A Computational Approach Using Ratio Statistics for Identifying Housekeeping Genes from cDNA Microarray Data

T. Sengupta, M. Bhushan, and P. P. Wangikar

Abstract—We predict housekeeping genes from replicate microarray gene expression data of human lymphoblastoid cells and liver tissue with outliers removed using a scoring scheme, by an algorithm based on statistical hypothesis testing, assuming that such genes are constitutively expressed. A few predicted genes were examined and found to be housekeeping.

Index Terms—Housekeeping genes, liver tissue, lymphoblastoid cells, microarrays, outliers

1 Introduction

A microarray experiment is a high-throughput approach for simultaneously inspecting the expression of a large number of genes (of the order of thousands) in a single experiment.

In this work, we consider two sample cDNA microarray experiments. In such experiments, mRNA is extracted from two samples (target and control) and reverse transcribed to cDNAs. Now, the two pools of cDNAs are labeled with different fluorescent dyes (red and green) and allowed to hybridize with probes present at specific locations called spots in a microarray slide. The probes are also cDNA molecules. At a spot, probes hybridize with copies of the same cDNA molecule (with sequence of bases complementary to the probes at the spot) from the two samples. This cDNA molecule provides a pointer to the gene which codes for the mRNA molecule from which the cDNA molecule was obtained by reverse transcription. Thus each spot can be associated with some gene. The expression ratio at a spot is the ratio of the representative red and green intensity values being emitted from the spot and thus provides a quantitative estimate of gene expression in the target relative to the control for the gene corresponding to the spot.

The first step in analyzing raw microarray experimental data is the identification and removal of gross measurement errors or outliers. Outliers can be conceptually identified in repeated measurements on the same sample by noting values which differ considerably from the rest of the measurements. In this work, we propose an extension of an outlier identification technique from microarray data with two replicates [1] to more than two replicates, where replicates connote multiple microarray experiments performed with the same samples.

After outlier removal, microarray data needs to be normalized. This is because, in any microarray experiment, systemic variations i.e., factors such as cross-hybridization, dye bias, etc. [2] introduce random errors and distort expression ratio values. Normalization is the process of assessing the quantitative effect of such factors. After normalization, expression ratios can be inspected to identify constitutively and differentially expressed genes.

Several normalization techniques exist in literature for two color microarray data. In fold change, a gene is considered differentially expressed if it's expression ratio value falls outside a certain range. However, it is not a statistical test and hence provides no information about typical statistical test parameters such as the significance level [3]. Statistical tests for normalization can be parametric in which data is assumed to follow some distribution or non-parametric in which no assumptions are made on the distribution of data. Examples of parametric tests are two sample t-tests [4]. For such tests, variance estimates for a gene are usually inaccurate due to low sample size [3]. Variance shrinkage resolves this issue by computing the variance for a gene from estimated values of the variance for the gene and that for the entire data. This results in a modified t-test [3]. Significance Analysis of Microarrays (SAM) [5] and Linear Models for Microarray Data (LIMMA) [6] package in R [7] are two examples of techniques using variance shrinkage. In SAM, model parameters are computed by resampling the data via permutations [4]. Another example of a resampling based technique is bootstrap analysis [3]. However, resampling based techniques are computationally intensive [3] as opposed to parametric tests. All the above methods assume that transcript levels for genes are independent. Methods, such as over-representation analysis, gene set enrichment analysis and optimal discovery procedure, have also been proposed which do not make such an assumption [4]. A recent paradigm in microarray data analysis is meta-analysis which concerns itself with integration and analysis of microarray data from distinct studies on the same set of conditions [8]. We note that a particular normalization technique is best suited for only certain data types [9].

In this work, we propose a modification of an existing parametric normalization strategy [10]. Using different probability density functions, we remove the requirement of the original strategy that the model parameter coefficient of variation (ratio of the standard deviation and the mean of the emitted intensities at a spot) be constrained to values lower than 0.3. The proposed modification can thus be applied to data with high coefficient of variation.

Finally, we propose an algorithm for determining housekeeping genes assuming that such genes are constitutively expressed across all the experimental conditions under consideration. We note that although expression levels of housekeeping genes vary depending on factors such as the tissue sample [11], it is always possible to define a constitutively expressed gene set under a finite set of experimental conditions.

2 Materials and Methods

2.1 A Scoring Scheme for Identifying Outliers from Replicate Microarray Experimental Data

We now describe an extension of an existing strategy for identifying outliers from replicate microarray experimental data. Replicate microarray experimental data implies expression data obtained from several microarray slides with cDNAs (reverse transcribed from mRNAs) from the same pair of samples being hybridized to each slide. The entire gamut of data obtained from all the slides is said to form a dataset and the data from a single slide is said to form a "replicate dataset". We first describe the original strategy [1].

2.1.1 Identifying Outliers from a Duplicate Dataset [1]

Let $R_{ij}, i = 1, 2, \ldots, G, j = 1, 2$ denote expression ratio for the i^{th} gene in the j^{th} duplicate (i.e., two replicates are performed) dataset. First, the Log Ratios $R_{Kj_1j_2} = log_2(R_{Kj_1}/R_{Kj_2})$, $j_1 = 1$, $j_2 = 2$, $K = 1, 2, \ldots, G$, are computed for each gene. Thereafter, the mean ($\mu_{j_1j_2}$) and the standard deviation ($\sigma_{j_1j_2}$) of the Log Ratios $R_{Kj_1j_2}$ across all genes $K = 1, 2, \ldots, G$, are obtained. Now, if any one of the two data points R_{Kj_1} and R_{Kj_2} is an outlier, then $R_{Kj_1j_2}$ would deviate significantly from zero. Otherwise, the Log Ratio $R_{Kj_1j_2}$ would be close to zero since the expression ratios R_{Kj_1} and R_{Kj_2} should be nearly equal in the absence of errors. Thus, if for the

• *The authors are with the Department of Chemical Engineering, Indian Institute of Technology Bombay, Mumbai 400076, India.*
E-mail: {tirthankar.s, mbhushan, wangikar}@iitb.ac.in.

Manuscript received 10 June 2014; revised 20 Dec. 2014; accepted 4 Feb. 2015. Date of publication 26 Feb. 2015; date of current version 4 Dec. 2015.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2015.2407399

TABLE 1
A Typical Dataset

	Exp 1	Exp 2	Exp 3		Exp n
Gene 1	MDP	R_{12}	R_{13}	$\cdots$	R_{1n}
Gene 2	R_{21}	R_{22}	R_{23}	$\cdots$	MDP
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
Gene G	R_{G1}	R_{G2}	R_{G3}	$\cdots$	R_{Gn}

TABLE 2
Comparison Table for Two Sets of Replicate Data

	Exp 1	Exp 2	Log Ratios	C(1,2)
Gene 1	MDP	R_{12}	X	X
Gene 2	R_{21}	R_{22}	R_{212}	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Gene G	R_{G1}	R_{G2}	R_{G12}	0

TABLE 3
Base Table for a Gene for Four Replicate Experiments

C(1,2)	C(1,3)	C(1,4)	C(2,3)	C(2,4)	C(3,4)
X	X	X	1	1	0

K^{th} gene, $|R_{Kj_1j_2} - \mu_{j_1j_2}| > 2\sigma_{j_1j_2}$, then the two expression ratios R_{Kj_1} and R_{Kj_2} are regarded as outliers.

2.1.2 Identifying Outliers from a Dataset with Multiple Replicates

We extend the above strategy to datasets in which the number of replicates is more than two. Table 1 represents a typical dataset with G genes and n replicate experiments in which $R_{ij}, i = 1, 2, \ldots, G, j = 1, 2, \ldots, n$ are expression ratios and MDP represents a missing data point. We compare one pair of experiments at a time. For example, comparing experiments 1 and 2 gives rise to Table 2. A cross appears in the fourth and fifth columns of Table 2 whenever either or both R_{Kj_1} and R_{Kj_2} are missing. Otherwise, the Log Ratio $R_{Kj_1j_2}$ is computed as in Section 2.1.1. Now the mean ($\mu_{j_1j_2}$) and standard deviation ($\sigma_{j_1j_2}$) of all entries in the fourth column other than the crosses are computed. If a particular value in the fourth column in Table 2 differs from $\mu_{j_1j_2}$ by more than $2\sigma_{j_1j_2}$, then a score of one is assigned to the comparison (implying the two corresponding expression ratios are potential outliers). Otherwise a score of zero is assigned to the comparison (implying the two corresponding expression ratios are not potential outliers). The column header $C(1, 2)$ in the fifth column of Table 2 denotes that experiments 1 and 2 are being compared. This column stores the scores. Pairwise comparisons between all experiments would give rise to a table with ${}^{n}C_2$ columns of scores (each column akin to the fifth column of Table 2) and G rows. This table is referred to as the base table and is analyzed for identifying outliers. This is done one row at a time i.e. one gene at a time using an iterative procedure.

For the purpose of illustration, we describe the iterative procedure for a gene with four replicate experiments. The first iteration starts by considering Table 3 which is the base table for the gene. In Table 3, the data point from experiment 1 is missing giving rise to crosses wherever experiment 1 appears in the column header. From Table 3, the scoring table (Table 4) for the first iteration is generated by counting the number of times the entry for an experiment is one. In Table 4, S_{Ki} denotes the score for the K^{th} (current) gene from the i^{th} experiment. All data points with the maximum score in Table 4 are regarded as outliers in accordance with the principle of Ockham's razor [12]. It is expected that the number of outliers would be far fewer than the number of valid data points. The principle of Ockham's razor minimizes the number of outliers since it makes the fewest new assumptions on the data and is hence used for determining outliers. Thus, from Table 4, data point 2 becomes an outlier. Now, the new base table (Table 5) for the first iteration is obtained by eliminating experiments pertaining to missing data points (experiment 1) and outliers (experiment 2) from the previous base table of the current iteration (Table 3). It is observed that in the new base table all comparisons have a score of zero. This implies that there is no need for subsequent iterations for the current gene. Thus, Table 6 results in which a missing data point is represented as MDP, an outlier 'o' and all other entries are valid expression ratios. This procedure of identifying outliers in expression ratios is then performed for all genes. A flowchart describing the algorithm is shown in Fig. 1.

TABLE 4
Scoring Table for a Particular Gene for Four Replicate Experiments

S_{K1}	S_{K2}	S_{K3}	S_{K4}
MDP	2	1	1

TABLE 5
Base Table for a Gene with Columns Pertaining to Experiments 1 and 2 Removed

C(3,4)
0

TABLE 6
Table Showing Outliers for a Particular Gene ("DP" Refers to "Data Point")

DP 1	DP 2	DP 3	DP 4
MDP	o	R_{K3}	R_{K4}

2.1.3 Discussion on the Proposed Outlier Identification Strategy

In the proposed outlier detection approach, information across all genes for a given set of experiments is used to identify an outlier expression ratio. In principle, the outliers could have been conceptually identified from replicate measurements for a *particular gene* by simply identifying observations which differ significantly from the rest for that gene only. This would not have involved expression ratios of other genes. A statistical test can be formulated for such an identification. The performance of this alternate simple approach is however likely to be inferior compared to our approach due to the small sample size (number of replicates) available for such analysis. In our approach, the sample size is the number of genes which is almost always expected to be large due to the very nature of microarray experiments. Irrespective of the variation in the magnitude of expression ratios for various genes, the Log Ratios are expected to have values of the same order of magnitude and hence can be meaningfully compared as in our approach.

2.1.4 Importance of Replicates

We have considered technical replicates in this work. Replicates are important in microarray data analysis, since without replicates it is impossible to gauge the noise in microarray data due to systemic variations and also to identify outliers [3]. Although both technical as well as biological replicates are beneficial, technical replication is essential for investigations pertaining to quality

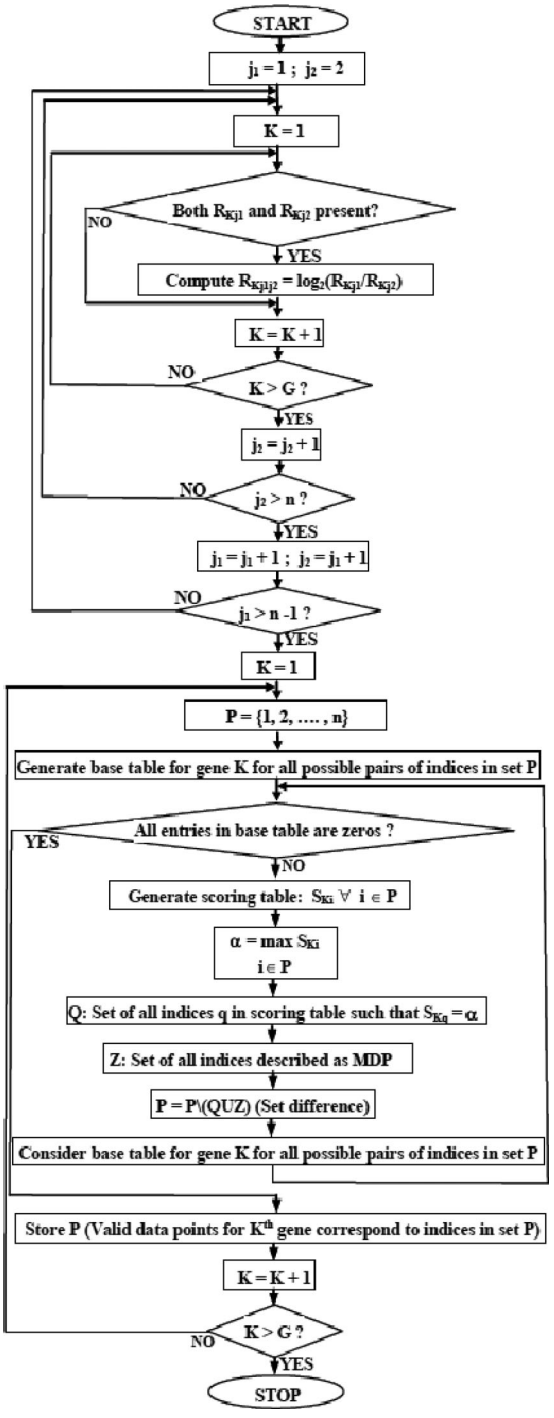

Fig. 1. Flowchart describing the algorithm for finding outliers.

control [3]. Also, when systemic variations introduce substantial noise in the data, technical replication is strongly advisable [13]. The larger the number of such replicates, the lesser is the estimation error, although at the expense of greater cost incurred [13]. Our outlier identification scheme thus provides an advantage in microarray data analysis over the original strategy, the latter being restricted to two replicates only.

2.2 A Ratio Statistics Based Normalization Strategy for cDNA Microarray Data

After outlier removal, the next step is to normalize the expression data. We first describe the original mathematical model used for normalization purposes [10] and then the proposed modification.

2.2.1 The Original Model [10]

In the original approach, the fluorescent intensity, designated as R_K, emanating at the red wavelength from the K^{th} spot is modeled as a Gaussian random variable i.e. $R_K \sim \mathcal{N}(\mu_{R_K}, \sigma^2_{R_K})$. Similarly, the intensity G_K emanating at the green wavelength is also modeled as a Gaussian random variable $G_K \sim \mathcal{N}(\mu_{G_K}, \sigma^2_{G_K})$. We note that while the mean and variance vary across genes, it is assumed that, for each gene the standard deviation is related to the corresponding mean as [10],

$$\left.\begin{aligned}\sigma_{G_K} &= c\mu_{G_K}\\ \sigma_{R_K} &= c\mu_{R_K}\end{aligned}\right\} \tag{1}$$

where c is the coefficient of variation assumed to be the same at all spots for both the wavelengths.

In the work of [10], it is assumed that for a constitutively expressed gene, the means of the intensities for the red and green wavelengths are related as $\mu_{R_K} = m\mu_{G_K} = m\mu_K$, where $\mu_K = \mu_{G_K}$ and m is an amplification factor assumed to be the same for all constitutively expressed genes. Constitutively expressed genes can now be detected by a hypothesis testing procedure with the null hypothesis being that the gene corresponding to the K^{th} spot is constitutively expressed, and the alternative hypothesis being that the gene is not constitutively expressed i.e. [10] :

Null Hypothesis is $\mathrm{H}_0 : \mu_{R_K} = m\mu_{G_K} = m\mu_K$

Alternate Hypothesis is $\mathrm{H}_1 : \mu_{R_K} \neq m\mu_{G_K}$

Based on (1), it can be seen that when the null hypothesis holds true, $\sigma_{G_K} = c\mu_{G_K} = c\mu_K$ and $\sigma_{R_K} = c\mu_{R_K} = cm\mu_K$.

The test statistic T_K for testing the null hypothesis is the ratio of the two intensities, i.e. $T_K = R_K/G_K$ with probability density function [10]:

$$f_{T_K}(t) = \int_0^\infty g f_{R_K}(tg) f_{G_K}(g) dg - \int_{-\infty}^0 g f_{R_K}(tg) f_{G_K}(g) dg \tag{2}$$

For c < 0.3, the tails of the density functions for R_K and G_K to the left of the origin are negligible [10]. This leads to the following simplification of (2) (the second integral in (2) becomes zero):

$$f_{T_K}(t) = \int_0^\infty g f_{R_K}(tg) f_{G_K}(g) dg \tag{3}$$

Now, since R_K and G_K are normally distributed with the density functions lying entirely in the positive domain, we have (when the null hypothesis is true):

$$f_{R_K}(tg) = \frac{1}{cm\mu_K\sqrt{2\pi}} \exp\left[\frac{-(tg - m\mu_K)^2}{2c^2m^2\mu_K^2}\right], 0 \leq tg < \infty \tag{4}$$

$$f_{G_K}(g) = \frac{1}{c\mu_K\sqrt{2\pi}} \exp\left[\frac{-(g - \mu_K)^2}{2c^2\mu_K^2}\right], 0 \leq g < \infty \tag{5}$$

Combining (3), (4) and (5), the following density function is arrived at:

$$f_T(t) = \frac{1}{2\pi mc^2} \int_0^\infty u \exp\left[\frac{-\left(\frac{t}{m}u - 1\right)^2}{2c^2}\right] \exp\left[\frac{-(u-1)^2}{2c^2}\right] du \tag{6}$$

where $u = g/\mu_K$. We note that the subscript K does not appear in $f_T(t)$, implying that the density function is the same at all spots. As mentioned, (6) is a valid density function as long as $c < 0.3$.

2.2.2 Proposed Model without Restricting the Value of the Coefficient of Variation

For $c > 0.3$, the tails of the density functions for R_K and G_K to the left of the origin are not negligible and hence normal density functions truncated at the origin must be used. It is in the use of these truncated density functions where the model proposed by us differs from the one developed by Chen et al. [10]. Equations (1), (2) and (3) are valid in our model also. The proposed density function is,

$$f_{R_K}(tg) = \begin{cases} \frac{1}{(1-A_1)} \frac{1}{\sigma_{R_K}\sqrt{2\pi}} \exp\left[\frac{-(tg-\mu_{R_K})^2}{2\sigma_{R_K}^2}\right], 0 \le tg < \infty, \\ 0, otherwise \end{cases} \quad (7)$$

where A_1 is the area of the non-truncated normal density function to the left of the origin:

$$A_1 = \frac{1}{\sigma_{R_K}\sqrt{2\pi}} \int_{-\infty}^{0} \exp\left[\frac{-(tg-\mu_{R_K})^2}{2\sigma_{R_K}^2}\right] d(tg) \quad (8)$$

Hence,

$$\int_0^{\infty} f_{R_K}(tg) d(tg) = 1 \quad (9)$$

Combining (1), (7) and (8),

$$f_{R_K}(tg) = \frac{1}{(1-A_1)} \frac{1}{cm\mu_K\sqrt{2\pi}} \exp\left[\frac{-(tg-m\mu_K)^2}{2c^2m^2\mu_K^2}\right], 0 \le tg < \infty \quad (10)$$

and

$$\begin{aligned} A_1 &= \frac{1}{cm\mu_K\sqrt{2\pi}} \int_{-\infty}^{0} \exp\left[\frac{-(tg-m\mu_K)^2}{2c^2m^2\mu_K^2}\right] \quad d(tg) \\ &= \frac{1}{c\sqrt{2\pi}} \int_{-\infty}^{0} \exp\left[\frac{-(z-1)^2}{2c^2}\right] dz, \end{aligned} \quad (11)$$

where $z = (tg/m\mu_K)$. Similarly,

$$f_{G_K}(g) = \begin{cases} \frac{1}{(1-A_2)} \frac{1}{c\mu_K\sqrt{2\pi}} \exp\left[\frac{-(g-\mu_K)^2}{2c^2\mu_K^2}\right], 0 \le g < \infty \\ 0, otherwise \end{cases} \quad (12)$$

with

$$A_2 = \frac{1}{c\sqrt{2\pi}} \int_{-\infty}^{0} \exp\left[\frac{-(q-1)^2}{2c^2}\right] dq, \quad (13)$$

where $q = g/\mu_K$. Comparing (11) and (13), it is observed that $A_1 = A_2$. Let $A = A_1$. Thus,

$$f_{R_K}(tg) = \frac{1}{(1-A)} \frac{1}{cm\mu_K\sqrt{2\pi}} \exp\left[\frac{-(tg-m\mu_K)^2}{2c^2m^2\mu_K^2}\right], 0 \le tg < \infty \quad (14)$$

$$f_{G_K}(g) = \frac{1}{(1-A)} \frac{1}{c\mu_K\sqrt{2\pi}} \exp\left[\frac{-(g-\mu_K)^2}{2c^2\mu_K^2}\right], 0 \le g < \infty \quad (15)$$

Combining (3), (14), (15) and performing the variable transformation: $u = g/\mu_K$, (16) is obtained which gives the probability density function $f_{T_K}(t)$ of the test statistic T_K. Noting that μ_K is the only parameter in the model which varies from spot to spot and that it appears only in the variable of integration with the limits of integration being constants, it is concluded that the density function $f_{T_K}(t)$ is the same for all spots. This density function is denoted by $f_T(t)$:

$$f_T(t) = \frac{1}{(1-A)^2} \frac{1}{2\pi mc^2} \int_0^{\infty} u \exp\left[\frac{-\left(\frac{t}{m}u-1\right)^2}{2c^2}\right] \exp\left[\frac{-(u-1)^2}{2c^2}\right] du \quad (16)$$

We note that when $c < 0.3$, $A \approx 0$ or equivalently $1/(1-A)^2 \approx 1$. Thus when $c < 0.3$, (16) reduces to (6) and the two normalization strategies give similar results.

2.2.3 Applicability of the Proposed Model

We now discuss the applicability of the proposed normalization strategy. First, as in the original strategy, we assume a constant coefficient of variation for all spots. This is true in microarray experiments when the signal to noise ratio is high [14]. High signal to noise ratios occur when high quality microarrays such as Agilent microarrays are used. Second, the pertinent question arises that whether high values of the coefficient of variation are possible in microarray experiments. A high coefficient of variation occurs when mRNA levels vary widely among genes. Such a scenario is expected in the case of, for example, human gene expression data for which expression patterns between individuals are highly variable [15]. In fact, in the results section, we report high values of coefficient of variation for an individual, whose expression data [16] is analyzed in this work. Third, our probability density functions are truncated normal and the normality assumption requires that the number of genes probed be large in accordance with the Central limit theorem.

2.2.4 Obtaining the Values of the Model Parameters c and m for Each Set of Replicate Data in a Dataset

We now discuss the procedure for obtaining c and m parameter values from datasets using the density function described by (16). Considering the dataset in Table 1, the following relationship holds true:

$$G = v^j + n_o^j + n_m^j,\ j = 1, 2, \ldots, n \quad (17)$$

In (17), v^j is the number of valid data points, n_o^j is the number of outliers and n_m^j is the number of missing data points for the j^{th} replicate experiment. Now, the data in each column (replicate) is analyzed separately. For the j^{th} replicate experiment, let S_j^V denote the set of genes with valid expression ratios. Let the valid expression ratio sample values for the genes in S_j^V be $t_i^j, i \in S_j^V$. Then, for the j^{th} replicate experiment, the likelihood function is defined as:

$$L_j = \prod_{i \in S_j^V} f_T(t_i^j) \quad (18)$$

For the j^{th} replicate experiment, a set of values c_j and m_j are obtained which maximize the function $\log L_j$. In our work, this maximization over c_j and m_j is performed using the in built function "fminsearch" in MATLAB which uses derivative free optimization methods.

2.3 Identifying a Set of Housekeeping Genes Using Our Normalization Strategy

The values of the model parameters c_j and m_j obtained for each set of replicate data in a dataset are now used to obtain a candidate set of housekeeping genes from the dataset. First, confidence intervals

are calculated for each replicate experiment. At a significance level α, the $100(1-\alpha)\%$ confidence interval is calculated for the j^{th} replicate experiment using the values c_j and m_j with the density function being defined by (16). Confidence intervals are calculated using an interval halving bisection search algorithm.

Next, from the set S_j^V, a set of genes S_j is obtained such that the expression ratio of each of the genes in the set S_j falls within the confidence interval calculated for the j^{th} replicate experiment. Thus,

$$S_j \subseteq S_j^V \tag{19}$$

The candidate set of housekeeping genes H is defined as the common intersection of all sets S_j, $j = 1 : n$. Hence,

$$H = \bigcap_{j=1}^{n} S_j \tag{20}$$

Using the above mentioned procedure, a candidate set of housekeeping genes can be obtained for an organism from several different experiments with the data from each experiment forming a dataset. Let the candidate set of housekeeping genes from the p^{th} dataset be the set H_p. Then, if the total number of datasets under consideration is N, the set of housekeeping genes for the organism (Hg) is defined as:

$$Hg = \bigcap_{p=1}^{N} H_p \tag{21}$$

Hence, for a gene to be labeled housekeeping gene with the given datasets, it has to be constitutively expressed for all the replicates across various datasets.

The inclusion of the non-housekeeping genes in training data will result in a calculated confidence interval to be longer than it actually should be if calculated with housekeeping genes only. Hence it is expected that false negatives will be predicted i.e. non-housekeeping genes will be predicted as housekeeping. However, the common intersection across several datasets would eliminate such false negatives. A non-housekeeping gene predicted as a candidate housekeeping gene in a dataset due to an expression ratio close to unity is expected to be up or down regulated in some other dataset thereby not being regarded as a candidate housekeeping gene in the other dataset. Hence this gene would not appear in the final housekeeping gene set.

2.4 Identifying Housekeeping Genes Using Chen's Normalization Strategy [10]

For comparison with our normalization strategy, the normalization strategy proposed by Chen et al. [10] was also used to arrive at a set of housekeeping genes. In this case, the c and m parameter values and the confidence intervals were calculated by the technique described by Chen et al. [10] for the case of uncalibrated signals. The rest of the procedure remained the same as in identifying housekeeping genes using our normalization strategy.

2.5 Artificially Generated Expression Data

For validation purposes, two sets of artificial data: one for which the null hypothesis was true for all genes and another in which the null hypothesis was true for only some genes, were generated and the two normalization strategies (proposed and Chen's) were applied to both the sets. Investigations on these datasets would reveal the efficacy of the two normalization strategies.

2.5.1 Data with all Genes as Housekeeping

We generated artificial data such that all genes are housekeeping as follows. First, values of the parameters c and m were chosen. These were kept constant in the repeated samplings for generating the data. A value for μ_{G_K} was obtained by sampling from a continuous uniform distribution with a range of 100 to 30,000 [10]. Next a value for G_K was generated by sampling from a truncated normal distribution with mean μ_{G_K}, standard deviation $\sigma_{G_K} = c\mu_{G_K}$ and a range of zero to infinity. Similarly, a value for R_K was generated by sampling from a truncated normal distribution with mean $\mu_{R_K} = m\mu_{G_K}$, standard deviation $\sigma_{R_K} = cm\mu_{G_K}$ and a range of zero to infinity. This resulted in a single datum for the expression ratio $T_K = R_K/G_K$. In all, 5,000 data points were generated.

The values of the parameter c used to generate the artificial data were taken to be 0.2, 0.4 and 0.6. For each of these values, two sets of artificial data were generated with the values of the parameter m being 0.8 and 1.2.

2.5.2 Data with Some Genes as Housekeeping

We also generated artificial data with some genes as housekeeping and some genes as non-housekeeping, thereby mimicking an actual dataset. We generated expression data for 4,500 housekeeping genes using the procedure described in Section 2.5.1 with the values of parameter c and m being 0.18 and 1.13 respectively [10]. Expression data for 500 non-housekeeping genes were then generated as follows. The value of the parameter c was again taken to be 0.18. First, a value for μ_{G_K} was sampled as described in Section 2.5.1. Thereafter, a single datum for G_K was generated by sampling from a truncated normal distribution with mean μ_{G_K}, standard deviation $\sigma_{G_K} = c\mu_{G_K}$ and a range of zero to infinity. Next, a single datum for R_K was generated by sampling from a truncated normal distribution with mean $\mu_{R_K} = m_K\mu_{G_K}$, standard deviation $\sigma_{R_K} = c\mu_{R_K} = cm_K\mu_{G_K}$ and a range of zero to infinity. The value of m_K was sampled from a piecewise uniform distribution with range 0.3 to 0.8 and 1.5 to 3. Thus, for the non-housekeeping genes, the value of the parameter m was taken to be different from 1.13 since for such genes the null hypothesis (Section 2.2.1) does not hold true.

2.6 Identifying Housekeeping Genes Using Method of Rodriguez-Lanetty et al. [17]

To compare with our housekeeping gene identification technique, we implemented the method of Rodriguez-Lanetty et al. [17] which also identifies housekeeping genes from cDNA microarray data. For this method, all expression ratios for a gene constitute a sample. A gene is regarded as constitutively expressed if the sample has low coefficient of variation, the sample mean is not significantly different from unity (tested using a one sample t-test), the mean expression in all the groups or treatments are the same (tested using ANOVA) and the transcripts for the gene are abundant in both target and control. For the sake of a proper comparison with our technique, we removed the requirement of abundance of transcripts and introduced the requirement that no expression ratios for a gene should be missing.

2.7 Normalization Using LIMMA [6]

We also normalized expression data [16] using the LIMMA [6] package in R [7], again for comparison. The methodology was similar to that described in Smyth and Speed [18] with some modifications. We used "loess" to normalize data within arrays, "avereps" to average over replicate spots and the functions "lmFit" and "ebayes" to compute statistics for the data.

2.8 Gene Expression Data

The data analysis techniques mentioned in Sections 2.1–2.4 and 2.6–2.7 were used to analyze cDNA microarray data pertaining to lymphoblastoid cells of 35 individuals reported by Cheung et al. [16]. The entire expression data consisted of 35 datasets with 4

TABLE 7
Table Showing Simulation Results on Artificial Data

Data Generation		Technique 1		Technique 2	
c	m	c	m	c	m
0.2	0.8	0.202	0.796	0.202	0.796
0.2	1.2	0.197	1.202	0.197	1.201
0.4	0.8	0.402	0.801	0.487	1.458
0.4	1.2	0.408	1.207	0.404	1.507
0.6	0.8	0.593	0.803	0.556	1.563
0.6	1.2	0.595	1.196	0.556	2.331

replicates for each dataset. File S1 in Supplementary data discusses results obtained on another set of gene expression data (this data is on human liver tissue), which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/ TCBB.2015.2407399.

3 Results

3.1 Results on Artificially Generated Data

Results obtained after normalizing artificial data generated with all genes as housekeeping are presented in Table 7. For each row in Table 7, the artificial data was generated only once and this data was used in all the analysis. Inspecting the c and m parameter values (columns 3 and 4) obtained from our normalization strategy (labeled Technique 1 in the table), it is observed that our strategy always predicts parameter values in close agreement with the true values used to generate the data (columns 1 and 2). This result validates our normalization strategy. Next, noting the differences in the parameter values (columns 5 and 6) predicted by Chen's normalization algorithm for uncalibrated signals [10] (labeled Technique 2 in the Table) and true parameter values (columns 1 and 2) for the last four rows of Table 7 (for these the values of the parameter c used to generate the data are greater than 0.3), we conclude that Chen's normalization technique is inadequate for data with a high coefficient of variation.

Results obtained by normalizing the artificial data generated with 4500 housekeeping genes and 500 non-housekeeping genes were as follows. Our normalization strategy (Technique 1) predicted c and m parameter values as 0.226 and 1.17 respectively and 4,889 candidate housekeeping genes of which 4,495 predictions were correct. Chen's normalization strategy (Technique 2) predicted c and m parameter values as 0.226 and 1.18 respectively and 4,808 candidate housekeeping genes of which 4,490 predictions were correct.

3.2 Results on Data of Cheung et al. [16]

3.2.1 Results from Techniques 1 and 2

Results from analyzing expression data reported by Cheung et al. [16] for an individual, Serial No. 12 in Table S2, (Table S2 in supplementary material lists data files pertaining to each individual, the latter being described by a Serial Number, available online) are presented in Table 8. We note that the 99 percent confidence intervals (lower and upper limits being denoted as LL and UL respectively) predicted using Chen's approach are erroneous since the lower limits are negative. The differences in the values (predicted by the two normalization strategies) in Table 8 are attributed to the high values of the coefficient of variation predicted by our normalization strategy for all the replicate datasets.

The area under our density function defined by (16) is guaranteed to always give a value of one. For comparison, we also numerically computed the area under Chen's density function defined by (6) with the c and m parameter values obtained from Table 8. The areas for replicate datasets 1, 2, 3 and 4 were 0.981, 0.98, 0.959 and 0.963 respectively, which are significantly different from one. Chen's strategy is thus inadequate for high coefficient of variation.

TABLE 8
The Table Shows the c and m Parameter Values and the 99 Percent Confidence Intervals for Each Replicate Dataset from an Individual

	Technique 1				Technique 2			
	c	m	LL	UL	c	m	LL	UL
RD 1	0.41	1.00	0.08	13.1	0.42	1.35	-0.07	8.24
RD 2	0.43	1.15	0.08	17.5	0.42	1.47	-0.08	9.06
RD 3	0.56	1.18	0.04	38.8	0.48	1.58	-0.34	13.6
RD 4	0.53	1.26	0.04	37.5	0.47	1.65	-0.30	13.3

RD Implies "Replicate Dataset".

As outlined in the methods section, we arrived at a set of housekeeping genes for the expression data of Cheung et al. [16] using Techniques 1 and 2. Technique 2 predicted 147 genes as housekeeping, whereas Technique 1 predicted an additional eight genes i.e. a total of 155 genes. A complete list of these eight additional genes is provided in Table S3 in the supplementary material , available online.

3.2.2 Functions of Four Housekeeping Genes

We now discuss the functions of four of the additional eight genes. These genes are MYT1L, HPX, HMGN3 and ATP5J2. The gene MYT1L is instrumental in building the nervous system [19] and there is strong evidence to suggest that the gene might contribute to diseases of the cognitive system such as intellectual disability [20], schizophrenia [19], major depressive disorder [19], etc. The gene HPX codes for the glycoprotein Hemopexin [21]. Hemopexin prevents the oxidative damage of cells [21] and is also responsible for preserving iron in the body [22]. The HMGN3 gene plays a central role in glucose homeostasis by regulating the levels of insulin [23] and glucagon [24] in blood. The gene ATP5J2 codes for one of the subunits of the protein complex ATP synthase which is instrumental in generating ATP. ATP is a very important metabolite required in cells for various purposes. Hence all the four genes are housekeeping genes.

3.2.3 Results from Method of Rodriguez-Lanetty et al. [17]

We applied the technique of Rodriguez-Lanetty et al. [17] (Section 2.6) to expression data of Cheung et al. [16]. In particular, we identified genes with coefficient of variation less than 0.5 (the range for the same in the data was approximately between 0.3 to 11.4), we set the significance level for ANOVA and the t-test at 0.05 and considered expression data from an individual as a group for ANOVA analysis. The approach identified 65 genes as housekeeping in which 50 genes were common with our list of 155 genes.

3.2.4 Results from LIMMA [6]

We also applied the LIMMA procedure (Section 2.7) to the expression data of Cheung et al. [16]. The expression data of Cheung et al. [16] consisted of 70 slides. We only considered spots for which data was present for all 70 slides. This resulted in 1024 spots pertaining to 782 genes. We ranked the 782 genes using the log2-fold-change statistic and also the moderated t-statistic. For each case, we computed cumulative ranks. First, we note that all the 155 genes identified as housekeeping by our approach, are present in the list of 782 genes. We define rank of a gene (out of the 155 genes) as the position at which it occurs in the sorted list of 782 genes (sorted using any one of the two statistics). Then, the cumulative rank for a x-coordinate value varying between 1 and 782 is the number of housekeeping genes (out of 155) with ranks less than or equal to x. These have been plotted in Fig. 2.

We would ideally expect the 155 genes to occur at the bottom of the list of 782 genes. Hence, a slope that is low in the beginning and rises sharply towards the end would be ideal. Observing

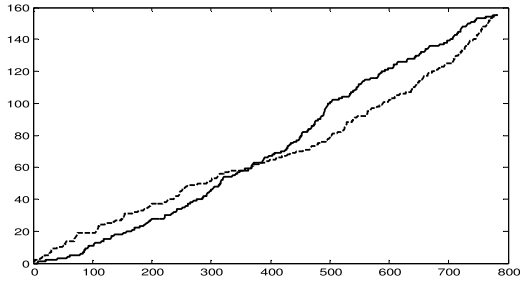

Fig. 2. Cumulative ranks for data sorted using the log2-fold-change statistic (solid line) and the moderated t-statistic (dashed line).

Fig. 2, we note that the predictions from the log2-fold-change statistic compare better. Nevertheless, based on the findings of Jeffery et al. [9], which asserts that different normalization strategies give strikingly different lists of differentially expressed genes when applied to the same expression data, we feel that the results of our housekeeping gene identification strategy compare well with results from LIMMA [6].

4 Conclusion

The proposed outlier detection technique is applicable to datasets with any number of replicates. The proposed normalization technique can successfully normalize data even with a high coefficient of variation. The predicted housekeeping genes have been validated by considering the functions of a few of the genes and also by comparison with other techniques of normalizing microarray data [6], [17]. Results indicate the efficacy of our approach.

Acknowledgments

The authors would like to thank the anonymous referees and Prof. Hongying Dai for their comments and suggestions. Mani Bhushan is the corresponding author.

References

[1] I. V. Yang, E. Chen, J. P. Hasseman, W. Liang, B. C. Frank, S. Wang, V. Sharov, A. I. Saeed, J. White, J. Li, N. H. Lee, T. J. Yeatman, and J. Quackenbush, "Within the fold: Assessing differential expression measures and reproducibility in microarray assays," *Genome Biol.*, vol. 3, no. 11, p. research0062, Oct. 2002.

[2] S. Russell, L. A. Meadows, and R. R. Russell, *Microarray Technology in Practice*. San Diego, CA, USA: Academic, 2009.

[3] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour, "Microarray data analysis: From disarray to consolidation and consensus," *Nat. Rev. Genetic*, vol. 7, no. 1, pp. 55–65, Jan. 2006.

[4] E. Bair, "Identification of significant features in DNA microarray data," *Wiley Interdisciplinary Rev.: Comput. Statist.*, vol. 5, no. 4, pp. 309–325, Jul. 2013.

[5] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 9, pp. 5116–5121, Apr. 2001.

[6] G. K. Smyth, "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments," *Stat. Appl. Genet. Mol. Biol.*, vol. 3, p. Article3, 2004.

[7] R Core Team. (2013).*R: A language and environment for statistical computing. R foundation for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing [Online]. Available: http://www.R-project.org/

[8] G. C. Tseng, D. Ghosh, and E. Feingold, "Comprehensive literature review and statistical considerations for microarray meta-analysis," *Nucleic Acids Res.*, vol. 40, no. 9, pp. 3785–3799, May 2012.

[9] I. B. Jeffery, D. G. Higgins, and A. C. Culhane, "Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data," *BMC Bioinformat.*, vol. 7, no. 1, p. 359, Jul. 2006.

[10] Y. Chen, E. R. Dougherty, and M. L. Bittner, "Ratio-based decisions and the quantitative analysis of cDNA microarray images," *J. Biomed. Opt.*, vol. 2, no. 4, pp. 364–374, Oct. 1997.

[11] A. Bas, G. Forsberg, S. Hammarström, and M.-L. Hammarström, "Utility of the housekeeping genes 18S rRNA, beta-actin and glyceraldehyde-3-phosphate-dehydrogenase for normalization in real-time quantitative reverse transcriptase-polymerase chain reaction analysis of gene expression in human T lymphocytes," *Scand. J. Immunol.*, vol. 59, no. 6, pp. 566–573, Jun. 2004.

[12] D. C. Montgomery, *Design and Analysis of Experiments*, 5th ed. Hoboken, NJ, USA: Wiley, 2000.

[13] G. A. Churchill, "Fundamentals of experimental design for cDNA microarrays," *Nat. Genetic*, vol. 32, no. Suppl, pp. 490–495, Dec. 2002.

[14] Y. Chen, V. Kamat, E. R. Dougherty, M. L. Bittner, P. S. Meltzer, and J. M. Trent, "Ratio statistics of gene expression levels and applications to microarray data analysis," *Bioinformatics*, vol. 18, no. 9, pp. 1207–1215, Sep. 2002.

[15] B. E. Stranger, M. S. Forrest, A. G. Clark, M. J. Minichiello, S. Deutsch, R. Lyle, S. Hunt, B. Kahl, S. E. Antonarakis, S. Tavaré, P. Deloukas, and E. T. Dermitzakis, "Genome-wide associations of gene expression variation in humans," *PLoS Genetics*, vol. 1, no. 6, p. e78, Dec. 2005.

[16] V. G. Cheung, L. K. Conlin, T. M. Weber, M. Arcaro, K.-Y. Jen, M. Morley, and R. S. Spielman, "Natural variation in human gene expression assessed in lymphoblastoid cells," *Nat. Genetics*, vol. 33, no. 3, pp. 422–425, Mar. 2003.

[17] M. Rodriguez-Lanetty, W. S. Phillips, S. Dove, O. Hoegh-Guldberg, and V. M. Weis, "Analytical approach for selecting normalizing genes from a cDNA microarray platform to be used in q-RT-PCR assays: A cnidarian case study," *J. Biochem. Biophys. Methods*, vol. 70, no. 6, pp. 985–991, Apr. 2008.

[18] G. K. Smyth and T. Speed, "Normalization of cDNA microarray data," *Methods*, vol. 31, no. 4, pp. 265–273, Dec. 2003.

[19] T. Wang, Z. Zeng, T. Li, J. Liu, J. Li, Y. Li, Q. Zhao, Z. Wei, Y. Wang, B. Li, G. Feng, L. He, and Y. Shi, "Common SNPs in myelin transcription factor 1-like (MYT1L): Association with major depressive disorder in the chinese han population," *PLoS ONE*, vol. 5, no. 10, p. e13662, 2010.

[20] S. J. C. Stevens, C. M. A. van Ravenswaaij-Arts, J. W. H. Janssen, J. S. Klein Wassink-Ruiter, A. J. van Essen, T. Dijkhuizen, J. van Rheenen, R. Heuts-Vijgen, A. P. A. Stegmann, E. E. J. G. L. Smeets, and J. J. M. Engelen, "MYT1L is a candidate gene for intellectual disability in patients with 2p25.3 (2pter) deletions," *Amer. J. Med. Genetics A*, vol. 155A, no. 11, pp. 2739–2745, Nov. 2011.

[21] F. I. Rosell, M. R. Mauk, and A. G. Mauk, "Effects of metal ion binding on structural dynamics of human hemopexin," *Biochemistry*, vol. 46, no. 32, pp. 9301–9309, Aug. 2007.

[22] W. T. Morgan and A. Smith, "Binding and transport of iron-porphyrins by hemopexin," in *Advances in Inorganic Chemistry*, vol. 51. San Diego, CA, USA: Academic Press, 2000, pp. 205–241.

[23] T. Ueda, T. Furusawa, T. Kurahashi, L. Tessarollo, and M. Bustin, "The nucleosome binding protein HMGN3 modulates the transcription profile of pancreatic β cells and affects insulin secretion," *Mol. Cell. Biol.*, vol. 29, no. 19, pp. 5264–5276, Oct. 2009.

[24] T. Kurahashi, T. Furusawa, T. Ueda, and M. Bustin, "The nucleosome binding protein HMGN3 is expressed in pancreatic alpha-cells and affects plasma glucagon levels in mice," *J. Cell. Biochem.*, vol. 109, no. 1, pp. 49–57, Jan. 2010.