Statistical Applications in Genetics and Molecular Biology

Volume 5, *Issue* 1 2006 *Article* 4

Statistical Selection of Maintenance Genes for Normalization of Gene Expressions

Yifan Huang[*] Jason C. Hsu[†]
Mario Peruggia[‡] Abigail A. Scott[**]

[*] H. Lee Moffitt Cancer Center and Research Institute, huangy@moffitt.usf.edu

[†] Ohio State University, jch@stat.ohio-state.edu

[‡] Ohio State University, peruggia@stat.ohio-state.edu

[**] National Council on Crime and Delinquency Children's Research Center, ascott@mw.nccd-crc.org

Statistical Selection of Maintenance Genes for Normalization of Gene Expressions[*]

Yifan Huang, Jason C. Hsu, Mario Peruggia, and Abigail A. Scott

Abstract

Maintenance genes can be used for normalization in the comparison of gene expressions. Even though the absolute expression levels of maintenance genes may vary considerably among different tissues or cells, a set of maintenance genes may provide suitable normalization if their expression levels are relatively constant in the specific tissues or cells of interest. A statistical procedure is proposed to select maintenance genes for normalization of gene expression data from tissues or cells of interest. This procedure is based on simultaneous confidence intervals for practical equivalence of relative gene expressions in these tissues or cells. As an illustration, the procedure is applied to the maintenance gene expression data from Vandesompele et al. (2002).

KEYWORDS: gene expressions, normalization, equivalence inference

[*] We sincerely thank Dr. Eirikur Steingrimsson of the Iceland Genomics Corporation for his helpful suggestions. Jason Hsu's research is supported by NSF Grant No. DMS-05055519. Abi A. Scott's participation in the Summer 2003 REU program of the Department of Statistics at the Ohio State University was supported with funds provided by the Department of Statistics and the College of Mathematical and Physical Sciences of the Ohio State University and by the American Statistical Association.

1 Motivation

In carrying out comparisons of gene expression data from RNA quantification, such as microarray data, normalization is of great importance, because during the sample preparation, the labelling process, the hybridization process, and the scanning process, there are large numbers of potential sources of systematic variation that need to be removed to make the gene expression data comparable. Normalization removes systematic, or non-biological, sources of variation so that differences in gene expression levels truly reflect biological variation.

Various gene sets have been used for gene expression normalization, such as all genes, maintenance genes and spiked controls (Yang *et al.* 2001). Among these, maintenance genes are frequently used. Maintenance genes, or housekeeping genes, are defined as those genes critical to the maintenance of cellular functions. These genes are presumably expressed in all tissues or cells. One major issue in maintenance gene normalization is that expression levels of maintenance genes vary considerably among different tissues or cells. Take a widely used maintenance gene, Glyceraldehyde-3-phosphate dehydrogenase (GAPDH), as an example. Bhatia *et al.* (1994) found that GAPDH gene expression levels are remarkably different between tumorigenic and metastatic cells.

However, for specific tissues or cells in the experiment of interest, it may be possible to find a set of maintenance genes for which the expression ratios of any two of them are approximately identical in the tissues or cells of interest. Such a set of maintenance genes can be used for normalization of gene expression data from the tissues or cells of interest. This study describes how to select, from preselected possible maintenance genes out of different functional classes, a set of maintenance genes with approximately constant expression ratios in given tissues or cells, statistically. The proposed methodology is general, in that it can be used to select experiment specific maintenance genes for particular tissues or cells of interest.

2 Formulation of the Problem

Suppose there are several tissues under consideration. Two maintenance genes ideal for normalization would have constant expression *ratios* over the tissues. A sample interaction plot of gene expression levels is shown in Figure 1 (a).

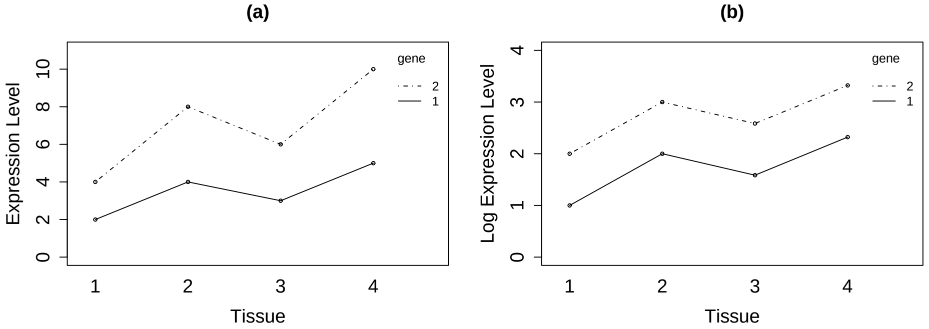


Figure 1: Interaction plots of expression levels and log expression levels of two maintenance genes in different tissues. (a) The expression ratios of the two maintenance genes are constant over the tissues. (b) The lines in the interaction plot are parallel to each other, indicating that the log expression differences of the two maintenance genes are constant over the tissues.

After logarithmic transformation of such gene expression data, the log expression *differences* of the two genes are constant over the tissues. Consequently, lines in the interaction plot of log expression levels are parallel to each other, as shown in Figure 1 (b). Parallelism of the lines indicates no interaction. Maintenance genes with small interaction should be selected for normalization of gene expression data from given tissues or cells.

To define a statistical measure of interaction, let Y_{ijr} denote the log (base 2) expression level of the rth observation on the ith gene in the jth tissue. We model Y_{ijr} as follows.

$$Y_{ijr} = \tau_{ij} + \epsilon_{ijr}, i = 1, \ldots, I, j = 1, \ldots, J, r = 1, \ldots, n_{ij} \tag{1}$$

where τ_{ij} represents the expected log expression level of the ith gene in the jth tissue, that is, the log expression level that would always be obtained for the ith gene in the jth tissue under identical experiment condition and without measurement error. The τ_{ij}'s are considered to be fixed parameters which are unknown but can be estimated from gene expression data.

The interaction contrasts

$$\theta_{ij}^{sk} = (\tau_{is} - \tau_{js}) - (\tau_{ik} - \tau_{jk}), \quad i < j, s < k \tag{2}$$

measure lack of parallelism of the lines in the interaction plot of log expression levels, that is, θ_{ij}^{sk} measures the interaction of gene i and gene j in tissue s and tissue k. Therefore, maintenance genes with small interaction in given tissues or cells are explicitly defined to be those with small values of $|\theta_{ij}^{sk}|$.

In (1), ϵ_{ijr} is the experimental error present in the rth observation on the ith gene in the jth tissue, which is a random variable with mean zero. We will determine what assumptions on ϵ_{ijr} would make inference on θ_{ij}^{sk} equivalent to inference on difference of expression ratios of gene i over gene j between tissue s and tissue k. The appropriateness of the assumptions can be examined by residual plots or other diagnostic methods.

Let X_{ij} and μ_{ij} denote the observed and expected expression level of the ith gene in the jth tissue. Similarly, let Y_{ij} and τ_{ij} denote the observed and expected log expression level of the ith gene in the jth tissue. Then $Y_{ij} = \log_2 X_{ij}$.

Consider gene i and gene j in tissue s and tissue k. If the distribution of Y_{is} differs from the distribution of Y_{js} by a location shift only, that is, Y_{is} and $Y_{js} + \delta_s$ are distributed as F_s, then $\tau_{is} - \tau_{js} = \delta_s$. Since $Y_{is} \overset{d}{=} Y_{js} + \delta_s$, $X_{is} \overset{d}{=} e^{\delta_s} X_{js}$. Thus, $\mu_{is}/\mu_{js} = e^{\delta_s}$. Similarly, If the distribution of Y_{ik} differs from the distribution of Y_{jk} by a location shift only, that is, Y_{ik} and $Y_{jk} + \delta_k$ are distributed as F_k, which may be different from F_s, then $\tau_{ik} - \tau_{jk} = \delta_k$. Consequently, $X_{ik} \overset{d}{=} e^{\delta_k} X_{jk}$ and $\mu_{ik}/\mu_{jk} = e^{\delta_k}$.

If we infer

$$\theta_{ij}^{sk} = (\tau_{is} - \tau_{js}) - (\tau_{ik} - \tau_{jk}) = \delta_s - \delta_k = 0,$$

that is, gene i and gene j have no interaction in tissue s and tissue k, then

$$\frac{\mu_{is}/\mu_{js}}{\mu_{ik}/\mu_{jk}} = e^{\delta_s - \delta_k} = 1,$$

that is, gene i and gene j have identical expression ratios over tissue s and tissue k.

Therefore, if the assumptions $\epsilon_{ijr} \sim F_j,\ j = 1, \ldots, J$, are satisfied, then inference on θ_{ij}^{sk} corresponds to inference on $\frac{\mu_{is}/\mu_{js}}{\mu_{ik}/\mu_{jk}}$. Also, if we write $\theta_{ij}^{sk} = (\tau_{is} - \tau_{ik}) - (\tau_{js} - \tau_{jk})$, then inference on θ_{ij}^{sk} corresponds to inference on $\frac{\mu_{is}/\mu_{ik}}{\mu_{js}/\mu_{jk}}$ with $\epsilon_{ijr} \sim G_i,\ i = 1, \ldots, I$.

So, inference on difference of expected log expression differences θ_{ij}^{sk} corresponds to inference on ratio of the ratios of expected expression levels if the distributions of log expression levels in the same *tissue* have identical shape and possibly shifted locations for different genes, or the distributions of log expression levels of the same *gene* have identical shape and possibly shifted locations for different tissues.

To be able to investigate the variability of the expression level of gene i in tissue j, that is, the distribution of ϵ_{ijr}, replicated observations on the expression of gene i in tissue j are necessary. Replication is not a waste of scientific resources. Instead, it is essential to obtain valid statistical inferences. It not only enables us to characterize the random variation in gene expression, but also gives us more accurate estimation of gene expression levels because the variance of the averaged observations is smaller than the variance of the individual observation.

3 Proposed Statistical Method

Our desired inference is to identify those θ_{ij}^{sk}'s with absolute values small enough to allow the corresponding maintenance genes to be used for normalization of gene expression data from the corresponding tissues. In addition, the contrast θ_{ij}^{sk} is the difference of $\tau_{is} - \tau_{js}$ and $\tau_{ik} - \tau_{jk}$.

Consider testing

$$H_0 : \theta_{ij}^{sk} = 0 \text{ vs. } H_a : \theta_{ij}^{sk} \neq 0 \tag{3}$$

at level-α and inferring that $|\theta_{ij}^{sk}|$ is small if $H_0 : \theta_{ij}^{sk} = 0$ is accepted. Suppose $\epsilon_{ijr} \overset{iid}{\sim} N(0, \sigma^2)$. If the two-sided size-$\alpha$ t-test is used, then $|\theta_{ij}^{sk}|$ is claimed to be small when

$$\frac{|\hat{\theta_{ij}^{sk}}|}{SE(\theta_{ij}^{sk})} = \frac{|(\bar{y}_{is.} - \bar{y}_{js.}) - (\bar{y}_{ik.} - \bar{y}_{jk.})|}{\hat{\sigma}\sqrt{(\frac{1}{n_{is}} + \frac{1}{n_{js}} + \frac{1}{n_{ik}} + \frac{1}{n_{jk}})}} \leq t_{\alpha/2,\nu} \tag{4}$$

where $\bar{y}_{is.} = \sum_{r=1}^{n_{is}} y_{isr}/n_{is}$ is the average log expression level of ith gene in sth tissue, and $t_{\alpha/2,\nu}$ is the upper $100\alpha/2$ percentile of a Student's t distribution with ν degrees of freedom. Regardless of the value of θ_{ij}^{sk}, the probability of (4) increases as the sample sizes decrease. That is, even when the true value of θ_{ij}^{sk} is away from zero, this test procedure might infer that $|\theta_{ij}^{sk}|$ is small if the sample sizes are small, which is usually the case in microarray experiments.

So one might consider reversing the null and alternative hypotheses, and test

$$H_0 : \theta_{ij}^{sk} \neq 0 \text{ vs. } H_a : \theta_{ij}^{sk} = 0.$$

But $\theta_{ij}^{sk} = 0$ requires infinitely many samples to verify. It is practically impossible and unnecessarily stringent. Therefore, this is an equivalence problem in which the usual test for (3) is not of interest (Berger and Hsu, 1996). We will test instead

$$H_0 : |\theta_{ij}^{sk}| \geq \delta \text{ vs. } H_a : |\theta_{ij}^{sk}| < \delta \tag{5}$$

where δ is a predetermined positive value, defining practical equivalence of the gene expression levels.

For example, if we infer that $|\theta_{23}^{12}|$, $|\theta_{23}^{13}|$ and $|\theta_{23}^{23}|$ are all less than δ, then gene 2 and gene 3 have small interaction in tissues 1, 2 and 3. These two maintenance genes can be used for gene expression normalization in experiments involving tissues 1, 2 and 3.

In our opinion, the determination of a δ value may be done by analogy to the choice of a clinically meaningful difference in clinical trials, where there are three types of controls: negative control, active control, and positive control (see the international guidance ICH E10).

Negative controls in clinical trials are placebos. Using maintenance genes to normalize gene expressions, as in Hsu *et al.* (2004), is analogous to using placebos as negative controls in clinical trials.

In clinical trials, positive controls are treatments known to be different from negative controls. They are sometimes used in safety trials to validate assay sensitivity. Spiking RNA samples is perhaps analogous to using positive controls in clinical trials.

Active controls in clinical trials are treatments known to be efficacious. Non-inferiority of a treatment is typically defined as a fraction of the improvement given by an active control over the placebo. Thus, by analogy, we believe the determination of a δ value may be done by observing differential expressions of genes that are known to be involved in the biological process. For example, in comparing normal human tissue and a certain type of cancer tissue, the p53 gene may serve as an active control if it is known to be involved in the development of this type of cancer.

The δ chosen should be smaller than the smallest log expression difference of the active control gene among the tissues. The value of δ may be determined based on previous experience, the purpose of the current analysis, and future use of the results. The larger the δ is, the more maintenance genes we can select for normalization.

The H_0 in (5) can be partitioned into two one-sided hypotheses

$$H_0 : \theta_{ij}^{sk} \leq -\delta \text{ vs. } H_a : \theta_{ij}^{sk} > -\delta \tag{6}$$

and

$$H_0 : \theta_{ij}^{sk} \geq \delta \text{ vs. } H_a : \theta_{ij}^{sk} < \delta. \tag{7}$$

Since $\theta_{ij}^{sk} \leq -\delta$ and $\theta_{ij}^{sk} \geq \delta$ cannot be true at the same time, we can make at most one mistake of incorrectly rejecting a true null hypothesis. Therefore, we can test (6) and (7) at level-α and reject H_0 in (5) if both H_0 in (6) and H_0 in (7) are rejected. This is a level-α test for H_0 in (5).

The least square estimator (LSE) for θ_{ij}^{sk} is

$$\hat{\theta}_{ij}^{sk} = (\bar{y}_{is.} - \bar{y}_{js.}) - (\bar{y}_{ik.} - \bar{y}_{jk.})$$

where $\bar{y}_{is.}$, $\bar{y}_{js.}$, $\bar{y}_{ik.}$ and $\bar{y}_{jk.}$ are similarly defined as in (4). If we assume $\epsilon_{ijr} \overset{iid}{\sim} N(0, \sigma^2)$, then $Y_{ijr} \overset{iid}{\sim} N(\tau_{ij}, \sigma^2)$. This is a special case of $\epsilon_{ijr} \overset{iid}{\sim} F_j$ (or $\epsilon_{ijr} \overset{iid}{\sim} G_i$), where all the F_j's (or G_i's) are normal distributions with zero mean and equal variance. Under this assumption, the variance of $\hat{\theta}_{ij}^{sk}$ is

$$Var(\hat{\theta}_{ij}^{sk}) = \sigma^2(\frac{1}{n_{is}} + \frac{1}{n_{js}} + \frac{1}{n_{ik}} + \frac{1}{n_{jk}}).$$

Then the standard error of $\hat{\theta}_{ij}^{sk}$ is

$$SE(\hat{\theta}_{ij}^{sk}) = \sqrt{\hat{\sigma}^2(\frac{1}{n_{is}} + \frac{1}{n_{js}} + \frac{1}{n_{ik}} + \frac{1}{n_{jk}})},$$

where $\hat{\sigma}^2$ is the mean square error (MSE). Since $(\hat{\theta}_{ij}^{sk} - \theta_{ij}^{sk})/SE(\hat{\theta}_{ij}^{sk})$ has a Student's t distribution with degrees of freedom $\nu = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} - IJ$, a level-$\alpha$ test for (6) will reject $H_0 : \theta_{ij}^{sk} \leq -\delta$ if

$$\frac{\hat{\theta}_{ij}^{sk} + \delta}{SE(\hat{\theta}_{ij}^{sk})} > t_{\alpha,\nu},$$

and a level-α test for (7) will reject $H_0 : \theta_{ij}^{sk} \geq \delta$ if

$$\frac{\hat{\theta}_{ij}^{sk} - \delta}{SE(\hat{\theta}_{ij}^{sk})} < -t_{\alpha,\nu}.$$

Therefore, a level-α test for (5) will reject $H_0 : |\theta_{ij}^{sk}| \geq \delta$ and infer $|\theta_{ij}^{sk}| < \delta$ if $T_{ij}^{sk} > t_{\alpha,\nu}$, where

$$T_{ij}^{sk} = \min\{\frac{\delta + \hat{\theta}_{ij}^{sk}}{SE(\hat{\theta}_{ij}^{sk})}, \frac{\delta - \hat{\theta}_{ij}^{sk}}{SE(\hat{\theta}_{ij}^{sk})}\}.$$

For I genes and J tissues, there are $N = \binom{I}{2}\binom{J}{2}$ θ_{ij}^{sk}'s. That is, there are N individual hypotheses to test. For multiplicity adjustment, a stepwise method can be used to achieve higher power. For example, Holm's step-down method (Holm, 1979) that is based on the Bonferroni inequality puts no restriction on the correlation structure of the test statistics. So it is suitable to test $H_0 : |\theta_{ij}^{sk}| \geq \delta, \ i < j, \ s < k$, because the correlation structure of the test statistics $T_{ij}^{sk}, \ i < j, \ s < k$, is complicated.

Let $T_{(j)}$ be the jth largest T_{ij}^{sk} among the N T_{ij}^{sk}'s. Holm's step-down procedure goes as follows.

Step 1: If $T_{(N)} > t_{\frac{\alpha}{N},\nu}$, infer the corresponding $|\theta_{ij}^{sk}| < \delta$ and go to step 2; else stop.

Step 2: If $T_{(N-1)} > t_{\frac{\alpha}{N-1},\nu}$, infer the corresponding $|\theta_{ij}^{sk}| < \delta$ and go to step 3; else stop.

...

Step N: If $T_{(1)} > t_{\alpha,\nu}$, infer the corresponding $|\theta_{ij}^{sk}| < \delta$, and stop.

If the assumption $\epsilon_{ijr} \overset{iid}{\sim} N(0, \sigma^2)$ is not satisfied, for example, if residual plots indicate unequal variance or departure from normality, then we may assume $\epsilon_{ijr} \sim F_j$ (or $\epsilon_{ijr} \sim G_i$) only and use a bootstrap method (Efron and Tibshirani, 1993) to obtain simultaneous confidence intervals for $\theta_{ij}^{sk}, \ i < j, \ s < k$.

Suppose $\epsilon_{ijr} \sim F_j$ with M possibly distinct F_j's, denoted by $F^m, m = 1, 2, \ldots, M$. If the F_j's are identical, then $M = 1$. If the F_j's are all different from each other, then $M = J$, the total number of tissues or cells. If some but not all of the F_j's are identical, then M is between 1 and J. The following bootstrap procedure adjusts multiplicity while taking advantage of the dependence between the genes.

1. Fit model (1) and obtain LSE $\hat{\tau}_{ij} = \bar{y}_{ij.}$ for each τ_{ij}.

2. Compute residuals $e_{ijr} = y_{ijr} - \hat{\tau}_{ij}$ and group them into $\boldsymbol{e}^1, \boldsymbol{e}^2, \ldots, \boldsymbol{e}^M$ where $\boldsymbol{e}^m = \{e_{ijr}\text{'s with } F_j \text{ identical to } F^m\}$, $m = 1, 2, \ldots, M$.

3. Let n_m be the number of residuals in $\boldsymbol{e}^m$, $m = 1, 2, \ldots, M$. Draw a random sample of size n_1 from $\boldsymbol{e}^1$ with replacement. Similarly, draw random samples of size $n_2, \ldots, n_M$ from $\boldsymbol{e}^2, \ldots, \boldsymbol{e}^M$, respectively. The resampled residuals are called bootstrap residuals and denoted by $e^*_{ijr}, i = 1, \ldots, I, j = 1, \ldots, J, r = 1, \ldots, n_{ij}$.

4. Compute bootstrap estimate for $\xi^{sk}_{ij} = \hat{\theta}^{sk}_{ij} - \theta^{sk}_{ij}$: $(\xi^{sk}_{ij})^* = (\bar{e}^*_{is.} - \bar{e}^*_{js.}) - (\bar{e}^*_{ik.} - \bar{e}^*_{jk.})$.

Repeat steps 3-4 B times (we use $B = 100,000$ in Section 4) to obtain B bootstrap estimates for ξ^{sk}_{ij}: $\{(\xi^{sk}_{ij})^*_1, (\xi^{sk}_{ij})^*_2, \ldots, (\xi^{sk}_{ij})^*_B\}$. For each set of $(\xi^{sk}_{ij})^*$, record the smallest value $\min(\xi^{sk}_{ij})^*$ and the maximum value $\max(\xi^{sk}_{ij})^*$.

Let d^L be the $100\frac{\alpha}{2}$th empirical percentile of the B $\min(\xi^{sk}_{ij})^*$ values. Let d^U be the $100(1 - \frac{\alpha}{2})$th empirical percentile of the B $\max(\xi^{sk}_{ij})^*$ values. Since

$$\begin{aligned}
& 1 - P\{\hat{\theta}^{sk}_{ij} - d^U < \theta^{sk}_{ij} < \hat{\theta}^{sk}_{ij} - d^L \text{ for } i < j, s < k\} \\
= \;& 1 - P\{d^L < \hat{\theta}^{sk}_{ij} - \theta^{sk}_{ij} < d^U \text{ for } i < j, s < k\} \\
= \;& 1 - P\{\min\{\hat{\theta}^{sk}_{ij} - \theta^{sk}_{ij}\} > d^L \text{ and } \max\{\hat{\theta}^{sk}_{ij} - \theta^{sk}_{ij}\} < d^U\} \\
= \;& P\{\min\{\hat{\theta}^{sk}_{ij} - \theta^{sk}_{ij}\} < d^L \text{ or } \max\{\hat{\theta}^{sk}_{ij} - \theta^{sk}_{ij}\} > d^U\} \\
\leq \;& P\{\min\{\hat{\theta}^{sk}_{ij} - \theta^{sk}_{ij}\} < d^L\} + P\{\max\{\hat{\theta}^{sk}_{ij} - \theta^{sk}_{ij}\} > d^U\} \\
\simeq \;& \alpha/2 + \alpha/2 \\
= \;& \alpha,
\end{aligned}$$

a $(1 - \alpha)100\%$ simultaneous confidence interval for θ^{sk}_{ij} is $(L^{sk}_{ij}, U^{sk}_{ij}) = (\hat{\theta}^{sk}_{ij} - d^U, \hat{\theta}^{sk}_{ij} - d^L)$. If $(L^{sk}_{ij}, U^{sk}_{ij})$ is contained in the interval $(-\delta, \delta)$, then $|\theta^{sk}_{ij}| \leq \delta$ is inferred. That is, the expression levels of genes i and j are considered to have small interaction in tissues s and k.

4 Example of Data Analysis

We will apply our statistical method to the maintenance gene expression data from Vandesompele *et al.* (2002). Vandesompele *et al.* obtained expression profiles of ten commonly used maintenance genes in 13 human tissues using

Table 1. Maintenance Genes in the Study by Vandesompele *et al.* (2002)

Symbol	Name	Function
ACTB	Beta actin	Structural constituent of Cytoskeleton
B2M	Beta-2-microglobulin	MHC class I receptor activity
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase	Oxidoreductase in glycolysis
HMBS	Hydroxymethyl-bilane synthase	Hydroxymethyl-bilane synthase activity
HPRT1	Hypoxanthine phosphoribosyl-transferase 1	Magnesium ion binding transferase activity
RPL13A	Ribosomal protein L13a	Structural constituent of ribosome
SDHA	Succinate dehydrogenase complex, subunit A	Electron transporter activity, oxidoreductase activity
TBP	TATA box binding protein	General RNA polymerase 2 transcription factor activity
UBC	Ubiquitin C	Protein degradation
YWHAZ	Tyrosine 3-monooxygenase/ trytophan 5-monooxygenase activation protein, zeta polypeptide	monooxygenase activity, protein domain specific binding

real-time RT-PCR that provides more precise quantification than microarray. They proposed an ad hoc procedure to select maintenance genes based on their gene-stability measure which is similar to our measure of interaction but without statistical justification. The ten maintenance genes are *ACTB, B2M, GAPDH, HMBS, HPRT1, RPL13A, SDHA, TBP, UBC* and *YWHAZ* (see Table 1 for full names and functions). We denote them by gene 1, gene 2, ..., and gene 10, respectively. There were replicated samples for four of the 13 tissues, so we will use maintenance gene expression data from these four tissues only (bone-marrow, fibroblast, leukocyte and neuroblastoma, denoted by tissue 1, tissue 2, tissue 3 and tissue 4). Therefore, $I = 10$, $J = 4$, and the number of parameters $N = 270$.

To decide which testing procedure, t-test or bootstrap procedure, is appropriate for the maintenance gene expression data from Vandesompele *et al.*, diagnostic plots are obtained (see Figures 2 and 3). The residual plots in Fig-

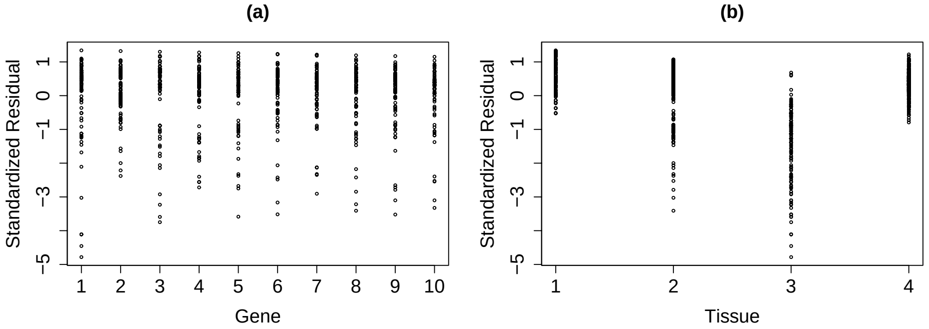


Figure 2: Residual plots for maintenance gene expression data. (a) Plot of standardized residuals against genes. (b) Plot of standardized residuals against tissues, indicating unequal variances among tissues.

ure 2 show unequal variances among tissues, so the assumption $\epsilon_{ijr} \overset{iid}{\sim} N(0, \sigma^2)$ is not satisfied. But the distribution of log expression levels in tissue 1 has a similar shape as in tissue 4 and the distribution of log expression levels in tissue 2 has a similar shape as in tissue 3 (see Figure 3). So the data seem to satisfy the assumption $\epsilon_{ijr} \sim F_j$, $j = 1, 2, \ldots, 4$, with $F_1 = F_4 = F^1$ and $F_2 = F_3 = F^2$. To further assess this issue, QQ-plot of residuals from tissue 4 against residuals from tissue 1 and QQ-plot of residuals from tissue 3 against residuals from tissue 2 are obtained in Figure 4. The QQ-plot in Figure 4(a) follows an approximate straight line along the 45-degree diagonal, indicating $F_1 = F_4$. The QQ-plot in Figure 4(b) does not look like a straight line along the 45-degree diagonal, so we cannot assume $F_2 = F_3$. Therefore, we will use the bootstrap procedure described in Section 3 with three distinct distributions ($F_1 = F_4$, F_2, and F_3) to obtain simultaneous confidence intervals.

For the maintenance gene expression data from Vandesompele *et al.*, we take $\alpha = 0.05$. If the confidence interval of θ_{ij}^{sk} is within $(-\delta,\ \delta)$, then θ_{ij}^{sk} has small absolute value. Based on the θ_{ij}^{sk}'s that are identified to have small absolute values, we can select the corresponding maintenance genes. It is possible to have multiple sets of maintenance genes selected and these sets have the same size. For example, when $\delta = 6$, the sets of genes selected are genes {1,3,4}, genes {6,7,10}, genes {6,9,10} and genes {7,8,10}. We can select any one of them for normalization. We can also select a smaller set, for example, genes {1,3} instead of genes {1,3,4}, for normalization. However,

Table 2. Number of Maintenance Genes Selected in the Largest Set

δ Value	< 5	5	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9
No. Genes	0	0	0	0	2	3	3	3	3	3	3
δ Value	6.0	6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8	6.9	7.0
No. Genes	3	3	3	4	4	4	4	4	4	4	4

we recommend selecting as many maintenance genes as possible. Generally, the larger the δ is, the more maintenance genes may be selected. To see how the value of δ would influence the results, we summarize the number of maintenance genes selected in the largest set and its corresponding δ value in Table 2 and Figure 5. For example, when $\delta = 5.3$, we can select as many as two maintenance genes. When $5.4 \leq \delta \leq 6.2$, we can select as many as three maintenance genes. Figure 5 could assist in the decision about the δ value. For instance, with only a 0.1 increase in δ (from 5.3 to 5.4), the number of maintenance genes selected increases from two to three. To select one more maintenance genes so that there are four maintenance genes for normalization, we have to increase the δ value from 5.4 to 6.3.

The information from positive control genes may serve as an upper bound of the δ values. There were no positive control genes included in the real-time RT-PCR by Vandesompele *et al.*, so we arbitrarily assume that the upper bound is 6. That is, the δ value should be less than 6. For all the δ values less than 6, we would choose $\delta = 5.4$ because it is the smallest value that gives as many maintenance genes as possible.

After determining the δ value to be 5.4, we can collate the results using the table and the diagram in Figure 6. In the table and the diagram in Figure 6, the numbers 1 through 10 represent gene 1 through gene 10. In the table in Figure 6(a), a symbol X means that the corresponding $H_0 : |\theta_{ij}^{sk}| \geq \delta$ cannot be rejected for all pairs of s and k. For example, there is no X in the sixth cell of the ninth row, which means that gene 6 and gene 9 have small interactions across all pairs of tissues. This is displayed in the diagram in Figure 6(b) with a line connecting gene 6 and gene 9. Similarly, genes 6 and 10, and genes 9 and 10 have small interactions across all pairs of tissues. Based on the diagram in Figure 6(b), maintenance genes $\{6,9,10\}$ can be selected for normalization of the data from the four tissues.

5 Conclusion

We have defined a set of maintenance genes for which all 2 x 2 gene by tissue/cell interactions are within $(-\delta, \delta)$ to be good candidates for expression level normalization. With this definition, the key to proper statistical inference is to formulate the problem as one of practical equivalence instead of significant difference. Provided that replicated sample measurements on candidate maintenance genes exist, the proposed statistical method can be used to select a suitable set for normalization purpose.

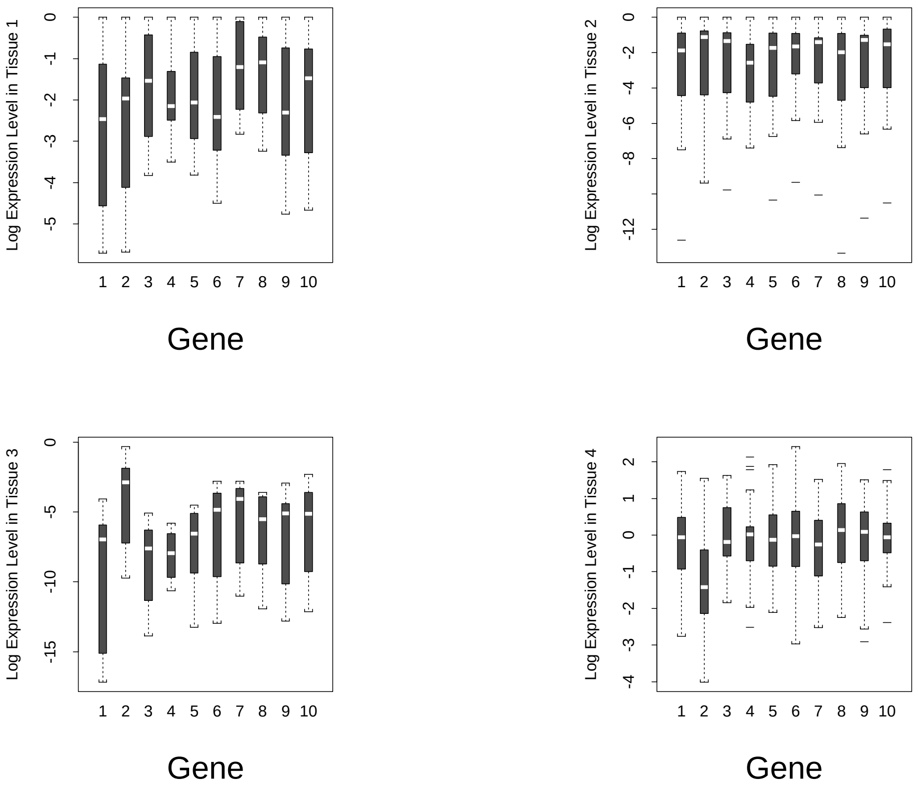


Figure 3: Side by side boxplots of log expression levels for each tissue.

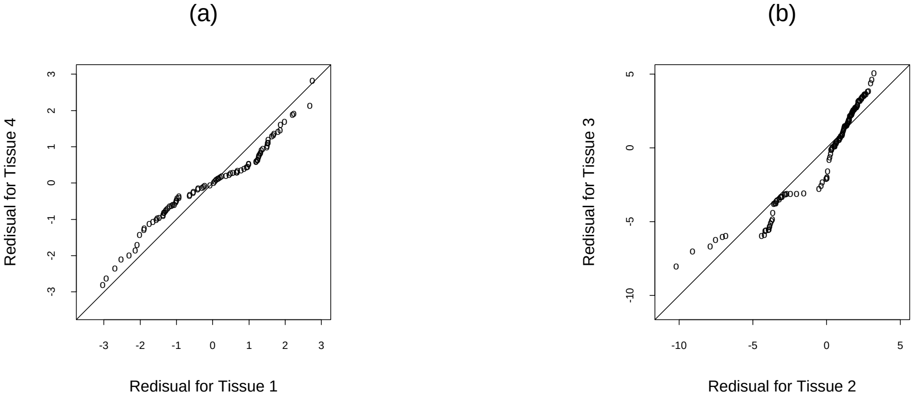


Figure 4: QQ-plots of residuals. (a) QQ-plot of residuals from tissue 4 against residuals from tissue 1. An approximately linear pattern along the 45-degree line indicates $F_1 = F_4$. (b) QQ-plot of residuals from tissue 3 against residuals from tissue 2. It does not look like a straight line along the 45-degree diagonal, so we cannot assume $F_2 = F_3$.

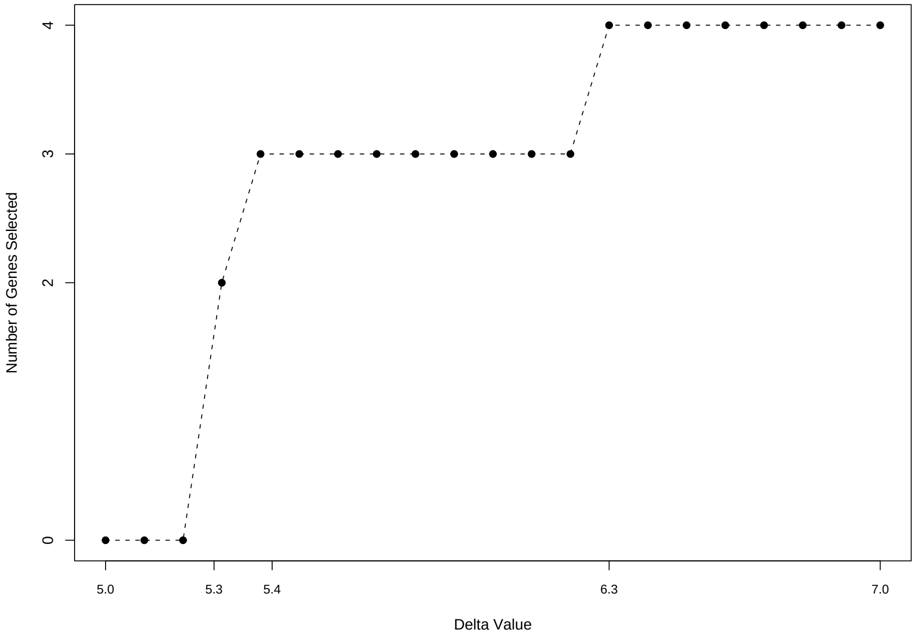
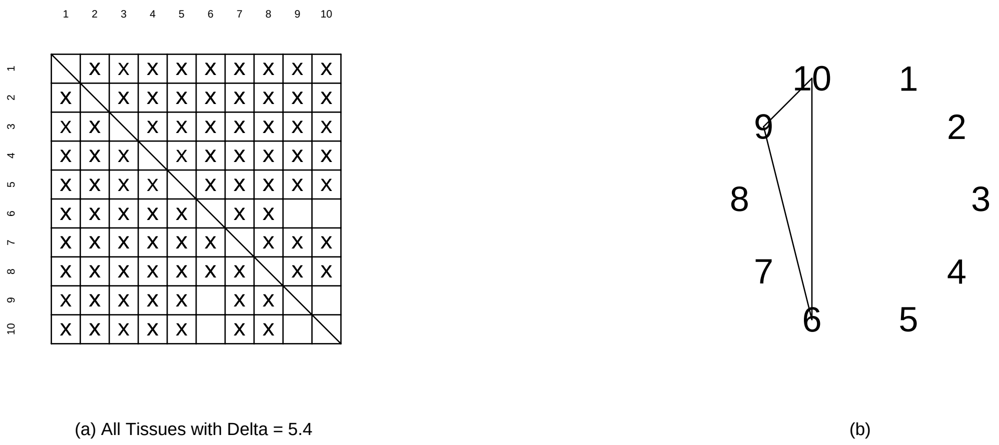


Figure 5: Plot of the number of maintenance genes selected in the largest set versus its corresponding δ value. When $\delta = 5.3$, we can select as many as two maintenance genes. When $5.4 \leq \delta \leq 6.2$, we can select as many as three maintenance genes. When $6.3 \leq \delta \leq 7$, we can select as many as four maintenance genes.



	1	2	3	4	5	6	7	8	9	10
1	\	X	X	X	X	X	X	X	X	X
2	X	\	X	X	X	X	X	X	X	X
3	X	X	\	X	X	X	X	X	X	X
4	X	X	X	\	X	X	X	X	X	X
5	X	X	X	X	\	X	X	X	X	X
6	X	X	X	X	X	\	X	X		
7	X	X	X	X	X	X	\	X	X	X
8	X	X	X	X	X	X	X	\	X	X
9	X	X	X	X	X		X	X	\	
10	X	X	X	X	X		X	X		\

(a) All Tissues with Delta = 5.4

(b)

Figure 6: Selection of maintenance genes for tissues 1, 2, 3 and 4 with $\delta = 6.4$. The numbers 1 through 10 represent gene 1 through gene 10. (a) Genes 6 and 9, genes 6 and 10, and genes 9 and 10 have small interactions across all pairs of tissues. (b) Genes {6,9,10} can be selected for normalization.

References

Berger, R. L. and Hsu, J. C. (1996). Bioequivalence trials, intersection-union tests, and equivalence confidence sets. *Statist. Sci.*, 11:283–302.

Bhatia, P., Taylor, W. R., Greenberg, A. H., and Wright, J. A. (1994). Comparison of Glyceraldehyde-3-Phosphate Dehydrogenase and 28S-Ribosomal RNA gene expression as RNA loading controls for northern blot analysis of cell lines of varying malignant potential. *Analyt. Biochem.*, 216:223–226.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapmans and Hall, London.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, 6:65–70.

Hsu, J. C., Chang, J. Y., and Wang, T. (2004). Simultaneous confidence intervals for differential gene expressions. To appear in *Journal of Statistical Planning and Inference.*

ICH (2001). ICH E10. Choice of control group and related issues in clinical trials. *Federal Register*, 66(93):24390–91.

Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., and Speleman, F. (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.*, 3(7):0034.1–0034.11.

Yang, Y. H., Dudoit, S., Luu, P., and Speed, T. P. (2001). Normalization for cDNA microarray data. San Jose, California. SPIE BiOS 2001.