Real-time PCR Statistics

PCR Encyclopedia (2005) 1: 101127-49
Joshua S. Yuan and C. Neal Stewart Jr.
Department of Plant Sciences and Genomics Hub, University of Tennessee, Knoxville, TN 37996, USA

Introduction

Real-time PCR is a powerful tool to compare relative transcriptional abundance, and has been broadly applied in biomedical sciences. However, methods for rigorous statistical analyses have lagged behind applications. Confidence interval and statistical significance considerations are not explicit in many of the current data analysis approaches. Based on the standard curve method and other useful data analysis methods, we propose four statistical approaches that can be useful in the analysis of real-time PCR data.

A PCR reaction has three phases: exponential, linear, and plateau, and real-time PCR analyzes the relative abundance of PCR products during the exponential phase, in which reagents are not limited (Gibson et al., 1996; Heid et al., 1996). During the exponential phase, PCR product will ideally double during each cycle if the amplification efficiency is perfect. It is possible to make the PCR amplification efficiency close to 100% in the exponential phases of PCR reactions, if the PCR conditions, primer characteristics, template purity, and amplicon lengths are optimal. Therein lies the ability to compare initial abundance of template, e.g., to compare transcript abundance between two different samples, since the PCR product quantity in the exponential phase correlates with the initial template abundance. The relative quantification can be achieved with analyzing so-called ‘Ct value’. In the real-time PCR data processing, a baseline and a threshold can be set for further analysis. The cycle number at the threshold level of log-based fluorescence is defined as Ct number, which is the observed value in most real-time PCR experiments, and therefore the primary statistical metric of interest.

Since relative quantification is the goal for most for real-time PCR experiments, several data analysis procedures have been developed. Two mathematical models are very widely applied: the efficiency Ct model (Livak and Schmittgen, 2001) and the Schmittgen, 2001). The experimental systems for both models are similar: including both control and treatment samples. For each sample, a target gene and a reference gene for internal control are included for PCR amplification from serially diluted aliquots. Typically, several replicates are used for each diluted concentration. The Ct model. Ct number efficiency-calibrated model is a more generalized is first plotted against cDNA input (or logarithm cDNA input), and the slope of the plot is calculated to determine the amplification Ct for each gene (target or reference) is then calculated efficiency. by subtracting the Ct number of target sample from that of control sample.

Ct models are Even though both the efficiency-calibrated and widely applied in gene expression studies, statistical considerations are not widely discussed for the analysis of the effect of each experimental factor as well as significance testing. One of the few studies that employed substantial statistical analysis used the REST® program (Pfaffl et al., 2002). The software presented in this article is based on the efficiency-calibrated model and employed randomization tests to obtain the significance level. However, the article did not provide a detailed model for the effects of different experimental factors involved. Another statistical study of real-time PCR data used a simple linear regression model to estimate the ratio through Ct calculation (Cook et al., 2004). However, the logarithm-based fluorescence was used as the dependent variable in the model, which we believe does not adequately reflect the nature of real-time PCR data. It follows that Ct should be the dependent variable for statistical analysis, because it is the outcome value directly influenced by treatment, concentration and sample effects. Both studies used the efficiency-calibrated models. Despite the publication of these two methods, many research articles published with real-time PCR data actually do not present P values and confidence intervals (Eleumi et al., 2004; Zenomì et al., 2004; Shen et al., 2005). We believe statistical analysis is desirable to facilitate robust interpretation of real-time PCR data.

Ct A priori, we consider the confidence interval and P value of data to be very important because these directly influence the interpretation of ratio. Without proper statistical modeling and analysis, the interpretation of real-time PCR data may lead the researcher to false positive conclusions, which is especially potentially troublesome in clinical applications. We hereby propose four statistical methodologies for processing real-time PCR data using Ct method with standard curves. The statistical a modified methodologies can be adapted to other mathematical models with modifications.

Data Quality and Experimental Design

Data quality standards are needed. Pfaffl (2001), who described the efficiency-calibrated method, assumed that the amplification efficiency for each gene (target and reference) is the same among different experimental samples (treatment and control). In contrast, the \( \Delta \Delta CT \) method is more stringent by assuming that all reactions should reach a 100% amplification efficiency. In other words, the amount of product should double during each cycle (Livak and Schmittgen, 2001). Moreover, the \( \Delta \Delta CT \) method assumes that the PCR amplification efficiency for each sample will be 100%, if PCR reactions for one set of the samples reaches 100% amplification efficiency. However, this assumption neglects the effects of different cDNA samples. We therefore recommend performing a standard curve experiment for each gene in each sample. The amplification efficiency can be estimated from the standard curve and the relative gene expression abundance can be calculated by \( \Delta \Delta CT \). In any standard curve experiment, PCR reactions of three replicates...
at three concentrations are recommended to derive enough statistical power. A minimal of two replicates at three concentrations are required. The concentration series helps to define the dynamic range of the experiment, and the replicates help to ensure the reproducibility of the data. The experimental design combining the standard curve and ∆∆Ct schema as well as the statistical analysis discussed later will ensure a robust system of highly reproducible real-time PCR.

**Simple Linear Regression Model for Data Quality Control**

Data quality could be examined through a regression model. Even though examining the association between Ct number and concentration can provide an effective quality control, a better approach might be to examine the correlation between Ct and the logarithm (base 2) transformed concentration of template, which should yield a significant simple linear relationship for each standard curve. For example, for a target gene in the control sample, the Ct number should correlate with the logarithm transformed concentration following the simple linear regression model: $\text{Ct} = \beta_0 + \beta_{G} X_{G} + \epsilon$. The acceptable real-time PCR data should have two features from the regression analysis. First, the slope should not be significantly different from -1. Second, the slopes for all four combinations of genes and samples should not be significantly different from one another. Many statistical software packages including SAS and SPSS can be used to test whether the slope is significantly different from -1.

**Multiple Regression Model for Data Analysis**

Several effects need to be taken into consideration in the ∆∆Ct method, namely, the effect of treatment, gene, concentration, and replicates. If we consider these effects as quantitative variables and have the Ct number relating to these multiple effects and their interactions, we can develop a multiple regression model as follows:

$$\text{Ct} = \beta_0 + \beta_{G} X_{G} + \beta_{C} X_{C} + \beta_{T} X_{T} + \beta_{G} \beta_{C} X_{G} X_{C} + \beta_{G} \beta_{T} X_{G} X_{T} + \beta_{C} \beta_{T} X_{C} X_{T} + \beta_{G} \beta_{C} \beta_{T} X_{G} X_{C} X_{T} + \epsilon$$

In this model, we are principally interested in the interaction between gene and treatment, which addresses the degree of the Ct differences between target gene and reference gene in treated vs. control samples; i.e., ∆∆Ct. ∆∆Ct can therefore be estimated from the different combinations values of βgenetreat. The goal here is to statistically test for differences between target and reference genes in treatment vs. control samples. Therefore, the null hypothesis is the Ct differences between target and reference genes will be the same in treatment vs control samples, which can be represented by combinatorial effect (CE) as: $\text{CET} = \text{CETargetGene_Treatmentsample-CERefGene_Targetsample} = \text{CETargetGene_Controlsample-CERefGene_Controlsample}$. An alternative formula will be: $\text{CETargetGene_Treatmentsample} - \text{CERefGene_Treatmentsample} - \text{CERefGene_Controlsample} + \text{CETargetGene_Controlsample} = 0$, which will yield an estimation of $\Delta \Delta \text{Ct}$. If the null hypothesis is not rejected, then the $\Delta \Delta \text{Ct}$ would not be significantly different from 0, otherwise, the $\Delta \Delta \text{Ct}$ can be derived from the estimation of the test. In this way, we can perform a test of different combinational effects of βgenetreat and estimate the $\Delta \Delta \text{Ct}$ from it. If $\Delta \Delta \text{Ct}$ is equal to 0, the ratio will be 1, which indicates no change in gene expression between control and treatment. The P value in the test therefore indicates how much confidence we have for the changes in gene expression.

**Analysis of Covariance for Data Analysis**

Another way to approach the real-time PCR data analysis is by using an analysis of covariance (ANCOVA), having the following model:

$$\text{Ct} = \beta_0 + \beta_{\text{group}} X_{\text{group}} + \beta_{\text{treat}} X_{\text{treat}} + \epsilon.$$  

We are interested in two questions here. First, are the covariance adjusted averages among the four groups equal? Second, what is the Ct difference of target gene value between treatment and control sample after corrected by reference gene? In this case, the null hypothesis will be $(\mu_{\text{targetGene\_Treatmentsample}} - \mu_{\text{RefGene\_Treatmentsample}}) - (\mu_{\text{targetGene\_Controlsample}} - \mu_{\text{RefGene\_Controlsample}}) = 0$, and the test will yield a parameter estimation of $\Delta \Delta \text{Ct}$.

**Simplified Alternatives – T-test and Wilcoxon Two Group Test for Data Analysis**

Many applications of real-time PCR do not require complex models. Therefore, more simplified alternatives can be used to analyze real-time data with biological replicates for each experiment. The primary assumption with this approach is that the additive effect of concentration, gene, and replicate can be adjusted by subtracting Ct number of target gene from that of reference gene, which will provide ∆Ct. The ∆Ct for treatment and control can therefore be subject to simple t-test, which will yield the estimation of $\Delta \Delta \text{Ct}$.

As a non-parametric alternative to the t-test, a Wilcoxon two group test can also be used to analyze the two pools of ∆Ct values. Two of the assumptions for t-test are that the both groups of ∆Ct will have Gaussian distributions and they will have equal variances. However, these assumptions are not valid in many real-time PCR experiments using realistically small sample sizes. Therefore a distribution-free Wilcoxon test will be a more robust and appropriate alternative in this case (Hollander and Wolfe 1973)

---

**Multiple Regression Model for Data Analysis**

Several effects need to be taken into consideration in the $\Delta \Delta \text{Ct}$ method, namely, the effect of treatment, gene, concentration, and replicates. If we consider these effects as quantitative variables and have the Ct number relating to these multiple effects and their interactions, we can develop a multiple regression model as follows:

$$\text{Ct} = \beta_0 + \beta_{\text{G}} X_{\text{G}} + \beta_{\text{T}} X_{\text{T}} + \beta_{\text{G}} \beta_{\text{T}} X_{\text{G}} X_{\text{T}} + \beta_{\text{G}} \beta_{\text{C}} X_{\text{G}} X_{\text{C}} + \beta_{\text{T}} \beta_{\text{C}} X_{\text{T}} X_{\text{C}} + \beta_{\text{G}} \beta_{\text{T}} \beta_{\text{C}} X_{\text{G}} X_{\text{T}} X_{\text{C}} + \epsilon$$

In this model, we are principally interested in the interaction between gene and treatment, which addresses the degree of the Ct differences between target gene and reference gene in treated vs. control samples; i.e., $\Delta \Delta \text{Ct}$. $\Delta \Delta \text{Ct}$ can therefore be estimated from the different combinations values of $\beta_{\text{genetreat}}$. The goal here is to statistically test for differences between target and reference genes in treatment vs. control samples. Therefore, the null hypothesis is the Ct differences between target and reference genes will be the same in treatment vs control samples, which can be represented by combinatorial effect (CE) as: $\text{CE}_{\text{TargetGene\_Treatmentsample}} - \text{CE}_{\text{RefGene\_Treatmentsample}} = \text{CE}_{\text{TargetGene\_Controlsample}} - \text{CE}_{\text{RefGene\_Controlsample}}$. An alternative formula will be: $\text{CE}_{\text{TargetGene\_Treatmentsample}} - \text{CE}_{\text{RefGene\_Treatmentsample}} + \text{CE}_{\text{RefGene\_Controlsample}} - \text{CE}_{\text{TargetGene\_Controlsample}} = 0$, which will yield an estimation of $\Delta \Delta \text{Ct}$. If the null hypothesis is not rejected, then the $\Delta \Delta \text{Ct}$ would not be significantly different from 0, otherwise, the $\Delta \Delta \text{Ct}$ can be derived from the estimation of the test. In this way, we can perform a test of different combinational effects of $\beta_{\text{genetreat}}$ and estimate the $\Delta \Delta \text{Ct}$ from it. If $\Delta \Delta \text{Ct}$ is equal to 0, the ratio will be 1, which indicates no change in gene expression between control and treatment. The P value in the test therefore indicates how much confidence we have for the changes in gene expression.

---

**Analysis of Covariance for Data Analysis**

Another way to approach the real-time PCR data analysis is by using an analysis of covariance (ANCOVA), having the following model:

$$\text{Ct} = \beta_0 + \beta_{\text{group}} X_{\text{group}} + \beta_{\text{treat}} X_{\text{treat}} + \epsilon.$$  

We are interested in two questions here. First, are the covariance adjusted averages among the four groups equal? Second, what is the Ct difference of target gene value between treatment and control sample after corrected by reference gene? In this case, the null hypothesis will be $(\mu_{\text{TargetGene\_Treatmentsample}} - \mu_{\text{RefGene\_Treatmentsample}}) - (\mu_{\text{TargetGene\_Controlsample}} - \mu_{\text{RefGene\_Controlsample}}) = 0$, and the test will yield a parameter estimation of $\Delta \Delta \text{Ct}$.

---

**Simplified Alternatives – T-test and Wilcoxon Two Group Test for Data Analysis**

Many applications of real-time PCR do not require complex models. Therefore, more simplified alternatives can be used to analyze real-time data with biological replicates for each experiment. The primary assumption with this approach is that the additive effect of concentration, gene, and replicate can be adjusted by subtracting Ct number of target gene from that of reference gene, which will provide $\Delta \text{Ct}$. The $\Delta \text{Ct}$ for treatment and control can therefore be subject to simple t-test, which will yield the estimation of $\Delta \Delta \text{Ct}$.

As a non-parametric alternative to the t-test, a Wilcoxon two group test can also be used to analyze the two pools of $\Delta \text{Ct}$ values. Two of the assumptions for t-test are that the both groups of $\Delta \text{Ct}$ will have Gaussian distributions and they will have equal variances. However, these assumptions are not valid in many real-time PCR experiments using realistically small sample sizes. Therefore a distribution-free Wilcoxon test will be a more robust and appropriate alternative in this case (Hollander and Wolfe 1973)
Synopsis

Choosing the appropriate model for statistical analyses is dependent on experimental design and scientific questions. However, performing comparative statistical analyses and presenting P values should be the norm in real-time PCR analyses for the future. For most simple experiments we prefer the non-parametric Wilcoxon test that can be performed in various statistical packages. This option will be provide a conservative test for significant differences between two sample types and answer the question if apparent differences in gene expression are statistically significant.

References


Pfaffl MW, Horgan GW, Dempfle L: Relative expression software tool (REST(C)) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. Nucl. Acids Res. 2002, 30: e36
