percent present calls of the probe sets defined by the two components. The percent present calls in the good-quality measurement groups were higher than those in the lower quality measurement groups. In the Illumina platform, probes are identified by types A, S, and I, where A represents the probes that detect all known transcripts for a gene, S represents the probes that detect a gene with a single transcript, and I represents the probes that detect a specific isoform of a gene. Figure 3B shows that the S or A probe sets had higher correlation coefficients compared to the I probe sets, and that the S probe sets were slightly better than the A probe sets. There were 5341 and 3720 good probes at the model-based threshold of 0.7110, corresponding to 3507 genes for further analysis in the Affymetrix and Illumina platforms. The different correlation patterns among different present calls and probe types indicate that the BMM was very effective in separating the probes according to two categories of quality by leveraging information across the platforms.

Functional annotation identified novel GO class. To identify enriched GO class, we applied the same statistical analysis as in Ref. 4. In particular, a modified *t*-test using empirical Bayes smoothing was fitted using the *Limma* package.[14] Benjamini and Hochberg correction method was used to adjust for multiple testing.[15] Enriched GO class was identified using hypergeometric test available in the *GOstats* package.[16] Table 2 shows the enriched GO class (adjusted *P* value <0.05 in at least one platform). Most GO classes defined by Maouche et al.[4] were also identified in our analysis. The exception was *protein metabolic process* (GO:0019538), which was not significant on any platform. However, the GO class *regulation of proteolysis* (GO:0030162), which turned out to be a child of the protein metabolic process (GO:0019538), was significant on the Affymetrix platform. A novel GO class *response to stress* (GO:0006950), which was not reported by Maouche et al.[4], was significant in both the Affymetrix and Illumina platforms. As indicated by Hume,[17] monocytes and macrophages are important players in the immune response. The identification

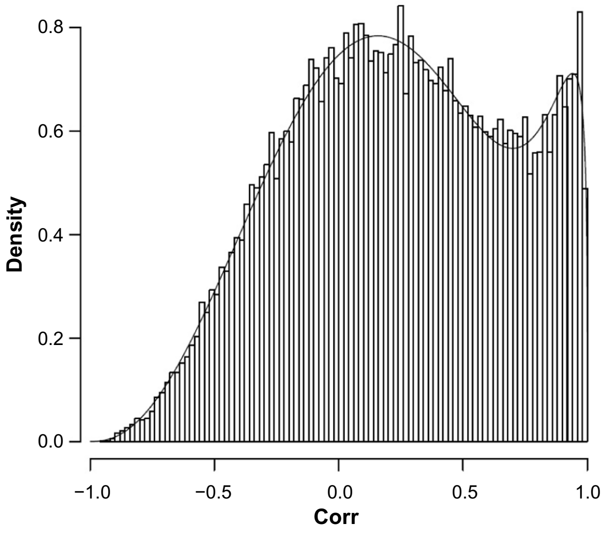

Figure 2. Histogram of the 42,491 Pearson's correlation coefficients of the matched probe sets between two platforms. The curve is the predicted density of the mixture of two beta distributions.

Table 1. Results of genes differentially expressed between macrophage and monocyte samples at different cutoff values.

THRESHOLD VALUE	PLATFORM	*P* ADJUST <0.001	COMMON GENES
0.71 (model based)	Affymetrix	1631 genes (2314 probes)	1108 genes
	Illumina	1443 genes (1500 probes)	
0.80 (stringent criterion 1)	Affymetrix	1472 genes (2095 probes)	1095 genes
	Illumina	1328 genes (1382 probes)	
0.90 (stringent criterion 2)	Affymetrix	1056 genes (1492 probes)	918 genes
	Illumina	1006 genes (1041 probes)	

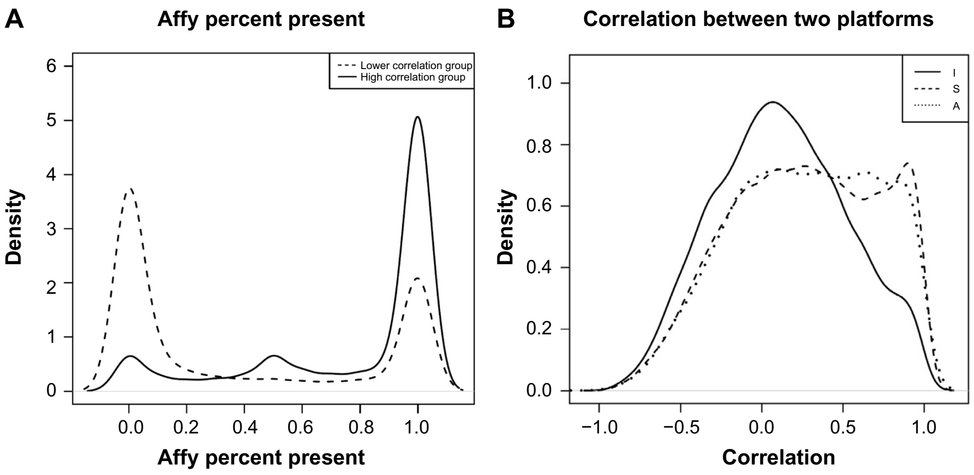

Figure 3. (**A**) Densities of the Affymetrix percent present calls within two groups defined by the beta-mixture model. (**B**) Densities of the 42,491 Pearson's correlation coefficients within the three Illumina annotation types. A: the probe detects all known transcripts for the gene. S: the probe detects the single known gene transcript for the gene. I: the probe detects a specific isoform of the gene (isoform specific).

of the response to stress as an enriched GO class supported our hypothesis that with the use of a small but high-quality probe list, we can obtain the same or even better biological findings than when using a larger probe list of mixed quality.

Application 2: analysis of lung cancer cell line expression data for developing EMT signature. *Data set and probe selection by BMM.* In this application, we used 54 matched lung cancer cell lines, 52 that were NSCLC, 1 small cell lung cancer, and 1 mesothelioma, and profiled them on Affymetrix U133A, B (U133A and U133B combined, GEO accession number GSE4824) and Illumina WG V2 (GEO accession number GSE32989).

Table 2. Enriched Gene Ontology using probes selected from model-based threshold (0.7110).

GO CLASS	AFFYMETRIX LIST	ILLUMINA LIST
Response to stress GO:0006950	0.0002	0.0011
Organic acid metabolic process GO:0006082	0.0002	0.0011
Carboxylic acid metabolic process GO:0019752	0.0002	0.0011
Oxoacid metabolic process GO:0043436	0.0002	0.0011
Cellular ketone metabolic process GO:0042180	0.0003	0.0011
Immune system process GO:0002376	0.0003	0.0011
Lipid metabolic process GO:0006629	0.0003	0.0011
Intracellular signaling cascade GO:0007242	0.0007	0.0162
Apoptosis GO:0006915	0.0014	0.0033
Carbohydrate metabolic process GO:0005975	ns	0.0446
Regulation of signal transduction GO:0009966	0.0328	0.0826
Regulation of proteolysis GO:0030162	0.0372	ns

Notes: GO classes that are significant (adjusted $P < 0.05$) on at least one platform are shown. Probes with adjusted $P < 0.001$ from differential expression analysis are used for GO enrichment analysis.

In order to develop a robust genomic EMT signature, we first examined the consistency of expression measurements within and between array platforms. Figure 4 illustrates the inconsistencies of expression measurements for fibronectin (FN1) within and between the two array platforms, only three highly correlated FN1 probe sets within Affymatrix platform. This analysis illustrates the need to use reliable probes to develop expression signatures that can be validated across platforms. As in the previous analysis, we computed the Pearson's correlation coefficients for each matched probe set pair based on the gene symbol. After filtering out the unmatched probes, there were 28,808 and 20,744 probes in Affymetrix and Illumina, respectively, and 42,491 pairs of probe sets that were matched between the two platforms, which corresponded to 15,637 unique common gene symbols.

We applied the BMM to transformed quantities of these correlation coefficients, the density of which is given in Figure 5. Parameter estimates of the fitted model were $(\alpha_1, \beta_1, \alpha_2, \beta_2, \pi) =$ (20.830, 18.237, 8.034, 2.443, 0.340). The means of the fitted beta distribution equaled 0.53 and 0.77, which corresponded to 0.07 and 0.53 on the correlation coefficient scale. The probe sets with good quality corresponded to the component with a mean correlation coefficient of 0.53; the probe sets with poor quality corresponded to the component with a mean correlation coefficient of 0.07. The model-based threshold was $\nu = 0.2718$. This threshold was much smaller than the threshold in application 1 ($\nu = 0.7110$), which reflects the fact that technical replicates were used in application 1, whereas tumor cell lines, which are more variable, were used in application 2. To demonstrate the

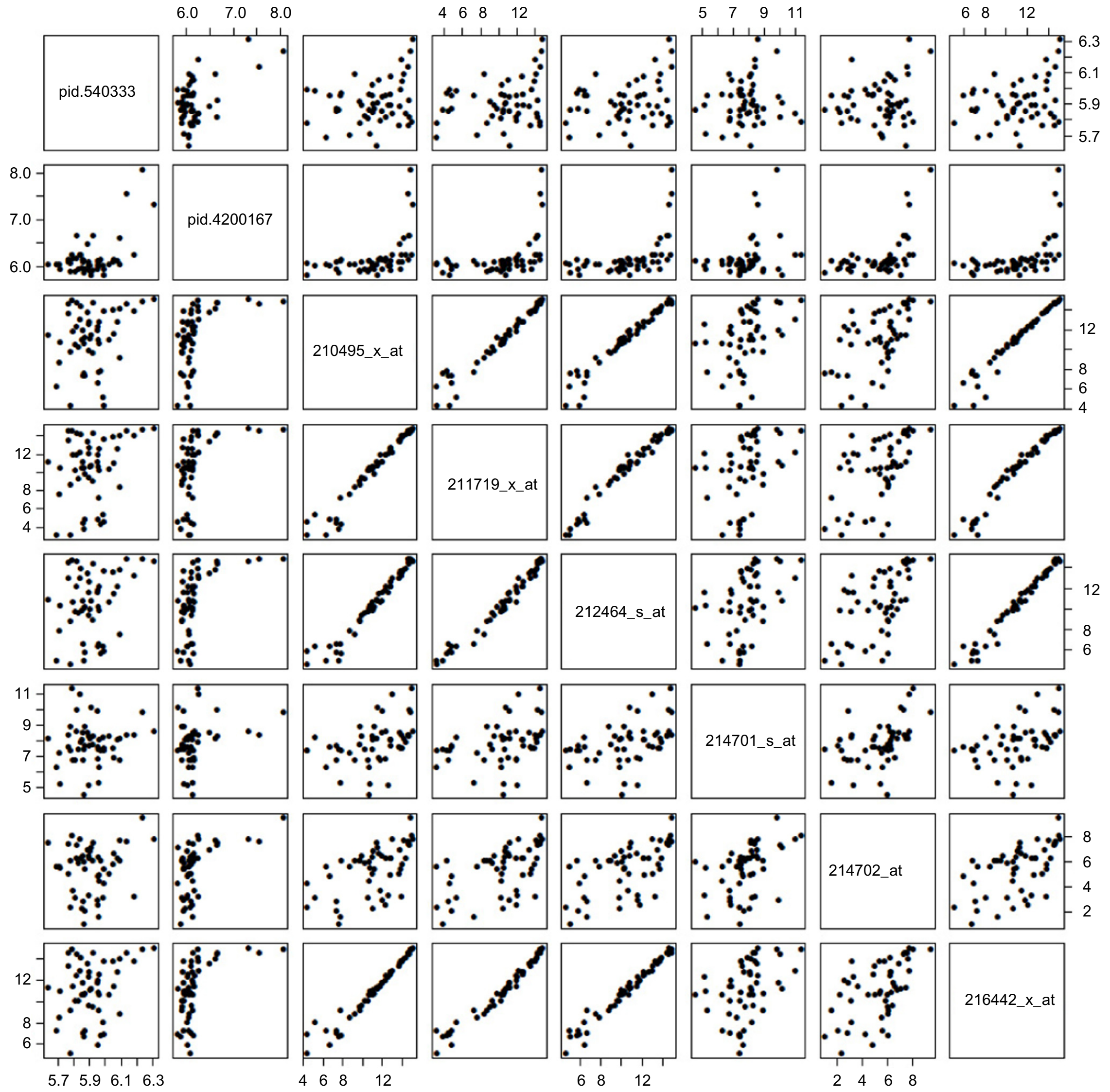

Figure 4. Pair scatter plot comparing multiple Affymetrix probe sets and Illumina V2 probes for fibronectin on common lung cancer cell lines.

effect of the threshold choice, we also considered more stringent cutoffs (0.4, 0.5, 0.6, 0.7, 0.8, and 0.9).

Figure 6A shows the densities of the Affymetrix percent present call rates between the two sets of probes defined by the model-based threshold *v*. The percent present call rates in the good-quality group were higher than those in the lower quality group. Similar observations were found in the Illumina platform. Figure 6B shows that the S and A probe sets had higher correlation coefficients compared to the I probe sets. These findings suggest that using the BMM can efficiently narrow the analyses to good-quality measurements and significantly reduce the noise level from that of the large feature set.

Developing EMT signature from good-quality probes consistently separates NSCLC cell lines into two distinct groups. The original EMT signature incorporated 76 unique genes corresponding to 146 probes,[13] which were identified on the basis of their correlation to known EMT markers and the bimodality index.[18,19] Here, we investigated whether the EMT signature

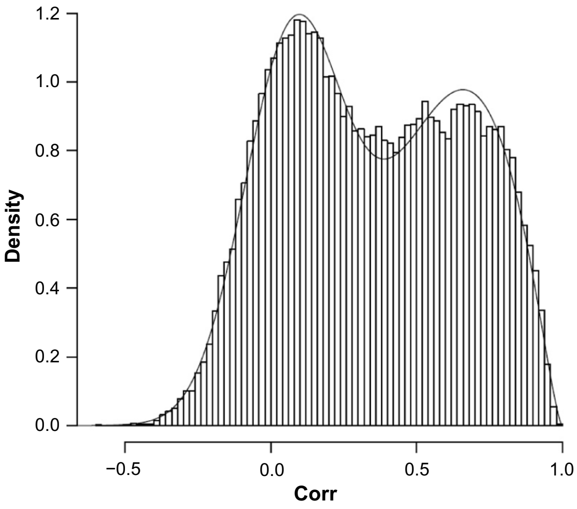

Figure 5. Histogram of the 42,491 Pearson's correlation coefficients of the matched probe sets between two platforms. The curve is the predicted density of the mixture of two beta distributions.

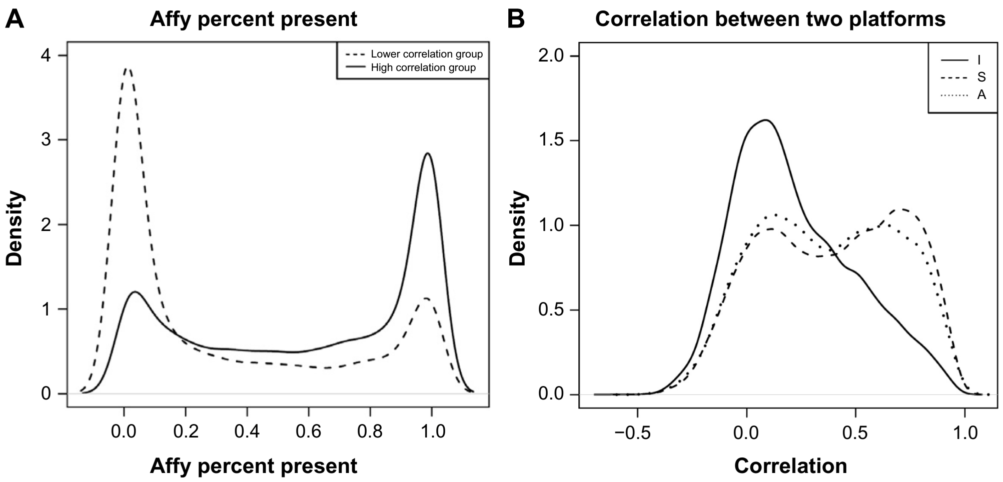

Figure 6. (**A**) Densities of the Affymetrix percent present calls within two groups defined by the beta-mixture model. (**B**) Densities of the 42,491 Pearson's correlation coefficients within the three Illumina annotation types. A: the probe detects all known transcripts for the gene. S: the probe detects the single known gene transcript for the gene. I: the probe detects a specific isoform of the gene (isoform specific).

can be reduced on the basis of our BMM approach by filtering out unreliable probes. We used the same data set and applied the same method for deriving the EMT signature as Byers et al.[13] but added a step to select reliable probes based on the BMM.

We first filtered the probe sets on both platforms using different correlation cutoffs, including the model-based threshold $\nu = 0.2718$ as well as other arbitrary cutoffs 0.40, 0.50, 0.60, 0.70, 0.80, and 0.90. With this criterion in addition to the procedures used by Byers et al.[13], the resulting probe sets on both platforms were used as an alternative EMT gene signature. We applied hierarchical clustering to the EMT signature genes to group the samples into two clusters assigned as epithelial-like and mesenchymal-like groups and compared this assignment with the assignment published by Byers et al.[13] Table 3 shows the number of probe sets, unique gene symbols, and consistent assignments on both platforms using different thresholds. Group assignment using the Illumina WG V2 platform was consistent for all 52 NSCLC cell lines, no matter the threshold choice. Group assignment using Affymetrix U133A, B was consistent for all 54 NSCLC cell lines at all thresholds except those lower than 0.60, for which the cell line HCC 2279 was assigned as mesenchymal like rather than the published assignment of epithelial like without filtering. The fact that HCC2279 shifted from epithelial like to mesenchymal like and then to epithelial like with increasing stringency suggests it may not preserve a clear epithelial or mesenchymal phenotype, which warrants further investigation.

Application 3: comparison of microarray and RNAseq data from The Cancer Genome Atlas. As a demonstration, we applied our proposed method to compare microarray and RNAseq data. We downloaded both Affymetrix HT-HG U133A microarray data and level 3 RNASeq data from glioblastoma multiforme samples from The Cancer Genome Atlas (TCGA) data portal (https://tcga-data.nci.nih.gov/tcga/). The downloaded RNAseq data were quantified and normalized using RSEM[20] by the TCGA team. A total of 150 samples with both Affymetrix and RNAseq data were used to compute 18,715 Pearson's correlations corresponding to 11,678 unique genes. Parameter estimates of the BMM model were $(\alpha_1, \beta_1, \alpha_2, \beta_2, \pi) = (27.830, 27.140, 6.377, 22.310, 0.806)$. The means of

Table 3. Consistent clustering of EMT signature genes after being filtered using different thresholds within Affymetrix U133A, B and Illumina WG V2 NSCLC cell lines.

CORRELATION THRESHOLD VALUE (MODEL BASED)	AFFYMETRIX U133A,B (54 CELL LINES)			ILLUMINA WG V2 (52 CELL LINES)		
	GENE SYMBOL	PROBE SETS USED	CONSISTENT CELL LINES	GENE SYMBOL USED	PROBES USED	CONSISTENT CELL LINES
0.2718	71	126	53 (HCC2279 mesenchymal)	71	93	52
0.40	70	117	53 (HCC2279 mesenchymal)	70	89	52
0.50	70	110	53 (HCC2279 mesenchymal)	70	87	52
0.60	67	99	54	67	81	52
0.70	66	95	54	66	75	52
0.80	55	73	54	55	61	52
0.90	35	40	54	35	36	52

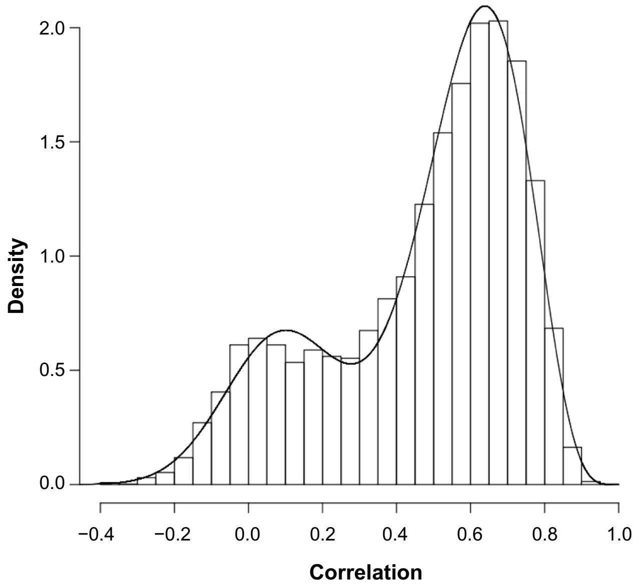

Figure 7. Histogram of the 18,715 Pearson's correlation coefficients between HT-HG Affymetrix and RNAseq from TCGA glioblastoma data.

the fitted beta distributions equaled 0.54 and 0.80, which corresponded to 0.08 and 0.60 on the correlation coefficient scale. The model-based threshold was $\nu = 0.307$. Figure 7 shows the predicted density of the correlations from the BMM. The fact that the predicted density agrees well with the empirical distribution of the correlations demonstrates the applicability of our proposed method to compare microarray data and RNAseq data.

Discussion

We have introduced an approach to reduce the probe set dimensions to reliable qualitative probes using correlation coefficients. We model the correlation coefficients with a two-component BMM that represents the populations of high-quality and low-quality measurements. This resembles similar efforts that modeled gene expression with either a normal mixture for microarray data or a mixture of discrete distributions for RNAseq data.[18,19] We applied the BMM to two real gene expression data sets. Reliable probes were first identified with a model-based threshold and then used for downstream analysis.

In both data sets, the densities of the Affymetrix percent present calls of the probe sets were defined by the two components. The percent present calls in good-quality measurement groups were higher than those in the lower quality measurement groups. In the Illumina platform, the type S or A probe sets had higher correlation coefficients than the type I probe sets. These findings support the hypothesis that good-quality measurements can be separated from the overall data by using the BMMs to avoid any ad hoc steps.

In the first application, we applied our model to a publicly available data set to identify genes that were differentially expressed between monocytes and monocyte-derived macrophages. Biologically, monocytes and macrophages are important in the immune response.[17] In this application, based on the selected reliable probe sets, we first identified genes differentially expressed between the two groups and then used the identified genes to perform GO analysis. We found that most GO classes defined in the original publication[4] were also identified in our analysis. However, the most significant GO class response to stress was not reported in the original investigation.[4] Research has found that the stress system is critical to homeostasis and the immune response.[21,22] The identification of response to stress as a GO biological process supports our hypothesis that the use of a small but highly qualified probe list can provide consistent or even more valuable biological findings.

In the second application, we narrowed the EMT gene signatures to the probes with reliable qualities. The separation into epithelial-like and mesenchymal-like groups was consistent using the high-quality probe sets even though some probe sets, such as CDH1, were missing. We found high correlation between the first principal component of the EMT signature and the E-cadherin protein expression level.[13] This supports our strategy that the dimension of the gene signature can be efficiently reduced using the correlation coefficients without losing the biological interpretation. Table 3 shows consistent clustering by EMT signature genes after being filtered using different correlation coefficient thresholds within each array platform. The ability to reduce the number of genes in the signature and retain an overall robust performance is critically important for translating the gene signature into clinical applications.

With rapid development of sequencing technology, microarray-based gene expression technology has become less popular in genomic studies. More investigations are using technologies based on sequencing, such as RNAseq for whole transcriptomic profiling and miRNAseq for microRNA sequencing. The proposed approach is not limited to microarray applications but can be applied broadly to compare arbitrary platforms such as microarray and RNAseq data or RNAseq and miRNAseq data. In the third application, we illustrated the use of the proposed algorithm to compare RNAseq and microarray mRNA expression profiling data. Many earlier studies have accumulated rich sets of microarray data that are still available for investigation and validation. Our proposed method will better leverage these data by identifying reliable measurements.

Acknowledgment

We thank LeeAnn Chastain for critical editing of the manuscript.

Author Contributions

Derived background method, designed the investigation, performed the analyses, and prepared the manuscript: PT, LD, LS, LL, KRC, JW. Participated in conducting lung cell line data for the study: LG, JDM, JVH, LB.

REFERENCES

1. Canales RD, Luo Y, Willey JC, et al. Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat Biotechnol.* 2006;24(9):1115–22.

2. MAQC Consortium, Shi L, Reid LH, et al. The microarray quality control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*. 2006;24(9):1151–61.
3. Couzin J. Microarray data reproduced, but some concerns remain. *Science*. 2006; 313(5793):1559–9.
4. Maouche S, Poirier O, Godefroy T, et al. Performance comparison of two microarray platforms to assess differential gene expression in human monocyte and macrophage cells. *BMC Genomics*. 2008;9(1):302.
5. Warnat P, Eils R, Brors B. Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics*. 2005;6(1):265.
6. Li Q, Birkbak NJ, Gyorffy B, Szallasi Z, Eklund AC. Jetset: selecting the optimal microarray probe set to represent a gene. *BMC Bioinformatics*. 2011;12(1):1.
7. Ji Y, Wu C, Liu P, Wang J, Coombes KR. Applications of beta-mixture models in bioinformatics. *Bioinformatics*. 2005;21(9):2118–22.
8. McLachlan G, Peel D. *Finite Mixture Models*. Wiley. com, Wiley, New York; 2004.
9. Huber MA, Kraut N, Beug H. Molecular requirements for epithelial–mesenchymal transition during tumor progression. *Curr Opin Cell Biol*. 2005;17(5):548–58.
10. Thiery JP, Acloque H, Huang RY, Nieto MA. Epithelial–mesenchymal transitions in development and disease. *Cell*. 2009;139(5):871–90.
11. Thiery JP. Epithelial–mesenchymal transitions in tumour progression. *Nat Rev Cancer*. 2002;2(6):442–54.
12. Hugo H, Ackland ML, Blick T, et al. Epithelial – mesenchymal and mesenchymal – epithelial transitions in carcinoma progression. *J Cell Physiol*. 2007;213(2):374–83.
13. Byers LA, Diao L, Wang J, et al. An epithelial–mesenchymal transition gene signature predicts resistance to EGFR and PI3 K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance. *Clin Cancer Res*. 2013;19(1):279–90.
14. Smyth GK. Limma: linear models for microarray data. In Bioinformatics and computational biology solutions using R and Bioconductor. Springer, New York; 2005:397–420.
15. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Methodol*. 1995;57:289–300.
16. Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. *Bioinformatics*. 2007;23(2):257–8.
17. Hume DA. The mononuclear phagocyte system. *Curr Opin Immunol*. 2006;18(1): 49–53.
18. Wang J, Wen S, Symmans WF, Pusztai L, Coombes KR. The bimodality index: a criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data. *Cancer Inform*. 2009;7:199.
19. Tong P, Chen Y, Su X, Coombes KR. SIBER: systematic identification of bimodally expressed genes using RNAseq data. *Bioinformatics*. 2013;29(5):605–13.
20. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12(1):1.
21. Webster JI, Tonelli L, Sternberg EM. Neuroendocrine regulation of immunity*. *Annu Rev Immunol*. 2002;20(1):125–63.
22. O'Connor T, O'Halloran D, Shanahan F. The stress response and the hypothalamic-pituitary-adrenal axis: from molecule to melancholia. *QJM*. 2000;93(6):323–33.