Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: http://www.tandfonline.com/loi/cjas20

Evaluation of different statistical methods using SAS software: an in silico approach for analysis of real-time PCR data

Mohammadreza Nassiri, Mahdi Elahi Torshizi, Shahrokh Ghovvati & Mohammad Doosti

To cite this article: Mohammadreza Nassiri, Mahdi Elahi Torshizi, Shahrokh Ghovvati & Mohammad Doosti (2017): Evaluation of different statistical methods using SAS software: an in silico approach for analysis of real-time PCR data, Journal of Applied Statistics, DOI: 10.1080/02664763.2016.1276890

To link to this article: http://dx.doi.org/10.1080/02664763.2016.1276890

JOURNAL OF APPLIED STATISTICS, 2017
http://dx.doi.org/10.1080/02664763.2016.1276890

Taylor & Francis
Taylor & Francis Group

Evaluation of different statistical methods using SAS software: an in silico approach for analysis of real-time PCR data

Mohammadreza Nassiri [a*], Mahdi Elahi Torshizi [b], Shahrokh Ghovvati [c*] and Mohammad Doosti [a]

[a]Department of Animal Science, Faculty of Agriculture, Ferdowsi University of Mashhad, Mashhad, Iran; [b]Department of Animal Science, Mashhad Branch, Islamic Azad University, Mashhad, Iran; [c]Department of Biotechnology, Faculty of Agriculture, University of Guilan, Rasht, Iran

ABSTRACT

Real-time polymerase chain reaction (PCR) is reliable quantitative technique in gene expression studies. The statistical analysis of real-time PCR data is quite crucial for results analysis and explanation. The statistical procedures of analyzing real-time PCR data try to determine the slope of regression line and calculate the reaction efficiency. Applications of mathematical functions have been used to calculate the target gene relative to the reference gene(s). Moreover, these statistical techniques compare C_t (threshold cycle) numbers between control and treatments group. There are many different procedures in SAS for real-time PCR data evaluation. In this study, the efficiency of calibrated model and delta delta C_t model have been statistically tested and explained. Several methods were tested to compare control with treatment means of C_t. The methods tested included *t*-test (parametric test), Wilcoxon test (non-parametric test) and multiple regression. Results showed that applied methods led to similar results and no significant difference was observed between results of gene expression measurement by the relative method.

ARTICLE HISTORY
Received 12 October 2015
Accepted 7 December 2016

KEYWORDS
Real-time PCR data; in silico; gene expression; statistical analysis; SAS procedures

1. Introduction

Real-time polymerase chain reaction (PCR) is one of the most sensitive, important and reliable quantitative technique for gene expression analysis. It is able to measure small amount of primary of a template sample specifically and sensitively. This technique could be considered as a proper substitute for other forms of PCR in which they determine the final quantification products [1,8,14]. All of the real-time methods are based on detection of a fluorescent signal. The increase in fluorescent signal is directly proportional to the increase in the amplified product during the PCR. The amplification curve in this method has three phases: exponential, linear and plateau (Figure 1). In the exponential phase, amplification

CONTACT Mohammadreza Nassiri nassiryr@um.ac.ir Department of Animal Science, Faculty of Agriculture, Ferdowsi University of Mashhad, P. O. Box 91775-1163, Mashhad, Iran; Shahrokh Ghovvati ghovvati@guilan.ac.ir Department of Biotechnology, Faculty of Agriculture, University of Guilan, P. O. Box 41635-1314, Rasht, Iran
*Both authors (Mohhammadreza Nassiri and Shahrokh Ghovvati) contributed equally to this study.

Supplemental data for this article can be accessed here. doi:10.1080/02664763.2016.1276890

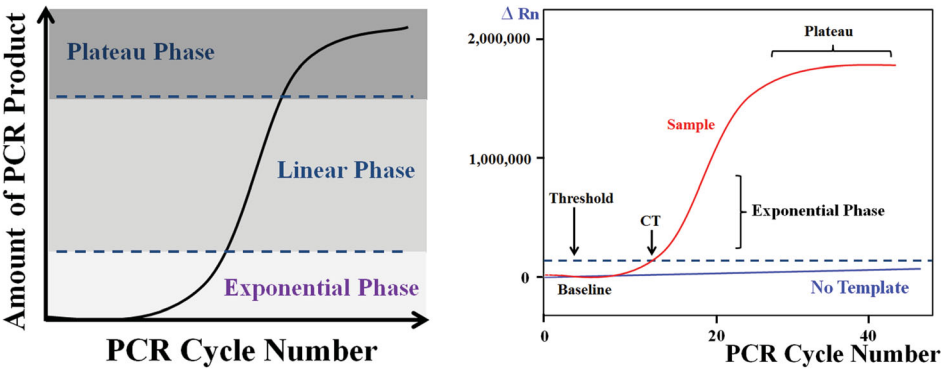

Figure 1. Amplification curve in real-time PCR.

increases exponentially. This means that in this phase of PCR, products will ideally become double during each cycle (if efficiency is around 100%). The following equation describes amplification of the exponential phase [12]:

$$X_n = X_0(1 + E)^n.$$

In which, X_0 and X_n are the amount of fluorescence signal after zero and n cycle, respectively, and E is the efficiency of reaction [5]. In linear phase, the amplification efficiency begins to taper off and in plateau phase, it decreases to minimum. The amplification curve was obtained by detecting fluorescence signal in each cycle [2].

In any given data analysis method, two parameters, C_t and PCR efficiency (E) must be determined. According to the so-called fit point method, the C_t is defined, as a fractional number of cycles, where the PCR kinetic curve reaches a defined threshold of fluorescence [10]. In other words, the C_t or cycle threshold value is the cycle number at which the fluorescence occurs within a reaction crosses the fluorescence threshold, a fluorescent signal significantly above the background fluorescence. Two different methods for analyzing real-time PCR data are the absolute and relative quantification. Absolute quantification determines the input number of the copy of the transcript of interest. This would be normally done by relating the PCR signal to a standard curve. Relative quantification describes the change in expression of the target gene relative to reference group such as an untreated control or a sample at time zero in a time-course study [5].

In relative quantification study, the following goals are (a) calculation of descriptive statistics of data, determination of slope and calculation of PCR efficiency, (b) using different mathematical functions for calculation of expression of a target gene relative to reference gene and (c) comparison of C_t numbers between employed treatment and control groups.

2. Materials and methods

The data used in this study were obtained from Yuan *et al.* [14] publication. Data includes two types of samples (treatment and control), two genes (reference and target) and four cDNA concentrations (10, 2, 0.4, 0.08 ng/μL). Reference gene is a gene whose expression

Table 1. Data of real-time assay.

rep	Sample	Gene	con	logcon	C_t	Class
1	Control	Target	10	1	23.1102	1
2	Control	Target	10	1	22.9003	1
3	Control	Target	10	1	22.8972	1
1	Control	Target	2	0.30103	26.5801	1
2	Control	Target	2	0.30103	26.2139	1
3	Control	Target	2	0.30103	26.0606	1
1	Control	Target	0.4	−0.39794	28.1125	1
2	Control	Target	0.4	−0.39794	28.1899	1
3	Control	Target	0.4	−0.39794	27.5949	1
1	Control	Target	0.08	−1.09691	30.2772	1
2	Control	Target	0.08	−1.09691	30.4667	1
3	Control	Target	0.08	−1.09691	30.7571	1
1	Treatment	Target	10	1	21.7813	2
2	Treatment	Target	10	1	21.7564	2
3	Treatment	Target	10	1	21.641	2
1	Treatment	Target	2	0.30103	23.7965	2
2	Treatment	Target	2	0.30103	23.7571	2
3	Treatment	Target	2	0.30103	23.7242	
1	Treatment	Target	0.4	−0.39794	26.3794	2
2	Treatment	Target	0.4	−0.39794	26.2542	2
3	Treatment	Target	0.4	−0.39794	25.9621	2
1	Treatment	Target	0.08	−1.09691	28.5479	2
2	Treatment	Target	0.08	−1.09691	28.3894	2
3	Treatment	Target	0.08	−1.09691	28.3416	2
1	Control	Reference	10	1	19.7415	3
2	Control	Reference	10	1	19.4943	
3	Control	Reference	10	1	19.3906	3
1	Control	Reference	2	0.30103	21.9838	3
2	Control	Reference	2	0.30103	22.4435	3
3	Control	Reference	2	0.30103	22.573	
1	Control	Reference	0.4	−0.39794	24.8109	3
2	Control	Reference	0.4	−0.39794	24.4327	3
3	Control	Reference	0.4	−0.39794	24.2342	3
1	Control	Reference	0.08	−1.09691	26.7319	3
2	Control	Reference	0.08	−1.09691	26.8206	3
3	Control	Reference	0.08	−1.09691	26.822	3
1	Treatment	Reference	10	1	18.4468	4
2	Treatment	Reference	10	1	18.8227	4
3	Treatment	Reference	10	1	18.3061	4
1	Treatment	Reference	2	0.30103	21.2568	4
2	Treatment	Reference	2	0.30103	21.0956	4
3	Treatment	Reference	2	0.30103	20.8473	4
1	Treatment	Reference	0.4	−0.39794	23.2322	4
2	Treatment	Reference	0.4	−0.39794	22.9577	4
3	Treatment	Reference	0.4	−0.39794	23.2415	4
1	Treatment	Reference	0.08	−1.09691	25.4817	4
2	Treatment	Reference	0.08	−1.09691	25.608	4
3	Treatment	Reference	0.08	−1.09691	25.5675	4

Note: rep, replicate; con, concentration; logcon, logarithm of concentration (base 10).

level is not different among samples, such as a housekeeping or maintenance gene. In this study, all combinations of gene and samples were used. In each concentration of primary target, there were four classes of sample gene combinations with three replicates. The data are shown in Table 1.

2.1. Calculation of descriptive statistics, slope and PCR efficiency

PCR efficiency is equivalent to suitability of the amplification. PCR efficiency of a reaction is an important criterion in determining the relative quantification and is defined as:

(1) percentage (from 0 to 1)
(2) time of increasing PCR product per cycle (from 1 to 2) [12].

Original methods of relative quantification were based on assuming an ideal amplification efficiency with a doubling of products in every cycle ($E = 2$). However, in most PCR reactions, the efficiency may not be full and it is essential to use a correction factor to avoid the overestimation. In other words, using $E = 2$ indicates that the efficiency has not been corrected.

Assay efficiency in a reaction with standard curve could be calculated using the following equation (1,2):

$$E = [10^{(-1/\text{slope})}]. \tag{1}$$

In this model, the slope of regression is calculated by fitting regression between observed C_t and logarithm transformed of concentration. A simple linear regression should be observed under the following equation:

$$C_t = \beta_0 + \beta_{con} X_{lcon} + \varepsilon,$$

where β_0 = intercept, β_{con} = coefficient of regression or slope of regression, X_{lcon} = logarithm transformed concentration with specific base, ε = residual.

The acceptable output of regression should have two features:

(a) If the base of logarithm transformation is 10 or 2, accordingly slope of regression should not be significantly different from either −3.32 or −1, respectively. It is possible to calculate different logarithm transformed of concentrations with the following command:

$$= \text{LOG (number; based)}.$$

(b) The slope of all of the genes and treatments should not be significantly different from each other [14]. The value of intercept indicates that whether the assay is perfect or not. In other words, intercept between 33 and 37 is equivalent to the slope of the regression of −3.32 and R^2 of 100%. Several studies have shown that as the slope of the line equation increases (become more negative), the efficiency decrease. By increasing the intercept, the sensitivity of the assay would decrease, too. For example, intercept value higher than 37 and lower than 33 indicates that the amount of template has not been correctly determined [2]. If logarithm-based two transformation of concentration should be used, then the amplification efficiency (E) must be calculated using Equation (2)

$$E = [2^{(-1/\text{slope})}]. \tag{2}$$

The best slope of regression with Equation (1) is −3.32, and with Equation (2) is −1 which both refers to the perfect efficiency. Using the regression parameters, the equation

will be obtained and the R^2 is accessible. R^2 is the coefficient of determination, a statistical term that explains how much variability of a factor can be explained by its relationship with another factor. In trend analysis, it is ranged between 0 (0%) and 1 (100%). The closer value to 1 indicates a better fit [11]. In many PCR methods, it is assumed that the efficiency is constant during exponential phase. This is possible if other factors such as primer specification, input amount of amplicon and the technical variation occurring during the reaction would be optimal. Studies have showed that the efficiency strongly influences the technically determined C_t value and small variation in C_t values have large effect on calculation of gene expression ratios [10].

2.2. Mathematical functions for calculation of target gene relative to the reference gene expression

Several mathematical models which determine the relative expression ratio have been developed. Two mathematical models that are widely being used are efficiency calibrated model and delta delta C_t model. The target in the first model is expression of target gene relative to reference gene(s) using the following models:

$$\text{Ratio} = \frac{(E_{\text{target}})^{\Delta C_{\text{ttarget}}}}{(E_{\text{reference}})^{\Delta C_{\text{treference}}}}, \quad \Delta C_{\text{ttarget}} = C_{\text{tcontrol}} - C_{\text{tsample}}, \tag{3}$$

$$\Delta C_{\text{treference}} = C_{\text{tcontrol}} - C_{\text{tsample}},$$

$$\text{Ratio} = \frac{(E_{\text{reference}})^{C_t \text{ sample}}}{(E_{\text{target}})^{C_t \text{ sample}}} \Big/ \frac{(E_{\text{reference}})^{C_t \text{ calibrator}}}{(E_{\text{target}})^{C_t \text{ calibrator}}}. \tag{4}$$

Delta C_t for each gene is calculated by subtracting the C_t number of target sample from that of the control sample. Then, the mean of these numbers for target and reference genes would be used in the model (3) or (4) for calculation of the ratio.

In delta delta C_t equation, real-time amplification efficiencies of target and reference gene for $E = 2$ are presumed as optimal and identical. This method is only applicable for a quick estimation of the relative expression ratio [5]. However, many PCRs do not have ideal amplification efficiencies, and calculation without an appropriate correction factor may overestimate the starting concentration. Under this condition $E_{\text{target}} = E_{\text{reference}} = 2$ and the ratios will be calculated by

$$\text{Ratio} = 2^{(\Delta C_{\text{treference}} - \Delta C_{\text{ttarget}})} \quad \text{or} \quad \text{ratio} = 2^{-\Delta\Delta C_t}. \tag{5}$$

Using this model, ΔC_t would be calculated as above (Equation (3)) and delta delta C_t would be obtained from the difference between means of ΔC_t of target and reference genes.

2.3. Comparison of C_t values between treatment and control group

Comparison between the mean of control group vs. treatment group is important in a real-time assay. To access the level of significance between any two groups of expression values, it is possible to perform different statistical tests like paired and unpaired *t*-test (parametric test), Wilcoxon signed-rank test (non-parametric test) and multiple regression [7].

Moreover, it is possible to calculate the correlation coefficient between the two groups. *t*-Test for evaluation of the difference between two groups is acceptable if the following assumptions are met: the C_t values and ΔC_t follow a normal distribution and they have equal variance [3,14], but these assumptions might not be valid in all of the real-time assays because of the small sample size. The minimum number of samples per group to find a statistical significant is optimum, but if the expected differential is high, three replicates can suffice for the test. The variability of the gene expression values between measurements from the same condition is an important factor, i.e. the lower this variability, the lower the number of required samples [3].

Non-parametric test does not have these assumptions so it is recommended that when normality is not proved, using a non-parametric test reduces the risk of misinterpretation of the results. However, using normal data, it is better to employ the parametric test to enhance and obtain reliable results. The Shapiro–Wilk test statistics W is a powerful test for normality for small to medium samples ($n < 2000$). Normality is rejected if W is sufficiently smaller than 1. W is similar to a correlation between the data and their normal scores. In a perfect normal population, there is a perfect correlation $W = 1$.

The objectives of this study were: (1) to calculate the descriptive statistics of data, (2) to calculate slope and regression equation of PCR efficiency, (3) to calculate ΔC_t and delta delta C_t ratio and finally (4) to evaluate the significant difference between control and treatment groups through *t*-test, Wilcoxon and Multiple regression analyses using different SAS programs [9].

3. Results and discussion

3.1. Calculation of descriptive statistics, slope and equation of regression

Calculation of different statistical items including regression analysis between C_t values and logarithm transformed concentration is possible in SAS through different procedures. For example, for descriptive statistics of data and regression analysis, Proc Univariate and Proc GLM or mixed are useful, respectively. If data is in the matrix format, then Proc IML would be used. Different descriptive statistical values would be calculated by the univariate procedure including mean, standard deviation, variance, standard error, the coefficient of determination, skewness and kurtosis. Output of the program 1 produces descriptive statistics of dataset.

In this program, the dependent variables are C_t values. For calculation of descriptive statistics of each class, command of (BY CLASS;) will be added after the PROC UNIVARIATE NORMAL. The output data of program 1 has been presented in Table 2.

Shapiro–Wilk index (W) shows whether the assumption of normality is met. Program 2 calculates regression analysis between logarithms transformed concentration and C_t values. The relationship between the log of target DNA and C_t values is linear [13].

In this program, the logarithm of base 10 was used for analysis of regression. Different equations for each class could be obtained. For this purpose, command of (BY CLASS;) can be added after PROC GLM [4]. The output of program 2 is as follows: also, the linear regression line between logarithm transformed concentration and C_t values is shown in Table 3.

Table 2. Summary descriptive statistics of the C_t values.

The UNIVARIATE procedure Variable: C_t Moments			
N	48	Sum weights	48
Mean	24.3298146	Sum observations	1167.8311
Std dev.	3.24566719	Variance	10.5343555
Skewness	0.02450174	Kurtosis	−0.6875194
Uncorrected SS	28,908.2288	Corrected SS	495.11471
Coeff.variation	13.3402874	Std error mean	0.46847171

Basic statistical measures			
Location		Variability	
Mean	24.32981	Std dev.	3.24567
Median	24.01535	Variance	10.53436
Mode		Range	12.45100
		Interquartile range	4.77345

Test statistic			*p*-Value	
Student's *t*	*t*	51.93444	Pr > |*t*|	< .0001
Sign	*M*	24	Pr ≥ |*M*|	< .0001
Signed rank	*S*	588	Pr ≥ |*S*|	< .0001

Test	Statistic		*p*-Value	
Shapiro–Wilk	*W*	0.978158	Pr < *W*	.5044
Kolmogorov–Smirnov	*D*	0.076168	Pr > *D*	> .1500
Cramer–von Mises	W-Sq	0.03961	Pr > W-Sq	> .2500
Anderson–Darling	A-Sq	0.256205	Pr > A-Sq	> .2500

Table 3. The results of regression analysis: GLM procedure output.

Dependent variable: C_t					
Source	DF	Sum of squares	Mean square	*F*-value	Pr > *F*
Model	1	330.5127546	330.5127546	92.37	< .0001
Error	46	164.6019550	3.5783034		
Corrected total	47	495.1147095			

R^2	Coeff. variation	Root MSE	$C_t$mean
0.667548	7.774989	1.891640	24.32981

Source	DF	Type I SS	Mean square	*F*-value	Pr > *F*
lcon	1	330.5127546	330.5127546	92.37	< .0001

Source	DF	Type III SS	Mean square	*F*-value	Pr > *F*
lcon	1	330.5127546	330.5127546	92.37	< .0001

Parameter	Estimate	Standard error	*t*-Value	*p*-Value
Intercept	24.16711044	0.27355912	88.34	< .0001
lcon	−3.35783961	0.34938513	−9.61	< .0001

This study results show that there is a significant relationship between the two variables. The equation and coefficient of determination for the whole data are as follows:

$$\hat{Ct} = 24.167 - 3.357X_{\text{lcon}} + \varepsilon \quad R^2 = 0.66.$$

3.2. Calculation of delta C_t (ΔC_t) and delta delta C_t ($\Delta\Delta C_t$)

Δc_t for each of the target and reference gene is calculated by subtracting the C_t number of the target gene from that of the reference gene. In other words, the difference between C_t

Table 4. ΔC_t calculation for target and reference genes.

Sample	gene	con	logcon	C_t	Sample	gene	con	logcon	C_t	ΔC_t
Control	Target	10	1.0000	23.1102	Control	Reference	10	1.0000	19.7415	3.3687
Control	Target	10	1.0000	22.9003	Control	Reference	10	1.0000	19.494	3.4063
Control	Target	10	1.0000	22.8972	Control	Reference	10	1.0000	19.3906	3.5066
Control	Target	2	0.3010	26.5801	Control	Reference	2	0.3010	21.9838	4.5963
Control	Target	2	0.3010	26.2139	Control	Reference	2	0.3010	22.4435	3.7704
Control	Target	2	0.3010	26.0606	Control	Reference	2	0.3010	22.57	3.4906
Control	Target	0.4	−0.3979	28.1125	Control	Reference	0.4	−0.3979	24.8109	3.3016
Control	Target	0.4	−0.3979	28.1899	Control	Reference	0.4	−0.3979	24.4327	3.7572
Control	Target	0.4	−0.3979	27.5949	Control	Reference	0.4	−0.3979	24.2342	3.3607
Control	Target	0.08	−1.0969	30.2772	Control	Reference	0.08	−1.0969	26.7319	3.5453
Control	Target	0.08	−1.0969	30.4667	Control	Reference	0.08	−1.0969	26.8206	3.6461
Control	Target	0.08	−1.0969	30.7571	Control	Reference	0.08	−1.0969	26.822	3.9351
									Mean = 3.6404	
Treatment	Target	10	1.0000	21.7813	Treatment	Reference	10	1.0000	18.4468	3.3345
Treatment	Target	10	1.0000	21.7564	Treatment	Reference	10	1.0000	18.8227	2.9337
Treatment	Target	10	1.0000	21.641	Treatment	Reference	10	1.0000	18.3061	3.3349
Treatment	Target	2	0.3010	23.7965	Treatment	Reference	2	0.3010	21.2568	2.5397
Treatment	Target	2	0.3010	23.7571	Treatment	Reference	2	0.3010	21.0956	2.6615
Treatment	Target	2	0.3010	23.724	Treatment	Reference	2	0.3010	20.8473	2.8767
Treatment	Target	0.4	−0.3979	26.3794	Treatment	Reference	0.4	−0.3979	23.2322	3.1472
Treatment	Target	0.4	−0.3979	26.2542	Treatment	Reference	0.4	−0.3979	22.9577	3.2965
Treatment	Target	0.4	−0.3979	25.9621	Treatment	Reference	0.4	−0.3979	23.2415	2.7206
Treatment	Target	0.08	−1.0969	28.5479	Treatment	Reference	0.08	−1.0969	25.4817	3.0662
Treatment	Target	0.08	−1.0969	28.3894	Treatment	Reference	0.08	−1.0969	25.608	2.7814
Treatment	Target	0.08	−1.0969	28.3416	Treatment	Reference	0.08	−1.0969	25.5675	2.7741
									Mean = 2.9556	

Table 5. The summary output of program 3.

Slope	R^2	ETARGET
Calculation of numerator in ratio method		
25.816562	0.9985328	0.9146716
−3.34882		1.9888987
Calculation of denumerator in ratio method		
22.517687	0.9991985	0.9027977
−3.366963		1.9815434

values of class 1 and 4 is ΔC_t of the target gene and the difference between C_t values of class 2 and 3 would be delta delta C_t of reference gene (Table 4). By taking the average of these two groups, delta delta C_t for target and reference genes would be produced. Therefore, the following values for the above terms are calculated: $\Delta C_{t\text{reference}} = 2.9556$, $\Delta C_{t\text{target}} = 3.6404$ and $\Delta\Delta C_t = 0.6848$. Using these values and by calculation of (E), the model (3) calculated of the ratios.

3.3. Calculation of ratio by models (3) and (5)

By using the following commands (Program 3) one can calculate the efficiency of target and reference gene, respectively. Therefore, first the regression line for each gene must be derived, and then by using model (3) the ratio could be calculated. This program was written using Proc IML in SAS software and this program output has been placed in Table 5.

Table 6. The summary output of program 4.

The *t*-test procedure statistics									
Variable	Treatment	*N*	Lower CL mean	Mean	Lower CL mean	Lower CL std dev.	Std dev.	Lower CL std dev.	Std error
deltaC_t	Control	12	3.4136	3.6404	3.8673	0.2529	0.357	0.6062	0.1031
deltaC_t	Treatment	12	2.7803	2.9556	3.1309	0.1954	0.2759	0.4684	0.0796
deltaC_t	Diff (1-2)		0.4147	0.6848	0.955	0.2468	0.3191	0.4516	0.1303

t-Tests					
Variable	Method	Variances	DF	*t*-Value	Pr > |*t*|
deltaC_t	Pooled	Equal	22	5.26	< .0001
deltaC_t	Satterthwaite	Unequal	20.7	5.26	< .0001
deltaC_t	Cochran	Unequal	11	5.26	.0003

Equality of variances					
Variable	Method	Num DF	Den DF	*F*-value	Pr > *F*
deltaC_t	Folded *F*	11	11	1.67	0.4057

The results show that the slope of regression in target gene is −3.3488 and efficiency of target gene is 1.9888 ($E_{\text{target}} = 1.9888$), too. The slope of regression in reference gene and its efficiency are −3.3669 and 1.9815, respectively ($E_{\text{reference}} = 1.9815$).

The ratio by using model (3) would be

$$\text{Ratio} = \frac{(1.9898)^{3.6404}}{(1.9815)^{2.9556}} = 1.621.$$

Calculation of ratio by delta delta C_t is also possible with model (5). In this model efficiency is equal for both target and reference genes ($E = 2$) [6].

3.4. Comparison of control and treatment groups with t-test and Wilcoxon test

There is a basic question in PCR assays as if there is any significant difference between target and reference genes in control sample vs. treatment sample. The difference of expression between these two samples is equivalent to comparison of two ΔC_t groups. The *t*-test or Wilcoxon tests are appropriate methods for finding the significance of this difference. If the goal is the comparison of several target genes with a reference gene, the best way for finding the significant difference is using one-way analysis of variance instead of *t*-test or Wilcoxon test. Program 4 shows *t*-test commands for the statistical comparisons.

The CLASS statement contains the variable that distinguishes between the groups being compared, i.e. in this case the treatment. Cochran requests the Cochran and Cox (1950) approximation of the probability level of the approximate *t* statistic for the situation where variances are unequal. The results of program 4 have been placed in Table 6.

For each class, the sample size, mean, standard deviation, standard error, maximum and minimum values are displayed. A group test statistic for the equality of means is reported for equal and unequal variances. The difference between means of two groups is shown as delta C_t difference (1–2) which is in fact the delta delta C_t. A group test statistic for the equality of means is reported for both equal and unequal variances. Before deciding which test is appropriate, one should look at the test for equality of variances, this test does not indicate a significant difference in the two variances ($F = 1.67, p = .4057$). Therefore,

Table 7. The results of Wilcoxon test: non-parametric procedure output.

The NPAR1WAY procedure Wilcoxon scores (rank sums) for variable DELTACT Classified by variable TREATMENT					
TREATMENT	*N*	Sum of scores	Expected under H0	Std dev.under H0	Mean score
Control	12	220.0	150.0	17.320508	18.333333
Treatment	12	80.0	150.0	17.320508	6.666667
Wilcoxon two-sample test					
Statistic				220.0000	
Normal approximation					
Z				4.0126	
One-sided Pr $>$ Z				$<$ .0001	
Two-sided Pr $>$ $\|Z\|$				$<$ .0001	
t-Approximation					
One-sided Pr $>$ Z				0.0003	
Two-sided Pr $>$ $\|Z\|$				0.0005	
Z includes a continuity correction of 0.5 Kurskul–Wallis test					
χ^2				16.3333	
DF				1	
Pr $>$ χ^2				$<$ .0001	

the pooled t statistic should be used. If the difference of variances of the two groups is significant, the appropriate way is using Cochran and Satterthwait method for expression of final results and p-value [4]. Although sometimes, distribution of data is not clear, the assumption of normality is important in most cases. If normality is not met, then in order to estimate the parameters, non-parametric methods could be used for data analysis [11]. One of the famous non-parametric tests is Wilcoxon test and program 5 compares two groups of data with this procedure (Table 7).

Again the result shows that there is a significant difference between control and treatment groups ($Z = 4.0126$, $p < .001$). As mentioned above, for normal data, non-parametric tests are not appropriate.

3.5. Comparison of control and treatment groups with multiple regression

Factors affecting C_t values can be analyzed by a multiple regression model in delta delta C_t method. These factors could be concentration, treatment, gene and their interaction. The following multiple regression model can be considered:

$$C_t = \beta_0 + \beta_{con}X_{icon} + \beta_{treat}X_{itreat} + \beta_{gene}X_{igene} + \beta_{contreat}X_{icon}X_{itrea} + \beta_{congene}X_{icon}X_{igene} + \beta_{genetreat}X_{igene}X_{itreat} + \beta_{congenetreat}X_{icon}X_{itreat}X_{igene} + \varepsilon.$$

In this model C_t is the dependent variable, β_0 and β_x are regression coefficients for the corresponding X terms, and ϵ is residual. In this analysis, the goal is to find a significant difference between target and reference genes in treatment vs. control sample. There is also, an interaction between gene and treatment, which addresses the degree of C_t differences between the target gene and reference gene in treatment vs. control or delta delta C_t. the significant effect of interaction shows that there is a relationship between treatment and gene. Program 6 shows the commands for multiple regression.

Table 8. The summary output of program 6.

The GLM procedure

Class	Level information			
Class	Levels	Values		
Treat	2	Control treatment		
Gene	2	Reference target		
Con	4	0.08	0.4	2	10

Number of observations read	48
Number of observations used	48

Estimation of treatment and gene name effects

The GLM procedure

Dependent variable: C_t

Source	DF	Sum of squares	Mean square	*F*-value	Pr > *F*
Model	15	493.76694		781.57	< .0001
Error	32	1.3477663	32.917796		
Corrected total	47	495.11471	0.0421177		

R^2	Coeff. variation	Root MSE	$C_t$mean
0.997278	0.843516	0.205226	24.32981

Source	DF	Type I SS	Mean square	*F*-value	Pr > *F*
Con	3	331.23	110.41	2621.46	< .0001
Treat	1	29.20554	29.20554	693.43	< .0001
Treat*Con	3	0.92112	0.30704	7.29	< .0007
Gene	1	130.5213	130.5213	3098.97	< .0001
Gene*Con	3	0.006192	0.002064	0.05	.9854
Treat*Gene	1	1.406956	1.406956	33.41	< .0001
Treat*Gene*Con	3	0.475859	0.15862	3.77	.0201

Source	DF	Type I SS	Mean square	*F*-value	Pr > *F*
Con	3	331.23	110.41	2621.46	< .0001
Treat	1	29.20554	29.20554	693.43	< .0001
Treat*Con	3	0.92112	0.30704	7.29	< .0007
Gene	1	130.5213	130.5213	3098.97	< .0001
Gene*Con	3	0.006192	0.002064	0.05	.9854
Treat*Gene	1	1.406956	1.406956	33.41	< .0001
Treat*Gene*Con	3	0.475859	0.15862	3.77	.0201

Estimation of treatment and gene effects
The GLM procedure
Least squares means
Adjustment for multiple comparisons: Tukey

Treat	Treatment	C_t LSMean	Std error	Pr > |*t*|
Control	Reference	23.2896417	0.0592436	< .0001
Control	Target	26.9300500	0.0592436	< .0001
Treatment	Reference	22.0719917	0.0592436	< .0001
Treatment	Target	25.0275750	0.0592436	< .0001

Least squares means for effect Treat*Gene
Pr > |*t*| for H0: LSMean(*i*) = LSMean(*j*)
Dependent variable: C_t

	1	2	3	4
1		< .0001	< .0001	< .0001
2	< .0001		< .0001	< .0001
3	< .0001	< .0001		< .0001
4	< .0001	< .0001	< .0001	

(*continued*).

Table 8. Continued.

Estimation of treatment and gene effects The GLM procedure				
			C_t	
Level of treat	Level of gene	*N*	Mean	Std
Control	Reference	12	23.2896417	2.80332850
Control	Target	12	26.9300500	2.86468070
Treatment	Reference	12	22.0719917	2.71102053
Treatment	Target	12	25.0275750	2.63575815

Estimation of treatment and gene effects The GLM procedure Dependent variable: C_t					
Contrast	DF	Contrast SS	*F*-value	Mean square	Pr > *F*
Treatment *gene	1	1.40695584	1.40695584	33.41	< .0001
Parameter		Estimate	Std error	*t*-Value	Pr > |*t*|
Treatment and gene effect		−0.68482500	0.11848726	−5.78	< .0001

In this program, there are four columns of data which are related to sample, gene, concentration and C_t values, respectively. Variables of treatment, gene and concentration are defined as class in the program. By using (|) sign, SAS can calculate all of the main, double and triple effects. Least-square means (LSM) for interactions with their standard error would also be calculated. If the interaction between gene and treatment is significant, the comparison of means could be done by Tukey test. The null hypothesis is that the C_t difference between target and reference genes in treatment vs. control is the same and has no significant difference. With the coefficient of orthogonal contrast, differences between values of delta delta C_t can be calculated. If the null hypothesis is accepted, then the Δc_t of reference and target genes are equal ($\Delta C_{treference} = \Delta C_{ttarget}$) and it can be concluded that $\Delta\Delta C_t = 0$. However, if the null hypothesis is rejected, the $\Delta\Delta C_t$ is not equal to zero and its value should be calculated through orthogonal contrast. Finally, if treat × gene interaction effect is significant, then null hypothesis would be rejected and $\Delta\Delta C_t$ could be calculated by multiple regression (Table 8).

As it is shown in the output, the regression model is significant with $R^2 = 0.9972$ which indicates the high precision of the model. Effect of gene × treatment interaction was also significant ($p < .001$). Therefore, the behavior of genes in two samples is different. LSM with their standard errors (LSM ± SE) for four groups of gene × sample interactions are shown in Figure 2.

The results show that there are significant differences among these four groups (SE = 0.059). The maximum and minimum expressions were related to CT and TR groups, respectively. The target gene shows the most expression in both control and treatment groups. Overall least-square means of control and treatment groups were 25.109 and 23.549, respectively, and the difference was significant. Considering the same replicate in each group, an increase in the mean of the control group could be related to the more expression of target gene in this group. Program 6 calculates the sum of square and estimates difference of target gene and reference gene in control vs. treatment groups which is exactly $\Delta\Delta C_t$. Therefore, the multiple regression could be a useful method for calculation of $\Delta\Delta C_t$ and statistical evaluation of factors affecting C_t values (Figure 3).

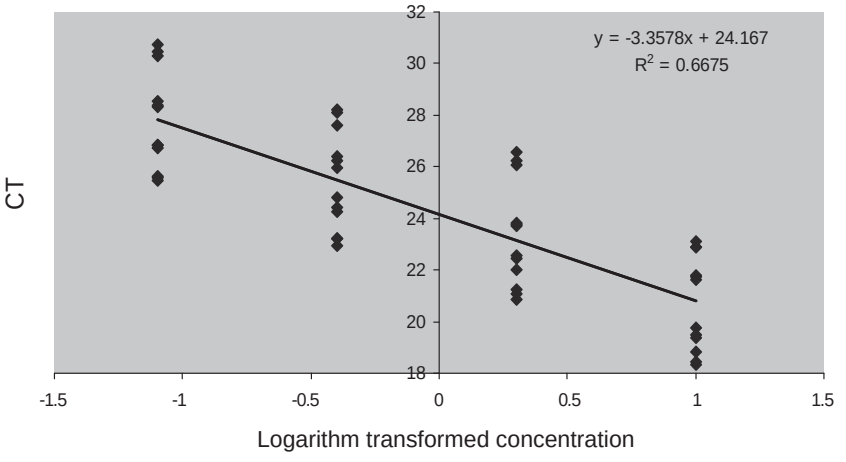

Figure 2. Linear regression line between logarithm transformed concentration and C_t values.

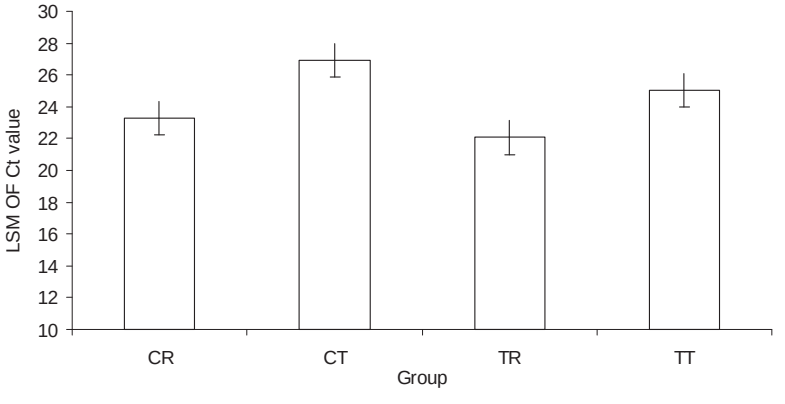

Figure 3. Least-square means for four combinations of gene × treatment, i.e. control-reference (CR), control-target (CT), treatment-reference (TR) and treatment-target (TT).

Acknowledgements

The authors would like to specially thank Excellent Center in Animal Science of Ferdowsi University of Mashhad for kindly providing necessary facilities and equipment.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Agricultural Faculty of Ferdowsi University of Mashhad, Iran [grant number 368P].

ORCID

Mohammadreza Nassiri http://orcid.org/0000-0001-7119-8155

Mahdi Elahi Torshizi http://orcid.org/0000-0002-1574-3865
Shahrokh Ghovvati http://orcid.org/0000-0002-2016-2184
Mohammad Doosti http://orcid.org/0000-0002-2728-8265

References

[1] R. Biassoni and A. Raso, *Quantitative Real-time PCR: Methods and Protocols*, Humana Press, New York, NY, 2014.
[2] M.T. Dorak, *Real-time PCR*, Taylor & Francis Group, New York, NY, 2006.
[3] R. Goni, P. Garcia, and S. Foissac, *The QPCR Data Statistical Analysis*, Integromics White Paper, 2009, pp. 1–9.
[4] M. Kaps and W.R. Lamberson, *Biostatistics for Animal Science: An Introductory Text*, 2nd ed., CABI Publishing, Wallingford, OX, 2009.
[5] J.K. Livak and D.T. Schmittgen, *Analysis of relative gene expression data using real-time quantitative PCR and the* $2^{-\Delta\Delta ct}$ *methods*, Methods 25 (2001), pp. 402–408.
[6] M.W. Pfaffl, *A new mathematical model for relative quantification in real-time RT-PCR*, Nucleic Acids Res. 29 (2001), pp. e45.
[7] M.W. Pfaffl, *Chapter 3: Quantification Strategies in Real-time PCR*, 1st ed., International University Line Publishing, La Jolla, CA, 2004.
[8] D. Rodriguez-Lazaro, *Real-time PCR in Food Science: Current Technology and Applications*, Caister Academic Press, Norfolk, 2013.
[9] SAS/STAT 9.3 User's Guide, SAS Inst. Inc., Cary, NC, 2011.
[10] J.H. Schefe, K.E. Lehmann, I.R. Buschmann, T. Unger, and H. Funke-Kaiser, *Quantitative real-time RT-PCR data analysis: Current concepts and the novel 'gene expression's CT differences' formula*, J. Mol. Med. 84 (2006), pp. 901–910.
[11] G.W. Snedecor and W.G. Cochran, *Statistical Methods*, 8th ed., Wiley & Sons Inc. Publishing, Ames, IA, 1989.
[12] N.P. Stuart, G.N. Butler, and G. Foster, *Experimental validation of novel and conventional approaches to quantitative real-time data analysis*, Nucleic Acids Res. 31 (2003), pp. 1–7.
[13] P. Vaerman, P. Saussoy, and I. Ingargiola, *Evaluation of real-time PCR data*, J. Biol. Regul. Homeost. Agents 18 (2004), pp. 212–214.
[14] J.S. Yuan, A. Reed, F. Chen, and N. Stewart, *Statistical analysis of real-time PCR data*, BMC Bioinform. 7 (2006), pp. 1–12.