Research Article

in Medicine

Received 12 June 2012, Accepted 29 January 2013 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5768

Mixed modeling and sample size calculations for identifying housekeeping genes

Hongying Dai,[a*†] Richard Charnigo,[b] Carrie A. Vyhlidal,[c] Bridgette L. Jones[c] and Madhusudan Bhandary[d]

Normalization of gene expression data using internal control genes that have biologically stable expression levels is an important process for analyzing reverse transcription polymerase chain reaction data. We propose a three-way linear mixed-effects model to select optimal housekeeping genes. The mixed-effects model can accommodate multiple continuous and/or categorical variables with sample random effects, gene fixed effects, systematic effects, and gene by systematic effect interactions. We propose using the intraclass correlation coefficient among gene expression levels as the stability measure to select housekeeping genes that have low within-sample variation. Global hypothesis testing is proposed to ensure that selected housekeeping genes are free of systematic effects or gene by systematic effect interactions. A gene combination with the highest lower bound of 95% confidence interval for intraclass correlation coefficient and no significant systematic effects is selected for normalization. Sample size calculation based on the estimation accuracy of the stability measure is offered to help practitioners design experiments to identify housekeeping genes. We compare our methods with geNorm and NormFinder by using three case studies. A free software package written in SAS (Cary, NC, U.S.A.) is available at http://d.web.umkc.edu/daih under software tab. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: housekeeping gene; normalization; RT-PCR; systematic effect; linear mixed-effects model (LMM); intraclass correlation coefficient (ICC)

1. Introduction

Quantitative real-time reverse transcription polymerase chain reaction (RT-PCR) is a technique that permits accurate quantification of steady-state mRNA levels and has become widely used for the expression profiling of regulated genes [1–3]. Normalization of gene expression data using reference genes that have biologically stable expression is an important process for analyzing RT-PCR data. Generally, target gene expression levels are divided by normalizing factors on the basis of expression levels from reference genes. This normalization process can remove the transcriptional variations in target gene expression, allow target genes to reflect biologically relevant interpretation, and make data from multiple experiments comparable [4].

Housekeeping genes [5] are those that are expressed at relatively constant levels regardless of gender, age, or treatments under investigation (defined as systematic effects in this paper). Although housekeeping genes are biologically expressed at relatively constant levels, the empirical measurements of their expression may vary depending on experimental conditions. For example, the density of cultured cells, variability in sample acquisition, RNA template isolation, and the presence of PCR inhibitors may introduce errors into the analysis process, which necessitates validation of housekeeping genes (i.e., the

[a]*Research Development and Clinical Investigation, Children's Mercy Hospital, 2401 Gillham Road, Kansas City, MO 64108, U.S.A.*
[b]*Department of Statistics, University of Kentucky, Lexington, KY 40536, U.S.A.*
[c]*Department of Pediatrics, Children's Mercy Hospital, 2401 Gillham Road, Kansas City, MO 64108, U.S.A.*
[d]*Department of Mathematics, Columbus State University, 4225 University Avenue, Columbus, GA 31907, U.S.A.*
**Correspondence to: Hongying Dai, Research Development and Clinical Investigation, Children's Mercy Hospital, 2401 Gillham Road, Kansas City, MO 64108, U.S.A.*
[†]*E-mail: hdai@cmh.edu*

verification that they can, in fact, be used for normalization) in each experiment [1, 6]. When target genes are confounded with systematic effects and random experimental effects, it is important to use housekeeping genes as reference genes in the normalization process to separate random experimental effects from systematic effects and to extract accurate, reproducible, and biologically relevant mRNA quantification.

Studies have shown that it might be inadequate to use a single housekeeping gene for normalization [1]. Several methods have been proposed to assist in the selection of multiple housekeeping genes, including [4, 7–14], etc. Among these, geNorm [7] and NormFinder [8] are the two most commonly applied methods.

geNorm [7] selects housekeeping genes by using standard deviation as a stability measure. Let Y_{ij} be the expression level from the jth ($j = 1, 2, \ldots, m$) candidate gene in the ith ($i = 1, 2, \ldots, n$) sample. The vector $A_{j_1 j_2} = \left\{\log_2\left(Y_{1j_1}/Y_{1j_2}\right), \log_2\left(Y_{2j_1}/Y_{2j_2}\right), \ldots, \log_2\left(Y_{nj_1}/Y_{nj_2}\right)\right\}$ is the $\log_2$-transformed expression ratio between gene j_1 and j_2 for $\forall\, j_1,\ j_2 \in \{1, 2, \ldots, m\}$ across all samples. Calculate the standard deviation of $A_{j_1 j_2}$, that is, $V_{j_1 j_2} = \text{st.dev.}\,(A_{j_1 j_2})$, and average the standard deviation for gene j_1 with respect to all other genes $M_{j_1} = \bar{V}_{j_1 \cdot}$. Large M_{j_1} value indicates a large variation in $\log_2$-transformed gene expression ratio when gene j_1 is compared with respect to other genes, which further indicates low stability for gene j_1. geNorm [7] suggests selecting genes with relatively small M values as suitable housekeeping genes. According to this rationale, geNorm adopts a step-down elimination approach to remove genes with the highest M value step by step. Every time after one gene is removed, the M values are recalculated for all remaining genes. This process is often repeated until two or three genes are remained.

NormFinder [8] introduces a group variable and takes between group variation into account in a two-way ANOVA model, $Y_{jgi} = \alpha_{jg} + \beta_{gi} + \varepsilon_{jgi}$, where α_{jg} is the amount of expression attributable to the jth gene within the gth group, β_{gi} is the amount of expression attributable to the ith sample in the gth group, and $\varepsilon_{jgi} \sim N\left(0, \sigma^2_{jg}\right)$. Let $z_{jg} = \bar{y}_{jg\bullet}$ be the average of the measured gene expressions for gene j in group g and $\theta_g = \bar{\beta}_{g\bullet}$ be the average sample level in group g. Then, the stability measure ρ_{jg} is the mean plus one standard deviation of $z_{jg} - \theta_g - \alpha_j$. Finally, combine ρ_{jg}, $g = 1, 2, \ldots, G$, into one value for gene j by taking average $\rho_j = \sum_{g=1}^{G} \rho_{jg}/G$.

Limitations in the current normalization approaches exist. (i) The existing methods cannot detect multiple systematic effects, systematic effects related to continuous variables, or interactions between genes and systematic effects. geNorm does not take systematic effects into account, which may lead to misspecification of housekeeping genes [8]. NormFinder addresses this issue by constructing a two-way ANOVA model but only incorporates *one categorical* group variable. For instance, assume that age has a potential systematic effect related to gene expression. geNorm is unable to take the age effect into account, whereas NormFinder has to analyze age as a categorical group variable. The possibility of a continuous age effect or gene by age interaction is not addressed by the existing methods. (ii) The existing methods do not allow missing data. The missing observations need to be imputed before analysis. Otherwise, a sample with a missing value will be excluded, leading to loss of valuable information. (iii) A sample size calculation procedure has not been established in the existing methods. Lack of confidence interval in stability measures poses a major challenge for practitioners to determine the cutoff of a stability measure in selection of housekeeping genes. When stability values from different gene combinations are close, there is no statistical testing to determine whether the scores are significantly different.

To address these issues, we propose a three-way linear mixed-effects model (LMM) and develop a procedure to determine housekeeping genes. In Section 2, we will lay out the model, provide confidence intervals for the stability measure, generate sample size calculation for experimental design, and propose a procedure for determination of housekeeping genes. In Section 3, we will illustrate our method and compare our method with geNorm and NormFinder in three case studies. The empirical assessment to compare our method versus geNorm and NormFinder is in Section 4. We will describe the advantages of the proposed method in Section 5.

2. Methods

2.1. Three-way linear mixed-effects model

We will construct a three-way LMM for selection and testing housekeeping genes with the strongest intraclass correlation coefficient (ICC). The three-way LMM is composed of (i) fixed gene effects,

(ii) random sample effects, and (iii) fixed systematic effects and/or their interactions with genes. Random sample effects are used to take the within-sample correlation into account as multiple genes are measured from the same sample. We propose using ICC among gene expression levels as the stability measure to select housekeeping genes that have high between-sample variation and low within-sample variation. Hypothesis testing is proposed to ensure that selected housekeeping genes are free of systematic effects or gene by systematic effect interactions.

Let $Y = (y_{11}, y_{12}, \ldots, y_{1m}, y_{21}, y_{22}, \ldots, y_{2m}, \ldots, y_{n1}, y_{n2}, \ldots, y_{nm})^t$ be the vector of gene expression levels, y_{ij}, of the jth ($j = 1, 2, \ldots, m$) candidate gene measured from the ith ($i = 1, 2, \ldots, n$) subject. Gene expressions can be log transformed if needed. We will construct a LMM,

$$\begin{aligned} Y &= \vec{1}\mu + (X_1\beta_1 + X_2\beta_2) + G\gamma + A\alpha + \varepsilon && (1.1) \\ &= (\vec{1}\mu + X_{12}\beta_{12} + G\gamma) + A\alpha + \varepsilon && (1.2) \\ &= X\beta + A\alpha + \varepsilon && (1.3) \end{aligned} \quad . \tag{1}$$

In model (1.1), μ is the global mean, and $\vec{1}$ is the vector of 1 with length mn. We express systematic effects in matrix notation where X_1 stands for variables and β_1 is a fixed effect parameter vector. Systematic effects include (i) the main effects from multiple continuous and/or categorical variables (i.e., treatment groups, genetic variants, demographic variables, and clinical variables), $X_1\beta_1$, and (ii) interactions between the systematic effect variables and genes, $X_2\beta_2$. The fixed gene effect, $G\gamma$, models the mean expression levels contributed by genes, where $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_m)^t$ and $G = 1_n \otimes I_m$. The random subject effect, $A\alpha$, models between-subject variation, where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)^t$ and $A = I_n \otimes 1_m$. The between-subject variation, α_i, identically and independently follows $N\left(0, \sigma_\alpha^2\right)$. The sample variation ε_{ij} identically and independently follows $N\left(0, \sigma_\varepsilon^2\right)$, and ε_{ij} and α_i are mutually independent.

We express model (1) into three equivalent formats as they are needed in different stages of model estimation. In model (1.2), the main systematic effects and interactions between systematic effects and genes are aggregated into $X_{12}\beta_{12}$, where $X_{12}\beta_{12} = X_1\beta_1 + X_2\beta_2$ and $\beta_{12} = \left(\beta_1^t, \beta_2^t\right)^t$. To ensure stable housekeeping genes free of systematic effects, we need to perform statistical inference regarding

$$\beta_{12} = \left(\beta_1^t, \beta_2^t\right)^t = \vec{0}. \tag{2}$$

We further combine all fixed effect components into $X\beta$, where $X\beta = \vec{1}\mu + X_{12}\beta_{12} + G\gamma$ and $\beta = \left(\mu, \beta_{12}^t, \gamma^t\right)^t$ in model (1.3), which has the standard format for the general LMM.

2.2. *Proposed housekeeping gene identification process*

We propose a step-up process to identify housekeeping genes:

Step 1: Start with exhaustive search of $k = 2$ gene combinations. Fit two genes j_1 and j_2 for $\forall\, j_1, j_2 \in \{1, 2, \ldots, m\}$ to model (1).

Step 2: Perform likelihood ratio test (LRT) to remove gene combinations with significant systematic effects. The LRT is the uniformly most powerful test and will be applied in the case studies. Two alternatives to the LRT, the Wald test and score test, are available. The details of hypothesis testing will be described in Section 2.3.

Step 3: Calculate ICC $\rho_{j_1 j_2} = \sigma_\alpha^2 / \left(\sigma_\alpha^2 + \sigma_\varepsilon^2\right)$ and 95% confidence interval of ICC for genes j_1 and j_2. See Section 2.4 for details.

Step 4: Repeat Steps 1–3 for $k = 3, 4, \ldots, m$ gene combinations. This process will terminate for $k < m$ if ICC stops improving.

Step 5: A gene combination with the highest lower bound of 95% confidence interval of ICC and no significant systematic effects (LRT p-value $\geqslant 0.05$) is the optimal reference for normalization.

After selection of housekeeping genes, one can perform the downstream analysis for target genes. A geometric mean from multiple housekeeping genes on each sample will be used as a normalizing factor. Target gene expression levels are divided by the normalizing factor from the same sample. Note the target genes will not be included in the selection of housekeeping genes as target genes usually have more variations potentially associated with diseases. Including target genes from the downstream analysis in the housekeeping gene selection process will increase the selection bias.

2.3. *Hypothesis testing of systematic effects*

Ideal housekeeping genes should be free of systematic main effects (β_1) and systematic effect by gene interactions (β_2). Therefore, it is important to perform statistical tests to assess overall systematic effects through $H_0 : \beta_{12} = \left(\beta_1^t, \beta_2^t\right)^t = \vec{0}$ versus $H_a : \beta_{12} \neq \vec{0}$. In LMM, the LRT is commonly used to fulfill this task. Under H_0, the LRT statistic converges in distribution to a chi-square distribution, that is, $\lambda \xrightarrow{d} \chi_v^2$, where the degrees of freedom v equals the length of β_{12}. One can also consider a Wald test. Let $\hat{\beta}_{12}$ be a maximum likelihood estimator of β_{12} and $I\left(\hat{\beta}_{12}\right)$ be the Fisher information matrix for β_{12}. Under H_0, $\hat{\beta}_{12}^t I\left(\hat{\beta}_{12}\right)\hat{\beta}_{12} \xrightarrow{d} \chi_v^2$. A score test, based on the derivatives of the log likelihood function l, can be formulated when the variability is difficult to estimate. Take the observed score $U\left(\beta_{12} = \vec{0}\right) = \frac{\partial l}{\partial \beta_{12}}\Big|_{\beta_{12}=\vec{0}}$ and the observed information $I\left(\beta_{12} = \vec{0}\right) = -\frac{\partial^2 l}{\partial \beta_{12}\partial \beta_{12}^t}\Big|_{\beta_{12}=\vec{0}}$, then the score test statistic $U\left(\beta_{12} = \vec{0}\right)^t I\left(\beta_{12} = \vec{0}\right)^{-1} U\left(\beta_{12} = \vec{0}\right) \xrightarrow{d} \chi_v^2$ under H_0.

In addition to the global test of systematic effects $H_0 : \beta_{12} = \vec{0}$, one can perform post hoc tests to assess each individual component of β_{12}. Let L be a vector with the length of β_{12}, one component as 1 and the reming components as 0. To test $H_0 : L^t\beta_{12} = 0$ versus $H_a : L^t\beta_{12} \neq 0$, let $\hat{C}_{\beta_{12}}$ be the estimated variance and covariance matrix for β_{12}. The t-test statistic $t = \frac{L^t\hat{\beta}_{12}}{\sqrt{L^t\hat{C}_{\beta_{12}}L}}$ approximately follows t-distribution, and the degrees of freedom can be estimated by [15]. If one is reluctant to employ formal hypothesis testing for identifying systematic variability, then another possibility is estimation. For example, one may compute the LRT (or Wald or score) statistic and subtract off the degrees of freedom. This will yield a point estimate of the underlying noncentrality parameter, which is essentially an effect size. One may declare as unsuitable for normalization candidate genes for which the estimated effect size is too large, even if formal hypothesis testing would not have rejected the corresponding null hypothesis (note that, in this context, a *type I error* is less serious than a *type II error*: declaring as unsuitable genes that would have been okay is better than declaring as suitable genes that would not have been okay).

2.4. *Confidence interval of intraclass correlation coefficient*

When genes are free of systematic effects or systematic effect by gene interactions, that is, $\beta_{12} = \vec{0}$, we can reduce model (1) to a two-way mixed-effects ANOVA model

$$Y = \vec{1}\mu + G\gamma + A\alpha + \varepsilon, \tag{3}$$

where μ is the overall fixed effect, γ is the fixed gene effect, and α is the random sample effect. Model (3) fits as Case 3A in [16]. Consider ANOVA for a complete randomized block design, let MS_{B} be the between-subject mean square, and let MS_{R} be the residual mean square. The ICC in model (3) can be estimated by $\hat{\rho} = (MS_{\mathrm{B}} - MS_{\mathrm{R}})/(MS_{\mathrm{B}} + (m-1)MS_{\mathrm{R}})$. An exact $100(1-c)\%$ confidence interval for ρ has

$$\text{lower limit} = (F_{\mathrm{L}} - 1)/(F_{\mathrm{L}} + n - 1) \text{ and upper limit} = (F_{\mathrm{U}} - 1)/(F_{\mathrm{U}} + n - 1), \tag{4}$$

where $F_{\mathrm{L}} = (MS_{\mathrm{B}}/MS_{\mathrm{R}})/F_{c/2;\upsilon_1,\upsilon_2}$, $F_U = (MS_B/MS_R)/F_{c/2;\upsilon_2,\upsilon_1}$, $\upsilon_1 = m-1$, and $\upsilon_2 = (m-1)(n-1)$.

2.5. *Sample size*

Algorithms to determine the number of samples in selection of housekeeping genes for RT-PCR data have not been specifically provided in literature. Our proposed method suggests that ideal reference genes for normalization do not have systematic effects or systematic effect by gene interactions. Therefore, one can utilize the ICC in model (3) to determine the sample size. According to [17], the width of the confidence interval of ICC derived in (4) can be approximated by $2z_{c/2}\sqrt{\mathrm{var}\hat{\rho}}$, where $\mathrm{var}\,(\hat{\rho}) = 2(1-\rho)^2(1+(m-1)\rho)^2/\{m(m-1)(n-1)\}$. Let w be a desired width of confidence interval and ρ be an expected ICC for housekeeping genes. Solving the equation $w = 2z_{c/2}\sqrt{\mathrm{var}\hat{\rho}}$, we obtain

Table I. Minimal sample sizes needed in experiments to detect m underlying housekeeping genes with ICC $= \rho$ and $w =$ width of 95% confidence interval of ρ.

	$w = 0.1$				$w = 0.2$			
ρ	$m = 2$	3	4	5	$m = 2$	3	4	5
0.90	57	42	37	4	15	12	10	10
0.89	68	49	43	40	18	13	12	11
0.88	80	58	50	47	21	16	14	13
0.87	92	66	58	54	24	18	16	15
0.86	106	76	66	61	28	20	18	16
0.85	120	86	74	68	31	23	20	18
0.84	135	96	83	76	35	25	22	20
0.83	150	106	92	84	39	28	24	22
0.82	166	117	101	93	43	30	26	24
0.81	183	128	110	101	47	33	29	26
0.8	201	140	120	110	51	36	31	29
0.79	219	152	130	119	56	39	34	31
0.78	237	164	140	128	60	42	36	33
0.77	256	176	150	137	65	45	39	35
0.76	276	189	160	146	70	48	41	38
0.75	296	202	171	155	75	52	44	40
0.74	316	214	181	164	80	55	46	42
0.73	337	227	191	174	85	58	49	45
0.72	358	241	202	183	91	61	52	47
0.71	379	254	213	192	96	65	54	49
0.70	401	267	223	201	101	68	57	51

the minimal total sample size as $n = 8z^2_{c/2}\{(1-\rho)^2(1+(m-1)\rho)^2\}/\{m(m-1)w^2\} + 1$. Because the underlying housekeeping genes are expected to have high ICC, we tabulate minimal sample sizes for two to five housekeeping genes with ICC ranging between 0.7 and 0.9 and two-sided confidence interval width = 0.1 and 0.2 in Table I. Here, m is the number of true housekeeping genes instead of the number of candidate genes in experiments.

3. Case studies

In this section, we will compare our proposed method with geNorm and NormFinder by using three case studies. In all three case studies, our proposed method included a gene fixed effect, tumor fixed effect, tumor by gene interaction, and sample random effect in full LMM (1). We performed LRT to remove candidate housekeeping genes with significant systematic effects (tumor effects or tumor by group interactions). The ICC and 95% confidence interval were calculated from the reduced LMM where insignificant systematic effects were removed. A gene combination with the highest lower bound in 95% CI of ICC was selected as the optimal housekeeping genes.

geNorm did not take the tumor effect or tumor by gene interaction into consideration. The algorithm measured stability of genes, with higher M values indicating lower stability. Genes with high M values were removed sequentially, and M values were updated until only two genes remained.

NormFinder took the tumor effect into consideration but did not consider tumor by gene interaction. All genes were fitted into one ANOVA model in one step, and the two most stable genes were selected as the optimal housekeeping genes.

3.1. Bladder cancer 1 study

Gene expression levels measured by RT-PCR were obtained for 14 genes on 28 subjects (Table II). The subjects were divided into three tumor groups: T_a ($n = 10$), T_1 ($n = 8$), and T_{2-4} ($n = 10$).

The results from the proposed method suggest that there were no significant systematic effects regarding the tumor group main effect and tumor group by gene interaction. The top three gene combinations with the highest ICCs in two-way, three-way, and four-way combinations are listed in Table III(a).

Table II. The dataset description.

Study name	Sample size	Number of candidate genes	Candidate genes	Systematic effect variable	Reference
Bladder cancer 1	28	14	ATP5B, HSPCB, S100A6, FLOT2, TEGT, UBB, TPT1, CFL1, ACTB, RPS13, RPS23, GAPD, UBC, FLJ20030	Tumor group	[8] supplementary data
Bladder cancer 2	26	8	CD14, FCN1, CCNG2, NPAS2, UBC, CFL1, ACTB, GAPD	Tumor group	[8] supplementary data
Colon cancer	40	13	UBC, UBB, SUI1, NACA, FLJ20030, CFL1, ACTB, CLTC, RPS13, RPS23, GAPD, TPT1, TUBA6	Cancer classification	[8] supplementary data

Our proposed method selected HSPCB, RPS13, and RPS23 as the most suitable reference genes for normalization (estimated ICC = 0.90, 95% CI: 0.817–0.95, LRT p-value = 0.53) (Table III(a)).

The selections of reference genes from NormFinder and geNorm were different. geNorm selected UBC and CFL1 (M-value = 0.358), whereas NormFinder selected HSPCB and RPS13 (combined stability value = 0.08) (Table IV). Because the proposed method selected optimal housekeeping genes from all combinations, whareas the software NormFinder only provides the optimal two-way combinations of housekeeping genes, the result from the proposed method is consistent with that of NormFinder in this case.

3.2. *Bladder cancer 2 study*

Gene expression levels measured by RT-PCR were obtained from eight genes on 26 subjects (Table II). The subjects were divided into two tumor groups: T_a ($n = 12$) and T_{2-4} ($n = 14$).

The proposed method tested the systematic effects, namely the main effect of tumor group and tumor group by gene interaction. The results of LRT show that four sets of gene combinations had significant systematic effects (LRT p-value < 0.05). Post hoc analysis indicates that the systematic effects were due to the interaction between gene and tumor group. These four sets of gene combinations were removed from further analysis. Our method suggests that UBC, CFL1, and GAPD are suitable housekeeping genes with the estimated ICC = 0.89, 95% CI = 0.80–0.94, and LRT p-value = 0.22 (Table III(b)).

The three genes with the lowest M-values are UBC ($M-$value = 0.46), CFL1 (M-value = 0.46) and GAPD (M-value = 0.496) according to geNorm (Table IV). The results from geNorm and our proposed methods are consistent. NormFinder suggests a different set of genes, CFL1 and ACTB, with the combined stability value = 0.088.

3.3. *Colon cancer study*

The RT-PCR measures of gene expression were obtained for 13 genes on 40 subjects (Table II). The subjects were classified into two cancer groups: Normal ($n = 10$) and Dukes ($n = 30$).

The proposed method tested the systematic effects regarding the cancer classification main effect and gene by cancer classification interaction. Three gene combinations had significant systematic effects (LRT p-value < 0.05) that were due to the interaction between gene and cancer classification according to the post hoc analysis. Our method suggests that CFL1, ACTB, and CLTC are suitable housekeeping genes with ESTIMATED ICC = 0.92, 95% CI: 0.864–0.95, LRT p-value = 0.42 (Table III(c)).

Our proposed method and geNorm have the same selections for two-way (RPS23 + TPT1) and three-way (RPS13 + RPS23 + TPT1) gene combinations, as the gene combination with the highest ICC by our proposed method matched the gene combination with the lowest M-values by geNorm. However, the results of LRT show that there was significant systematic effect (LRT p-value < 0.05) due to

Table III. Selection of housekeeping genes using the proposed method.

(a) Bladder 1 study top Top gene combinations	ICC	Systematic effects 95% CI of ICC	Gene	LRT $H_0 : \beta_{12} = \vec{0}$	Tumor group	Gene*tumor group
HSPCB + RPS23	0.91	(0.816, 0.96)	< 0.01	0.18	0.90	0.21
RPS13 + RPS23	0.91	(0.81, 0.96)	0.02	0.25	0.81	0.31
ATP5B + HSPCB	0.89	(0.79, 0.95)	< 0.01	0.48	0.99	0.49
HSPCB+RPS13+RPS23	0.90	(0.817, 0.95)	< 0.01	0.53	0.90	0.29
ATP5B + HSPCB + TEGT	0.88	(0.80, 0.94)	< 0.01	0.88	0.99	0.66
ATP5B + HSPCB + RPS23	0.88	(0.78, 0.94)	< 0.01	0.36	0.98	0.1
ATP5B + HSPCB + RPS13 + RPS23	0.87	(0.79, 0.93)	< 0.01	0.44	0.96	0.24
ATP5B + HSPCB + TEGT + RPS23	0.87	(0.78, 0.93)	< 0.01	0.69	0.97	0.46
HSPCB + TPT1 + RPS13 + RPS23	0.86	(0.77, 0.93)	< 0.01	0.32	0.82	0.17

(b) Bladder 2 study top Top gene combinations	ICC	Systematic effects 95% CI of ICC	Gene	LRT $H_0 : \beta_{12} = \vec{0}$	Tumor group	Gene*tumor group
UBC + GAPD	0.90	(0.79, 0.95)	0.02	0.27	0.86	0.11
CFL1 + GAPD	0.89	(0.76, 0.95)	0.16	0.78	0.54	0.75
UBC + CFL1	0.88	(0.75, 0.94)	0.30	0.11	0.77	0.04
UBC+CFL1+GAPD	0.89	(0.80, 0.94)	0.05	0.22	0.71	0.12
UBC + CFL1 + ACTB	0.80	(0.65, 0.89)	0.52	0.01	0.38	< 0.01
CFL1 + ACTB + GAPD	0.79	(0.64, 0.89)	0.50	0.17	0.28	0.14
UBC + CFL1 + ACTB + GAPD	0.81	(0.69, 0.90)	0.26	0.02	0.44	0.01
CCNG2 + UBC + CFL1 + GAPD	0.74	(0.60, 0.86)	0.35	< 0.01	0.68	< 0.01
CCNG2 + UBC + CFL1 + ACTB	0.65	(0.47, 0.80)	0.60	< 0.01	0.90	< 0.01

(c) Colon study top Top gene combinations	ICC	Systematic effects 95% CI of ICC	Gene	LRT $H_0 : \beta_{12} = \vec{0}$	Tumor group	Gene*tumor group
RPS23 + TPT1	0.94	(0.892, 0.97)	0.94	< 0.01	0.12	< 0.01
RPS13 + RPS23	0.94	(0.889, 0.97)	< 0.01	0.02	0.17	0.02
UBB + CFL1	0.92	(0.857, 0.96)	0.07	0.79	0.69	0.59
RPS13 + RPS23 + TPT1	0.93	(0.88, 0.96)	< 0.01	0.03	0.19	0.02
SUI1 + RPS13 + RPS23	0.92	(0.87, 0.95)	< 0.01	0.10	0.16	0.1
CFL1+ACTB+CLTC	0.92	(0.864, 0.95)	< 0.01	0.42	0.46	0.32
SUI1 + RPS13 + RPS23 + TPT1	0.92	(0.87, 0.95)	< 0.01	0.09	0.17	0.09
NACA + RPS13 + RPS23 + TPT1	0.91	(0.86, 0.95)	< 0.01	0.09	0.19	0.09
CFL1 + ACTB + CLTC + TUBA6	0.91	(0.857, 0.95)	< 0.01	0.11	0.60	0.07

Genes with the highest lower bound for 95% CI of ICC and LRT p-value > 0.05 are selected as reference genes for normalization and highlighted in red. Gene combinations with LRT p-value < 0.05 have systematic effects and thus are removed from consideration. Stop testing higher-order gene combinations if ICC does not increase.

gene by cancer classification. Therefore, these genes were removed from consideration in our method. NormFinder suggests a different set of genes, TPT1 and TUBA6 with combined stability value = 0.061 for normalization.

4. Empirical assessment

An empirical assessment was performed to compare our proposed method with competitors, geNorm, and NormFinder. We focus on two scenarios regarding genes with and without a group effect to illustrate the major differences among the three approaches (Table V). For each scenario, sample sizes of 25 and 50 subjects were assessed, respectively. We used bladder 2 data in Section 3 as a reference for parameter values in the following simulation models.

We first considered scenario I where expression levels for three candidate genes were simulated from the model $\ln(y_{ij}) = \mu_j + \alpha_i + v_{ij} + \varepsilon_{ij}$ for the jth gene in the ith subject. The mean gene expression

Table IV. Selection of housekeeping genes using geNorm and NormFinder.

	geNorm (no group)		NormFinder (disease group)	
Study name	Gene	*M* value	Gene	Stability value
Bladder 1	UBC	0.358	HSPCB	0.107
	CFL1	0.358	TEGT	0.136
	ATP5B	0.441	ATP5B	0.138
	HSPCB	0.465	UBC	0.141
	GAPD	0.550	RPS23	0.148
	TEGT	0.568	RPS13	0.149
	RPS23	0.634	CFL1	0.185
	RPS13	0.634	FLJ20030	0.185
	TPT1	0.695	TPT1	0.187
	FLJ20030	0.732	UBB	0.196
	FLOT2	0.770	FLOT2	0.205
	UBB	0.776	GAPD	0.236
	ACTB	0.813	S100A6	0.239
	S100A6	0.931	ACTB	0.242
Bladder 2	UBC	0.46	CFL1	0.109
	CFL1	0.46	ACTB	0.171
	GAPD	0.496	UBC	0.259
	ACTB	0.708	GAPD	0.262
	CCNG2	1.069	CCNG2	0.68
	CD14	2.134	CD14	0.87
	NPAS2	2.338	NPAS	0.889
	FCN1	2.551	FCN1	1.022
Colon	RPS23	0.388	UBC	0.088
	TPT1	0.388	GAPD	0.099
	RPS13	0.491	TPT1	0.128
	SUI1	0.545	RPS13	0.143
	UBC	0.566	TUBA6	0.147
	TUBA6	0.581	NACA	0.177
	UBB	0.587	UBB	0.178
	GAPD	0.594	SUI1	0.218
	NACA	0.6	CFL1	0.222
	CLTC	0.637	FLJ20030	0.228
	ACTB	0.645	ACTB	0.247
	CFL1	0.647	RPS23	0.265
	FLJ20030	0.811	CLTC	0.278

The best combination of two suitable genes is in red.

level was set as $\mu_j = 4$ for all three genes ($j = 1, 2, 3$). The random effect $\alpha_i \sim N(0, 0.64)$ induced the correlations among gene expression levels in the ith subject. Every subject had an independent random residual $\varepsilon_{ij} \sim N(0, 0.16)$. We assumed that genes 1 and 2 were the true housekeeping genes that were stable over subjects with variability $v_{ij} = 0$. We assumed that gene 3 was not a true housekeeping gene, and its expression level varied across subjects with $v_{ij} \sim \text{uniform}(0, 3)$.

Counts of correct selections of housekeeping genes over 10 simulations appear in Table V. For both $n = 25$ and $n = 50$, our method and geNorm correctly selected genes 1 and 2 as housekeeping genes with 100% accuracy. NormFinder misclassified gene 3 as one of the housekeeping genes in two cases when $n = 25$ and in four cases when $n = 50$.

The selection error of NormFinder in scenario I could arise from selection bias due to violation of the NormFinder model assumptions. Indeed, Anderson *et al*. noted in [8] (page 5247, requirements) that the candidate genes are assumed to have no prior expectation of expression difference between groups. More specifically, in the appendix of [8] (pages 3 and 4), the authors indicated the confounding of mean parameters. To address this confounding, NormFinder assumes that the average expression level of all genes is independent of the group. Furthermore, NormFinder does not perform filtration to remove genes

Table V. Empirical comparison of three methods for selection of housekeeping genes.

	Models (gene expression level for the jth gene of the ith subject in the gth group)	Sample size	Proposed method	NormFinder	geNorm
Scenario 1: no group effect	$\ln(y_{ij}) = \mu_j + \alpha_i + v_{ij} + \varepsilon_{ij}$ Genes 1 and 2: $v_{ij} = 0$ Gene 3: $v_{ij} \sim$ uniform$(0, 3)$ For all three genes, $\mu_j = 4$, $\alpha_i \sim N(0,\ 0.64)$, and $\varepsilon_{ij} \sim N(0,\ 0.16)$	25	10/10	8/10	10/10
		50	10/10	6/10	10/10
Scenario 2: group effect	Genes 1 and 2: $\ln(y_{ij}) = \mu_j + \alpha_i + v_{ij} + \varepsilon_{ij}$, $v_{ij} = 0$ Gene 3: $\ln(y_{ijg}) = \mu_j + \alpha_i + v_{ij} + \gamma_g + \varepsilon_{ij}$ $v_{ij} \sim$ uniform$(0, 3)$, $\gamma_g = -2$ for $g = 1$ and $\gamma_g = 0$ for $g = 2$ Gene 4: $\ln(y_{i4g}) = 2 * \ln(y_{i3g}) + \xi_{i4}$, $\xi_{i4} \sim N(0, 0.01)$ For all genes, $\mu_j = 4$, $\alpha_i \sim N(0,\ 0.64)$, $\varepsilon_{ij} \sim N(0,\ 0.16)$	25	10/10	0/10	0/10
		50	10/10	0/10	0/10

Counts of correct selections of housekeeping genes over 10 simulations are listed. Genes 1 and 2 are true housekeeping genes with variability $v_{ij} = 0$.

with group effects. The prescription $v_{i3} \sim$ uniform$(0, 3)$ in our assessment implies that, for some of the simulated data sets, there might have been problematic imbalances across groups in the *sample* average expression level of all genes, even though scenario I did not have any group effect in the underlying *population*.

Next, we simulated scenario II with a group effect. We assume genes 1 and 2 are the true housekeeping genes with the model $\ln(y_{ij}) = \mu_j + \alpha_i + v_{ij} + \varepsilon_{ij}$, where $\mu_j = 4$, $\alpha_i \sim N(0,\ 0.64)$, $\varepsilon_{ij} \sim N(0,\ 0.16)$, and $v_{ij} = 0$. We assume that genes 3 and 4 are not true housekeeping genes with variability $v_{ij} \sim uniform(0, 3)$. For $n = 25$, 10 subjects were randomly assigned to group 1, and the remaining subjects are in group 2. For $n = 50$, 25 subjects were randomly assigned to each group. For genes 3 and 4, a group effect $\gamma_g = -2$ is added to group 1. Gene 4 has twice the gene expression level of gene 3 plus a residual $\xi_{i4} \sim N(0,\ 0.01)$. The explicit expressions for genes 3 and 4 under groups $g = 1,\ 2$ are listed in Table V.

In scenario II, our method correctly selected the housekeeping genes with 100% accuracy, whereas geNorm and NormFinder were unable to select the correct housekeeping genes. geNorm compares the ratios between expression levels for two genes and selects genes with the lowest variability, ignoring the potential group effect among genes. As a result, geNorm mistakenly selected genes 3 and 4 as housekeeping genes with a selection error rate of 100%. NormFinder was unable to identify the correct housekeeping genes, as this method is based on a null hypothesis of no group effect.

5. Discussion and conclusion

We have proposed a three-way LMM and ICC to determine reference genes for normalization. The mixed-effects model can take multiple continuous or categorical systematic effect variables into account and ensure that the selected housekeeping genes are free of systematic effects. The proposed method offers a 95% confidence interval for the stability measure. Sample size calculation is offered on the basis of the proposed method framework, which fills in the gap in the existing methods. One of attractive features for the LMM is that it can accommodate data that are missing at random. Our method inherits this nice feature.

geNorm is one of the most commonly used approaches for normalization in RT-PCT data with more than 4000 citations by research articles. Our proposed method provided consistent results when compared with geNorm in two of our case studies while addressing the limitations in geNorm. We constructed the three-way LMM to analyze fixed gene effects, random sample effects, and fixed

systematic effects. Multiple tests of individual systematic effects would inflate the probability of type I error. We address this issue by performing a global test on $H_0 : \beta_{12} = \vec{0}$ with the LRT or the Wald test. Condensing multiple effects into one global test offers a powerful test without inflating the probability of type I error. Gene combinations with LRT p-value < 0.05 will be removed from analysis.

The ICC is suitable to serve as a stability measure in search of reference genes with low variability. The ICC $\rho = \sigma_\alpha^2 / \left(\sigma_\alpha^2 + \sigma_\varepsilon^2\right)$ is the ratio of between-subject variation to sum of between-subject variation and residual variation. Maximizing ICC is equivalent to minimizing relative residual variation when taking the sum of variations into account as $\rho = \sigma_\alpha^2 / \left(\sigma_\alpha^2 + \sigma_\varepsilon^2\right) = 1 - \sigma_\varepsilon^2 / \left(\sigma_\alpha^2 + \sigma_\varepsilon^2\right)$. The rationale of using ICC to identify reference genes is in agreement with the rationales adopted by existing methods. For instance, geNorm minimizes the standard deviation for the ratios between gene expression levels, although it is heuristic to average the standard deviations and combine them into an M-value. The BestKeeper method [10] suggests the use of Pearson correlation coefficients for all pairwise gene combinations to select reference genes with high Pearson correlation coefficients. However, Pearson correlation coefficients cannot measure the correlations among three or more genes.

The proposed method utilizes a step-up algorithm to search for gene combinations with high ICC and no systematic effects. The algorithm stops if the lower bound of 95% confidence interval of ICC does not increase for higher-order gene combinations. Using the lower bound of 95% confidence interval for ICC will take the variation of the stability measure into account and avoid selection of genes with ICC estimated so imprecisely that we cannot be confident of a high value. geNorm performs a step-down algorithm to remove genes with the highest M-value step by step and recalculates M-values for remaining genes. There is no objective cutoff point to determine when to stop the process. It is difficult to justify implementing an arbitrary cutoff such as $M < 0.5$ or $M < 1.5$ to determine reference genes in all experiments.

The rationales of the step-up and step-down algorithms for the proposed method and geNorm are consistent in that both methods try to exclude noisy genes that are inappropriate to serve as reference genes for normalization. All genes are analyzed by NormFinder in a two-way ANOVA model. Not removing irrelevant and noisy genes might inflate the variation in the model. As a result, geNorm and NormFinder did not select same genes in the three case studies.

Anderson *et al*. [8] pointed out the importance of analyzing systematic effects and constructed a two-way ANOVA model to measure the stability value after adjusting for a group effect. However, the method in NormFinder only works when the candidates are chosen from a set of genes with no prior expectation of expression difference between groups. In other words, the stability value and selection of housekeeping genes will be biased in NormFinder analysis if any candidate genes have significant group effects [8].

We compare our methods with geNorm and NormFinder by using three case studies. In three case studies (bladder cancer 1, bladder cancer 2, and colon cancer), there was no agreement in the selection of housekeeping genes between geNorm and NormFinder. In bladder cancer 1 study when there was no significant systematic effect, our selection was consistent with that of NormFinder. In bladder 2 and colon cancer case studies, the significant systematic effects due to group by gene interaction led to selection bias in NormFinder. As a result, the findings from our proposed method and geNorm were consistent, and they were different from those of NormFinder. We also provide sample size calculation for experiments to identify housekeeping genes. The existing methods do not provide sample size calculation formula. Our formula suggests increasing sample sizes as the expected ICC approaches $(m-2)/(2m-2)$ or a higher confidence level is desired. Moreover, sample size decreases as the number of true housekeeping genes increases. As indicated by [18], the effective sample size, using the notation from our manuscript, is $mn/(1+\rho(m-1))$. Making m larger does increase the effective sample size, which in turn provides narrower confidence intervals at a fixed n(or, as in Table I, permits a smaller m at a fixed confidence interval width). However, there is a lower bound $8z^2(1-\rho)^2\rho^2/w^2$ as $m \to \infty$. Thus, the sample size n cannot be made arbitrarily small by making m sufficiently large. A free software package written in SAS is available at http://d.web.umkc.edu/daih under software tab for practitioners to apply the proposed method.

Acknowledgements

There are no competing interests to this work. Special thanks to two reviewers for instructive comments to help us improve the manuscript.

References

1. Tricarico C, Pinzani P, Bianchi S, Paglierani M, Distante V, Pazzagli M, Bustin SA, Orlando C. Quantitative real-time reverse transcription polymerase chain reaction: normalization to rRNA or single housekeeping genes is inappropriate for human tissue biopsies. *Analytical Biochemistry* 2002; **309**(2):293–300.
2. Gibson UE, Heid CA, Williams PM. A novel method for real time quantitative RT-PCR. *Genome Research* 1996; **6**(10):995–1001.
3. Heid CA, Stevens J, Livak KJ, Williams PM. Real time quantitative PCR. *Genome Research* 1996; **6**(10):986–994.
4. Mane VP, Heuer MA, Hillyer P, Navarro MB, Rabin RL. Systematic method for determining an ideal housekeeping gene for real-time PCR analysis. *Journal of Biomolecular Techniques* 2008; **19**(5):342–347.
5. Butte AJ, Dzau VJ, Glueck SB. Further defining housekeeping, or "maintenance," genes focus on "A compendium of gene expression in normal human tissues". *Physiological Genomics* 2001; **7**(2):95–96.
6. Greer S, Honeywell R, Geletu M, Arulanandam R, Raptis L. Housekeeping genes; expression levels may change with density of cultured cells. *Journal of Immunological Methods* 2010; **355**(1-2):76–79.
7. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biology*; **3**(7). RESEARCH0034.
8. Andersen CL, Jensen JL, Orntoft TF. 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 Research* 2004; **64**(15):5245–5250.
9. Haller F, Kulle B, Schwager S, Gunawan B, von Heydebreck A, Sultmann H, Fuzesi L. Equivalence test in quantitative reverse transcription polymerase chain reaction: confirmation of reference genes suitable for normalization. *Analytical Biochemistry* 2004; **335**(1):1–9.
10. Pfaffl MW, Tichopad A, Prgomet C, Neuvians TP. Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper–Excel-based tool using pair-wise correlations. *Biotechnology Letters* 2004; **26**(6):509–515.
11. Huang Y, Hsu JC, Peruggia M, Scott AA. Statistical selection of maintenance genes for normalization of gene expressions. *Statistical Applications in Genetics and Molecular Biology* 2006; **5**. Article4.
12. Rodriguez-Lanetty M, Phillips WS, Dove S, Hoegh-Guldberg O, Weis VM. Analytical approach for selecting normalizing genes from a cDNA microarray platform to be used in q-RT-PCR assays: a cnidarian case study. *Journal of Biochemical and Biophysical Methods* 2008; **70**(6):985–991.
13. Mar JC, Kimura Y, Schroder K, Irvine KM, Hayashizaki Y, Suzuki H, Hume D, Quackenbush J. Data-driven normalization strategies for high-throughput quantitative RT-PCR. *BMC Bioinformatics* 2009; **10**:110.
14. Hruz T, Wyss M, Docquier M, Pfaffl MW, Masanetz S, Borghi L, Verbrugghe P, Kalaydjieva L, Bleuler S, Laule O, Descombes P, Gruissem W, Zimmermann P. RefGenes: identification of reliable and condition specific reference genes for RT-qPCR data normalization. *BMC Genomics* 2011; **12**:156.
15. McLean RA, Sanders WL. Approximating degrees of freedom for standard errors in mixed linear models. *Proceedings of the Statistical Computing Section, American Statistical Association*, New Orleans, 1988; 50–59.
16. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychological Methods* 1996; **1**(1):30–46.
17. Bonett DG. Sample size requirements for estimating intraclass correlations with desired precision. *Statistics in Medicine* 2002; **21**(9):1331–1335.
18. Killip S, Mahfoud Z, Pearce K. What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. *The Annals of Family Medicine* 2004; **2**(3):204–208.