Inference for Quantitation Parameters in Polymerase Chain Reactions via Branching Processes with Random Effects

Bret Hanlon and Anand N. Vidyashankar *

Abstract

The quantitative polymerase chain reaction (qPCR) is a widely used tool for gene quantitation and has been applied extensively in several scientific areas. The current methods used for analyzing qPCR data fail to account for multiple sources of variability present in the PCR dynamics, leading to biased estimates and incorrect inference. In this paper, we introduce a branching process model with random effects to account for within reaction and between reaction variability in PCR experiments. We describe, in terms of the observed fluorescence data, new statistical methodology for gene quantitation. Using simulations, PCR experiments, and asymptotic theory we demonstrate the improvements achieved by our methodology compared to existing methods. This article has supplemental materials online.

KEYWORDS: Between Reaction Variability, Generalized Method of Moments, Martingale Limits, Quantitative Polymerase Chain Reaction, Relative Quantitation, Within Reaction variability.

*Bret Hanlon is an Assistant Professor, Department of Statistics, University of Wisconsin, Madison, WI 53706 (email: hanlon@stat.wisc.edu). Anand N. Vidyashankar is a Visiting Professor, Department of Statistics, George Mason University, Fairfax, VA 22030 (email: avidyash@gmu.edu). Dr. Vidyashankar's research was supported in part by a grant from NSF DMS 000-03-07057 and also by grants from the NDCHealth Corporation. The authors thank the editor, associate editor, and two anonymous referees for a careful reading of the manuscript and several useful suggestions. All computations were performed in Matlab.

1. INTRODUCTION

The polymerase chain reaction (PCR) is a biochemical tool used to amplify the number of copies of a specific DNA sequence. Quantitative PCR (qPCR) is an enhancement of PCR which allows gene quantitation. A gene expression experiment begins with extraction of mRNA from a tissue sample. The amount of mRNA is typically too small to be detected by standard instruments. Reverse transcription is then used to convert the mRNA to cDNA. Finally, qPCR is used to amplify the cDNA to a detectable level (Speed, 2004). To assess gene expression, the initial number of cDNA molecules, or *copy number*, must then be estimated using the data from the amplified product. This process is called *quantitation*, with the quantitation parameters denoting the mean number of DNA molecules in the initial product and the efficiency of the process (defined formally in Section 3.2 below). Quantitation via qPCR is one of the most commonly used methods for gene expression analysis with applications in diverse areas such as forensic science, virology, and parasitology (Ferré, 1998). The use of qPCR as an accurate tool for gene expression depends on the statistical validity of the estimator of the copy number.

In this article we focus on the statistical analysis of data from a qPCR experiment used as a gene expression tool. We propose a new design and a novel inferential methodology for quantitation, using branching processes with random effects. We establish the asymptotic validity of our methodology via simulation and demonstrate its superiority using data from PCR experiments from two distinct scientific disciplines. Compared to the currently used procedures, our method yields point estimates with smaller relative bias and confidence intervals with accurate coverage.

The article is organized as follows. Section 2 provides background for PCR and qPCR. In Section 3 we describe the data structures encountered in a typical PCR experiment and the sources of variability associated with them. Additionally, we develop a hierarchical, non-homogenous binary branching process model for describing the dynamics of PCR experiments. Section 4 contains the main results of the paper, namely, generalized method of moments estimators for quantitation parameters and their limit distributions. We emphasize that the asymptotic framework used in this paper is not intended to understand the behavior of the reactions when infinitely many cycles are run; rather, we seek to explain features that arise in a typical PCR from the behavior of the limits of statistical quantities encountered in the data analyses. Thus, the asymptotic framework

serves to set up standards, that are biologically meaningful, for comparing various estimators and confidence intervals. In Sections 5 and 6, we use simulation studies data analysis to demonstrate our procedure's robustness to model assumptions and it's superiority compared to other existing methodologies. Section 7 contains concluding remarks. Proofs of technical results are in the supplemental material.

2. SCIENTIFIC BACKGROUND FOR PCR AND QUANTITATION

In this section, we summarize the critical details of PCR, qPCR, and the related question of quantitation. More detailed explanations of these concepts can be found in the literature (Mullis et al., 1994; Ferré, 1998; Speed, 2004). PCR is a biochemical experiment used to amplify the number of DNA molecules in a genetic material until data from 30-50 cycles are available. Theoretically, the number of molecules doubles in every cycle (until saturation); however, in practice only a fraction of the molecules within a given cycle duplicates. Hence, a supercritical Galton-Watson branching process with a Bernoulli offspring distribution provides a natural model to describe the dynamics of PCR (Kimmel and Axelrod, 2002; Lalam, 2009; Follestad et al., 2010). The probability of a molecule duplicating is known as the *efficiency* of the reaction; one plus this probability gives the *amplification rate*.

The amount of genetic material from a qPCR is measured via the intensity of the fluorescence signal computed for every cycle of the reaction. The cycles of a PCR can be classified into three phases: the initial phase, the exponential phase, and the plateau phase. Fluorescence data from the initial phase is noisy and hence it is customary in the PCR literature to ignore data from these initial cycles and only use data from the exponential phase. Figure 1, which plots fluorescence data on the log scale from 32 reactions, illustrates these three phases. The relationship between the number of DNA molecules (N) and the fluorescence intensity (F) is given by the formula

$$N = \left(\frac{c^\star \times 9.1 \times 10^{11}}{A_S} \right) F \tag{1}$$

where A_S is the amplicon size, defined to be the size (measured in base-pairs) of the target DNA sequence, and $c^\star$ is the calibration factor, which represents the number of nanograms of double-stranded DNA per fluorescence unit (Rutledge, 2004).

We discuss two types of quantitation: *absolute quantitation* and *relative quantitation.* Absolute quantitation refers to estimation of the copy number of a target gene, which requires accurate knowledge of the parameter $c^\star$. Frequently, scientists are concerned with relative quantitation, where the interest is understanding how much of a target gene is expressed relative to a so-called house keeping gene (Skern et al., 2005). More generally, relative quantitation is concerned with the estimation of the ratio of the copy numbers of a target gene to that of a reference gene, which we call a calibrator. Under the assumption that $c^\star$ is the same for both the amplifications, the estimate of the ratio does not explicitly involve this parameter. Statistical questions addressed in this paper concern inference for this ratio.

Existing techniques for quantitation involve a linear or non-linear regression model for the fluorescence data. These methods are based on a statistic called C_T, which essentially gives the cycle at which the fluorescence crosses a user-specified threshold. In fact, the value of C_T is not an integer, but is computed by linear interpolation on the log-scale (Livak and Schmittgen, 2001). These techniques, used in the PCR literature, are empirical and ad hoc as opposed to the mechanistic approach adopted in this paper. Furthermore, these methods do not use all of the available data in the exponential phase for statistical analyses. The estimates of the standard error provided by these techniques are usually incorrect because they do not account for all sources of variation.

The branching process model proposed in this paper is a mechanistic model and is based on the observed fluorescence intensity only and takes into account some of the shortcomings of the existing methodologies.

3. DATA STRUCTURES AND STATISTICAL MODELS

A typical qPCR experiment can produce data from either 96 or 384 separate reactions, each reaction occurring in a separate well. Repetition of the same PCR experiment in multiple wells, under identical conditions, yields data that can be represented as $\{F_{k,j} : k \geq 1, j \geq 1\}$; that is, $F_{k,j}$ represents the fluorescence intensity from the j^{th} cycle of the k^{th} replicate. We will use the words replicate and reaction interchangeably throughout the paper. Let $N_{k,j}$ denote the number of DNA molecules at cycle j of replicate k associated with the fluorescence intensity $F_{k,j}$ and $\{N_{k,j}, j \geq 0\}$

denote the associated branching process. Let $m_{a,k} = E(N_{k,0})$ and $m_{e,k} = E(N_{k,1}|N_{k,0} = 1)$.

3.1 Sources of variability and non-homogeneous models

Accuracy of inference (concerning the quantitation parameters) depends on identifying the sources of variability and accounting for them in the methodology. In a PCR experiment, variability is introduced at several stages of the experiment. First, we never know exactly the number of DNA molecules for a given genetic material. Second, the variability in the accumulated product at each cycle between replicates causes variability in the estimates of efficiency. Third, there are differences in efficacy between various cycles within a replicate (Saha et al., 2007).

While there are several causes for between replication variability, one of the dominant factors is the pipetting error (Curry et al., 2002; Super-Array, 2010). Pipetting error is caused due to the changes in the volume of PCR supplies that is pipetted from the master mix into the wells. The changes in efficiency within a reaction, observed at every cycle, are due to various biochemical reasons. Not much information concerning this variability can be obtained from fluorescence intensities alone. Hence while non-homogeneous models can be proposed, it is difficult to ascertain the practical effect of such modeling. For alternate approaches when more information is available see (Saha et al., 2007) and the references therein.

To address these issues, we represent the initial number of molecules as a random variable whose variance determines the precision of the estimator of $m_{a,k}$. Since $N_{k,0}$'s are unobservable, and estimators are based on the accumulated product, variability in the efficiency of the reaction affects inference. Information concerning between reaction variability in efficiency can be extracted by using either replicated or dilution data (see 3.2 below for more details on these terms). In this paper, we account for between reaction variability, by modeling the efficiency of the exponential phase of the k^{th} reaction as a random variable p_k with some distribution G on (0,1). During the exponential phase, experimental evidence suggests that the support of G lies in the interval $(1-\epsilon, 1)$ for some "small" ϵ (Livak and Schmittgen, 2001). Thus, the model proposed for describing PCR

dynamics is,

$$p_k \overset{i.i.d.}{\sim} G(.), \quad 1 \le k \le r(n), \tag{2}$$

$$N_{k,0} \overset{independent}{\sim} H_k(.), \quad 1 \le k \le r(n), \tag{3}$$

where given $N_{k,0}$ and p_k, we have that $\{N_{k,j} : j \ge 1\}$ is a binary branching process initiated by $N_{k,0}$ ancestors with splitting probability (i.e. efficiency) p_k. That is,

$$N_{k,j+1}|N_{k,j}, p_k \sim N_{k,j} + Bin(N_{k,j}, p_k), \text{ for all } j \ge 0, \tag{4}$$

where $Bin(N_{k,j}, p_k)$ is a binomial random variable with parameters $N_{k,j}$ and p_k. The sequence $\{p_k : k \ge 1\}$ representing the splitting probability is assumed to be independent of the sequence $\{N_{k,0} : k \ge 1\}$ of initial number of molecules. We call this model a branching process incorporating random effects. Let $p_e = E(p_k)$ and $m_e = p_e + 1$. We call p_e the marginal efficiency of PCR.

3.2 Designs and Data Structures

Assumptions on $N_{k,0}$ depend on the data set. In a replicated design, where the concentration of the genetic material is fixed across replicates, $\{N_{k,j}, 1 \le j \le r(n)\}$ are i.i.d. random variables and $m_{a,k} = m_a$ for all $k \ge 1$. In the case of dilution data, the initial concentration is diluted in a systematic way by certain amounts denoted by the constants d_k. In this case, $N_{k,j}$ are conditionally independent branching processes initiated by *independent* initial random variables, not i.i.d. initial random variables.

We will use the notation $Var(N_{k,0}) = \sigma_a^2 d_k^2$, $E(N_{k,0}^t) = m_{t,0} d_k^t$, $t = 3, 4$, for all $k \ge 1$. The constants d_k appear in our analysis through $D_t(n) = (r(n))^{-1} \sum_{k=1}^{r(n)} d_k^t$ for $t = 1, 2, 3, 4$. The following condition concerns the stability of the sequence $D_t(n)$ as n increases.

Condition 1 (*Regularity of dilution constants*)**.** *Assume that $D_t(n) \to D_t > 0$ as $n \to \infty$ for all $t = 1, 2, 3, 4$. Furthermore, assume that $\sum_{k \ge 1} k^{-2} d_k^{2t} < \infty$ for t=1, 2.*

We will assume that condition 1 holds throughout the paper. The condition on $D_4(n)$ and the convergence of $\sum_{k \ge 1} k^{-2} d_k^4$ are needed for studying the consistency of the variance estimate. For replicated data, $D_t(n) = 1$ for all t=1,2, 3, 4.

3.3 Inferential Problem

Our primary objective is inference concerning the quantitation parameters, (m_a, p_e). We also consider estimation of σ_a^2 since it is needed for standard error calculations. As discussed previously, it is assumed, and experimentally verified (Goll et al., 2006), that the amount of fluorescence is proportional to the number of DNA molecules; that is $N_j = cF_j$, where, using (1), $c = \frac{c^\star \times 9.1 \times 10^{11}}{A_S}$. The methodology described in this paper uses the observed fluorescence F_j.

4. INFERENCE FOR COPY NUMBER

In this section we describe our inferential results for the quantitation parameters. Section 4.1 describes the moment calculations and the role of martingales in our analysis. Sections 4.2 and 4.3 discuss absolute and relative quantitation, respectively. Section 4.4 is concerned with bias correction while Section 4.5 contains inferential results concerning efficiency. Finally, Section 4.6 deals with estimation of variances.

As explained before, asymptotic theory helps to identify features that occur in PCR when the number of cycles and the number of replicates increase. We emphasize that if efficiency is known, then one can allow only the number of replicates to increase to infinity, as is standard in statistical large sample theory. However, since PCR efficiency is unknown and is critical for inference, we can extract its behavior assuming we have a large number of DNA molecules. This is satisfied in the PCR problem since even after ten cycles, the sample size for inference about PCR efficiency is substantial. This phenomenon is studied in the branching process literature for estimation of the offspring mean via large deviations (Ney and Vidyashankar, 2003).

4.1 Moments and Martingales

It is instructive to express the model using the standard branching process recursion (Athreya and Ney, 1972)

$$N_{k,j+1} = \sum_{l=1}^{N_{k,j}} \xi_{k,j,l}, \tag{5}$$

where the binary random variable $\xi_{k,j,l}$ represents whether the l^{th} molecule in the j^{th} cycle of the k^{th} reaction duplicates or not. In terms of the random variables $\xi_{k,j,l}$, our assumption states that for every fixed k and p_k, the random variables are i.i.d. with distribution $P(\xi_{k,j,l} = 2|p_k) = p_k$, and

$P(\xi_{k,j,l} = 1|p_k) = 1 - p_k$. Thus, $E(\xi_{k,j,l}|p_k) = 1 + p_k \equiv m_k$ and $Var(\xi_{k,j,l}|p_k) = p_k(1 - p_k) \equiv \sigma_k^2$. We will use the notation $E_k(.)$ to denote the conditional expectation $E(.|p_k)$ and $Var_k(.)$ to denote the conditional variance $Var(.|p_k)$. Hence, it follows that $E_k(N_{k,j+1}) = E_k(E_k(N_{k,j+1}|N_{k,j})) = m_k E_k(N_{k,j})$. Iterating the identity, we have $E_k(N_{k,j+1}) = m_k^{j+1} m_a d_k$. In the case of fluorescence data, using (1) it follows that $E_k(F_{k,j+1}) = c^{-1} m_k^{j+1} m_a d_k$. Thus, conditioned on the random effect p_k, $V_{k,j} \equiv m_k^{-j} N_{k,j}$ is a positive martingale sequence with respect to the sigma field containing information up to $(j-1)$ cycles and the value of the random effect. Hence, as $j \to \infty$, $V_{k,j}$ converges to $V_k^\star$, where $V_k^\star > 0$ (see Athreya and Ney (1972)). Furthermore, since $V_{k,j}$ has uniformly bounded marginal and conditional moments of at least order four (see the supplemental material), the sequence $\{V_{k,n}^2 : n \geq 1\}$ is uniformly integrable and hence $V_{k,n}$ converges in L_2 to $V_k^\star$.

The moments of the random variable $V_k^\star$ are needed for deriving the generalized method of moments estimator and for comparing the estimators in terms of their variances. It is easy to see that the marginal and the conditional means of $V_k^\star$ coincide and are given by $m_a d_k$. However, the marginal and conditional variances are different. The conditional variance of $V_k^\star$ is $Var_k(V_k^\star) = m_a d_k \frac{\sigma_k^2}{m_k(m_k-1)} + \sigma_a^2 d_k^2$ while the marginal variance is given by $\omega_k^2 = m_a d_k E(\frac{1-p_1}{1+p_1}) + \sigma_a^2 d_k^2$. The marginal variance of the limiting random variable $V_k^\star$ depends on the reaction only via the dilution factor used in that reaction.

4.2 Absolute Quantitation

Information about m_a is contained both in $N_{k,0}$ and in $V_k^\star$. Let us assume, for the moment that $d_k = 1$. If one can obtain a random sample of size r from $N_{k,0}$, then the resulting sample mean is an unbiased estimator of m_a. The variance of this estimator is then $r^{-1}\sigma_a^2$ and the problem would be completely resolved. However, since it is not possible to obtain observable samples from $N_{k,0}$, one could use instead the sample mean of a random sample from $V_k^\star$ to estimate m_a. The variance of this estimator would be $r^{-1}(\sigma_a^2 + m_a E(\frac{1-p_1}{1+p_1}))$. Since $V_k^\star$ are unobservable, once again this recipe is not feasible. The discussion however suggests that if one were to average observable data over replicates then consistent estimators of m_a may exist. Since one can obtain fluorescence data at every cycle, it is natural to use data from the cycles in the exponential phase to obtain estimators of m_a. Thus, the first step is to identify cycles belonging to the exponential phase i.e.,

cycles beyond C_T.

Let τ_k and n_k denote the first and last cycles of the exponential phase, respectively, in the k^{th} reaction. Then, the cycles in the exponential phase of that reaction can be denoted by $\tau_k, \tau_{k+1}, \cdots n_k$. To make the conditions more transparent when studying asymptotics, we will take $n_k = n$ and $\tau_k = \tau$. This does not entail any loss of generality and also minimizes cumbersome notation. In our data analysis, we do not make this assumption.

Since more than one cycle is involved during the exponential phase, we consider the total accumulated fluorescence during the exponential phase, namely

$$Y_{k,n} = \sum_{j=\tau}^{n} F_{k,j}. \tag{6}$$

Our formulation of the inference problem in terms of the generalized method of moments technique will involve the behavior of $Y_{k,n}$ and not $F_{k,n}$. The proposition below describes the asymptotic behavior of $Y_{k,n}$ for every reaction k.

Proposition 1. *Under the assumptions of our model, conditioned on the random effect p_k, with probability one*

$$\lim_{n\to\infty} \frac{Y_{k,n}}{m_k^n} = c^{-1}(\frac{m_k}{m_k - 1})V_k^{\star}. \tag{7}$$

Motivated by the above proposition, we consider the following generalized method of moments estimator of m_a, given by

$$\tilde{m}_{a,n} = \frac{c}{r(n)D_1(n)} \sum_{k=1}^{r(n)} \frac{\tilde{p}_{k,n}}{\tilde{m}_{k,n}^{n+1}} Y_{k,n}, \tag{8}$$

where $\tilde{p}_{k,n}$ is an estimator of the efficiency of the reaction for the k^{th} reaction and $\tilde{m}_{k,n} = 1 + \tilde{p}_{k,n}$. The estimator takes into account the variability in amplification rates between cycles and scales the product from the k^{th} replicate by the amplification rate of that reaction. The factor c is needed to convert the fluorescence information into number of molecules. As one would expect, the asymptotic properties of $\tilde{m}_{a,n}$ depend on the properties of the estimator of efficiency. While several estimators for efficiency are available, we use the weighted conditional least squares estimator of the reaction efficiency since it is based on the total accumulated fluorescence during the exponential phase; in fact, it is the non-parametric maximum likelihood estimator (MLE) (Guttorp, 1991). The

estimator is given by,

$$\tilde{p}_{k,n} = \frac{\sum_{j=\tau}^{n-1}(F_{k,j+1} - F_{k,j})}{\sum_{j=\tau}^{n} F_{k,j}} = \frac{Y_{k,n} - F_{k,\tau} - Y_{k,n-1}}{Y_{k,n-1}}. \tag{9}$$

Theorem 1. *Assume that the dilution constants satisfy condition 1. Let the number of replicates* $r(n)$ *be such that* $r(n)n^{-1} \to 0$ *as* $n \to \infty$. *Then,* $\tilde{m}_{a,n}$ *is a strongly consistent estimator of* m_a. *Furthermore,*

$$\sqrt{r(n)D_1(n)}(\tilde{m}_{a,n} - m_a) \xrightarrow{d} H, \tag{10}$$

where $H \sim N(0, \sigma_L^2)$, *with* $\sigma_L^2 = m_a E(\frac{1-p_1}{1+p_1}) + D_L\sigma_a^2$ *and* $D_L = D_1 D_2^{-1}$.

Thus, it follows from the theorem that

$$\tilde{m}_{a,n} \dot{\sim} N(m_{a,n}, \frac{\sigma_L^2}{r(n)D_1(n)}). \tag{11}$$

When $p_1 \equiv 1$, then the number of molecules exactly doubles in each cycle, and the only variation in the estimation comes from the variability in the initial amount of genetic material. If $\sigma_a^2 = 0$, then one can quantitate exactly and the results reduce to classical results from the PCR literature. Of course, neither of these are feasible and the above theorem shows the precise nature of the variability in the quantitation process, providing a decomposition along the lines of classical analysis of variance. Finally, for the replicated data structure we have $D_1(n) = 1$ and $D_L = 1$; and hence the limiting variance does not involve the dilution parameters.

4.3 Relative Quantitation

In relative quantitation, we have two sets of genetic material, the calibrator and the target. We will add a subscript C and T to our notation to distinguish between data collected from calibrator and target materials. Hence $F_{k,j,C}$ and $F_{k,j,T}$ will represent the fluorescence from the j^{th} cycle of the k^{th} reaction from the calibrator and target materials, respectively. The unobservable branching process associated with these fluorescence data are denoted by $N_{k,j,C}$ and $N_{k,j,T}$ respectively. Let $E(N_{k,0,C}) = m_{a,C}d_k$ and $E(N_{k,0,T}) = m_{a,T}d_k$. Let $\sigma_{a,C}^2 d_k^2$ and $\sigma_{a,T}^2 d_k^2$ denote the variance of $N_{0,C}$ and $N_{0,T}$, respectively. To complete the description of the model, we assume that, for $I = C, T$,

$\{p_{k,I} : k \geq 1\}$ is a collection of independent random variables with distribution $G_I(.)$ and support $(1-\epsilon_I, 1)$.

In relative quantitation, the object of interest is R, where $R = \frac{m_{a,T}}{m_{a,C}}$. Analogous to the absolute quantitation case (see (9)), we define the non-parametric MLE of the reaction efficiency as follows:

$$\tilde{p}_{k,n,C} = \frac{Y_{k,n,C} - F_{k,\tau,C} - Y_{k,n-1,C}}{Y_{k,n-1,C}}, \quad \tilde{p}_{k,n,T} = \frac{Y_{k,n,T} - F_{k,\tau,T} - Y_{k,n-1,T}}{Y_{k,n-1,T}}. \tag{12}$$

This yields, for $I = C, T$, $\tilde{m}_{k,n,I} = 1 + \tilde{p}_{k,n,I}$. Hence, one can now estimate the ratio R using $\tilde{R}_n = \frac{\tilde{m}_{a,n,T}}{\tilde{m}_{a,n,C}}$, where

$$\tilde{m}_{a,n,T} = \frac{1}{r(n)} \sum_{k=1}^{r(n)} \frac{\tilde{p}_{k,n,T}}{\tilde{m}_{k,n,T}^{n+1}} Y_{k,n,T}, \quad \tilde{m}_{a,n,C} = \frac{1}{r(n)} \sum_{k=1}^{r(n)} \frac{\tilde{p}_{k,n,C}}{\tilde{m}_{k,n,C}^{n+1}} Y_{k,n,C}. \tag{13}$$

Theorem 2. *(Relative Quantitation) Under the assumptions of Theorem 1, $\tilde{R}_n$ is a strongly consistent estimator of R. Furthermore,*

$$\sqrt{r(n)D_1(n)}(\tilde{R}_n - R) \xrightarrow{d} G_2, \tag{14}$$

where $G_2 \sim N(0, \sigma_R^2)$. The limiting variance σ_R^2 is given by $\sigma_R^2 = R^2(\sigma_{L,T}^2 + \frac{\sigma_{L,C}^2}{m_{a,C}^2})$, where

$$\sigma_{L,I}^2 = m_{a,I} E(\frac{1 - p_{1,I}}{1 + p_{1,I}}) + D_L \sigma_{a,I}^2, \quad I = C, T, \tag{15}$$

and D_L is as in Theorem 1.

4.4 Bias Correction

The estimator of m_a proposed in Section 4.2 can be improved by accounting for the cycles during the noisy initial phase. To address this issue, we observe that the mean fluorescence during the exponential phase of the k^{th} reaction is given by $c^{-1} m_k^{n+1} (m_k - 1)^{-1} (1 - m_k^{\tau-(n+1)})$. Since, $(1 - \tilde{m}_{k,n}^{\tau-(n+1)})$ converges to one exponentially fast, one can show that the bias corrected estimator

$$\tilde{m}_{a,n}^{(b)} = \frac{c}{r(n)D(n)} \sum_{k=1}^{r(n)} \frac{\tilde{p}_{k,n}}{\tilde{m}_{k,n}^{n+1}} (1 - \tilde{m}_{k,n}^{\tau-(n+1)})^{-1} Y_{k,n}, \tag{16}$$

inherits the asymptotic properties of $\tilde{m}_{a,n}$. For this reason, we use and recommend this estimator for data analysis.

4.5 Inference for PCR Efficiency

Sections 4.2 and 4.3 show that inference for quantitation depends critically on the estimator of the experiment efficiency, both conditional and marginal. The conditional efficiency is useful for quantitation purposes and is estimated as the conditional weighted least squares estimator given in (9). The following proposition describes the asymptotic limit distribution of $\tilde{p}_{k,n}$.

Proposition 2. *Under the assumptions of our model, for every fixed k,*

$$\sqrt{Y_{k,n-1}}(\tilde{p}_{k,n} - p_k) \xrightarrow{d} H_2, \tag{17}$$

where $P(H_2 \leq x) = \int_{1-\epsilon}^{1} \Phi(\frac{x}{t(1-t)})dG(t)$.

The marginal efficiency, which is helpful in determining the efficiency of the PCR equipment and the related question of design of experiments, is defined to be Ep_1. The estimator of marginal efficiency, is obtained by averaging the reaction efficiencies and is given by

$$\tilde{p}_{n,pool} = \frac{1}{r(n)} \sum_{k=1}^{r(n)} \tilde{p}_{k,n}. \tag{18}$$

Theorem 3. *Under the assumptions of Theorem 1, $\tilde{p}_{n,pool}$ is a strongly consistent estimator of the overall efficiency of the PCR, namely $E(p_1)$. Furthermore,*

$$\sqrt{r(n)}(\tilde{p}_{n,pool} - E(p_1)) \xrightarrow{d} H_1, \tag{19}$$

where $H_1 \sim N(0, \sigma_G^2)$, where σ_G^2 is the variance of the random variable p_1.

4.6 Estimation of Variances

We now focus on the estimation of limiting variances σ_L^2 and σ_a^2. Set

$$\tilde{\sigma}_{L,n}^2 = \frac{c^2}{r(n)D_1(n)} \sum_{k=1}^{r(n)} (\frac{\tilde{p}_{k,n}}{\tilde{m}_{k,n}^{n+1}} Y_{k,n} - \tilde{m}_{a,n} d_k)^2, \tag{20}$$

and

$$\tilde{\theta}_{1,n} = \frac{1}{r(n)} \sum_{k=1}^{r(n)} \frac{1-\tilde{p}_{k,n}}{1+\tilde{p}_{k,n}} \qquad \tilde{\theta}_{2,n} = \frac{1}{r(n)} \sum_{k=1}^{r(n)} \frac{\tilde{p}_{k,n}}{1+\tilde{p}_{k,n}}. \tag{21}$$

Theorem 4. *Under the assumptions of Theorem 1, $\tilde{\sigma}^2_{L,n}$ is a consistent estimator of σ^2_L. Furthermore, $\tilde{\theta}_{1,n}$ and $\tilde{\theta}_{2,n}$ are consistent estimators of $E(\frac{1-p_1}{1+p_1})$ and $E(\frac{p_1}{1+p_1})$, respectively.*

An immediate consequence of Theorem 4 is the following corollary concerning consistent estimation of σ^2_a.

Corollary 1. *Define*

$$\tilde{\sigma}^2_{a,n} = \frac{\tilde{\sigma}^2_{L,n} - \tilde{m}_{a,n}\tilde{\theta}_{1,n}}{D_{L,n}}, \tag{22}$$

where $D_{L,n} = D_2(n)D_1^{-1}(n)$. Then, $\tilde{\sigma}^2_{a,n}$ is a consistent estimator of σ^2_a.

5. SIMULATION EXPERIMENT

In this section we describe simulation results to evaluate the performance of the proposed methodology and compare it with other procedures studied in the literature. The results in this section are based on 5000 simulations. Table 2, which summarizes these results, includes both the mean value of the point estimate and the variance of the point estimate over the 5000 simulations; the Monte Carlo error of the point estimate can either be evaluated as this variance value or its square root. Similarly, the Monte Carlo error for the coverage probabilities is readily assessed. For example, the estimated coverage probability of a true 95% confidence interval will have a simulation accuracy of approximately 0.6% ($1.96\sqrt{.05(.95)/5000} = 0.006$)

5.1 Data Generation

We generate data from three different models. We use the notation $X \sim Bern(p)$ to refer to a binary random variable on $\{1, 2\}$, with $P(X = 2) = 1 - P(X = 1) = p$. The first model is an example of the random effect model proposed in the paper; specifically we use a beta distribution to describe the random effect.

Model 1. *(Random effects). For $I = C, T$, let $F_{k,j,I} = N_{k,j,I}$, where $N_{k,j,I}$ has offspring distribution $Bern(p_{k,I})$, with $p_{k,I} \sim^{iid} Beta(90, 10)$.*

We also address the robustness of our procedure to model assumptions, namely the constancy of efficiency across cycles and the constancy of c. The PCR literature argues that the efficiency of reactions changes between cycles (Saha et al., 2007). We study the impact of varying efficiency with the following model.

Model 2. *(Random environments). For* $I = C, T$. $F_{k,j,I} = N_{k,j,I}$, *where* $N_{k,j,I}$ *has offspring distribution* $Bern(p_{k,j,I})$, *with* $p_{k,j,I} \sim^{iid} Beta(90, 10)$.

It is difficult to quantify the magnitude of variability in c within and between reactions. To identify how this variability can affect our results, we consider the following model.

Model 3. *(Random fluorescence coefficient). For* $I = C, T$. $F_{k,j,I} = c_{k,j,I} N_{k,j,I}$, *where* $N_{k,j,I}$ *has offspring distribution* $Bern(p_{k,I})$, *with* $p_{k,I} \sim^{iid} Beta(90, 10)$. *And* $c_{k,j,I} \sim gamma(1, 10^{-3})$, *i.e.* $Ec_{k,j,I} = 1$ *and* $var(c_{k,j,I}) = 10^{-3}$.

In all three models, $N_{k,0,T} \sim Poiss(10^3)$ and $N_{k,0,C} \sim Poiss(10^2)$; hence the true value for relative quantitation is 10. We use the $Beta(90, 10)$ distribution for the reaction efficiencies; this distribution has mean .9 and variance $\approx 8.9109 \times 10^{-4}$. These values are similar to values for means and variances of computed efficiencies across different experimental reactions. We considered simulation experiments over a wide variety of parameters and the specific parameters of the beta distribution do not seem to impact the overall conclusions (results not presented).

All of the results are based on $n = 20$ cycles and $r(n) = 20$ replicates. For the branching process estimator, generations 15 to 20 are used. The standard curve method requires the use of standards. In each simulation three replicates of a five fold dilution series were used to form the standard curve. Here, the initial number for the dilution series is Poisson distributed with means $80, 400, 2000, 10,000$ and $50,000$.

5.2 Estimators and Discussion of Results

We compare four estimators: branching process estimator proposed in this paper, the standard curve based estimator, the comparative C_T estimator, and the adjusted comparative C_T estimator. Table 2 contains a summary of these experiments. Both the standard curve and comparative C_T estimators are described in the ABI User's Manual (ABI, 2001). We briefly discuss the adjusted

comparative C_T estimator. The comparative C_T method assumes perfect doubling for both the target and calibrator and estimates R using the formula $\hat{R} = 2^{C_{T,C}-C_{T,T}}$, where $C_{T,C}$ is the C_T value for the calibrator and $C_{T,T}$ is the C_T value for the target. In the presence of replicates, we use the averaged values of $C_{T,C}$ and $C_{T,T}$ in the above formula for $\hat{R}$. Researchers (Guescini et al., 2008) have criticized the assumption of perfect doubling and have suggested replacing 2 with estimated efficiency. Accordingly, we consider an adjusted comparative C_T estimator, which estimates the efficiency of the reaction from the observed fluorescence data. For details, regarding these C_T-based estimators, see the supplemental material.

Since the asymptotic behavior of the estimators is unknown for the C_T-based methods, we use the bootstrap method (by resampling replicates) to construct confidence intervals for relative quantitation. All confidence intervals presented are 95% confidence intervals; all bootstrap confidence intervals are based on 2000 bootstrap samples. The bootstrap sample size was taken to be $r(n)$.

The comparative C_T method does not perform well under any of the three models. Interestingly, this comparative C_T method is still one of the most frequently used estimators for relative quantitation (Guescini et al., 2008). In contrast, the branching process, standard curve, and adjusted comparative C_T methods perform well under all three models in terms of bias. However, the mean square error of the branching process method is smaller than all other methods. The increased variability present in Model 2 and Model 3 is reflected in increased variance of the point estimate and increased length of the confidence intervals.

In the presented simulations, the adjusted C_T estimator performs nearly as well as the branching process method. We note that this is the case because the parameters were chosen so that the bias in the estimates are not magnified. We reiterate that raising the estimated efficacies to the respective powers of $C_{T,C}$ and $C_{T,T}$ is too simplistic. While simulation experiments can be constructed to emphasize this point, analysis of PCR data from parasitology, given in Section 6, actively illustrates the issue.

Overall, the branching process method performs well, in terms of point estimates, mean squared error, and confidence interval coverage, in all three models. These results are important because they suggest that our methodology is robust to model assumptions. In particular it is robust to the assumption of a constant splitting probability (in a given replicate, across cycles). The confidence

intervals based on the t distribution have an accurate coverage (closest to the nominal 95%).

It is informative to notice that when $p_k = 1$ for all k, then $Beta(\alpha, \beta)$ distribution with parameters (1, 0) or a "very small" $\beta\alpha^{-1}$ will all yield " doubling of DNA" between cycles. Indeed, the comparative C_T method is based on this assumption. The adjusted comparative C_T and our branching process estimators are based on the assumption that $Var(c) = 0$. However, when $Var(c) > 0$ the results show that the branching process estimator is better behaved (in terms of mean square error) than the estimator based on adjusted comparative C_T.

6. ANALYSIS OF EXPERIMENTAL DATA

In this section we consider the analysis of two experimental data sets, both generated from a ABI Prism 7700 Sequence Detection System. In both experiments, we know the true value for relative quantitation (up to experimental error). We devised the following simple and convenient idea for comparing multiple estimators for relative quantitation: start with a target and dilute it by a known factor R, and call the resulting product the calibrator. Then the 'true' answer for relative quantitation is given by the factor R. As before, all bootstrap confidence intervals are computed based on 2000 bootstrap samples (where the replicates are re-sampled). Below we give a more detailed description of how the branching process estimator is computed for the experimental data. We then describe the two experiments and discuss the results of the data analysis.

6.1 Branching Process Method

Our methodology requires identification of the exponential phase. As discussed above, we allow the starting cycle ($\tau_{k,I}$) and the ending cycle ($n_{k,I}$) to differ for each replicate (for both the target and calibrator). The strategy is to choose those cycles which yield a fluorescence of at least $F^\star$ and per-cycle amplification of at least m_c. The following algorithm identifies the cycles of data belonging to the exponential phase, for each replicate k for target and calibrator. For the algorithm, we define $\gamma_{k,I} \equiv \inf\left\{j : F_{kjI} > F^\star\right\}$, for $I = T, C$. Then for each replicate, for both the target and calibrator, we run the following algorithm. Initialization: 1) Add cycles $\gamma_{k,I}$, $\gamma_{k,I} + 1$, 2) Set $i = \gamma_{k,I} + 1$. While loop: While $F_{k,i+1,I}/F_{k,i,I} > m_c$ 1) Add cycle $i+1$, 2) Update i to $i+1$. In our analysis, we use $m_c = 1.5$ and $F^\star = 0.2$, which is the default for the ABI machine.

Let r_T and r_C denote the number of replicates for the target and the calibrator. After

identifying the exponential phase, we compute the bias corrected estimator (see (16)) $\tilde{R} = \left(\frac{1}{r_T}\sum_{k=1}^{r_T}\hat{V}_{k,T}\right)\left(\frac{1}{r_C}\sum_{k=1}^{r_C}\hat{V}_{k,C}\right)^{-1}$, where the quantities $\hat{V}_{k,I}$ represent a scaled value of the total fluorescence in the exponential phase for each replicate, namely

$$\hat{V}_{k,I} = \frac{\tilde{p}_{k,n_{kI},I}}{\tilde{m}_{k,n_{kI},I}^{n_{kI}+1}}\left(1-\tilde{m}_{k,n_{kI},I}^{\tau_{kI}-(n_{kI}+1)}\right)^{-1}\left(\sum_{j=\tau_{kI}}^{n_{kI}} F_{k,j,I}\right).$$

6.2 Luteinizing hormone

We first consider a qPCR experiment where the target material is luteinizing hormone (LH) taken from a mouse pituitary gland; the importance of LH in the study of the human menstrual cycle is discussed in Tien et al. (2005). This data set consists of 16 replicates of two dilution sets, denoted LH_1 and LH_2; the LH_2 product was obtained by diluting the LH_1 product by a factor of 2.9505. In this data analysis we proceed as if LH_1 is the target group and LH_2 is the calibrator group. Thus, the desired answer for relative quantitation is 2.9505. To compute the estimator for the standard curve, fluorescence data from another product is required. In this case, the standard curve was computed using an eight point dilution series of the hormone prolactin (PRL) (this was also obtained from a mouse pituitary gland).

Graphs of the fluorescence data, both for each individual reaction and for the mean over all reactions, are displayed in Figure 1. We recall that the branching process method (as described in the above subsection) involves obtaining a scaled value of the sum of the fluorescence from each replicate, and then averaging these numbers. These scaled values $\hat{V}_k$ are also plotted in the figure. In our analysis we excluded data from one of the reactions for LH_2, since it did not reach a detectable level. Similarly, we excluded data from an LH_1 replicate since its C_T value was much larger than those of other replicates. The results of this analysis, which are based on 15 replicates from both LH_1 and LH_2, are summarized in Table 1.

6.3 Strongylus vulgaris

We consider a second experimental data set, where the target material is Strongylus vulgaris (S.vulgaris) ribosomal DNA; the importance of designing PCR assays for S.vulgaris in the field of equine parasitology is discussed in Nielsen et al. (2008). This data set consists of ten replicates of

two dilution sets, denoted SV_1 and SV_2; the SV_2 product was obtained by diluting the SV_1 product by a factor of ten. Thus, the desired answer for relative quantitation is 10. A figure analogous to Figure 1 for this data set is in the supplemental material. For this data set we compare the analysis using the branching process method and the comparative C_T method (we do not have appropriate data for computing the standard curve estimator). The results of this analysis are summarized in Table 1.

6.4 Discussion of the Results

From the data analysis, it is clear that the proposed branching process method using replicate data yields point estimates with smaller bias than other methods. Furthermore, even though the confidence intervals are wider, the simulation results suggest that they provide nominal coverage. To wit, the confidence intervals for the C_T method have poor coverage properties in our simulation studies; this is because the methodology under estimates the variability in the data. In the S.vulgaris example, the confidence interval based on the C_T method and the adjusted C_T method does not include the true value of the parameter. It should be noted that there is no asymptotic justification for the use of bootstrap confidence intervals for the C_T method or the adjusted C_T method. The S. vulgaris example is telling, because the efficiency of the target and the calibrator groups are very different; namely, the average efficiency for the target group is ≈ 0.70 compared to an average efficiency for the calibrator group of $\approx .94$. Because of these differences in efficiency, the adjusted comparative C_T method performs poorly. The branching process estimator is unaffected by the differences in efficiency between the two groups, because it scales each reaction by a separate estimate of efficiency.

7. DISCUSSION

In this work we utilized the branching process approximation of PCR, which provides a natural model for the dynamics of the reaction. We also used the availability of replicates, provided by a multiple well plate, to propose a new replicated design for quantitation experiments. We incorporated a random effect model to account for the between replicate variability. Finally, combining these components, we developed a novel generalized method of moments approach for inference concerning the quantitation parameters. We established strong consistency and asymptotic normality

of the resulting estimators. The simulation studies evaluated the behavior of our methodology under scenarios that are considerably different from the assumed model and illustrated its robust behavior. More importantly, in two controlled dilution experiments, our methodology outperforms existing estimators.

qPCR is considered the "gold standard" for gene expression tools (Lefever et al., 2009) and, in addition to its frequent stand-alone usage, it is used to validate the results from other expression tools such as microarrays and next generation sequencing. In fact, it has been used successfully for the important question of discovering up/down-regulated genes (Pridgeon et al., 2010). Thus, proper data analysis for qPCR experiments is of great scientific importance. Quantitation methods based on a branching process model with random effects offer great promise relative to C_T based methods, which fail to account for the variability in the dynamics of the process. For a given reaction, assuming a branching process model with constant efficiency is admittedly somewhat unrealistic. However, this assumption is made for theoretical developments only. Establishing that the proposed methods work without such an assumption seems to be difficult.

We emphasize that assumption of constant efficiency is approximately true within the exponential phase and, like most quantitation methods, our estimator uses data from only this phase. Based on the accuracy of our estimator using experimental data and simulation results, it is clear that the statistical methods presented here are useful.

SUPPLEMENTAL MATERIALS

Proofs and Numerical Results: This document includes a detailed discussion of the comparative C_T estimators, proofs for the asymptotic results, and additional figures for the data analysis. (InfQuantSuppl.pdf).

REFERENCES

ABI (2001), *ABI Prism 7700 Sequence Detection System.User Bullentin Number 2* PE Applied Biosystems.

Athreya, K., and Ney, P. (1972), *Branching Processes*, Berlin: Springer-Verlag.

Curry, J. D., McHale, C., and Smith, M. T. (2002), "Factors Influencing Real-Time RT-PCR Results: Application of Real-Time RT-PCR for the Detection of Leukemia Translocations," *Molecular Biology Today*, 3(1), 79–84.

Ferré, F. (1998), *Gene Quantification*, Boston: Birkhauser.

Follestad, T., Jorstad, T. S., Erlandsen, S. E., Sandvik, A. K., Bones, A. M., and Langaas, M. (2010), "A Bayesian Hierarchical Model for Quantitative Real-Time PCR Data," *Statistical Applications in Genetics and Molecular Biology*, 9(1), 3.

Goll, R., Olsen, T., Cui, G., and Florholmen, J. (2006), "Evaluation of absolute quantitation by nonlinear regression in probe-based real-time PCR," *BMC Bioinformatics*, 7, 107–118.

Guescini, M., Sisti, D., Rocchi, M., Stocchi, L., and Stocchi, V. (2008), "A new real-time PCR method to overcome significant quantitative inaccuracy due to slight amplification inhibition," *BMC Bioinformatics*, 9(1), 326.

Guttorp, P. (1991), *Statistical Inference for Branching Processes*, New York: Wiley.

Kimmel, M., and Axelrod, D. E. (2002), *Branching processes in biology*, New York: Springer-Verlag.

Lalam, N. (2009), "A quantitative approach for polymerase chain reactions based on a hidden Markov model," *Journal of Mathematical Biology*, 59, 517–533.

Lefever, S., Hellemans, J., Pattyn, F., Przybylski, D. R., Taylor, C., Geurts, R., Untergasser, A., Vandesompele, J., and RDML consortium (2009), "RDML: structured language and reporting guidelines for real-time quantitative PCR data," *Nucleic Acids Research*, 37(7), 2065–2069.

Livak, K. J., and Schmittgen, T. D. (2001), "Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2-[Delta][Delta]CT Method," *Methods*, 25(4), 402 – 408.

Mullis, K. B., Ferre, F., and Gibbs, R. A. (1994), *The Polymerase chain reaction*, Boston: Birkhauser.

Ney, P., and Vidyashankar, A. (2003), "Harmonic moments and large deviation rates for supercritical branching processes," *The Annals of Applied Probability*, 13, 475–489.

Nielsen, M. K., Peterson, D. S., Monrad, J., Thamsborg, S. M., Olsen, S. N., and Kaplan, R. M. (2008), "Detection and semi-quantification of Strongylus vulgaris DNA in equine faeces by real-time quantitative PCR," *International Journal for Parasitology*, 38(3-4), 443 – 453.

Pridgeon, J. W., Russo, R., Shoemaker, C. A., and Klesius, P. H. (2010), "Identification of in vitro upregulated genes in a modified live vaccine strain of Edwardsiella ictaluri compared to a virulent parent strain," *Comparative Immunology, Microbiology and Infectious Diseases*, In Press, Corrected Proof, –.

Rutledge, R. (2004), "Sigmoidal curve-fitting redefines quantitative real-time PCR with the prospective of developing automated high-througput applications," *Nucleic Acids Research*, 32, 178–186.

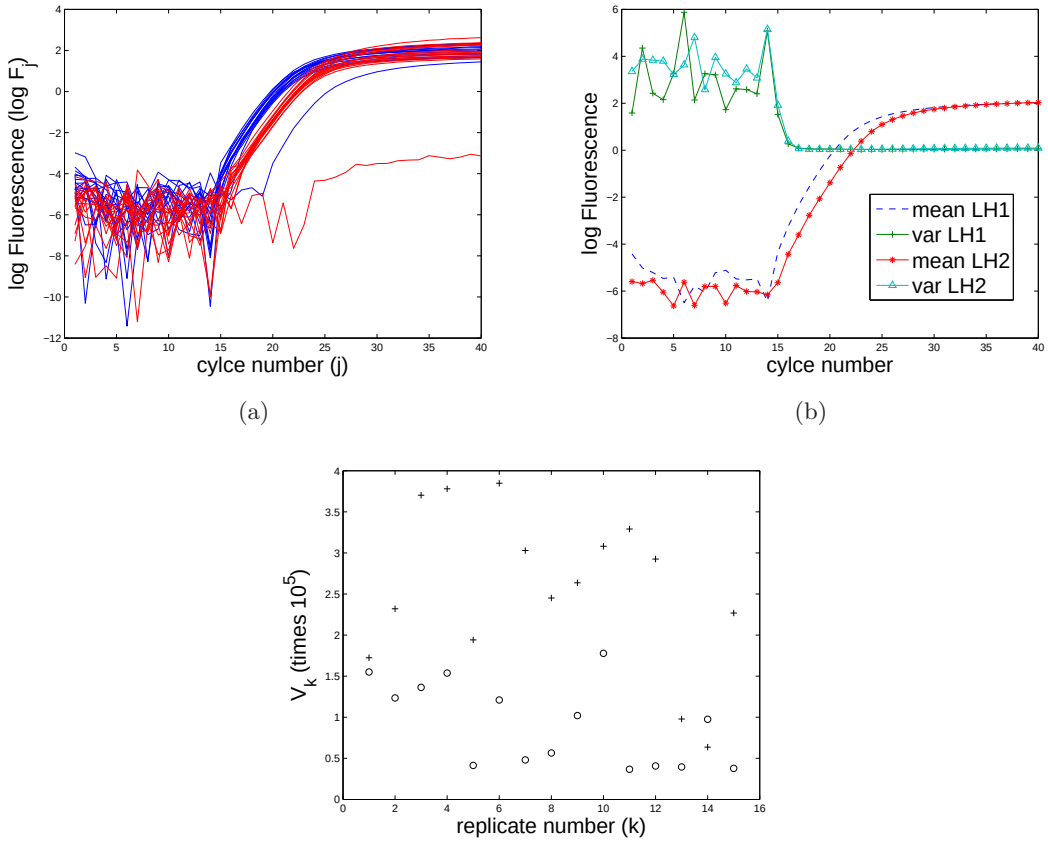
Saha, N., Watson, L. T., Kafadar, K., Ramakrishnan, N., Onufriev, A., Mane, S. P., and Vasquez-Robinet, C. (2007), "Validation and Estimation of Parameters for a General Probabilistic Model of the PCR Process," *Journal of Computational Biology*, 14(1), 97–112.

Skern, R., Frost, P., and Nilsen, F. (2005), "Relative transcript quantification by Quantitative PCR: Roughly right or precisely wrong?," *BMC Molecular Biology*, 6(1), 10.

Speed, T. (2004), Statistics and Gene Expression Analysis,, Technical report, University of California at Berkeley. www.proba.jussieu.fr/bulletin/ArticleSpeed.pdf.

Super-Array (2010), "Technical Note: Does Pipetting Error Affect the Consistency of PCR Array Results?,", www.superarray.com.

Tien, J. H., Lyles, D., and Zeeman, M. L. (2005), "A potential role of modulating inositol 1,4,5-trisphosphate receptor desensitization and recovery rates in regulating ovulation," *Journal of Theoretical Biology*, 232(1), 105 – 117.



(a) (b) (c)

Figure 1: LH experiment. (a) plot of log fluorescence versus cycle number ($\log F_j$ vs. j) for all 16 replicates of LH_1 (in blue) and all 16 replicates of LH_2 (in red); (b) the mean and variance (taken over the 15 replicates, which were retained for the data analysis, at each cycle) of log fluorescence versus cycle number for LH_1 (blue) and LH_2 (red); (c) plot of $\hat{V}_k$ ($\times 10^5$) versus replicate number for LH_1 ($+$) and LH_2 (o).

Luteinizing Hormone

	BP	C_T method	Std. Curve	Adj C_T
$\hat{R}$	2.8221	3.3108	3.5558	3.9567
GCI	[1.8624, 3.7817]	.	.	.
tCI	[1.7719, 3.8722]	.	.	.
BCI	[1.6870, 3.6013]	[2.7935, 3.7477]	[2.8646, 4.1355]	[2.9024, 5.4157]

S.vulgaris

	BP	C_T method	Adj C_T
$\hat{R}$	10.1793	6.3622	303.1662
GCI	[2.8686, 17.4900]	.	.
tCI	[1.7414, 18.6172]	.	.
BCI	[1.0446, 15.7493]	[5.2733, 7.3176]	[29.7583, 892.6178]

Table 1: Summary of data analysis for the luteinizing hormone (LH) and S.vulgaris (SV) experiments. $\hat{R}$ gives the point estimate for relative quantitation; the target value is 2.9505 for the LH experiment and 10 for the SV experiment. All confidence intervals have nominal 95% coverage. BCI refers to the bootstrap confidence interval; GCI refers to the confidence interval based on the asymptotic Gaussian limit; and tCI refers to the confidence interval based on the t distribution. Adj C_T represents the adjusted comparative C_T method.

Model 1

	BP	SC	CT	A-CT
Point mean	10.00	9.92	12.15	10.07
Point var	0.06	0.66	0.82	0.06
B Cov	0.93	0.85	0.37	0.91
B Mean	0.91	2.47	3.44	0.93
G Cov	0.94	.	.	.
G Mean	0.93	.	.	.
t Cov	0.95	.	.	.
t Mean	1.00	.	.	.

Model 2

	BP	SC	CT	A-CT
Point mean	10.03	9.98	12.09	10.06
Point var	0.38	0.14	0.15	0.43
B Cov	0.93	0.85	0	0.93
B Mean	2.28	1.13	1.48	2.50
G Cov	0.94	.	.	.
G Mean	2.34	.	.	.
t Cov	0.95	.	.	.
t Mean	2.49	.	.	.

Model 3

	BP	SC	CT	A-CT
Point mean	10.08	9.89	12.15	9.93
Point var	1.20	0.66	0.83	3.35
B Cov	0.93	0.84	0.38	0.92
B Mean	4.17	2.48	3.47	6.98
G Cov	0.94	.	.	.
G Mean	4.25	.	.	.
t Cov	0.96	.	.	.
t Mean	4.54	.	.	.

Table 2: Summary of the results for the simulation experiment, based on the four estimators: branching process (BP), standard curve (SC), comparative C_T (CT), and adjusted comparative C_T (A-CT). Point mean and Point var give the mean and variance of the point estimates, over the 5000 simulations. The target value for Point mean is 10. All confidence intervals have nominal 95% coverage. Here Cov gives the simulated coverage and mean gives the mean length of the confidence interval over the 5000 simulations. B is for the bootstrap confidence interval; G is for the confidence interval based on asymptotic normality; t is for the confidence interval based on asymptotic normality, using the t distribution.

Supplemental Material for: Inference for Quantitation Parameters in Polymerase Chain Reactions via Branching Processes with Random Effects

Bret Hanlon and Anand N. Vidyashankar

Supplemental material to follow consists of a detailed discussion of the comparative C_T estimators used in the paper, proofs for the asymptotic results stated in the paper, and additional figures for the data analysis (which are shown on the last pages).

APPENDIX A. COMPARATIVE C_T METHODS

In this section we describe the algorithms used in our paper for computing both the unadjusted and adjusted comparative C_T estimators. There are several resources which describe the use of the comparative C_T method for relative quantitation, including ABI (2001), Livak and Schmittgen (2001), Skern et al. (2005), Pfaffl (2006), and Guescini et al. (2008). These works discuss the computation of C_T based on linear interpolation on the log scale. Here we focus on the calculation of relative quantitation, denoted R, assuming C_T has been computed. Our presentation follows the material in Livak and Schmittgen (2001), which coincides with the material in ABI User's Manual.

We begin by describing the (implied) model assumptions for a single reaction. The comparative C_T method assumes deterministic dynamics, with initial exponential growth. Specifically, if N_0 is the number of initial molecules and p is the amplification efficiency, then the number of molecules present after j cycles is denoted N_j and given by $N_j = N_0(1+p)^j$. It is assumed that the observed fluorescence intensity is proportional to the number of molecules so that the model for the fluorescence intensity at cycle j, denoted by F_j, is given by

$$F_j = cN_0(1+p)^j, \tag{A.1}$$

where c is the fluorescence constant. Recall that C_T essentially gives the cycle at which the fluorescence crosses a user-specified threshold, $F^\star$. In fact, the value of C_T is not an integer, but is computed by linear interpolation on the log-scale. The default setting on the ABI machines is $F^\star = .2$; but users can also set the value manually. At "cycle" C_T, (A.1) becomes

$$F^\star = F_{C_T} = cN_0(1+p)^{C_T}. \tag{A.2}$$

It is implicitly assumed that p is constant through the exponential phase, at least until the calculation of C_T. In fact, the comparative C_T method assumes perfect doubling so that $p = 1$; however, the efficiency can instead be estimated from the data. The data is typically noisy for cycles $j < C_T$; thus the first available cycle of "clean" data is $j = \lceil C_T \rceil$. Hence, a natural way to estimate p from the observed fluorescence (remembering that the desired p approximately describes the efficiency through cycle C_T) is via the following estimator

$$\hat{p} = \frac{F_{\lceil C_T \rceil + 1}}{F_{\lceil C_T \rceil}}. \tag{A.3}$$

Building on the above concepts for a single reaction, we now describe the comparative C_T method for computing an estimate of relative quantitation from multiple reactions of a target and calibrator group. The notation is naturally extended by adding a second subscript, T or C, to denote target or calibrator. These methods implicitly assume that all the reactions in the target (or calibrator group) start with the same number of initial molecules, $N_{0,T}$ (or $N_{0,C}$). Accordingly, the method purports to estimate the ratio $R \equiv \frac{N_{0,T}}{N_{0,C}}$. Following ABI (2001), we simply average the C_T values from each replicate to compute a single value for the target and calibrator; similarly we average the efficiency values computed from (A.3). Denote the averaged C_T values for the target and calibrator as $\bar{C}_{T,T}$ and $\bar{C}_{T,C}$, respectively. Similarly, denote the average estimator of efficiency (using (A.3) for each replicate and then averaging across the replicates) as $\hat{p}_T$ for the target group and $\hat{p}_C$ for the calibrator group.

Under the assumption that the fluorescence constant is equal for both the target and calibrator groups, (A.2) suggests the following estimators for R,

$$\hat{R}_1 = \frac{2^{\bar{C}_{T,C}}}{2^{\bar{C}_{T,T}}} = 2^{\bar{C}_{T,C} - \bar{C}_{T,T}}, \quad \hat{R}_2 = \frac{(1+\hat{p}_C)^{\bar{C}_{T,C}}}{(1+\hat{p}_T)^{\bar{C}_{T,T}}}, \tag{A.4}$$

where $\hat{R}_1$ is the (unadjusted) comparative C_T estimator and $\hat{R}_2$ is the adjusted comparative C_T estimator.

APPENDIX B. PROOFS

In this section we present the proofs of our main theorems. Without loss of generality we will assume that $c = 1$ since otherwise all our estimates hold with a factor of c. Under this simplification, $Y_{k,n}$ represents total number of molecules in the reaction during the exponential phase, namely $Y_{k,n} = \sum_{j=\tau}^{n} N_{k,j}$. In the following C (or C_ϵ) denotes a generic constant that could change between successive lines and between successive inequalities.

Proof of Proposition 1. Conditioned on the random effect p_k, $N_{k,n}$ is a Galton Watson process with finite conditional and marginal second moments. The proof then follows using the Toeplitz lemma and Theorem 8.1 of Harris (2002).

The proof of Theorem 1 involves several steps and hence we proceed by proving several lemmas. Our first lemma is concerned with the behavior of the inverse moment raised to the n^{th} power.

Lemma 1. *Under the assumptions of our model,*

$$E(\frac{1}{N_{k,n}^r}) \leq (1 - \frac{1}{2}E(p_1))^n. \tag{A.5}$$

Proof: It is sufficient to consider the case $N_{k,0} = 1$ and $r = 1$, since $N_{k,n}^r \geq N_{k,n}$ for all $r \geq 1$ and $N_{k,n} = \sum_{l=1}^{N_{k,0}} N_{k,n,l} \geq N_{k,n,1}$, where $N_{k,n,j}$ is the number of DNA templates in the n^{th} cycle initiated by the j^{th} template in the 0^{th} cycle of the k^{th} reaction. Now,

$$\begin{aligned} E(\frac{1}{N_{k,n}}|N_{k,0} = 1) &= E(\frac{1}{N_{k,n-1}}E(\frac{1}{N_{k,n-1}^{-1}N_{k,n}}|N_{k,n-1})|N_{k,0} = 1) &\text{(A.6)} \\ &\leq E(\frac{1}{N_{k,n-1}}|N_{k,0} = 1)E(\frac{1}{N_{k,1}}|N_{k,0} = 1), &\text{(A.7)} \end{aligned}$$

where the last step follows using the inequality concerning the arithmetic mean and harmonic mean. Now iterating the above inequality, it follows that

$$E(\frac{1}{N_{k,n}}|N_{k,0} = 1) \leq (E(\frac{1}{N_{k,1}}|N_{k,0} = 1))^n. \tag{A.8}$$

Now, observe that

$$\begin{aligned} E(\frac{1}{N_{k,1}}|N_{k,0} = 1) &= E(E(\frac{1}{N_{k,1}}|N_{k,0} = 1, p_k)) &\text{(A.9)} \\ &= E(1 - p_k + \frac{1}{2}p_k)) = (1 - \frac{1}{2}E(p_1)) < 1, &\text{(A.10)} \end{aligned}$$

where the last inequality follows from $E(p_1) > 0$. This completes the proof of Lemma 1.

Our next lemma is concerned with the bound on the $E(\sqrt{Y_{k,n-1}}(\tilde{m}_{k,n} - m_k))^{2r}$.

Lemma 2. *Under the assumptions of our model, there exists a universal constant C such that*

$$E(\sqrt{Y_{k,n-1}}(\tilde{m}_{k,n} - m_k))^4 \leq Cn^4. \tag{A.11}$$

Proof: We note that

$$\begin{aligned} \sqrt{Y_{k,n-1}}(\tilde{m}_{k,n} - m_k) &= \sqrt{Y_{k,n-1}}(\frac{Y_{k,n} - F_{k,\tau}}{Y_{k,n-1}} - m_k) && \text{(A.12)} \\ &= \sum_{j=\tau}^{n} \frac{N_{k,j+1} - m_k N_{k,j}}{\sqrt{N_k, j}} w_{k,n,j}, && \text{(A.13)} \end{aligned}$$

where

$$w_{k,n,j}^2 = \frac{N_{k,j}}{Y_{k,n-1}}. \tag{A.14}$$

Thus, setting $X_{k,j} = \frac{N_{k,j+1} - m_k N_{k,j}}{\sqrt{N_k, j}}$, we have that

$$\begin{aligned} (\sqrt{Y_{k,n-1}}(\tilde{m}_{k,n} - m_k))^4 &\leq n^4 (\frac{1}{n-\tau} \sum_{j=\tau}^{n-1} |X_{k,j}|)^4 && \text{(A.15)} \\ &\leq n^4 \left(\frac{1}{n-\tau} \sum_{j=\tau}^{n-1} X_{k,j}^4 \right), && \text{(A.16)} \end{aligned}$$

where the last inequality follows from Jensen's inequality for convex functions. Now, conditioned on the random effect p_k and $N_{k,j-1}$, the numerator of $X_{k,j}$ is $Bin(N_{k,j-1}, p_k) - E_k(Bin(Z_{k,j-1}, p_k)|Z_{k,j-1})$. Now, using the formula for the fourth central moment of a binomial random variable, it follows that

$$E_k(X_{k,j}^4 | N_{k,j-1})^4 = N_{k,j-1}^{-1} p_k q_k (3N_{j,k-1} p_k q_k - 6 p_k q_k + 1), \tag{A.17}$$

where $q_k = 1 - p_k$. Hence, since $(1 - \epsilon) \leq p_k \leq 1$, it follows that

$$\begin{aligned} E_k(X_{k,j}^4 | N_{k,j-1}, p_k) &\leq 3(p_k(1-p_k))^2 + 1 && \text{(A.18)} \\ &\leq 3\epsilon^2 + 1. && \text{(A.19)} \end{aligned}$$

Now taking expectation with respect to $N_{k,j-1}$ and with respect to the distribution of the random effect, it follows that

$$E(X_{k,j}^4) \leq 3\epsilon^2 + 1. \tag{A.20}$$

Finally, taking expectation in (A.16), and using (A.20) it follows that

$$E(\sqrt{Y_{k,n-1}}(\tilde{m}_{k,n} - m_k))^4 \leq n^4(1 + 3\epsilon^2). \tag{A.21}$$

This completes the proof of the lemma.

Our next lemma is concerned with the almost sure behavior of $\vee_{k=1}^{r(n)}(\frac{\tilde{m}_{k,n}-1}{m_{k-1}-1} - 1)$.

Lemma 3. *Under the conditions of Theorem 1, it happens with probability one that*

$$\lim_{n\to\infty} \sqrt{r(n)D_1(n)} \max_{1\leq k\leq r(n)} |\frac{\tilde{m}_{k,n} - 1}{m_k - 1} - 1| = 0. \tag{A.22}$$

Proof: It is sufficient to show that for all $\eta > 0$

$$\sum_{n\geq 1} r(n) max_{1\leq k\leq r(n)} P(|\frac{\tilde{m}_{k,n} - m_k}{p_k}| > \frac{\eta}{\sqrt{r(n)D_1(n)}}) < \infty. \tag{A.23}$$

By Markov's inequality,

$$\begin{aligned} P(|\frac{\tilde{m}_{k,n} - m_k}{p_k}| > \frac{\eta}{\sqrt{r(n)D_1(n)}}) &\leq (\frac{\sqrt{r(n)D_1(n)}}{\eta})^2 E|\frac{\tilde{m}_{k,n} - m_k}{p_k}|^2 & \text{(A.24)} \\ &\leq (\frac{\sqrt{r(n)D_1(n)}}{\eta(1-\epsilon)})^2 E|\tilde{m}_{k,n} - m_k|^2 & \text{(A.25)} \\ &\leq (\frac{\sqrt{r(n)D_1(n)}}{\eta(1-\epsilon)})^2 d_n(1)d_n(2), & \text{(A.26)} \end{aligned}$$

where

$$d_n(1) = (E(|\sqrt{Y_{k,n-1}}(\tilde{m}_{k,n} - m_k))^4|)^{1/2}, \text{ and } d_n(2) = E(\frac{1}{Y_{k,n-1}^2})^{1/2}, \tag{A.27}$$

and the last inequality follows by first multiplying and dividing by $\sqrt{Y_{k,n-1}}$ inside the expectation in (A.25) and then applying the Cauchy-Schwarz inequality. Now by Lemma 2, $d_n(1) \leq Cn^2$, where C is a deterministic constant. By Lemma 1, it follows that $E(d_n(2)) \leq C\gamma^n$ where $0 < \gamma < 1$. Thus,

$$P(|\frac{\tilde{m}_{k,n} - m_k}{p_k}| > \frac{\eta}{\sqrt{r(n)}}) \leq C(\frac{\sqrt{r(n)D_1(n)}}{\eta(1-\epsilon)})^2 n^2\gamma^n. \tag{A.28}$$

Thus, it follows from the regularity of the dilution constants and that $r(n)n^{-1} \to 0$ that

$$\begin{aligned} \sum_{n\geq 1} r(n) max_{1\leq k\leq r(n)} P(|\frac{\tilde{m}_{k,n} - m_k}{p_k}| > \frac{\eta}{\sqrt{r(n)}}) &\leq C\sum_{n\geq 1} r^2(n)D_1(n)n^2\gamma^n & \text{(A.29)} \\ &\leq C\sum_{n\geq 1} n^4\gamma^n < \infty, & \text{(A.30)} \end{aligned}$$

where the finiteness is established using the ratio test.

Lemma 4. *Under the conditions of Theorem 1, with probability one*

$$\lim_{n\to\infty}\sqrt{r(n)D_1(n)}\max_{1\le k\le r(n)}|(\frac{m_k^n}{\tilde{m}_k^n}-1)|=0. \tag{A.31}$$

Proof: It is sufficient, using Borel-Cantelli, to show that for any $\eta>0$,

$$\sum_{n\ge 1} r(n)\max_{1\le k\le r(n)} P(|(\frac{\tilde{m}_{k,n}}{m_k})^n-1)|>\frac{\eta}{\sqrt{r(n)D_1(n)}})<\infty. \tag{A.32}$$

We will now obtain estimates on $P(|(\frac{\tilde{m}_{k,n}}{m_k})^n-1)|>\frac{\eta}{\sqrt{r(n)D_1(n)}})$. To this end, it is easy to see that

$$P(|(\frac{\tilde{m}_{k,n}}{m_k})^n-1)|>\frac{\eta}{\sqrt{r(n)D_1(n)}}) \quad = \quad J_n(1)+J_n(2), \tag{A.33}$$

where

$$J_n(1)=P(\tilde{m}_{k,n}-m_k>m_k a_1(n)) \tag{A.34}$$

$$J_n(2)=P(\tilde{m}_{k,n}-m_k<m_k a_2(n)), \tag{A.35}$$

$a_1(n)=(1+\frac{\eta}{\sqrt{r(n)D_1(n)}})^{\frac{1}{n}}-1$ and $a_2(n)=(1-\frac{\eta}{\sqrt{r(n)D_1(n)}})^{\frac{1}{n}}-1$. We will deal with $J_n(1)$ as the proof of the other term is similar. By Markov's inequality,

$$\begin{aligned} J_n(1) &\le E(\frac{E_k(|\tilde{m}_{k,n}-m_k|)}{m_k a_1(n)}) && \text{(A.36)}\\ &\le \frac{1}{(2-\epsilon)a_1(n)}E|m_{k,n}-m_k| && \text{(A.37)}\\ &\le \frac{(E(|\sqrt{Y_{k,n-1}}|\tilde{m}_{k,n}-m_k|)^2)^{\frac{1}{2}}}{(2-\epsilon)a_1(n)}(E(Y_{k,n-1}^{-1}))^{\frac{1}{2}} && \text{(A.38)}\\ &\le \frac{C}{(2-\epsilon)a_1(n)}n^2\gamma^n && \text{(A.39)} \end{aligned}$$

Using the mean value theorem and $r(n)\le n$, one can show that $a_1^{-1}(n)\le Cn^2$. Using this estimate and the ratio test it follows that $\sum_{n\ge 1}J_n(1)<\infty$. A similar calculation for $J_n(2)$ then yields the lemma.

Lemma 5. *Under the conditions of Theorem 1, for l=1, 2, with probability one,*

$$\lim_{n\to\infty}\frac{1}{r(n)D_1(n)}\sum_{k=1}^{r(n)}|\frac{Y_{k,n}}{(1+p_k)^n}-V_k)(\frac{p_k}{1+p_k})|^l=0, \tag{A.40}$$

where $V_k=V_k^\star(\frac{m_k}{m_k-1})$.

Proof: Let $\theta_k = \frac{p_k}{1+p_k}$. We begin by developing an estimate of $Var[(\frac{Y_{k,n}}{(1+p_k)^n} - V_k)\theta_k]$. Using $V_k = V_k^\star \sum_{j\geq 0} m_k^{-j}$ and a change of variables, it follows that

$$\begin{aligned} \frac{Y_{k,n}}{(1+p_k)^n} - V_k &= \sum_{j=0}^{n-\tau}(V_{k,n-j} - V_k^\star)m_k^{-j} - V_k^\star \sum_{j\geq n+1-\tau} m_k^{-j} && \text{(A.41)} \\ &= J_n(1,k) - J_n(2,k) && \text{(A.42)} \end{aligned}$$

Thus,

$$Var[(\frac{Y_{k,n}}{m^n} - V_k)\theta_k] = Var(J_n(1,k)\theta_k) + Var(J_n(2,k)\theta_k) - 2Cov(J_n(1,k)\theta_k, J_n(2,k)\theta_k). \quad \text{(A.43)}$$

Now, setting $S(k,n,j) = \theta_k \sum_{j\geq n+1-\tau} m_k^{-j}$

$$\begin{aligned} Var(J_n(2,k)\theta_k) &= Var(E(V_k^\star S(k,n,j)|p_k)) + E(Var(V_k^\star S(k,n,j)|p_k)) && \text{(A.44)} \\ &\leq E(S^2(k,n,j)(m_a^2 d_k^2 + Var_k(V_k^\star))). && \text{(A.45)} \end{aligned}$$

Now, using $m_k \geq (2-\epsilon)$ and $\theta_k \leq 1$, it follows that $S^2(n,k,j) \leq ((1-\epsilon)(2-\epsilon)^n)^{-1}$. Using this estimate in (A.45) it follows that

$$Var(J_n(2,k)\theta_k) \leq ((1-\epsilon)(2-\epsilon)^n)^{-1}(m_a^2 d_k^2 + \omega_k^2). \quad \text{(A.46)}$$

We next study the behavior of $Var(J_n(1,k)\theta_k)$. Now, using conditioning it follows that

$$Var(J_n(1,k)\theta_k) = E(Var(\sum_{j=0}^{n-\tau}(V_{k,n-j} - V_k^\star)m_k^{-j}\theta_k)|p_k)). \quad \text{(A.47)}$$

Now,

$$Var(\sum_{j=0}^{n-\tau}(V_{k,n-j} - V_k^\star)m_k^{-j}\theta_k|p_k) = J_n(1,1,k) + J_n(1,2,k), \quad \text{(A.48)}$$

where

$$J_n(1,1,k) = \sum_{j=0}^{n-\tau} Var(V_{k,n-j} - V_k^\star|p_k)m_k^{-2j}\theta_k^2, \quad \text{(A.49)}$$

and

$$J_n(1,2,k) = \sum_{j=0}^{n-\tau}\sum_{j\neq l=0}^{n-\tau} \frac{\theta_k^2}{m_k^{j+l}} Cov(V_{k,n-j} - V_k^\star, V_{k,n-l} - V_k^\star|p_k). \quad \text{(A.50)}$$

Using the branching property it follows that,

$$Var(V_{k,n-j} - V_k^\star|p_k) \leq C_\epsilon m_a d_k (2-\epsilon)^{n-j}, \quad \text{(A.51)}$$

where C_ϵ is a finite positive constant independent of k. Now, using this estimate and that $\theta_k \leq 1$ it follows that

$$J_n(1,1,k) \leq C_\epsilon m_a d_k (2-\epsilon)^{-n}. \tag{A.52}$$

Now, we deal with $J_n(1,2,k)$. Using the Cauchy-Schwarz inequality and (A.51) it follows that

$$|Cov(V_{k,n-j} - V_k^\star, V_{k,n-l} - V_k^\star | p_k)| \leq C_\epsilon m_a d_k (2-\epsilon)^{n-(j+l)/2}. \tag{A.53}$$

Using this estimate and $\theta_k \leq 1$ it follows that

$$J_n(1,2,k) \leq C_\epsilon m_a d_k (2-\epsilon)^{-n}. \tag{A.54}$$

Now, combining the estimates for $J_n(1,1,k)$ and $J_n(1,2,k)$ we get

$$Var(J_n(1,k)\theta_k) \leq C_\epsilon d_k (2-\epsilon)^{-n}. \tag{A.55}$$

Again using the Cauchy-Schwarz inequality and $\theta_k \leq 1$, it follows that

$$Cov(J_n(1,k)\theta_k, J_n(2,k)\theta_k | p_k) \leq C_\epsilon (2-\epsilon)^{-n} (C_{1,\epsilon} d_k^2 + C_{2,\epsilon} d_k^3)^{1/2}. \tag{A.56}$$

Thus combining all the estimates, taking expectation with respect to the distribution of p_k, summing over k and using the Cauchy-Schwarz inequality, one can show, using the regularity of the dilution constants, that

$$\sum_{k=1}^{r(n)} Var[(\frac{Y_{k,n}}{m_k^n} - V_k)\theta_k] \leq C_{3,\epsilon} r(n) (2-\epsilon)^{-n}. \tag{A.57}$$

Next, we obtain an estimate of $|E[(\frac{Y_{k,n}}{m_k^n} - V_k)\theta_k]|$. Again, using the decomposition (A.42) and $E(J_n(1,k)\theta_k) = 0$, it follows that

$$|E[(\frac{Y_{k,n}}{m_k^n} - V_k)\theta_k]| \leq |E(\theta_k m_a d_k \sum_{j \geq n+1} m_k^{-j})| \leq C_{4,\epsilon} (2-\epsilon)^{-n} d_k. \tag{A.58}$$

Now, using (A.57), (A.58), and the regularity of the dilution constants it follows that

$$\begin{aligned} E(\sum_{k=1}^{r(n)} \theta_k (\frac{Y_{k,n}}{m_k^n} - V_k))^2 &= \sum_{k=1}^{r(n)} Var[(\frac{Y_{k,n}}{m_k^n} - V_k)\theta_k] + (\sum_{k=1}^{r(n)} E[(\frac{Y_{k,n}}{m_k^n} - V_k)\theta_k])^2 & \text{(A.59)} \\ &\leq C_{5,\epsilon} r(n) (2-\epsilon)^{-n}, & \text{(A.60)} \end{aligned}$$

where $0 < C_{5,\epsilon} < \infty$ is some constant depending on ϵ. Finally, using Markov's inequality and (A.60) it follows that for $l = 1, 2$,

$$P(\frac{1}{r(n)}|\sum_{k=1}^{r(n)}\theta_k(\frac{Y_{k,n}}{(1+p_k)^n} - V_k)|^l > \eta) \leq \frac{1}{\eta^2 r(n)}E|\sum_{k=1}^{r(n)}\theta_k(\frac{Y_{k,n}}{m_k^n} - V_k)|^2 \quad (A.61)$$

$$\leq C_{5,\epsilon}(2-\epsilon)^{-n}. \quad (A.62)$$

Since the RHS of (A.62) is summable, (A.40) follows using the Borel-Cantelli lemma.

Our next lemma is concerned with the moment behavior of the limit random variable $V_k^\star$ when the process is initiated by a single ancestor.

Lemma 6. *Let $N_{k,0} = 1$ for all $k \geq 1$. Then there exists a finite positive constant C such that $E(V_1^{\star 4}) \leq C$.*

Proof: First note that for all k and j $E(V_{k,j}) = 1$. Also using the representation $N_{k,j+1} = N_{k,j} + Bin(N_{k,j}, p_k)$, where $Bin(N_{k,j}, p_k)$ is a binomial random variable (given $N_{k,j}$ and p), one can show that

$$E(V_{k,j}^2) \leq E(V_{k,j-1}^2) + E(\frac{1}{m_k^j}). \quad (A.63)$$

Now, iterating the above and using Tonelli's theorem, it follows that

$$E(V_{k,j}^2) \leq \sum_{l\geq 0} E(\frac{1}{m_k^l}) = E(\frac{1+p_k}{p_k}) \equiv C < \infty. \quad (A.64)$$

We next show that $E(V_{k,j}^3)$ is uniformly bounded. Using the representation of $V_{k,j}$ alluded to above and the uniform boundedness of the first and second moments it follows that

$$E(V_{k,j}^3) \leq E(V_{k,j-1}^3) + E(\frac{C}{m_k^j}). \quad (A.65)$$

The uniform boundedness follows by iteration and summing as before. Now, using the uniform boundedness of $V_{k,j}^3$ and using the fourth moment of a binomial random variable one can show that

$$E(V_{k,j}^4) \leq E(V_{k,j-1}^4) + E(\frac{C}{m_k^j}). \quad (A.66)$$

Iterating and summing, it follows that $E(V_{k,j}^4)$ is uniformly bounded. Now it follows using Jensen's inequality, uniform boundedness of the fourth moment of $V_{k,n}$, and that $V_k^\star - V_{k,n}$ are identically distributed in k that

$$E(V_k^{\star 4}) \leq 4(\sup_{n\geq 1} E(V_{k,n}^4) + E|V_1^\star - V_{1,n}|^4) \leq C + E|V_1^\star - V_{1,n}|^4, \quad (A.67)$$

where C is some finite positive constant. Thus, to complete the proof of the lemma, it is sufficient to show that the second term of the RHS of (A.67) is bounded in n. We will actually show that $E|V_1^\star - V_{1,n}|^4 \to 0$ as $n \to \infty$. To this end, it is sufficient to show that $\{V_{1,n} : n \geq 1\}$ is a Cauchy sequence in L_4 space. Now, using conditioning, the Marcinkiewicz-Zygmund inequality for independent random variables (Chow and Teicher, 1997) and the branching property, it can be seen that

$$\begin{aligned} E(|V_{1,k+n} - V_{1,n}|^4|p_1) &\leq (2\sqrt{2})^4 E(N_{1,n}^{1/2}) m_1^{-4n} E|V_{1,k} - 1|^4 \quad (A.68) \\ &\leq (2\sqrt{2})^4 E|V_{1,k} - 1|^4 (2-\epsilon)^{-7n/2}. \quad (A.69) \end{aligned}$$

Now, using the uniform boundedness of the fourth moments of $V_{1,k}$ and that $0 < \epsilon < 1$, it follows first by taking expectations with respect to the distribution of p_1 and then taking the supremum over k that

$$\sup_{k\geq 1} E|V_{1,k+n} - V_{1,n}|^4 \leq C(2-\epsilon)^{-7n/2}, \quad (A.70)$$

establishing the L_4 convergence of $V_{k,n}$ to $V_k^\star$.

Lemma 7. *Under the conditions of Theorem 1, with probability one,*

$$\lim_{n\to\infty} \frac{1}{r(n)D_1(n)} \sum_{k=1}^{r(n)} V_k^\star = m_a, \quad (A.71)$$

and

$$\frac{1}{\sqrt{r(n)D_1(n)}} \sum_{k=1}^{r(n)} (V_k^\star - m_a d_k) \xrightarrow{d} G_1, \quad (A.72)$$

where $G_1 \sim N(0, \sigma_L^2)$ and σ_L^2 is defined in Theorem 1.

Proof. Note that the random variables $V_k^\star$ are independent with mean $m_a d_k$ and variance ω_k^2. Thus, by regularity of the dilution constants, it follows that

$$\sum_{k\geq 1} \frac{E(V_k - m_a d_k)^2}{k^2} = \sum_{k\geq 1} \frac{\omega_k^2}{k^2} < \infty. \quad (A.73)$$

Hence, by Loeve's generalization of Kolmogorov's laws of large numbers (Chow and Teicher, 1997), it follows that $\frac{1}{r(n)} \sum_{k=1}^{r(n)} V_k^\star$ converges almost surely to m_a. To establish the asymptotic normality, we will verify the Lyapunov condition for independent random variables. To this end, we consider

$E|V_k^\star - m_a d_k|^3$. By the branching property and using $E(N_{k,0}) = m_a d_k$, it follows that

$$E|V_k^\star - m_a d_k|^3 = E|(\sum_{j=1}^{N_{k,0}}(V_{k,j}^\star - 1) + (N_{k,0} - E(N_{k,0})|^3 \quad (A.74)$$

$$\leq 4(E|\sum_{j=1}^{N_{k,0}}(V_{k,j}^\star - 1)|^3 + E|(N_{k,0} - E(N_{k,0})|^3), \quad (A.75)$$

where $V_{k,j}^\star$ are independent random variables (and independent of $N_{k,0}$) with $E(V_{k,j}^\star) = 1$. Now, by first conditioning on $N_{k,0}$ and then using conditional Jensen's inequality it follows that

$$E|\sum_{j=1}^{N_{k,0}}(V_{k,j}^\star - 1)|^3 \leq E(N_{k,0}^3 E(|\frac{1}{N_{k,0}}|\sum_{j=1}^{N_{k,0}}(V_{k,j}^\star - 1)|^3 | N_{k,0})). \quad (A.76)$$

Now, using the independence of $V_{k,j}^\star$ and $N_{k,0}$ and that for each fixed k, $EV_{k,j}^{\star 3} = EV_{k,1}^{\star 3}$, it follows that

$$E|\sum_{j=1}^{N_{k,0}}(V_{k,j}^\star - 1)|^3 \leq E(N_{k,0}^3)E(V_{k,1}^{\star 3}) \quad (A.77)$$

$$\leq CE(N_{k,0}^3) = Cm_{3,0}d_k^3, \quad (A.78)$$

where the last inequality follows from Lemma 6 and the parametrization for the third moment. Hence,

$$(\frac{1}{r(n)})^{\frac{3}{2}}\sum_{k=1}^{r(n)} E|\sum_{j=1}^{N_{k,0}}(V_{k,j}^\star - 1)|^3 \leq C(\frac{1}{r(n)})^{\frac{1}{2}} D_3(n). \quad (A.79)$$

Now, by the regularity of the dilution constants, $\{D_3(n) : n \geq 1\}$ is a bounded sequence. This implies that

$$\lim_{n\to\infty}(\frac{1}{r(n)})^{\frac{3}{2}}\sum_{k=1}^{r(n)} E|\sum_{j=1}^{N_{k,0}}(V_{k,j}^\star - 1)|^3 = 0. \quad (A.80)$$

Now, using the fact that

$$\lim_{n\to\infty}\left(\frac{1}{r(n)D_1(n)}\sum_{k=1}^{r(n)}\omega_k^2\right) = \sigma_L^2, \quad (A.81)$$

the lemma follows.

Proof of Theorem 1. First we express $\tilde{m}_{a,n}$ as

$$\tilde{m}_{a,n} - m_a = T_n(1) + (T_n(2) - m_a), \quad (A.82)$$

where

$$T_n(1) = \frac{1}{r(n)D_1(n)} \sum_{k=1}^{r(n)} \frac{Y_{k,n}}{(1+p_k)^n} (\frac{p_k}{1+p_k})((\frac{\tilde{p}_{k,n}}{p_k})(\frac{(1+p_k)^{n+1}}{(1+\tilde{p}_{k,n})^{n+1}}) - 1), \quad \text{and}$$

$$T_n(2) = \frac{1}{r(n)D_1(n)} \sum_{k=1}^{r(n)} \frac{Y_{k,n}}{(1+p_k)^n} (\frac{p_k}{1+p_k}).$$

We begin with a decomposition for $T_n(2)$ to obtain an expression for $T_n(2) - m_a$.

$$T_n(2) - m_a = T_n(3) + T_n(4), \tag{A.83}$$

where

$$T_n(3) = \frac{1}{r(n)D_1(n)} \sum_{k=1}^{r(n)} (\frac{Y_{k,n}}{m_k^n} - V_k)(\frac{p_k}{m_k}), \text{and} \tag{A.84}$$

$$T_n(4) = \frac{1}{r(n)D_1(n)} \sum_{k=1}^{r(n)} (V_k^\star - m_a d_k). \tag{A.85}$$

Returning to $T_n(1)$ we have

$$|T_n(1)| \leq \max_{1 \leq k \leq r(n)} |(\frac{\tilde{p}_{k,n}}{p_k})(\frac{(1+p_k)^{n+1}}{(1+\tilde{p}_{k,n})^{n+1}} - 1)| T_n(2). \tag{A.86}$$

Now by Lemma 5, $T_n(3)$ converges to zero with probability one and by Lemma 7, $T_n(4)$ converges to 0 with probability one. Combining the results we get that $|T_n(2) - m_a|$ converges to zero with probability one. Also, we obtain the convergence to zero of $|T_n(1)|$ using Lemma 3 and Lemma 4. This yields the strong consistency of $\tilde{m}_{a,n}$. To establish the asymptotic normality, first note that by Lemma 7,

$$(r(n)D_1(n))^{1/2} T_n(4) \xrightarrow{d} N(0, \sigma_L^2). \tag{A.87}$$

Define $\theta_k \equiv \frac{p_k}{1+p_k}$. For any $\eta > 0$, using Chebyshev's inequality

$$\begin{aligned} P(|(r(n)D_1(n))^{1/2} T_n(3)| > \eta) &\leq \frac{1}{\eta^2 r(n) D_1(n)} (E \sum_{k=1}^{r(n)} \theta_k (\frac{Y_{k,n}}{m_k^n} - V_k))^2 & \text{(A.88)} \\ &\to 0, & \text{(A.89)} \end{aligned}$$

where the last convergence follows form (A.60). Finally, using Lemma 3 and Lemma 4, it follows that $(r(n)D_1(n))^{1/2} T_n(1)$ converges to zero in probability. Combining the above, asymptotic normality follows using Slutsky's lemma.

Proof of Theorem 2. Theorem 2 follows by an application of delta method to the function $f(x,y) = \frac{x}{y}$.

Proof of Proposition 2. Conditioned on the random effect, the process $\{N_{k,n} : n \geq 1\}$ is a branching process with offspring distribution $1 + X$, where $X \sim Ber(p_k)$ denotes a Bernoulli random variable with $P(X = 1|p_k) = p_k$. Hence, it follows that

$$\lim_{n\to\infty} P(\sqrt{Y_{k,n-1}}(\tilde{m}_{k,n} - m_k) \leq x|p_k) = P(N(0, p_k(1-p_k)) \leq x|p_k). \tag{A.90}$$

Thus by the bounded convergence theorem, it follows that

$$\lim_{n\to\infty} E(P(\sqrt{Y_{k,n-1}}(\tilde{m}_{k,n} - m_k) \leq x|p_k)) = \int_{1-\epsilon}^{1} \Phi(\frac{x}{t(1-t)})dG(t). \tag{A.91}$$

The proposition follows since the random variables p_k are identically distributed.

Proof of Theorem 3. First we rewrite

$$\begin{aligned} \frac{1}{r(n)}\sum_{k=1}^{r(n)}(\tilde{p}_{k,n} - E(p_1)) &= \frac{1}{r(n)}\sum_{k=1}^{r(n)}(\tilde{p}_{k,n} - p_k) + \frac{1}{r(n)}\sum_{k=1}^{r(n)}(p_k - E(p_1)) &\text{(A.92)}\\ &= T_n(1) + T_n(2) &\text{(A.93)} \end{aligned}$$

and verify that $T_n(1) \to 0$ with probability 1. Now by Chebyshev's inequality and the independence of $(\tilde{p}_{k,n} - p_k)$ in k,

$$\begin{aligned} P(|T_n(1)| > \eta) &\leq \eta^{-2}E(T_n^2(1)) = \eta^{-2}(Var(T_n(1)) + (E(T_n(1)))^2) &\text{(A.94)}\\ &\leq \frac{C}{r^2(n)}(\sum_{k=1}^{r(n)}(E(\tilde{p}_{k,n} - p_k)^2 + (E(\tilde{p}_{k,n} - p_k)^2)^{1/2}), &\text{(A.95)} \end{aligned}$$

where the last inequality follows by bounding the variance term by the second moment and using the Cauchy-Schwarz inequality on the expectation term. Now,

$$\begin{aligned} E(\tilde{p}_{k,n} - p_k)^2 &= E(\tilde{m}_{k,n} - m_k)^2 &\text{(A.96)}\\ &= E((\tilde{m}_{k,n} - m_k)^2 Y_{k,n} Y_{k,n}^{-1}) &\text{(A.97)}\\ &\leq (E(\tilde{m}_{k,n} - m_k)^4 Y_{k,n}^2))^{1/2}(E(Y_{k,n}^{-2}))^{1/2}, &\text{(A.98)} \end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality. Now applying Lemma 1 and Lemma 2 it follows that for some $0 < C < \infty$ and $0 < \gamma < 1$

$$E(\tilde{p}_{k,n} - p_k)^2 \leq Cn^2\gamma^{n/2}. \tag{A.99}$$

Now using this estimate in (A.95) it follows that $P(|T_n(1)| > \eta)$ is bounded above by $Cn^2\gamma^{n/4}$. By ratio test, the above probability sums there by yielding the almost sure convergence to 0 of $T_n(1)$. Since $T_n(2)$ is a sum of i.i.d. random variables with finite second moments, the theorem follows via the law of large numbers and central limit theorem for i.i.d. random variables.

Proof of Theorem 4. The estimator of variance can be expressed as

$$\tilde{\sigma}^2_{L,n} = \frac{1}{r(n)D_1(n)} \sum_{k=1}^{r(n)} (T_n(1,k) + T_n(2,k) + T_n(3,k))^2, \tag{A.100}$$

where

$$T_n(1,k) = (\frac{Y_{k,n}\tilde{p}_{k,n}}{\tilde{m}^{n+1}_{k,n}} - V^\star_k), \tag{A.101}$$

$$T_n(2,k) = V^\star_k - m_a d_k \quad \text{and} \quad T_n(3,k) = (m_a - \tilde{m}_{a,n} d_k). \tag{A.102}$$

One can show using the Cauchy-Schwarz inequality and Lemma 5 that the cross-product terms in the expansion of (A.100) converge to zero with probability one. Furthermore, normalized sums of squares of $T_n(3,k)$ converges to zero with probability one by regularity of the dilution constants and strong consistency of $\tilde{m}_{a,n}$. Also the normalized sums of squares of $T_n(1,k)$ converges to zero by Lemma 5. Finally, by using the arguments in Lemma 7 and the regularity of the dilution constants it follows that normalized sums of squares $T_n(2,k)$ converges to σ^2_L. This yields the strong consistency of $\tilde{\sigma}^2_{L,n}$. Strong consistency of $\tilde{\theta}_{1,n}$ and $\tilde{\theta}_{2,n}$ follow from Lemma 5 and the strong law of large numbers for i.i.d. random variables $(1-p_k)^{-1}(1+p_k)$.

Proof of Corollary 1. The proof follows from the strong consistency of $\sigma^2_{L,n}$ $\tilde{m}_{a,n}$, $\tilde{\theta}_{1,n}$, the regularity of the dilution constants , and the definition of σ^2_L.

REFERENCES

ABI (2001), *ABI Prism 7700 Sequence Detection System.User Bullentin Number 2* PE Applied Biosystems.

Chow, Y. S., and Teicher, H. (1997), *Probability theory*, Springer Texts in Statistics, third edn, New York: Springer-Verlag.

Guescini, M., Sisti, D., Rocchi, M., Stocchi, L., and Stocchi, V. (2008), "A new real-time PCR method to overcome significant quantitative inaccuracy due to slight amplification inhibition," *BMC Bioinformatics*, 9(1), 326.

Harris, T. E. (2002), *The Theory of Branching Processes*, second edn, New York: Dover Publications.

Livak, K. J., and Schmittgen, T. D. (2001), "Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2-[Delta][Delta]CT Method," *Methods*, 25(4), 402 – 408.

Pfaffl, M. W. (2006), *A new real-time PCR method to overcome significant quantitative inaccuracy due to slight amplification inhibition* In. Real-time PCR (ed. T. Dorak).

Skern, R., Frost, P., and Nilsen, F. (2005), "Relative transcript quantification by Quantitative PCR: Roughly right or precisely wrong?," *BMC Molecular Biology*, 6(1), 10.

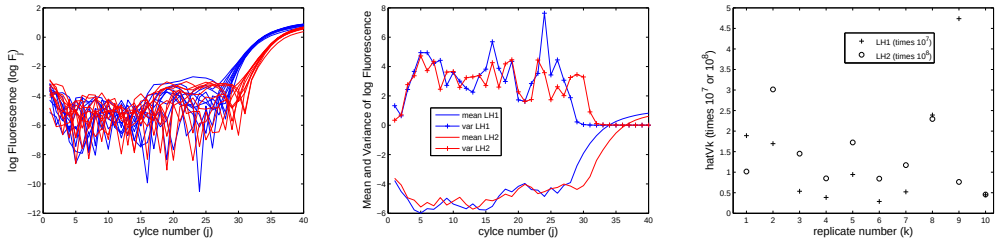


Figure 1: S. Vulgaris experiment. (L) plot of log fluorescence versus cycle number ($\log F_j$ vs. j) for all 10 replicates of SV_1 (in blue) and all 10 replicates of SV_2 (in red); (C) the mean and variance (taken over the 10 replicates at each cycle) of log fluorescence versus cycle number for SV_1 (blue) and SV_2 (red); (R) plot of $\hat{V}_k$ ($\times$ factor) versus replicate number for SV_1 ($+$, factor $= 10^7$) and SV_2 (o, factor $= 10^8$).