Genomics 94 (2009) 146–152

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Methods

A powerful and flexible linear mixed model framework for the analysis of relative quantification RT-PCR data

Juan Pedro Steibel [a,b,*], Rosangela Poletto [c], Paul M. Coussens [a], Guilherme J.M. Rosa [d]

[a] Department of Animal Science, Michigan State University, East Lansing, MI 48824, USA
[b] Department of Fisheries and Wildlife, Michigan State University, East Lansing, MI 48824, USA
[c] Department of Animal Sciences, Purdue University, West Lafayette, IN 47907, USA
[d] Department of Dairy Science, University of Wisconsin – Madison, Madison, WI 53706 USA

ARTICLE INFO

Article history:
Received 24 January 2009
Accepted 28 April 2009
Available online 5 May 2009

Keywords:
Gene expression
qRT-PCR
Linear mixed models
Hypothesis testing

ABSTRACT

Quantitative reverse transcription polymerase chain reaction (qRT-PCR) is currently viewed as the most precise technique to quantify levels of messenger RNA. Relative quantification compares the expression of a target gene under two or more experimental conditions normalized to the measured expression of a control gene. The statistical methods and software currently available for the analysis of relative quantification of RT-PCR data lack the flexibility and statistical properties to produce valid inferences in a wide range of experimental situations. In this paper we present a novel method for the analysis of relative quantification of qRT-PCR data, which consists of the analysis of cycles to threshold values (C_T) for a target and a control gene using a general linear mixed model methodology. Our method allows testing of a broader class of hypotheses than traditional analyses such as the classical comparative C_T. Moreover, a simulation study using plasmode datasets indicated that the estimated fold-change in pairwise comparisons was the same using either linear mixed models or a comparative C_T method, but the linear mixed model approach was more powerful. In summary, the method presented in this paper is more accurate, powerful and flexible than the traditional methods for analysis of qRT-PCR data. This new method is especially useful for studies involving multiple experimental factors and complex designs.

Introduction

Reverse transcription (RT), followed by quantitative polymerase chain reaction (qPCR), is currently the method of choice to quantify levels of messenger (m)RNA [1]. At present, there are several instrumentations and chemistries available for implementation of this technique, all of which rely on the same fundamental principle [2]. This principle consists of the specific amplification of cDNA from a target transcript in several cycles of PCR, coupled with measurement of a fluorescence intensity that is assumed to be directly proportional to the amount of product in each cycle [3]. This methodology has been extensively validated, and its accuracy and specificity have been proved for the different chemistries available [4].

The quantitative output of the qRT-PCR consists of an amplification curve, which is composed of a set of cycle numbers and associated fluorescence intensities that are ulteriorly summarized in a single value called the cycles to threshold (C_T). The C_T is a unitless value defined as the fractional cycle number at which the sample fluorescence signal passes a fixed threshold above the baseline. Because the threshold is set within the exponential amplification phase, the C_T is proportional to the (negative) log of the initial transcript copy number (or log-transcript concentration) of the assayed sample. The constant of proportionality of the C_T to the log-concentration is the amplification efficiency (E).

Absolute and relative quantification strategies can be applied to measure mRNA abundance using qRT-PCR [3,5]. Relative quantification compares the expression of a target gene under various conditions (treatments) normalized to the measured expression of an internal control [6] (assumed to be constantly expressed across samples). In general, the numerous mathematical expressions available for such calculation [7–16] may be summarized by the equation below [6]:

$$\mathrm{FC}_{\mathrm{trt}_1:\mathrm{trt}_2} = \frac{\left(E_{\mathrm{Target}}\right)^{\Delta C_{\mathrm{T(target)}}(\mathrm{trt}_2 - \mathrm{trt}_1)}}{\left(E_{\mathrm{Control}}\right)^{\Delta C_{\mathrm{T(control)}}(\mathrm{trt}_2 - \mathrm{trt}_1)}}, \quad (1)$$

where, $\mathrm{FC}_{\mathrm{trt}_1:\mathrm{trt}_2}$ is the relative expression (fold-change) of the target gene in a sample from treatment 1 compared to a sample from treatment 2, E_{Target} and E_{Control} are the amplification efficiencies of the target and the control genes, respectively, and $\Delta C_{\mathrm{T(target)}}(\mathrm{trt}_2 - \mathrm{trt}_1)$

* Corresponding author. Department of Animal Science, Michigan State University, East Lansing, MI 48824, USA. Fax: +1 517 353 1699.
E-mail address: steibelj@msu.edu (J.P. Steibel).

0888-7543/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2009.04.008

and $\Delta C_{T(Control)}(trt_2 - trt_1)$ correspond to the C_T of the treatment 2 minus the C_T of the treatment 1, for the target and control genes, respectively. If both amplification efficiencies take the maximum possible value ($E=2$), expression (1) becomes the familiar $2^{-\Delta\Delta CT}$ expression [9]. Moreover, almost any other mathematical expression or method available in the literature to calculate fold-change is a variant of Eq. (1). The differences among variants of Eq. (1) refer mainly to estimation of the efficiency either from a relative standard curve [5] or from individual amplification curves [7,13].

The methods based on expression (1) are mathematical equations devised to calculate fold-change between two samples. In some cases, these equations, however, lack the statistical formalism needed to draw valid inferences, especially when multiple levels of biological replicates from each experimental group are assayed [7,10,17]. Moreover, many ad-hoc approaches associated with formulas similar to Eq. (1) have been used with the objective of generating a set of "companion" *P*-values or standard errors [9,11]. However, few of them are valid in the presence of both biological and technical replication. Currently, the REST® software [17] is one of the few programs that implements a valid statistical analysis to test hypotheses and estimate the fold-changes using Eq. (1). However, such software is limited to the analysis of pairwise comparisons with respect to a control group, under a completely randomized design.

Even if a valid statistical test can be implemented for pairwise contrasts based on Eq. (1), comparing two treatments at a time in the context of a large experiment may be inefficient and lead to reduced power. In such case, a linear model could be used to analyze data from all treatment groups simultaneously in the same fashion of classical analysis of variance (ANOVA) techniques.

A linear mixed model [18] was recently proposed for the implementation of the so-called analytical method [10]. Such a model is potentially more flexible than the existing alternatives, but it makes the strong assumption that there is a common random effect for the control and test genes in each biological replicate. Assuming no gene-specific biological effects is not realistic, as it is expected that constitutively expressed genes will have more similar expression pattern across biological replicates than regulated genes. Violation of such assumption, in turn, may lead to wrong inferences. Consequently, there is a need for a formal statistical method for analysis of the relative quantification RT-PCR data that allows accommodation of more complex experimental designs (such as blocking factors) and testing of general hypotheses (including interactions, pairwise and group contrasts).

The objective of this paper is to present a novel, flexible method for analysis of relative quantification RT-PCR data using linear mixed models. The main advantage of the model is that it can be used to compute valid *P*-values associated with any general linear hypothesis of interest. Additionally, the model allows proper accounting of all sources of variation and it is expected to be more powerful than methods based on individual pairwise comparisons. In this paper we use a variety of approaches to validate the proposed methodology, to compare it with existing methods, and to illustrate its flexibility. First, our model is compared to other alternatives using a real dataset. Second, a model-free simulation based on the same dataset is used for comparative validation of the methodology. Lastly, several datasets are analyzed and different linear models are compared.

Results

Motivating example

Quantitative RT-PCR was used to study expression of the gene diazepam binding inhibitor (DBI) in the brain of piglets subject to weaning and social isolation treatments [19]. The experimental layout followed a randomized complete block design ($n=3$ litters) and the treatments consisted of a 2×2 factorial combination of weaning (early-weaned or non-weaned) and social isolation (isolated or control).

Preliminary assays indicated that *Sus scrofa* 18S ribosomal RNA (18S) was suitable for use as an endogenous control gene and that the amplification efficiency for primers of the two genes (18S and DBI) was close to two [19]. All reactions were performed in triplicate but some observations were excluded from the analysis because of evidence of non-specific amplifications (as revealed by dissociation curve analyses) [20]. The following model (denoted as Model I) was used for the analysis of the joint expression of DBI and 18S:

$$y_{gijkr} = TG^*_{gi} + l_{gj} + B_{gijk} + D_{ijk} + e_{gijkr},$$

where y_{gijkr} is the C_T obtained from the thermocycler software for the *g*th gene (18S or DBI) from the *r*th well, corresponding to the *k*th animal in the *j*th litter subjected to the *i*th treatment, TG^*_{gi} is the effect of treatment *i* in the expression of gene *g*, $l_{gj}\sim N(0,\sigma^2_{lg})$ is a gene-specific random effect of the *j*th litter, $B_{gijk}\sim N(0,\sigma^2_{Bg})$ is a gene-specific random effect of the *k*th piglet in the *j*th litter, $D_{ijk}\sim N(0,\sigma^2_D)$ is a random sample-specific effect (common to both genes), and $e_{gijkr}\sim N(0,\sigma^2_e)$ is a residual term. The sample-specific effect, D_{ijk}, captures differences among samples that are common to both genes, particularly those that affect total mRNA concentration, such as differential extraction or amplification efficiencies among samples. The treatments consisted of the combination of two factors, and the sub-index $i=1, 2, 3, 4$ corresponds to: early weaning + control (EWC), early weaning + isolation (EWI), non-weaning + control (NWC), and non-weaning + isolation (NWI), respectively.

Model I was fit to the data using the SAS mixed procedure [21] and a residual analysis was performed to check model assumptions. Tests of differential expression among groups were performed for the interaction of weaning by isolation and for pairwise treatment differences (simple effects). Point and interval estimates of fold-changes were approximated from the linear contrasts (in the log scale) by back transformation. The fold-changes were also estimated with the $2^{-\Delta\Delta CT}$ method [9] (ΔC_T) using a procedure presented in the original work [9].

In addition, an alternative linear model (denoted as Model II) was also used to analyze the data:

$$y_{gijkr} = TG^*_{gi} + l_j + D_{ijk} + \varepsilon_{gijkr}.$$

Model II is a simplified version of Model I, without the gene-specific sample and litter effects, and is equivalent to a previously published model for analysis of amplification curve data [18]. We anticipate that Model II is under parameterized and it may lead to wrong inferences.

Testing and estimating differential expression

Contrarily to the ΔCT procedure, Model I yielded a formal test for the interaction between isolation and weaning. There was no evidence of interaction effect between isolation and weaning on the expression of DBI ($P=0.829$), but there was a significant three-fold decrease in DBI gene expression due to isolation ($P=0.003$). As mentioned, the traditional analysis method (ΔCT) does not allow testing of this interaction, but it may be still used to estimate the fold-change of pairwise comparisons (Fig. 1).

While the estimates of fold-change were similar using Models I, II and ΔCT, the confidence intervals for the fold-changes based on ΔCT were wider than those based on Models I and II, and the general conclusions were not equivalent. For example, Models I and II indicated a significant decrease in the expression of DBI in response to social isolation in both early-weaned and non-weaned animals ($P=0.013$ and $P=0.019$ respectively from Model I), while ΔCT only detected the contrast EWI – EWC as significant ($P=0.03$). At a