Validation of oligonucleotide microarray data using microfluidic low-density arrays: a new statistical method to normalize real-time RT-PCR data

Lynne V. Abruzzo[1], Kathleen Y. Lee[2], Alexandra Fuller[2], Alan Silverman[2], Michael J. Keating[1], L. Jeffrey Medeiros[1], and Kevin R. Coombes[1]

BioTechniques 38:785-792 (May 2005)

Profiling studies using microarrays to measure messenger RNA (mRNA) expression frequently identify long lists of differentially expressed genes. Differential expression is often validated using real-time reverse transcription PCR (RT-PCR) assays. In conventional real-time RT-PCR assays, expression is normalized to a control, or housekeeping gene. However, no single housekeeping gene can be used for all studies. We used TaqMan® Low-Density Arrays, a medium-throughput method for real-time RT-PCR using microfluidics to simultaneously assay the expression of 96 genes in nine samples of chronic lymphocytic leukemia (CLL). We developed a novel statistical method, based on linear mixed-effects models, to analyze the data. This method automatically identifies the genes whose expression does not vary significantly over the samples, allowing them to be used to normalize the remaining genes. We compared the normalized real-time RT-PCR values with results obtained from Affymetrix Hu133A GeneChip® oligonucleotide microarrays. We found that real-time RT-PCR using TaqMan Low-Density Arrays yielded reproducible measurements over seven orders of magnitude. Our model identified numerous genes that were expressed at nearly constant levels, including the housekeeping genes PGK1, GAPD, GUSB, TFRC, *and 18S rRNA. After normalizing to the geometric mean of the unvarying genes, the correlation between real-time RT-PCR and microarrays was high for genes that were moderately expressed and varied across samples.*

INTRODUCTION

Gene expression profiling studies using microarray technology to measure messenger RNA (mRNA) expression frequently yield long lists of genes that appear to be differentially expressed. Differential expression of a few genes is subsequently validated using an alternative technology, often real-time reverse transcription PCR (RT-PCR) assays. Compared to other methods to assess RNA expression, such as Northern blot analysis or conventional RT-PCR, real-time RT-PCR assays are rapid, reproducible, and have a wide dynamic range. All these technologies, however, are relatively labor-intensive. So, they are rarely used to validate the differential expression of more than 10 or 20 genes. Applied Biosystems (Foster City, CA, USA) has recently introduced TaqMan® Low-Density Arrays, a medium-throughput method for real-time RT-PCR that uses microfluidics cards. With the TaqMan Low-Density Arrays, researchers can simultaneously assay the RNA expression levels of up to 384 genes on a single card. In principle, this technology makes it possible to validate rapidly the differential expression of all genes detected in a microarray study.

In conventional real-time RT-PCR assays, expression of the target gene of interest is normalized to an endogenous control, or housekeeping gene (1). The purpose of normalization is to remove or eliminate differences due to sampling; that is, differences in total RNA quantity and quality. The most commonly used housekeeping genes in real-time RT-PCR assays are β-actin (*ACTB*), glyceraldehyde-3-phosphate dehydrogenase (*GAPD*), and 18S and 28S ribosomal RNA (rRNA) (2). Other commonly used housekeeping genes include the transferrin receptor (*TFRC* or *CD71*), β-glucuronidase (*GUSB*), β-2 microglobulin (*B2M*), phosphoglycerate kinase (*PGK1*), and hypoxanthine phosphoribosyltransferase (*HPRT*) (3,4). The underlying assumption when using housekeeping genes to normalize data is that they are expressed at constant levels across the samples and that their expression does not vary in response to the experimental manipulation. Because the expression of the target gene is measured relative to the housekeeping gene, it is critical that these assumptions hold. It has become clear, however, that no single housekeeping gene can be used for all studies and that the choice of the control gene depends on the tissue type and the experimental conditions (5–11).

[1]The University of Texas M.D Anderson Cancer Center, Houston, TX and [2]Applied Biosystems, Foster City, CA, USA

Although many studies have demonstrated that the expression of housekeeping genes may vary considerably, relatively few studies have been performed to address this issue directly. Tricarico and colleagues (3) used real-time RT-PCR to study the levels of vascular endothelial growth factor (*VEGF*) mRNA in breast cancer and colon cancer. They compared normalization to 11 different housekeeping genes with normalization to total RNA. No single housekeeping gene produced values that matched the VEGF protein levels (as measured by immunohistochemistry) or the results of functional assays (microvessel density) as accurately as normalization to total RNA. However, accurate quantification of the total amount of RNA may be difficult in some circumstances, such as with very small tissue biopsies. Moreover, the total amount of RNA may itself vary, reflecting important differences in the biological activity of cells under different conditions. Vandesompele and colleagues (12) showed that normalization based on a single housekeeping gene led to erroneous quantifications of up to 3-fold in 25% of cases and 6.4-fold in 10% of cases, with sporadic cases showing errors greater than 20-fold. They recommended normalizing to the geometric mean of several housekeeping genes. Arguing that it was impractical to quantify eight control genes in order to study a few target genes, they also proposed an iterative method for finding a minimal set of housekeeping genes to include in the normalization set. Tricarico and colleagues (3) viewed the necessity of identifying an appropriate combination of housekeeping genes as a nontrivial practical limitation of this method.

Using TaqMan Low-Density Arrays, we can now rapidly assay the expression of up to 384 genes on a single microfluidics card, which largely eliminates these practical limitations. In their place, we are faced with the statistical challenge of finding a good combination of genes to use for normalization. In this paper, we describe the results of an experiment using TaqMan Low-Density Arrays on nine samples from patients with chronic lymphocytic leukemia (CLL). We introduce a novel statistical method, based on the established theory of linear mixed models, for analyzing the data. This method automatically identifies a subset of genes that do not change significantly over the samples, allowing them to be used to normalize the expression levels of the remaining genes. Finally, we compare the normalized real-time RT-PCR values with the results obtained from GeneChip® oligonucleotide microarray (Affymetrix, Santa Clara, CA, USA) experiments on the same samples.

MATERIALS AND METHODS

Sample Collection and RNA Preparation

CLL samples were collected from the peripheral blood of nine untreated patients at the University of Texas M.D. Anderson Cancer Center (Houston, TX, USA) after obtaining informed consent. Total RNA was prepared from CD19-positive CLL cells and was used to determine the somatic hypermutation status of the immunoglobulin heavy chain variable region genes, as previously described (13). The RNA was also used for hybridization to oligonucleotide arrays (U133A GeneChip) and for real-time RT-PCR on TaqMan Low-Density Array microfluidics cards (Applied Biosystems).

Oligonucleotide Microarray Analysis

Each U133A GeneChip microarray contains 22,215 noncontrol probe sets that correspond to more than 18,400 distinct transcripts, including 14,593 well-characterized human genes. The list of probe sets and corresponding genes is available from the Affymetrix web site (www.affymetrix.com/support/technical/libraryfilesmain.affx). Hybridization of biotin-labeled cRNA to the oligonucleotide arrays and image analysis were performed in the DNA Microarray Core Facility at the M.D. Anderson Cancer Center, according to protocols available on their web site (www.mdanderson.org/departments/dnamicroarray). The microarray image data were quantified and normalized using the DNA Chip Analyzer (www.dchip.org) as previously described (14).

TaqMan Low-Density Array

The TaqMan Immune Profiling Low-Density Array consists of 96 TaqMan Gene Expression Assays (Applied Biosystems) preconfigured in a 384-well format and spotted on a microfluidic card (4 replicates per assay). Each TaqMan Gene Expression Assay consists of a forward and reverse primer at a final concentration of 900 nM and a TaqMan MGB probe (6-FAM dye-labeled; Applied Biosystems), 250 nM final concentration. The assays are gene specific and have been designed to span an exon-exon junction. Each assay and its assay ID number are available at docs.appliedbiosystems.com/pebiodocs/00112893.pdf.

First, 500 μL of cDNA from each sample (20 ng total input RNA/μL) were combined with an equal volume of TaqMan Universal PCR Master Mix (Applied Biosystems), mixed by inversion, and spun briefly in an Eppendorf 5415D microcentrifuge (Brinkmann Instruments, Westbury, NY, USA). After the cards reached room temperature, 100 μL of each sample were loaded into each of 8 ports on the TaqMan Low-Density Array. The cards were placed in Sorvall®/Heraeus® Custom Buckets (Applied Biosystems) and centrifuged in a Sorvall Legend™ centrifuge (Kendro Scientific, Asheville, NC, USA) for 1 min at 331× *g*. Cards with excess sample in the fill reservoir were spun for an additional 1 min. Immediately following centrifugation, the cards were sealed with a TaqMan Low-Density Array Sealer (Applied Biosystems) to prevent cross-contamination. The final volume in each well after centrifugation was less than 1.5 μL; thus, the final concentration was approximately 15 ng per reaction. The real-time RT-PCR amplifications were run on an ABI PRISM® 7900HT Sequence Detection System (Applied Biosystems) with a TaqMan Low Density Array Upgrade. Thermal cycling conditions were as follows: 2 min at 50°C [to activate uracil-DNA glycosylase (UNG)], 10 min at 95°C (activation), 40 cycles of denaturation at 95°C for 15 s, and annealing and extension at 60°C for 1 min. Each CLL sample was processed on a separate card.

Statistical Analysis of Real-Time RT-PCR Data

Real-time RT-PCR data were quantified using the SDS 2.1 software package (Applied Biosystems). Results from each card were quantified separately, using an automatic baseline and a manual threshold of 0.10 to record the cycle thresholds (C_ts). Assays that did not yield a cycle threshold for less than 40 cycles were treated as missing data. For our analysis, we used the base-two logarithm of the relative abundance of RNA in each sample, which we computed as $y = a - C_t$ where we chose the constant a to make the minimum of y equal 0 over the entire data set. Both the real-time RT-PCR data and the Affymetrix data were imported into S-Plus® (Insightful, Seattle, WA, USA) for statistical analysis. Linear mixed-effects models were fit using the lme package. Spearman rank correlation coefficients were computed to compare data from the two platforms, and beta distributions were used to assess the null hypothesis of no significant correlation between platforms.

RESULTS

Real-Time RT-PCR Data

We randomly selected nine CLL samples (four with unmutated immunoglobulin heavy chain variable region genes and five with mutated variable region genes) that we had previously analyzed using U133A GeneChip arrays and analyzed them using the TaqMan Immune Profiling Low-Density Array. Real-time RT-PCR data were acquired and quantified as described in Materials and Methods. Figure 1 illustrates the dynamic range and variability of these measurements within and between samples on six genes; comparable results were obtained for the other genes (data not shown). The coefficient of variation (CV) in the observed C_t of the replicate wells was less than 3.5% for all but 13 of the 62 × 9 = 558 gene-sample combinations. The cutoff of 3.5% was chosen because that was the largest CV observed for a replicate involving 18S rRNA, which had the smallest mean C_t (near 8). All 13 outliers had mean C_t greater than 30.9 cycles. Only one gene-sample combination (ACE in CLL37) had CV > 8%. So, we concluded that replicate wells gave reproducible values across seven orders of magnitude ($2^{25} \approx 3.4 \times 10^7$), which is consistent with previously published, conventional real-time RT-PCR data (15). Further, the four replicate wells were nearly indistinguishable in most cases. As expected, the expression levels of some genes (including 18S rRNA, *GAPD*, *CD71*, and *Stat3*) appeared relatively constant between samples; the expression levels of other genes (including *CD38* and *CD152*) appeared highly variable.

Data Filtering

We performed a bioinformatic analysis to determine which genes were measured by both the TaqMan Low-Density Array and the Affymetrix U133A GeneChip oligonucleotide microarray. First, all analyses were restricted to genes that were expressed at detectable levels (in at least one of four replicate wells) for all nine CLL samples on the TaqMan Immune Profiling Low-Density Array. This filtering step reduced the number of genes to 62 and the total number of wells providing measurements to 2142. All 62 genes were used in analyses of the real-time RT-PCR data. Of the 96 genes on the Low-Density Array, 88 were represented by at least one probe set on the U133A GeneChip array. Three of the eight genes not represented on the U133A array had been removed because they provided no useful measurements on the Low-Density Array. After removing the other five genes, we were left with 57 genes on the Low-Density Array represented by at least one probe set. Because of redundancy on the U133A GeneChip, we could compare the real-time RT-PCR data for these 57 genes with 87 probe sets.

Modeling Real-Time RT-PCR Data

Two recent papers have introduced similar statistical models for the normalization of real-time RT-PCR data (16,17). Both models (Equation A in Reference 16 and Model 1a in Reference 17) used fixed effects for genes and samples and an error model, accounting for gene-specific variability. We evaluated seven different models for their ability to describe our data (Table 1). Models 1A and 1B are identical to those introduced in Reference 17; for the ease of comparison, we have retained the same alphanumeric labels. Model 1A uses a gene-specific error model, while model 1B assumes that the variability is the same for all genes. Models 2A and 2B have parallel error structures; model 2 differs from model 1 by adding a fixed effect γ_{ij}, which represents different expression levels for gene j in sample i. Models 3–5 use only a fixed effect for the average expression of each gene; they incorporate random effects to account for sample differences (18). Model 3 uses a simple error structure that attributes a random effect, $B_i \sim N(0, \sigma_B^2)$, with a common variance to sample i. Model 4 extends this error structure by attributing an additional random effect, $C_{ij} \sim N(0, \sigma_C^2)$, to gene j in sample i, again assuming a common variance. Model 5, similar to models 1A and 2A, allows for gene-specific variability in the error model. All random effects and errors are assumed to be independent and normally distributed with mean zero.

All seven models were evaluated both on the subset of 6 genes displayed in Figure 1 and on the full data set of 62 genes. Models were fit in S-Plus using the functions lm (1B and 2B), gls (1A and 2A), and lme (models 3–5). The validity of the model assumptions was assessed graphically (Supplementary Figures S1–S14; all Supplementary Materials are available at bioinformatics.mdanderson.org/Supplements/Microfluidics/index.html). From these graphs, we concluded that the assumptions were violated for models 1A, 1B, and 3: the residuals were not centered at zero, they were heteroscedastic, and they exhibited marked departures from normality. The assumptions were reasonable for the remaining models, with the only departure from normality being heavier tails in the distributions.

Random-effects models were fit using maximum likelihood (ML) rather than restricted maximum likelihood (REML) for better comparison with models using different fixed effects.

Table 1. Statistical Models Used to Analyze Real-Time RT-PCR Data

ID	Model	Random Effects	Number of Parameters[a]		AIC		BIC	
			Six	All	Six	All	Six	All
1A	$Y_{ijk} = \mu + \alpha_j + \beta_i + E_{ijk}$	$E_{ijk} \sim N(0, \sigma_j^2)$	2G + S - 1					
			20	132	379	5240	447	5989
1B	$Y_{ijk} = \mu + \alpha_j + \beta_i + E_{ijk}$	$E_{ijk} \sim N(0, \sigma^2)$	G + S					
			15	69	798	7468	848	7860
2A	$Y_{ijk} = \mu + \alpha_j + \beta_i + \gamma_{ij} + E_{ijk}$	$E_{ijk} \sim N(0, \sigma_j^2)$	GS + G					
			60	620	-275	-197	-72	3318
2B	$Y_{ijk} = \mu + \alpha_j + \beta_i + \gamma_{ij} + E_{ijk}$	$E_{ijk} \sim N(0, \sigma^2)$	GS + 1					
			55	559	-188	2527	-3	5696
3	$Y_{ijk} = \mu + \alpha_j + B_i + E_{ijk}$	$B_i \sim N(0, \sigma_B^2)$, $E_{ijk} \sim N(0, \sigma^2)$	G + 2					
			8	64	806	7488	833	7851
4	$Y_{ijk} = \mu + \alpha_j + B_i + C_{ij} + E_{ijk}$	$B_i \sim N(0, \sigma_B^2)$, $C_{ij} \sim N(0, \sigma_C^2)$, $E_{ijk} \sim N(0, \sigma^2)$	G + 3					
			9	65	110	4295	40	4663
5	$Y_{ijk} = \mu + \alpha_j + C_{ij} + E_{ijk}$	$C_{ij} \sim N(0, \sigma_j^2)$, $E_{ijk} \sim N(0, \sigma^2)$	2G + 1					
			13	125	35	3865	78	4574

RT-PCR, reverse transcription PCR; ID, unique identifier for the model in this paper; AIC = Akaike Information Criterion; BIC= Bayes Information Criterion.
[a] Number of parameters in the model.

To compare models, we computed both the Akaike Information Criterion (AIC) and the Bayes Information Criterion (BIC); the results are shown in Table 1. Both criteria reward models based on maximum likelihood and penalize models that use more parameters; models with smaller values for AIC or BIC are preferred (18). Based on either AIC or BIC, model 2A provides the best fit to the data. Among the random-effects models, model 5 provides the best fit.

Identifying Endogenous Control Genes

Next, we wanted to identify genes that were expressed at essentially constant levels in our samples to use as endogenous controls. A fundamental characteristic of the random-effects model 5 is that it allows the data to inform us which genes are constant across the set of experimental samples. In particular, σ_j is an explicit estimate of the between-sample standard error of each gene (Figure 2). Because the method also provides an estimate of the within-replicate standard error (σ), the structure of the data is similar to a classical analysis of variance (ANOVA). Under the null hypothesis that the expression of a gene does not change across the samples, the estimated ratio σ_j^2/σ^2 should have an F-distribution with N - 1 and M - N

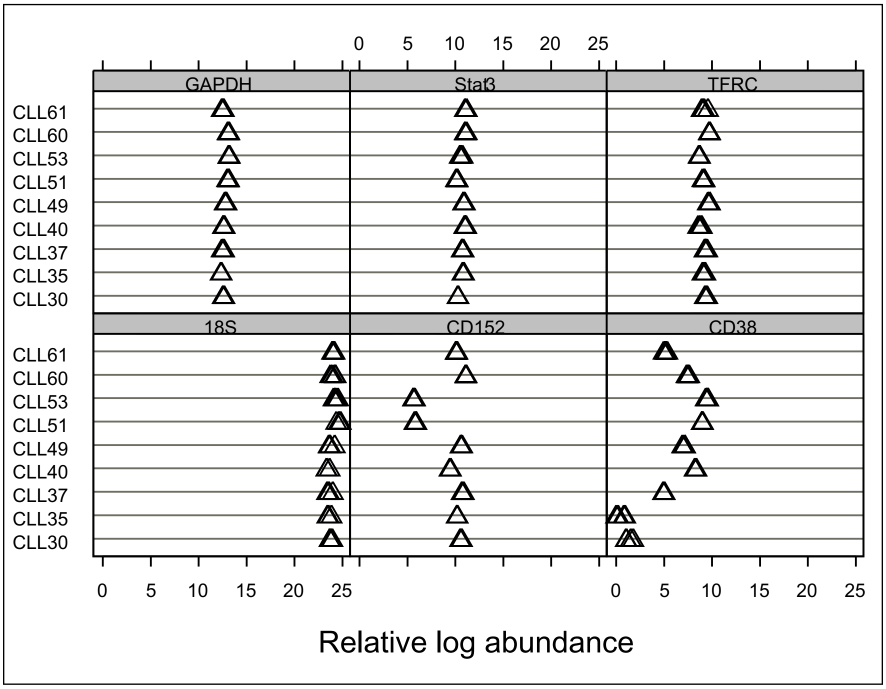

Figure 1. Dot plot of the base-two logarithm of the relative abundance of RNA present for six genes in nine CLL samples. The log relative abundance is computed by adding a constant to the negative cycle threshold (C_t) values to set the minimum value equal to 0. The four replicates for each gene are highly consistent within samples. CLL, chronic lymphocytic leukemia.

degrees of freedom (where N is the number of samples and M is the total number of observed wells). By using the properties of the F-distribution, we can determine statistically whether the variation of a gene between samples is greater than the variation within samples.

For our data, the residual standard error is estimated to be $\sigma = 0.3907$, and, under the null hypothesis, the estimated ratios σ_j^2 / σ^2 should have an F-distribution with 8 and 2017 degrees of freedom. The 90th percentile of this F-distribution is bounded by $\sigma_j^2 / \sigma^2 < 1.673$ or by $\sigma_j < 0.5054$ (*BAX*); the 95th percentile, by $\sigma_j < 0.5446$ (*CD68*); and the 99th percentile, by $\sigma_j < 0.6202$ (*TNFβ*) (Figure 2). Common housekeeping genes such as 18S rRNA, *PGK1*, *CD71*, *GUSB*, and *GAPD* had small standard errors (less than 0.4). This level of variability of housekeeping genes is consistent with other published results (12,17). In contrast, *CD38*, which is known to vary and is reported to have prognostic significance in CLL (19) had one of the largest standard errors (greater than 3.2).

Normalization

Writing $y_{ij\cdot}$ for the average over replicate wells of the measurements of gene j in sample i, the comparative C_t ($\Delta\Delta C_t$) method chooses a calibrator sample $i = 0$ and an endogenous control $j = 0$ and uses the values

$$\Delta\Delta C_t(i,j) = (y_{ij\cdot} - y_{i0\cdot}) - (y_{0j\cdot} - y_{00\cdot}). \quad \text{[Eq. 1]}$$

Under model 5, a simple algebraic calculation shows that this is equivalent to

$$\Delta\Delta C_t(i,j) = (C_{ij} - C_{i0}) - (C_{0j} - C_{00}). \quad \text{[Eq. 2]}$$

As a calibrator, we used the average over all experimental samples. We evaluated several choices of endogenous control. First, we normalized to the geometric mean of nonvarying genes, based on the 90th percentile of the F-distribution. Next, we looked at single-gene normalization with five separate housekeeping genes: *GUSB*, *PGK1*, GAPD, 18S, and *TFRC*. To compare methods, we computed the standard deviation across samples of the normalized values ($\Delta\Delta C_t$) for each method (Figure 3). Normalizing to the geometric mean gave smaller standard deviations for almost all genes compared to any single-gene normalization. The best single-gene normalizations were obtained using *PGK1*, which also had the smallest standard error in model 5. Finally, we normalized to the geometric means of different number of genes (Supplementary Figure S15). Using the first 5 genes gave results comparable to using *PGK1* alone; using 10 or 15 genes gave results comparable to using all 20 genes.

Comparison of U133A GeneChip Microarrays and Low-Density Arrays

In order to compare the results from the U133A chips and the TaqMan Immune Profiling cards, we

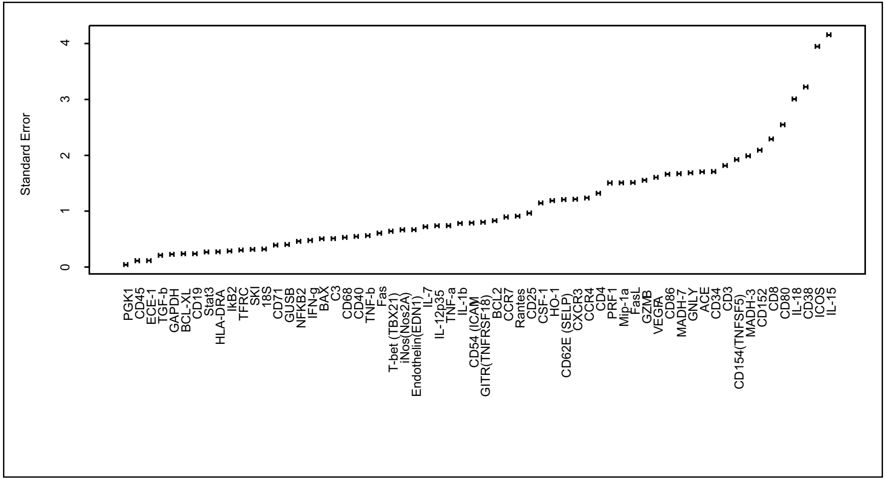

Figure 2. Estimates of the standard error across CLL samples of individual genes based on the mixed-effects model 5. CLL, chronic lymphocytic leukemia.

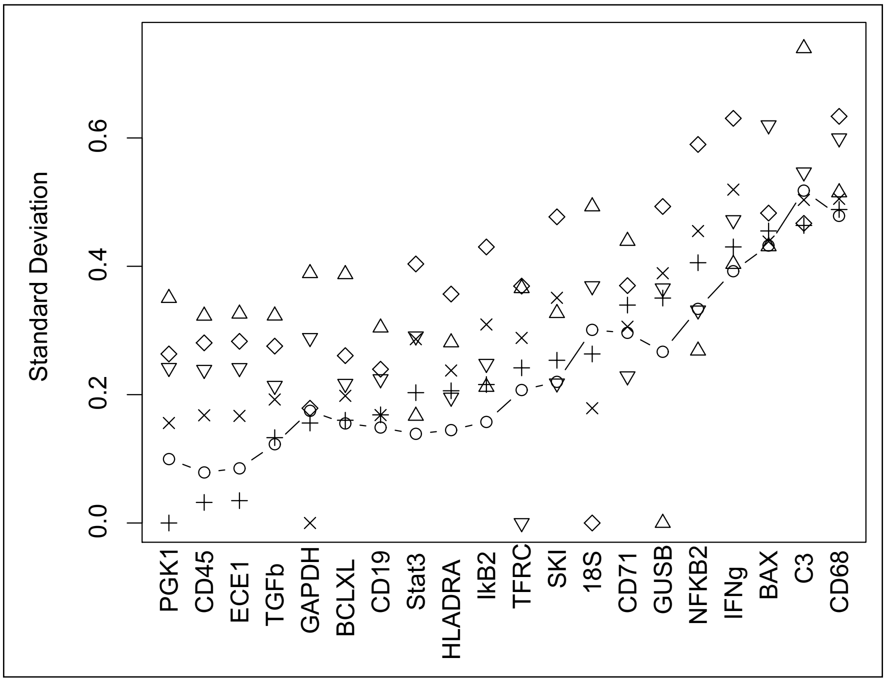

Figure 3. Plots of the standard deviation across samples of the 20 least variable genes. The genes included were based on the mixed-effects model 5 after normalizing to the geometric mean of all 20 genes (ο) or to individual housekeeping genes (+, *PGK1*; ×, *GAPD*; ∇, *TFRC*; ◊, 18S; Δ, *GUSB*).

computed Spearman rank correlation coefficients over the CLL samples (*n* = 9) for each of the 87 probe sets on the U133A chips that matched one of 57 genes on the TaqMan Immune Profiling Low-Density Array. Under the null hypothesis that the probe sets are independent and no gene shows a significant correlation across platforms, the correlation coefficients would have a symmetric beta distribution, *Beta*((*n* - 2)/2, (n - 2)/2), shifted and scaled to lie on the interval from -1 to 1 (20). Justification for the comparison of the observed rank correlations with the beta distribution is provided in the supplementary material available at bioinformatics.mdanderson.org/Supplements/MicroFluidics. A histogram of the observed correlations shows substantial enrichment of highly correlated genes compared with the null distribution; that is, bars that extend above the curve on the right end of the graph represent groups of genes whose expression is more highly correlated between platforms than would be expected by chance (Figure 4).

We next tried to explain why some probe sets were poorly correlated. The observed distribution of correlations in Figure 4 is bimodal, with the highest mode peaking near 0.8. We defined a gene to be poorly correlated if its observed correlation fell significantly below the expected value if the true correlation were 0.8. By simulating 10,000 pairs of length 9 vectors with correlation 0.8, we determined that 95% of the observed rank correlations would be expected to lie above 0.35, and 99% would be expected to lie above 0.08. In our data, 24 of the 27 probe sets with correlation less than 0.08, and 35 of the 42 probe sets with correlation less than 0.35, matched genes that either showed little variation (defined as a standard error less than 0.5054, the 90th percentile of the F-distribution) or were expressed at low levels by real-time RT-PCR (defined as a relative log abundance less than 10). Thus, the correlation between real-time RT-PCR and microarrays was poor for genes that were expressed at low levels or did not vary across samples.

DISCUSSION

In conventional real-time RT-PCR experiments, the observed C_t values of the target genes in an experimental sample are normalized both to a calibrator sample (e.g., total RNA prepared from a standard cell line) and to an endogenous control gene (e.g., 18S rRNA) using the $\Delta\Delta C_t$ method (1). In such experiments, the calibrator and the endogenous control are processed in parallel with the experimental samples and the target genes on the same 96-well plate. Because the calibrator and the experimental samples are processed under identical thermal cycling conditions, normalization to the calibrator automatically adjusts for minor differences in those conditions between runs. Normalization of the target genes to the endogenous control (housekeeping) gene adjusts for differences in the RNA quantity and quality across samples.

Experiments using the TaqMan Low-Density Array microfluidics cards differ from conventional real-time RT-PCR assays in two ways. First, depending on the experiment, each sample may be run on a separate card. Using a calibrator sample from a separate card may introduce bias or increase variability because it may be processed under different thermal cycling conditions. Second, because there may be as many as 384 genes per card, there is a wider choice of endogenous controls. The Immune Profiling Low-Density Array contains probes and primers for several common housekeeping genes, including *GAPD*, *GUSB*, *PGK1*, *TFRC*, and 18S rRNA. The best choice of housekeeping gene to use as an endogenous control varies, depending on the kinds of tissue samples used in the experiment. Both differences in the experimental method must be accounted for when analyzing real-time RT-PCR data from TaqMan Low-Density Arrays.

In our study, we replaced the external calibrator with the mean across all of the samples. A key advantage to using the mean is that it gives valid measurements for all genes that are detected in the experimental samples; an external calibrator can only be used to normalize a gene if it expresses that gene in a detectable amount. While it might be possible to select a calibrator that expresses many of the target genes, it seems unlikely that one could find a calibrator that expresses all of the genes on the Immune Profiling card. In previous conventional real-time RT-PCR experiments, we used a Burkitt lymphoma cell line, GA-10,

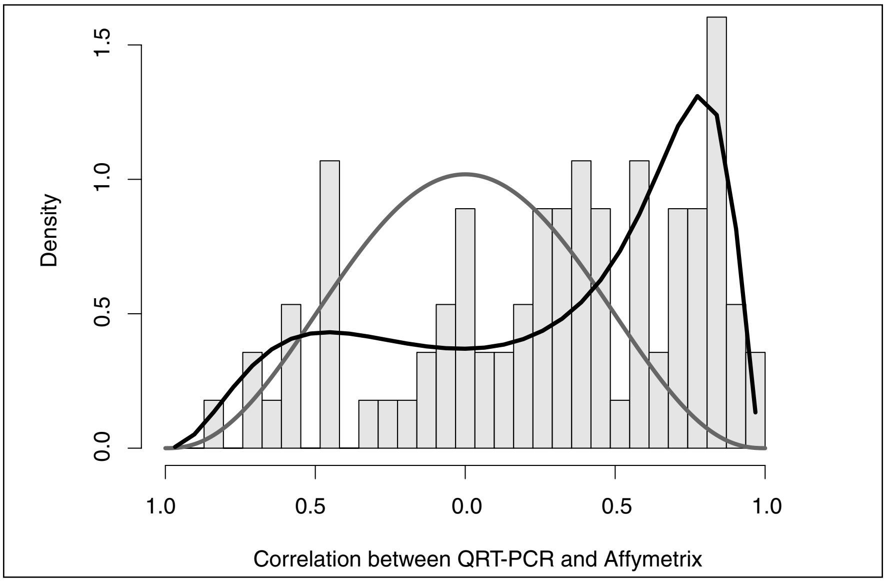

Figure 4. Histogram of observed correlation coefficients between genes on Affymetrix U133A GeneChip microarrays and TaqMan Low-Density Arrays. The overlaid gray curve is the expected null distribution if no genes are significantly correlated; the overlaid black curve is a smooth estimate of the observed distribution, obtained by a B-spline fit after log transformation. QRT-PCR, real-time reverse transcription PCR.

as the calibrator (13). However, 11 of the 62 genes that were expressed in all nine CLL samples were unexpressed in GA-10 (data not shown). One tempting alternative is to designate one experimental sample as a "baseline" and normalize the other samples to the baseline. This alternative would introduce an asymmetry, and might introduce a bias; that depends on the choice of baseline. By using the mean, we avoid this asymmetry by averaging over all possible choices of baseline.

By now, there is ample evidence that no single housekeeping gene will work well in all studies (5–11) and that better results will be obtained using multiple housekeeping genes (3,12,16,17). The key question, then, is how to identify valid endogenous controls for a specific study. Andersen et al. (16) and Szabo et al. (17) have shown that statistical modeling of the data can help solve this problem. In our study, however, model 1A, as proposed in References 16 and 17, proved inferior both to the fixed-effects model 2A and to the random-effects model 5. Better models were obtained by explicitly including a term (γ_{ij} or C_{ij}) that allowed genes to be expressed at different levels in different samples. Without this term, the biological variability between samples is confounded with the technological variability of the assay. In spite of the AIC and BIC results, we contend that the random-effects model 5 is better than the fixed-effects model 2A for identifying housekeeping genes. In order to decide if gene *j* is a housekeeping gene using model 2A, one must test the hypothesis $\gamma_{ij} = 0$ for all *i*. Technically, this hypothesis only applies to the specific samples included in the study and does not extend to the population of similar samples that might be included in future studies. By contrast, the random effect $C_{ij} \sim N(0, \sigma_j^2)$ in model 5 is used to estimate the variability of gene *j* in the population, allowing us to draw inferences about the entire population of CLL patients and not just the specific samples used in this study. This property of the random-effects model is a major motivation behind their development (18).

One can argue that, because of the AIC and BIC results, model 2A is better than model 5 for normalization. However, it is unclear whether researchers would use two models on the same data set (one to select housekeeping genes and another to normalize). To determine how much the models differ, we looked at the predicted values from each model for all 9 × 62 = 558 gene-sample combinations. The differences in the predictions had a mean of -0.00055 and a standard deviation of 0.0787, and these differences were approximately normally distributed. So, one would expect 99% of the differences to be less than 0.2034 in absolute value. There were only a few more outliers than expected; 18 of the 558 predicted values differed by more than 0.2034, and the largest difference was 0.41. To put these numbers in perspective, our estimate of the residual standard error was 0.39, suggesting that the variability attributable to the choice of model is roughly the same as that in replicate wells. For these reasons, we prefer model 5.

There are two reasons why we did not observe large positive correlations for all probe sets in Figure 4. First, as we have shown above, many genes on the Immune Profiling card are effectively expressed at a constant level across all of the CLL samples. For these genes, the technological variation in expression levels dominates the biological variation. Not surprisingly, the sources of technological variation are very different on GeneChip microarrays than they are on the Immune Profiling cards. So, there is no reason to expect them to correlate. Second, the dynamic range of the real-time RT-PCR measurements is much wider than the dynamic range of GeneChip microarrays. As we have shown above, the dynamic range of real-time RT-PCR is approximately seven orders of magnitude (also see Supplementary Figure S16). By contrast, analyses of spike-in experiments performed by Affymetrix suggest that the dynamic range of the U133A GeneChip is about three orders of magnitude (21,22). Similar dynamic ranges have been reported for cDNA microarrays (23). Thus, genes whose mean expression level is low, based on real-time RT-PCR, are less likely to be measured accurately on the U133A GeneChip, where they will be obscured by relatively large noise levels. When one restricts the analysis to higher intensity genes whose expression varies significantly between samples, then the results are well correlated between the two technology platforms.

ACKNOWLEDGMENTS

The authors thank Lynn L. Barron and Candy Cromwell for expert technical assistance. This work was supported by a grant from the Commonwealth Foundation for Cancer Research to L.V.A. and K.R.C. and Mr. and Mrs. William H. Goodwin, Jr. The Microarray Core Facility is supported by Cancer Center Support grant no. 16672.

COMPETING INTERESTS STATEMENT

K.Y.L., A.F., and A.S. are employed by Applied Biosystems Group (Applera Corporation), the manufacturer of a number of products used in this study. The other authors declare no competing interests.

REFERENCES

1. **Livak, K.J. and T.D. Schmittgen.** 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods *25*:402-408.
2. **Suzuki, T., P.J. Higgins, and D.R. Crawford.** 2000. Control selection for RNA quantitation. BioTechniques *29*:332-337.
3. **Tricarico, C., P. Pinzani, S. Bianchi, M. Paglierani, V. Distante, M. Pazzagli, S.A. Bustin, and C. Orlando.** 2002. Quantitative real-time reverse transcription polymerase chain reaction: normalization to rRNA or single housekeeping genes is inappropriate for human tissue biopsies. Anal. Biochem. *309*:293-300.
4. **Lee, P.D., R. Sladek, C.M. Greenwood, and T.J. Hudson.** 2002. Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies. Genome Res. *12*:292-297.
5. **Schmittgen, T.D. and B.A. Zakrajsek.** 2000. Effect of experimental treatment on housekeeping gene expression: validation by real-time, quantitative RT-PCR. J. Biochem. Biophys. Methods *46*:69-81.
6. **Goidin, D., A. Mamessier, M.J. Staquet, D. Schmitt, and O. Berthier-Vergnes.** 2001. Ribosomal 18S RNA prevails over glyceraldehyde-3-phosphate dehydrogenase and beta-actin genes as internal standard for quantitative comparison of mRNA levels in invasive and noninvasive human melanoma cell subpopulations. Anal. Biochem. *295*:17-21.
7. **Gorzelniak, K., J. Janke, S. Engeli, and A.M. Sharma.** 2001. Validation of endogenous controls for gene expression studies in human adipocytes and preadipocytes. Horm. Metab. Res. *33*:625-627.
8. **Deindl, E., K. Boengler, N. van Royen, and W. Schaper.** 2002. Differential expression of GAPDH and beta3-actin in growing collateral arteries. Mol. Cell Biochem. *236*:139-146.
9. **Prieto-Alamo, M.J., J.M. Cabrera-Luque, and C. Pueyo.** 2003. Absolute quantitation of normal and ROS-induced patterns of gene expression: an in vivo real-time PCR study in mice. Gene Expr. *11*:23-34.
10. **Schmid, H., C.D. Cohen, A. Henger, S. Irgang, D. Schlondorff, and M. Kretzler.** 2003. Validation of endogenous controls for gene expression analysis in microdissected human renal biopsies. Kidney Int. *64*:356-360.
11. **Aerts, J.L., M.I. Gonzales, and S.L. Topalian.** 2004. Selection of appropriate control genes to assess expression of tumor antigens using real-time RT-PCR. BioTechniques *36*:84-91.
12. **Vandesompele, J., K. De Preter, F. Pattyn, B. Poppe, N. Van Roy, A. De Paepe, and F. Speleman.** 2002. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol. *3*:research0034.1-0034.11.
13. **McCarthy, H., W.G. Wierda, L.L. Barron, C.C. Cromwell, J. Wang, K.R. Coombes, R. Rangel, K.S. Elenitoba-Johnson, et al.** 2003. High expression of activation-induced cytidine deaminase (AID) and splice variants is a distinctive feature of poor prognosis chronic lymphocytic leukemia. Blood *101*:4903-4908.
14. **Gold, D., K.R. Coombes, D. Medhane, A. Ramaswamy, Z. Ju, L. Strong, J.S. Koo, and M. Kapoor.** 2004. A comparative analysis of data generated using two different target preparation methods for hybridization to high-density oligonucleotide microarrays. BMC Genomics *5*:2.
15. **Malarstig, A., T. Tenno, S. Jossan, M. Aberg, and A. Siegbahn.** 2003. A quantitative real-time PCR method for tissue factor mRNA. Thromb. Res. *112*:175-183.
16. **Andersen, C.L., J.L. Jensen, and T.F. Orntoft.** 2004. Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. Cancer Res. *64*:5245-5250.
17. **Szabo, A., C.M. Perou, M. Karaca, L. Perreard, J.F. Quackenbush, and P.S. Bernard.** 2004. Statistical modeling for selecting housekeeper genes. Genome Biol. *5*:R59.
18. **Pinheiro J.C. and D.M. Bates.** 2000. Mixed-Effects Models in S and S-PLUS. Springer-Verlag, New York.
19. **Damle, R.N., T. Wasil, F. Fais, F. Ghiotto, A. Valetto, S.L. Allen, A. Buchbinder, D. Budman, et al.** 1999. Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. Blood *94*:1840-1847.
20. **Fisher, R.A.** 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika *10*:507-521.
21. **Irizarry, R.A., B.M. Bolstad, F. Collin, L.M. Cope, B. Hobbs, and T.P. Speed.** 2003. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. *31*:e15.
22. **Zhang, L., M.F. Miles, and K.D. Aldape.** 2003. A model of molecular interactions on short oligonucleotide microarrays. Nat. Biotechnol. *21*:818-821.
23. **Worley, J., K. Bechtol, S. Penn, D. Roach, D. Hanzel, M. Trounstine, and D. Barker.** 2000. A systems approach to fabricating and analyzing DNA microarrays, p. 65-86. *In* M. Schena (Ed.), Microarray Biochip Technology. Eaton Publishing, Natick, MA.

Received 2 August 2004; accepted 11 January 2005.

Address correspondence to:

Kevin R. Coombes
Department of Biostatistics and Applied Mathematics
University of Texas M.D. Anderson Cancer Center
1515 Holcombe Blvd., Box 447
Houston, TX 77030, USA
e-mail: kcoombes@mdanderson.org