Biomolecular Detection and Quantification 9 (2016) 1–13

Contents lists available at ScienceDirect

Biomolecular Detection and Quantification

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

Research Paper

Flexible analysis of digital PCR experiments using generalized linear mixed models

Matthijs Vynck[a,*], Jo Vandesompele[b,c,d], Nele Nijs[d], Björn Menten[b,c], Ariane De Ganck[d], Olivier Thas[a,e]

[a] Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Coupure Links 653, 9000 Ghent, Belgium
[b] Center for Medical Genetics, Ghent University, De Pintelaan 185, 9000 Ghent, Belgium
[c] Bioinformatics Institute Ghent N2N, Ghent University, De Pintelaan 185, 9000 Ghent, Belgium
[d] Biogazelle, Technologiepark 3, 9052 Zwijnaarde, Belgium
[e] National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong, NSW 2522, Australia

ARTICLE INFO

Article history:
Received 11 December 2015
Received in revised form 2 June 2016
Accepted 7 June 2016

Keywords:
Digital PCR
Statistics
Data analysis
Mixed models
Replicates
Quantification

ABSTRACT

The use of digital PCR for quantification of nucleic acids is rapidly growing. A major drawback remains the lack of flexible data analysis tools. Published analysis approaches are either tailored to specific problem settings or fail to take into account sources of variability. We propose the generalized linear mixed models framework as a flexible tool for analyzing a wide range of experiments. We also introduce a method for estimating reference gene stability to improve accuracy and precision of copy number and relative expression estimates. We demonstrate the usefulness of the methodology on a complex experimental setup.

1. Introduction

The number of publications on digital PCR (dPCR) have markedly increased during the last decade, with a rapid growth of publications in the field of biomedical sciences in recent years. This adoption has in part been possible due to an increase of commercially available, user-friendly instruments [1,2] and is further stimulated by positive reports on dPCR demonstrating the advantages over quantitative PCR (qPCR) [3], particularly for applications such as low-level quantification [4,5], absolute quantification [4,5] and copy number variation (CNV) determination [6].

Despite the advantages and increasing popularity of dPCR and as a consequence of the technique still being in its infancy, one major drawback of dPCR remains the lack of dedicated data analysis tools taking full advantage of the specific digital nature of the data. Most published papers rely on data-analysis software provided by hardware manufacturers. These software suites are typically black-box tools providing the user with a limited amount of information on the algorithms. They furthermore do not allow the user to analyze more complicated experimental setups such as the correct use of technical replicates or the use of multiple reference loci for determining CNVs, even though such approaches may be advisable [7–9].

Although several papers have been published that propose data analysis methods, these methods have been developed to analyze very specific experimental setups. For example, Whale et al. [6] and Dube et al. [10] developed ad hoc methods for calculating CNVs, but these methods can only be used to calculate CNVs using a single reference locus and do not take into account interreplicate variability. Extending these methods to cope with other experimental setups would require significant work, tailored to each of these specific designs. A major difficulty is the correct estimation of standard errors and confidence intervals.

In this paper, we detail how the established generalized linear mixed model (GLMM) framework [11] can be used to analyze dPCR data from a wide range of experimental setups, ranging from simple experiments such as absolute quantification to complicated studies such as CNV estimation with multiple reference loci normalization and handling of variable numbers of technical replicates, while correctly accounting for various sources of variability. The basis of this GLMM framework has recently also been described by Dorazio and Hunter [12]. We argue that known sources of variability should be

* Corresponding author.
E-mail address: matthijs.vynck@ugent.be (M. Vynck).

http://dx.doi.org/10.1016/j.bdq.2016.06.001

accounted for and that the approach of pooling counts of technical replicates used for analysis by Dorazio and Hunter [12] (among others, e.g. Yu et al. [13]) may lead to incorrect estimation of standard errors and confidence intervals.

Further, a novel approach for selecting stable reference loci for CNV studies from a pool of candidate reference loci is developed and successfully applied. An approach for reference gene selection in relative expression experiments is also suggested.

To demonstrate the flexibility of the approach, our methodology is used to analyze a dataset consisting of droplet digital PCR (ddPCR) data for 14 individuals who have been screened for chromosomal abnormalities using 14 genes on 6 chromosomes. The performance in terms of accuracy and precision is evaluated for calculating CNVs using both a single reference locus and multiple reference loci.

2. Materials and methods

2.1. Absolute quantification

dPCR splits a sample mixture into partitions. Each of these partitions is subsequently called as containing target nucleic acid, or having no target nucleic acid. A positive signal thus indicates that one or more target copies may be present. As a consequence of the random partitioning of copies, the number of copies in a partition is assumed to follow a Poisson distribution with parameter λ which has the interpretation of the average number of copies per partition. If Y_j^* denotes the unobserved number of copies in partition j ($j = 1, \ldots, J$, with J the number of partitions), then we can write the observed digital outcome as the binary variable Y_j:

$$Y_j = \min(Y_j^*, 1) = \begin{cases} 0 & \text{if } Y_j^* = 0 \\ 1 & \text{otherwise.} \end{cases} \tag{1}$$

Having observed the digital outcomes, the λ parameter of the Poisson distribution can be estimated from the probability of zero copies, relying on the probability mass function of the Poisson distribution (Eqs. (2) and (3)):

$$\mathrm{P}\{Y_j^* = 0\} = \frac{\lambda^0}{0!}\exp(-\lambda) = \exp(-\lambda) \tag{2}$$

$$\lambda = -\log \mathrm{P}\{Y_j^* = 0\} = -\log \mathrm{P}\{Y_j = 0\} \tag{3}$$

The final equality in Eq. (3) follows from the construction of the binary outcomes (Eq. (2)). Since a probability of a binary event can be estimated from simple counts, an estimate of λ is given by

$$\hat{\lambda} = -\log\left(\frac{\text{number of negative partitions}}{\text{total number of partitions}}\right). \tag{4}$$

$\hat{\lambda}$ can also be obtained using a Generalized Linear Model (GLM). The GLM for the unobserved counts Y_j^* is specified by a Poisson distribution with mean λ related to a parameter β_0 through a log-link function,

$$\log \lambda = \beta_0. \tag{5}$$

Using Eq. (3), the observed binary outcomes Y_j can be described by a binomial distribution with probabilities

$$\begin{aligned} \mathrm{P}\{Y_j = 0\} &= \mathrm{P}\{Y_j^* = 0\} = \exp(-\lambda) \\ &= \exp(-\exp(\beta_0)) \\ \mathrm{P}\{Y_j = 1\} &= \mathrm{P}\{Y_j^* > 0\} = 1 - \exp(-\lambda) \\ &= 1 - \exp(-\exp(\beta_0)). \end{aligned} \tag{6}$$

Eqs. (6) state a GLM for a binomial distribution with a complementary log-log link. The more conventional model formulation is:

$$\log(-\log(\mathrm{P}\{Y_j = 0\})) = \beta_0, \tag{7}$$

where β_0 is the same as in Eq. (5). Since the digital outcomes Y_j are observed, GLM software can be used for estimating β_0. If $\hat{\beta}_0$ denotes the estimate, an estimate of λ is then given by

$$\hat{\lambda} = \exp(\hat{\beta}_0). \tag{8}$$

Using Eq. (4) or Eq. (8) will result in the same estimate for λ.

Assuming a constant volume of the partitions, say $V_{\text{partition}}$, the concentration can be estimated from the average number of copies per partition (Eq. (9)):

$$\hat{c} = \frac{\hat{\lambda}}{V_{\text{partition}}}. \tag{9}$$

To obtain a reliable estimate of the concentration, an experiment is typically replicated. We now define Y_{ij}^* as the number of copies in partition j of replicate i ($j = 1, \ldots, J_i$, with J_i the number of partitions in replicate i, $i = 1, \ldots, I$, with I the number of replicates). As before, the counts are not observable, but upon applying equation (1), binary outcomes Y_{ij} can be calculated. To take the replicate variability into account, we introduce a random effect for the replicate in the Poisson model. Within a replicate, the counts are still Poisson distributed. The statistical model is formulated hierarchically. In particular, within a replicate:

$$Y_{ij}^* \mid R_i \sim \text{Poisson}(\lambda_i) \tag{10}$$

where

$$\log \lambda_i = \beta_0 + R_i, \tag{11}$$

with R_i the effect of replicate i on the Poisson mean. These replicate effects R_i are described by a normal distribution,

$$R_i \sim N(0, \sigma^2). \tag{12}$$

This model implies that the random effect terms are exchangeable, which is warranted if replicates are considered as a random sample from a larger population of potential replicates (see Supplementary Material 4, Section 4).

The model results again in a binomial regression model with a complementary log-log link for the observed digital outcomes. In particular, within a replicate

$$\log(-\log(\mathrm{P}\{Y_{ij} = 0 \mid R_i\})) = \beta_0 + R_i, \tag{13}$$

with β_0 and R_i as before. The model is a special case of a GLMM [11]. Statistical software is available for estimating the model parameters (e.g. R [14], an environment often used for analysis of PCR experiments [15]), including random effect variances [16].

The objective is to estimate the mean number of copies, averaged over all replicates, i.e. $\mathrm{E}\{Y_{ij}^*\}$ is the quantity of interest for absolute quantification. Statistical theory (Supplementary Material 4, Section 1) gives

$$\mathrm{E}\{Y_{ij}^*\} = \exp(\beta_0 + 0.5\sigma^2). \tag{14}$$

From the estimate of β_0 (say $\hat{\beta}_0$), the estimate of the variance σ^2 of the random effect (say $\hat{\sigma}^2$) and from Eq. (9) a concentration estimate can subsequently be calculated as

$$\hat{c} = \frac{\exp(\hat{\beta}_0 + 0.5\hat{\sigma}^2)}{V_{\text{partition}}}. \tag{15}$$

The statistical software also gives the estimated standard errors of the estimates $\hat{\beta}_0$ which can be used for the calculation of an approximate confidence interval of the concentration