SAS programs for real-time RT-PCR having multiple independent samples

Peyton Cook, Chunxiao Fu, Morgen Hickey, Eun-Soo Han, and Kenton S. Miller
University of Tulsa, Tulsa, OK, USA

BioTechniques 37:990-995 (December 2004)

Relative real-time reverse transcription PCR (RT-PCR) has become an important tool for quantifying changes in messenger RNA (mRNA) populations following differential development or stimulation of tissues or cells. However, the best methods for conducting such experiments and analyzing the resultant data remain an issue of discussion. In this report we describe an appropriate experimental methodology and the computer programs necessary to generate a meaningful statistical analysis of the combined biological and experimental variability in such experiments. Specifically, logarithmic transformations of raw fluorescence data from the log-linear portion of real-time PCR growth curves for both target and reference genes are analyzed using a SAS/STAT Mixed Procedure program specifically designed to give a point estimate of the relative expression ratio of the target gene with associated 95% confidence interval. The program code is open-source and is printed in the text.

INTRODUCTION

Relative real-time reverse transcription PCR (RT-PCR) has become an important method for determining messenger RNA (mRNA) expression levels, both in its own right (1,2) and as a tool for the validation of microarray experiments (3). However, the best approach for the analysis of such data remains a topic of discussion (4,5). Recently, we published a procedure for determining the relative expression ratio for a target mRNA under two different treatment conditions directly from the raw real-time RT-PCR data without the use of a standard curve (6). Our method exploited a modification of the procedure originally proposed by Gentle et al. (7) for determining PCR efficiency and produces results that are both accurate and statistically verified. In the process of testing our method we demonstrated small but statistically significant day effects, which may call into question the use of a standard curve to determine efficiency when that efficiency will be applied to PCRs run on a different day. Because our procedure determines relative expression using slope and intercept data from each experiment individually rather than using a cycle threshold (C_t) and a general efficiency, this problem is effectively circumvented (6).

However, the statistical method used in our previously published procedure assumes that there are no confounding correlations between any pair of target and reference gene values. While this simple assumption is valid for the analysis of sample replicates, it is not valid when analyzing data from multiple independent samples (e.g., when each RT-PCR derives from a distinct RNA, perhaps isolated from individual mice or tissue samples that have been exposed to distinct treatment conditions), a situation frequently encountered in gene expression studies and important for establishing the true biological variability in the system. In this case, each target and reference gene pair is entangled in a “sample effect” that should be taken into account when calculating variance.

Here we present a new method for calculating the point estimate of a relative gene expression ratio between two treatment populations and for assigning appropriate 95% confidence intervals to the estimate. Our procedure exploits Mixed Procedure algorithms built-in to the commercially available statistical program SAS, and thus should be of general applicability to anyone performing such experiments.

MATERIALS AND METHODS

Animals

Parental mice were purchased from Jackson Laboratories (Bar Harbor, ME, USA) and bred at the Animal Core of the Nathan Shock Center at the University of Texas Health Science at San Antonio. All mice were fed ad libitum (AL) Harlan Teklad LM-485 mouse/rat sterilizable diet 7912 (Madison, WI, USA) until 6 weeks of age. At 6 weeks, half of the mice were allowed to continue on this diet until sacrificed. The remaining mice were calorie-restricted (CR) by limiting them to 60% of the mean food intake of group AL until sacrificed. All procedures involving the use of mice were approved by the Institutional Animal Care and Use Committee of the University of Texas Health Science Center and the Subcommittee for Animal Studies at the Audie L. Murphy Memorial Veterans Hospital.

Tissue Collection and RNA Preparation

Livers from 4- to 6-month-old male mice were collected. The tissues were quickly frozen in liquid nitrogen and stored at -80°C until RNA extraction. Total RNA was extracted from each liver as previously described (8).

Real-Time RT-PCR

The reverse transcription reaction was performed using 1 µg of DNase I (Invitrogen, Carlsbad, CA, USA) digested total RNA, random primers, and SuperScript® II RT (Invitrogen) in a total volume of 20 µL according to the protocol of the manufacturer. The cDNA was diluted to 6-fold for the real-time RT-PCR. Primers were designed using the OligoPerfect™ Designer (Invitrogen) and purchased from Invitrogen. The *18S* ribosomal RNA (rRNA) was used as the reference gene for the target gene (NF κ light chain; *NFK*) normalization. PCR was carried out using a Smart Cycler® thermal cycler (Cepheid, Sunnyvale, CA, USA). Each PCR included 3.0 µL diluted cDNA, 2.5 µL 10× PCR buffer without $MgCl_2$ (Sigma, St. Louis, MO, USA), 1.0 µL

Table 1. PCR Primers Used in this Study

Gene	Primer	Sequence
18S	Forward	5′-TCAAGAACGAAAGTCGGAGGTT-3′
	Reverse	5′-GGACATCTAAGGGCATCACAG-3′
NFK	Forward	5′-GAGGATGAGGTGAGTGTTCC-3′
	Reverse	5′-CACCAGGCTGTAGGAGTTTC-3′

18S, ribosomal RNA (rRNA), the reference gene; *NFK*, NF κ light chain, the target gene.

25 mmol/L $MgCl_2$, 0.5 µL 10 mmol/L dNTPs (Invitrogen), 1.0 µL 0.5 µmol/L each primer, 0.25 µL *Taq* DNA polymerase (Sigma), 0.2 µL 300 g/L bovine serum albumin (BSA; Sigma), 2.5 µL SYBR® Green I (Molecular Probes, Eugene, OR, USA) at a concentration of 1:4000 of the commercial stock, and 13.05 µL AccuGENE® Molecular Biology Grade water (Cambrex Bio Science, Rockland, ME, USA). The thermal cycling parameters included a 94°C heating step for 1 min at the beginning of every run. The tubes were then cycled 40 times at 94°C for 30 s, annealed at the optimal annealing temperature of each primer set (18S = 62°C; NFK = 59°C) for 60 s, and extended at 72°C for 60 s. Optical data were collected during the annealing step. The specificity of the reaction was monitored by melting curve analysis to avoid nonspecific signals.

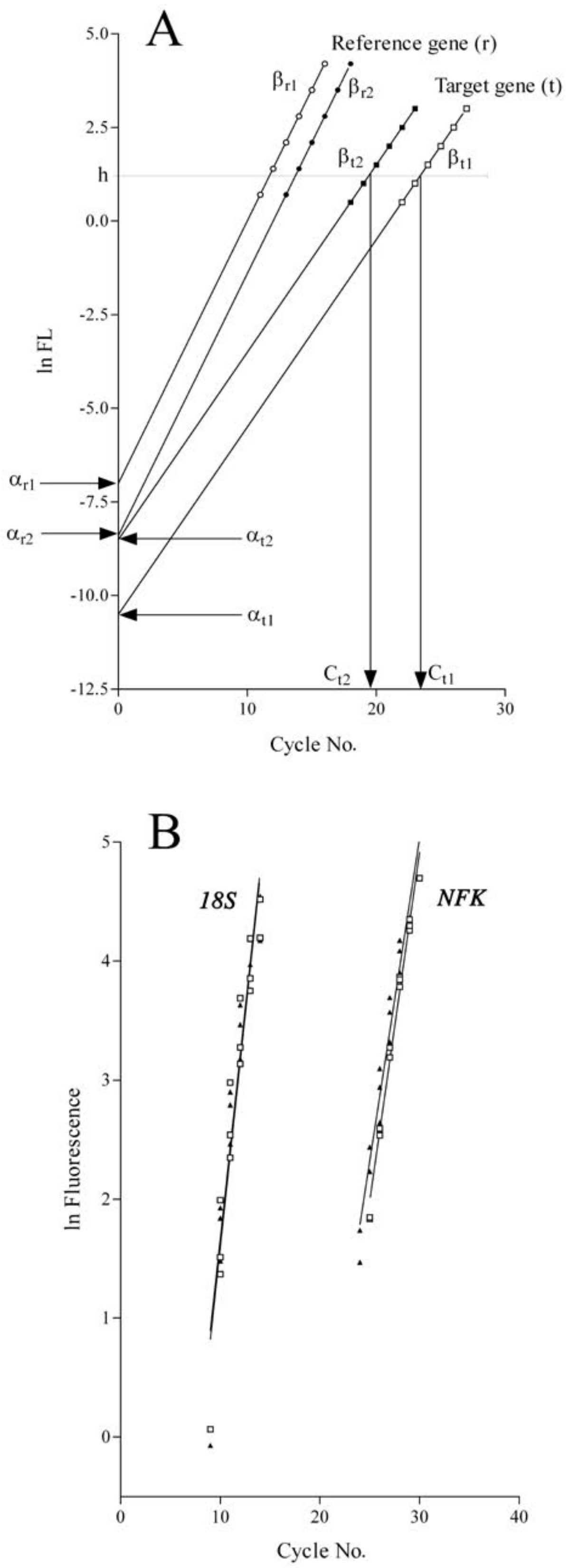
Data Analysis

Optics data was exported from the Cepheid Smart Cycler as comma separated values files (*.csv) and imported into an Microsoft® Excel® spreadsheet. We have written a Visual Basic Excel macro that facilitates determination and conversion of the appropriate subset of Smart Cycler

Table 2. Typical PCR Data Set Format for Import into SAS[a]

Treatment	Sample	Cycle	Gene	Y	Treatment	Sample	Cycle	Gene	Y
AL	1	10	*18S*	1.99243	AL	1	26	*NFK*	2.584725
AL	1	11	*18S*	2.978925	AL	1	27	*NFK*	3.275488
AL	1	12	*18S*	3.688879	AL	1	28	*NFK*	3.863938
AL	1	13	*18S*	4.189655	AL	1	29	*NFK*	4.350482
AL	1	14	*18S*	4.518159	AL	1	30	*NFK*	4.697861
AL	2	10	*18S*	1.369486	AL	2	25	*NFK*	1.835244
AL	2	11	*18S*	2.348195	AL	2	26	*NFK*	2.535962
AL	2	12	*18S*	3.135494	AL	2	27	*NFK*	3.191456
AL	2	13	*18S*	3.750288	AL	2	28	*NFK*	3.784622
AL	2	14	*18S*	4.195697	AL	2	29	*NFK*	4.258648
AL	3	9	*18S*	0.064538	AL	3	25	*NFK*	1.847487
AL	3	10	*18S*	1.511458	AL	3	26	*NFK*	2.594338
AL	3	11	*18S*	2.538974	AL	3	27	*NFK*	3.274559
AL	3	12	*18S*	3.275886	AL	3	28	*NFK*	3.846071
AL	3	13	*18S*	3.855805	AL	3	29	*NFK*	4.297548
CR	4	10	*18S*	1.926678	CR	4	26	*NFK*	2.645232
CR	4	11	*18S*	2.901421	CR	4	27	*NFK*	3.320283
CR	4	12	*18S*	3.634071	CR	4	28	*NFK*	3.903647
CR	4	13	*18S*	4.182559	CR	4	29	*NFK*	4.367307
CR	4	14	*18S*	4.54542	CR	4	30	*NFK*	4.698167
CR	5	9	*18S*	-0.06899	CR	5	24	*NFK*	1.470377
CR	5	10	*18S*	1.840549	CR	5	25	*NFK*	2.23284
CR	5	11	*18S*	2.793208	CR	5	26	*NFK*	2.942775
CR	5	12	*18S*	3.466777	CR	5	27	*NFK*	3.572641
CR	5	13	*18S*	3.971549	CR	5	28	*NFK*	4.08742
CR	6	10	*18S*	1.481606	CR	6	24	*NFK*	1.737691
CR	6	11	*18S*	2.462434	CR	6	25	*NFK*	2.436686
CR	6	12	*18S*	3.180828	CR	6	26	*NFK*	3.09946
CR	6	13	*18S*	3.762749	CR	6	27	*NFK*	3.695611
CR	6	14	*18S*	4.175413	CR	6	28	*NFK*	4.174711

[a] Please note that in a typical experiment we would use six samples per treatment condition (i.e., sample 1–12), but in the interests of brevity, we have only listed three for each condition. AL, ad libitum; CR, calorie-restricted; *18S*, 18S ribosomal RNA (rRNA), the reference gene; *NFK*, NF κ light chain, the target gene; Y, ln fluorescence at listed cycle number.



optics data to a logarithmic format for subsequent analysis. The resultant data (see Table 2) was then imported into SAS (SAS/STAT software version 8; SAS Institute, Cary, NC, USA) for subsequent analysis with the described programs (see Figure 2).

RESULTS AND DISCUSSION

As first proposed by Gentle et al. (7), the efficiency (E) of a PCR may be determined from the slope (β) of a linear regression line through a natural logarithmic transform of the log-linear portion of the PCR growth curve and $E = e^{\beta}$. The C_t number is also easily determined from such a plot by extrapolating an x-value from the point at which the PCR growth curve crosses some arbitrarily chosen y-value (the threshold), and C_t is frequently calculated in this fashion (Figure 1). C_t can also be expressed in terms of the threshold value (h), the y-intercept (α), and the slope (β) as

$$C_t = \frac{h - \alpha}{\beta},$$

and the difference between two C_t values determined for a given gene can be expressed as:

$$\mathrm{C}_{t1} - \mathrm{C}_{t2} = \frac{h - \alpha_1}{\beta_1} - \frac{h - \alpha_2}{\beta_2}$$

(6). Assuming that the efficiency of the PCR reaction is the same for the two C_t determinations (i.e., the regression lines are parallel), then this equation reduces to:

$$C_{t1} - C_{t2} = \frac{\alpha_2 - \alpha_1}{\beta} = \Delta C_t.$$

In any relative PCR experiment, the natural logarithm of the ratio (R) of two starting populations (N_{01} and N_{02}) of a target gene (i.e., $R = N_{01}/N_{02}$) is equal to natural logarithm of the

Figure 1. Analysis the log-linear region of a relative real-time RT-PCR experiment. (A) Graph of a model experiment using data chosen to clearly define the variables used in a typical analysis (α = y-intercept; β = slope; C_t = cycle threshold; h = threshold value). (B) A graph of actual data obtained from a typical reverse transcription PCR (RT-PCR) experiment (data taken from Table 1). *18S*, 18S ribosomal RNA (rRNA), the reference gene; *NFK*, NF κ light chain, the target gene.

efficiency of the reaction multiplied by the difference in C_t values for the two treatment conditions (i.e., 1n $R = E\Delta C_t$) (6). But since 1n $E = \beta$ and

$$\Delta C_t = \frac{\alpha_2 - \alpha_1}{\beta}, \text{then } 1\text{n}\, R = \beta\frac{\alpha_2 - \alpha_1}{\beta} = \alpha_2 - \alpha_1,$$

the difference between the intercepts.

The ratio (R^*) of the two starting populations of a target gene (t) normalized to the ratio of the starting populations of a reference gene (r), which is the number being sought in a relative real-time RT-PCR experiment, is simply: $R^* = R_t/R_r$, or in terms of logarithms: 1n R^* = 1n R_t - 1n R_r. Thus, substituting from above we have: 1n $R^* = (\alpha_{t2} - \alpha_{t1}) - (\alpha_{r2} - \alpha_{r1})$.

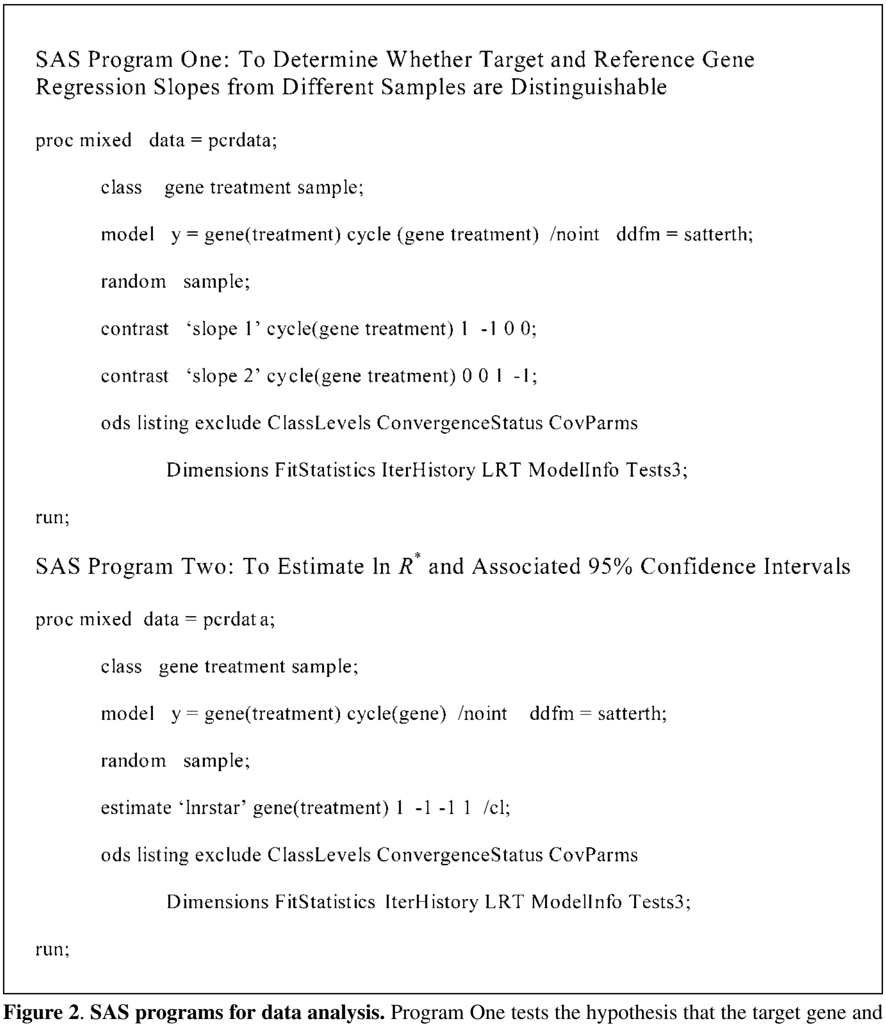
To account for "sample effects" as described in the Introduction, we have modeled the experimental measurement of fluorescent product accumulation during real-time PCR with the following equation:

$$1\text{n}\, Fl_{gdsl} = \alpha_{g(d)} + \upsilon_s + C_{gdsl}\beta_g + \varepsilon_{\text{gdsl}}.$$

The indices for this equation are gene (g) = 1, 2; treatment (d) = 1, 2; and sample (s) = 1, 2, ···, $\text{n}_{\text{s.}}$ The number of cycles (C) and the values of C may vary across genes, treatments, and samples. If we let l = 1, 2, ···, n_{gds}, then 1n Fl_{gdsl} is the natural logarithm of the measured fluorescence where $\alpha_{g(d)}$ is the true y-intercept for a particular gene/treatment combination, υ_s is a random effect on the intercept value due to the particular sample from which the cDNA was made, C_{gdsl} is the particular PCR cycle number in which the measurement is made, β_g is natural logarithm of the true efficiency of the PCR for the gene in question (i.e., the slope of the logarithmic transform of growth curve), and ε_{gdsl} is the experimental error associated with each particular measurement.

Using the model above, we have written a SAS program (Figure 2) to report P values for the hypothesis that the calculated slopes of either the target gene pair (β_{t1} and β_{t2} in Figure 1 and slope 1 in Figure 3) or the reference gene pair (β_{r1} and β_{r2} in Figure 1 and slope 2 in Figure 3) are the same between the two treatment conditions. If either of these values is <0.05, then the two treatment conditions under consideration cannot be usefully compared, because the PCR efficiencies of either the target or reference genes (or both) are distinguishable and thus violate an important assumption of the PCR model written above. If both P values are >0.05, then the second SAS program shown in Figure 2 may be run to estimate the logarithm of the relative expression ratio of the two target populations (*lnrstar*).

There is good reason to believe that under reasonably controlled conditions, the true "efficiencies" of two PCRs run using identical primers should, in fact, be identical. If they differ significantly, it is because some aspect of the PCR analysis has not been properly controlled. Indeed, Peirson et al. (8) compared both individual and mean efficiency corrections for both target and reference genes, and they note: "Applying individual corrections appears unjustified based upon these findings and rather than improving accuracy, introduces systematic errors which exaggerate the difference in expression and increase the assay noise." After assaying dozens of different genes in multiple samples using our method, this has been our experience as well. Calculating relative expression ratios from data in which either target or reference gene amplification efficiencies differ significantly between samples may, in fact, lead to misinterpretations of the true relative expression levels of the genes concerned and should, in our opinion, be avoided. This highlights the importance of SAS program one, which



SAS Program One: To Determine Whether Target and Reference Gene Regression Slopes from Different Samples are Distinguishable

```
proc mixed   data = pcrdata;

        class   gene treatment sample;

        model   y = gene(treatment) cycle (gene treatment)  /noint   ddfm = satterth;

        random   sample;

        contrast   'slope 1' cycle(gene treatment) 1  -1 0 0;

        contrast   'slope 2' cycle(gene treatment) 0 0 1  -1;

        ods listing exclude ClassLevels ConvergenceStatus CovParms

                Dimensions FitStatistics IterHistory LRT ModelInfo Tests3;

run;
```

SAS Program Two: To Estimate ln R^* and Associated 95% Confidence Intervals

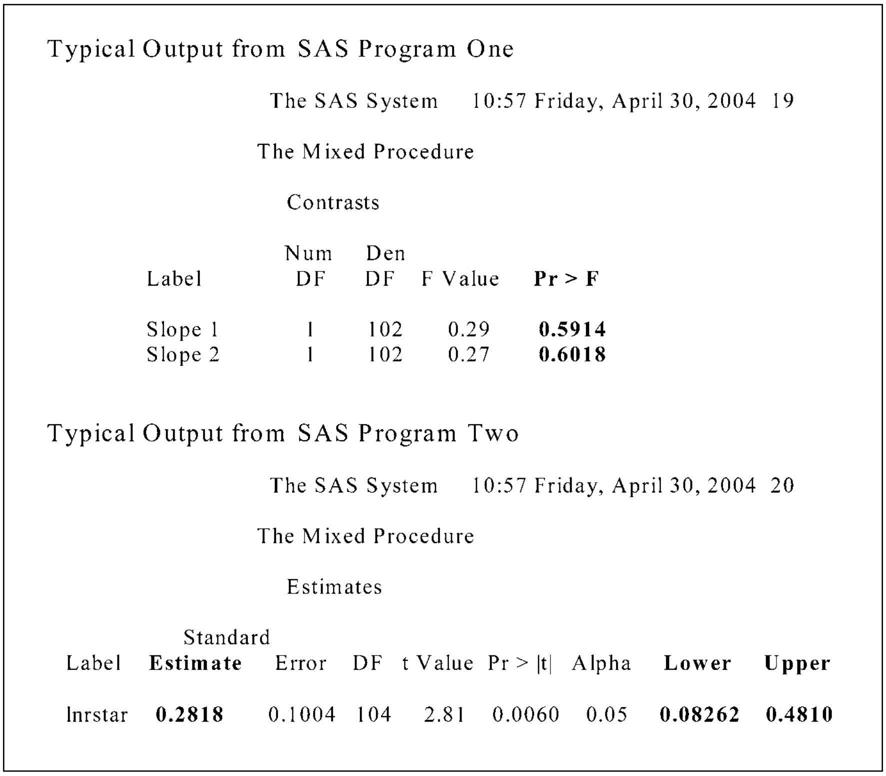
```
proc mixed  data = pcrdat a;

        class  gene treatment sample;

        model   y = gene(treatment) cycle(gene)  /noint    ddfm = satterth;

        random   sample;

        estimate 'lnrstar' gene(treatment) 1  -1 -1 1  /cl;

        ods listing exclude ClassLevels ConvergenceStatus CovParms

                Dimensions FitStatistics  IterHistory LRT ModelInfo Tests3;

run;
```

Figure 2. **SAS programs for data analysis.** Program One tests the hypothesis that the target gene and reference gene slopes are identical between treatment conditions. Program Two generates a point estimate for relative expression of the target gene between two different treatment conditions and give associated 95% confidence intervals for the measurement.

allows the user to validate that amplification efficiencies are not statistically distinguishable between the samples, and thus it should be run before running program two for each analysis.

In these programs "pcrdata" is the SAS database imported from an Excel spreadsheet containing the appropriately arranged and labeled PCR data (Table 2). As shown by Gentle et al. (7) and confirmed by Marino et al. (6), at least six samples with five data points each are required for good statistics. Typical output from these programs is shown in Figure 3, and the data of interest is shown in bold. Significantly more output relating to the model can be obtained by suppressing the "ods listing exclude" statement in either program. However, this output is not normally required for real-time RT-PCR analysis and has thus been suppressed. To determine the normalized target gene expression ratio R*, it is only necessary to exponentiate *lnrstar* ($e^{lnrstar}$) and its associated confidence intervals (CI). This could be done either by hand or by adding a few additional lines to the SAS code. We find it most convenient, however, to compare lnR^* in most situations, because genes that do not change with treatment will have 95% CI spanning zero, while for up-regulated genes the lower CI will be positive, and for down-regulated genes the upper CI will be negative. As shown in Figure 3, there is a slight but statistically significant increase in *NFK* gene expression between AL fed and calorie-restricted mice (lnR^* = 0.08 - 0.48 and R^* = 1.1 - 1.6). It should be noted that the program can be used with data generated by means other than SYBR Green fluorescence, including TaqMan®, molecular beacon, and multiplex probes.

Although gaining access to and processing the raw fluorescence data from a real-time PCR can be more or less labor-intensive depending upon which machine is being used to collect the data, there are several advantages to taking such an approach. One of these, as noted in our previous paper (6), is that statistically significant differences can be detected between relative expression ratios determined from RT-PCRs run on different days. Although these differences tend to be small where extreme care is taken to replicate experimental conditions, they nonetheless contribute to an overall error between experiments and thus, if possible, should be avoided. Another difficulty, which is corrected with the SAS procedure reported here, is the use of a threshold value to determine the C_t. It can easily be shown that a calculated ΔC_t between two experimental conditions can change significantly, as the threshold for the C_t determination is raised or lowered. This is because while the efficiency for the two reactions being measured may be statistically indistinguishable and therefore technically identical, random variation in the actual determinations may make the two regression lines nonparallel. Thus, ΔC_t will change as the threshold is moved. This problem is eliminated in our method because, having tested the slopes for identity with the "contrast" algorithm, we force equal slopes in the "estimate" algorithm.



```
Typical Output from SAS Program One

                         The SAS System     10:57 Friday, April 30, 2004  19

                        The Mixed Procedure

                          Contrasts

                          Num    Den
        Label             DF     DF    F Value     Pr > F

        Slope 1            1     102     0.29      0.5914
        Slope 2            1     102     0.27      0.6018

Typical Output from SAS Program Two

                         The SAS System     10:57 Friday, April 30, 2004  20

                        The Mixed Procedure

                          Estimates

                  Standard
Label    Estimate    Error   DF   t Value  Pr > |t|  Alpha    Lower     Upper

lnrstar   0.2818    0.1004  104    2.81    0.0060    0.05    0.08262   0.4810
```

Figure 3. **Typical Output for SAS Programs One and Two.** (Top) Output from SAS Mixed Procedure "contrast" algorithm applied to total data set (i.e., six AL mice and six CR mice) from experiment shown as partial data set in Table 2. The *P* values for target (slope 1) and reference (slope 2) genes are shown in bold. (Bottom) Output form SAS Mixed Procedure "estimate" algorithm applied to total data set from experiment shown as partial data set in Table 2. The point estimate for ln R^* (lnrstar) as well as the upper and lower 95% confidence intervals of the estimate are shown in bold. AL, ad libitum; CR, calorie-restricted; *18S*, 18S ribosomal RNA (rRNA), the reference gene; *NFK*, NF κ light chain, the target gene.

Finally, for a particular treatment condition, if the data from both reference and target genes are collected on the same day, it is possible to accumulate a database for later relative gene expression studies in which any one treatment condition can be usefully compared to any other treatment condition collected on any other day, as long as it is shown that the target and reference gene amplification efficiencies are the same (i.e., $P > 0.05$ for both genes in SAS Program One). Under these conditions, real-time RT-PCR becomes as generally useful as microarray analysis when only a limited number of genes (a few dozen or so) are to be examined.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health grant no. AI-41164 to K.S.M., grant nos. AG-00746 *and AG-14674-04S1 to E.-S.H., P20 RR-16478-03 to E.-S.H. and M.H., and by grants from the Mervin Bovaird Center for Molecular Cell Biology and Biotechnology to K.S.M.*

COMPETING INTERESTS STATEMENT

The authors declare no conflicts of interest.

REFERENCES

1. **Ginzinger, D.G.** 2002. Gene quantification using real-time quantitative PCR: an emerging technology hits the mainstream. Exp. Hematol. *30*:503-513.
2. **Mocellin, S., C.R. Rossi, P. Pilati, D. Nitti, and F.M. Marincola.** 2003. Quantitative real-time PCR: a powerful ally in cancer research. Trends Mol. Med. *9*:189-195.
3. **Wiguins, E., M.H. Meyer, J. Peppers, and R.A. Meyer, Jr.** 2004. Comparison of mRNA gene expression by RT-PCR and DNA microarray. BioTechniques *36*:618-626.
4. **Muller, P.Y., H. Janovjak, A.R. Miserez, and Z. Dobbie.** 2002. Processing of gene expression data generated by quantitative real-time RT-PCR. BioTechniques *32*:1372-1379.
5. **Pfaffl, M.W., G.W. Horgan, and L. Dempfle.** 2002. Relative expression software tool (REST©) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. Nucleic Acids Res. *30*:e36.
6. **Marino, J.H., P. Cook, and K.S. Miller.** 2003. Accurate and statistically verified quantification of relative mRNA abundances using SYBR Green I and real-time RT-PCR. J. Immunol. Methods *283*:291-306.
7. **Gentle, A., F. Anastasopoulos, and N.A. McBrien.** 2001. High-resolution semi-quantitative real-time PCR without the use of a standard curve. BioTechniques *31*:502-508.
8. **Peirson, S.N., J.N. Butler, and R.G. Foster.** 2003. Experimental validation of novel and conventional approaches to quantitative real-time PCR data analysis. Nucleic Acids Res. *32*:e73.
9. **Sambrook, J., E.F. Fritsch, and T. Maniatis.** 1989. Molecular Cloning: A Laboratory Manual. CSH Laboratory Press, Cold Spring Harbor, NY.

Received 9 June 2004; accepted 12 July 2004.

Address correspondence to Kenton S. Miller, Faculty of Biological Science, 600 S. College Ave., The University of Tulsa, Tulsa, OK 74104, USA. e-mail: Kenton-Miller@utulsa.edu