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
Available online 24 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

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

http://dx.doi.org/10.1016/j.bdq.2016.06.001
2214-7535/

GLMM framework has recently also been described by Dorazio and Hunter [12]. We argue that known sources of variability should be 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:

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

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

(Supplementary Material 4, Section 3). Example analyses are given in Sections 2.1 and 3.2 of Supplementary Material 1.

2.2. Copy number variation

For the estimation of CNV, data on both a target and at least one reference must be available. Several experimental designs are appropriate for obtaining target and reference measurements. Fig. 1 shows six examples, ranging from single reference settings with single channel experiments (panel A) or duplex experiments (panel B) to multiple reference studies with single channel (panel C) or duplex (panel D, E) or multiplex (panel F) experiments. In this section, a GLMM methodology is outlined that is applicable to all of these designs, also in the presence of replicates.

The general guideline for obtaining valid statistical estimation, error propagation and hypothesis testing, is that the data analysis method should account for dependencies and sources of variability implied by the experimental setup. For example, as in Section 2.1, random replicate effects should be included in the model to take care of the dependence between droplets within the same replicate.

2.2.1. Single reference designs

The same notation (Y_{ij} and Y_{ij}^*) as before is used, but the partition index j may now refer to a measurement which can be from a target or a reference. The distinction between target and reference is made by a dummy regressor X_{ij} which is defined as zero when partition (i, j) comes from the target and one when it comes from the reference. For the designs A and B (Fig. 1), the model for the unobservable number of copies is written as

$$Y_{ij}^* \mid R_i \sim \text{Poisson}(\lambda_{ij}) \tag{16}$$

where

$$\log \lambda_{ij} = \beta_0 + X_{ij}\beta_1 + R_i \tag{17}$$

and

$$R_i \sim N(0, \sigma^2). \tag{18}$$

Thus within replicate i, the mean number of target copies per partition again equals $\exp(\beta_0 + 0 \times \beta_1 + R_i)$, and the mean number of reference copies per partition equals $\exp(\beta_1 + 1 \times \beta_1 + R_i)$.

Let $c_{\text{target},i}$ and $c_{\text{ref},i}$ denote the concentrations of target and reference in replicate i, respectively, and N_b the ploidy of the organism. For design A (Fig. 1), the CNV based on replicate i for the target and replicate i' for the reference, is given by

$$\begin{aligned}\text{CNV}_{i,i'} &= \frac{c_{\text{target},i}}{c_{\text{ref},i'}} N_b = \frac{\exp(\beta_0 + 0 \times \beta_1 + R_i)/V_{\text{partition}}}{\exp(\beta_1 + 1 \times \beta_1 + R_{i'})/V_{\text{partition}}} N_b \\ &= \exp(-\beta_1 + R_i - R_{i'})N_b. \end{aligned} \tag{19}$$

The overall CNV is then given be the average of $\text{CNV}_{i,i'}$ over all replicates (see Supplementary Material 4, Section 2 for details), resulting in

$$\text{CNV} = \text{E}\{\text{CNV}_{i,i'}\} = \exp(-\beta_1 + \sigma^2)N_b. \tag{20}$$

As before, the model parameters may be estimated by reformulating the model for the digital outcomes Y_{ij}. In particular, a GLMM with a complementary log-log link is obtained:

$$\log(-\log(\text{P}\{Y_{ij} = 0 \mid R_i\})) = \beta_0 + X_{ij}\beta_1 + R_i, \tag{21}$$

with β_0, β_1 and R_i as in model (17).

For design B (Fig. 1), the CNV based on replicate i, which now contains droplets with both target and reference (duplex), is given by

$$\begin{aligned}\text{CNV}_i &= \frac{c_{\text{target},i}}{c_{\text{ref},i}} N_b = \frac{\exp(\beta_0 + 0 \times \beta_1 + R_i)/V_{\text{partition}}}{\exp(\beta_1 + 1 \times \beta_1 + R_i)/V_{\text{partition}}} N_b \\ &= \exp(-\beta_1)N_b. \end{aligned} \tag{22}$$

Note that the random effect cancels out and that the CNV does not depend on i. Hence, an overall CNV estimate is given by $\exp(-\hat{\beta}_1)N_b$, with the estimates again calculated from the GLMM with a complementary log-log link. The random effect can however not be omitted altogether, as it influences the variance on the fixed effect parameters, and thus the inclusion of the random effect is essential for a correct error propagation.

2.2.2. Multiple reference designs

The model can be further extended to contain multiple reference loci. The number of copies and the deduced binary outcome for partition (i, j) are denoted by Y_{ijk}^* and Y_{ijk}, respectively, in which the index k refers to the reference $k = 1, \ldots, K$, with K the number of reference loci and with $k = 0$ referring to the target. Consider the dummy X_{ijk}, which is defined as one when the signal belongs to the kth reference and zero when the signal comes from the target. Reference-to-reference differences are allowed by making use of nested random effects.

For designs C and D, for a given replicate i and for a given target or reference k, the Poisson model for the unobserved counts Y_{ijk}^* has log-mean

$$\log \text{E}\{Y_{ijk}^* \mid S_k, R_{i(k)}\} = \log \lambda_{ijk} = \beta_0 + \beta_1 X_{ijk} + S_k X_{ijk} + R_{i(k)} \tag{23}$$

with S_k the effect of reference k on the log-mean, and $R_{i(k)}$ the effect of the ith replicate of the experiment with the PCR mix containing reference k (or $k = 0$ for target in design C). The variability of these two random effects are described by independent normal distributions:

$$S_k \sim N(0, \sigma_1^2) \quad \text{and} \quad R_{i(k)} \sim N(0, \sigma_2^2). \tag{24}$$

Hence, S_k is a random effect for the between reference locus variation and $R_{i(k)}$ is a random effect for the interreplicate variation nested within a given target or reference. Note that the model formulation assumes that the random effects are exchangeable (see Supplementary Material 4, Section 4 for more information).

The same model applies to design E, except that the index k in $R_{i(k)}$ should be replaced by an index k^* which is an indicator of the unique PCR mix (each row in panel E of Fig. 1 represents a unique PCR mix). The model for design F is also similar, except that the replicate effect $R_{i(k)}$ does not depend on reference k, because in this multiplex experiment all references are potentially included in all partitions, i.e. in each replicate all references are included in the PCR mix. Hence, the nested random effect $R_{i(k)}$ in model (23) has to be replaced by R_i.

As before, the model parameters can be estimated from the corresponding GLMM for the binary outcome:

$$\log(-\log(\text{P}\{Y_{ijk} = 0 \mid S_k, R_{i(k)}\})) = \beta_0 + \beta_1 X_{ijk} + S_k + R_{i(k)}. \tag{25}$$

For design C the CNV is first given for target versus a single reference k, based on replicates i and i':

$$\begin{aligned}\text{CNV}_{i,i';k} &= \frac{\exp(\beta_0 + R_{i(0)})}{\exp(\beta_1 + \beta_1 + S_k + R_{i'(k)})} N_b \\ &= \exp(-\beta_1 - S_k + R_{i(0)} - R_{i'(k)})N_b. \end{aligned} \tag{26}$$

The overall CNV is obtained by averaging over all replicates and all references (see Supplementary Material 4, Section 2 for details):

$$\text{CNV} = \text{E}\{\text{CNV}_{i,i';k}\} = \exp\left(-\beta_1 + \frac{1}{2}\sigma_1^2 + \sigma_2^2\right) N_b. \tag{27}$$

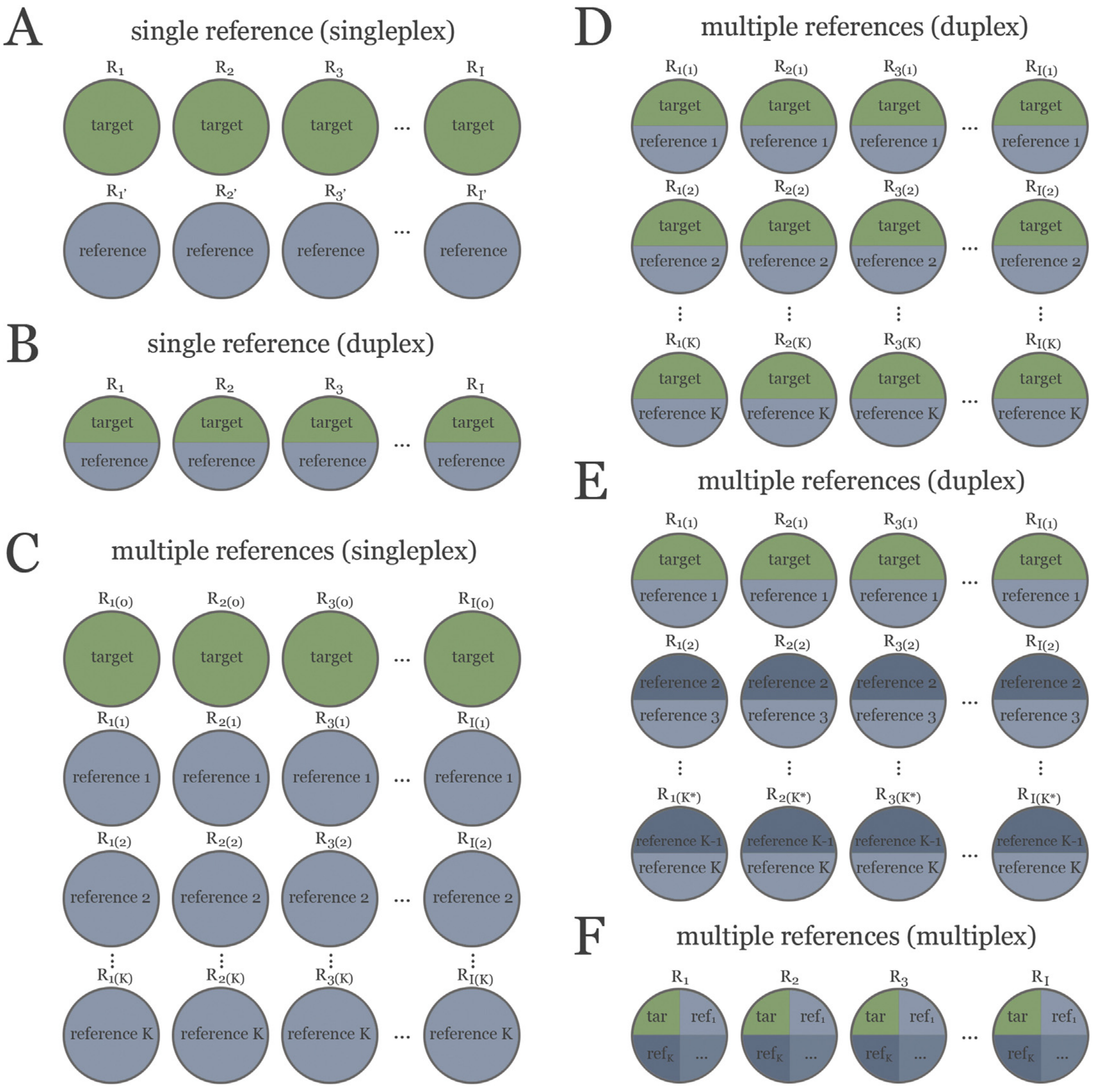

Fig. 1. Experimental designs for calculating copy numbers using digital PCR. (A) Singleplex experiments with a single reference gene. (B) Duplex experiments with a single reference gene. (C) Singleplex experiments with multiple reference genes. (D) Duplex experiments with multiple reference genes and repeated analysis of the target gene. (E) Efficient duplex experiments with multiple reference genes. (F) Multiplex experiments with multiple reference genes. For each design, a single droplet is shown for each replicate. The circles represent partitions and they show whether a target or a reference is included in either singleplex, duplex or multiplex. The replicates are indicated with the letter *R*, the total number of replicates is given by I. Setup-specific subscripts (prime, asterix) as used in the model specifications of Section 2.

For design D, which allows for CNV calculation w.r.t. a single reference *k* within a replicate *i* (duplex), we first give

$$\mathrm{CNV}_{ik} = \frac{\exp\left(\beta_0 + R_{i(k)}\right)}{\exp\left(\beta_1 + \beta_1 + S_k + R_{i(k)}\right)} N_b = \exp(-\beta_1 - S_k) N_b. \qquad (28)$$

Note that the replicate effect has been eliminated (due to combining target and reference *k* in the duplex). After averaging over references, the overall CNV becomes

$$\mathrm{CNV} = \mathrm{E}\{\mathrm{CNV}_{ik}\} = \exp\left(-\beta_1 + \frac{1}{2}\sigma_1^2\right) N_b. \qquad (29)$$

For design E, the CNV w.r.t. a single reference measured in duplex with the target is given by equation (28). CNVs for the other $k-1$ references measured in duplex with respect to one another are given by equation (26). An overall CNV is obtained as

$$\mathrm{CNV} = \exp\left(-\beta_1 + \frac{1}{2}\sigma_1^2\right) N_b \frac{(1 + \exp(\sigma_2^2)(k-1))}{k}. \qquad (30)$$

For design F (multiplex) we also start with the CNV calculation w.r.t. a single reference *k* within a replicate *i*:

$$\mathrm{CNV}_{ik} = \frac{\exp(\beta_0 + R_i)}{\exp(\beta_1 + \beta_1 + S_k + R_i)} N_b = \exp(-\beta_1 - S_k) N_b. \qquad (31)$$

Also here the replicate effect is eliminated. After averaging over references, the overall CNV becomes

$$\mathrm{CNV} = \mathrm{E}\{\mathrm{CNV}_{ik}\} = \exp\left(-\beta_1 + \frac{1}{2}\sigma_1^2\right) N_b. \qquad (32)$$

Detailed derivations are available in Supplementary Material 4. Example analyses in R are given in Sections 2.2 and 3.3 of Supplementary Material 1.

2.3. *Reference gene stability*

Reference locus stability for CNV estimation is calculated for each reference locus *k* (*k* = 1, . . ., *K*) as

$$\mathrm{Stability}_k = \frac{\sum_{l \in S_k} \left(\hat{\beta}_{1kl}^2 + \mathrm{Var}(\hat{\beta}_{1kl})\right)}{K-1} \qquad (33)$$

where S_k is the set $\{1, \ldots, K\} \setminus k$. If only data from one sample are available, the $\hat{\beta}_{1kl}$ are parameter estimates that result from fitting the model

$$\log \lambda_{ij} = \beta_{0kl} + \beta_{1kl} X_{ij} + R_i \tag{34}$$

with $R_i \sim N(0, \sigma^2)$ and X_{ij} defined as zero when partition (i, j) comes from reference k and one when it comes from reference l. For fitting model (34) only the data from references k and l are used. An example analysis is given in Section 2.3 of Supplementary Material 1.

For across-sample stability in the CNV case, the model needs to account for variability between samples: similarly to the introduction of the between-reference gene random effect in Section 2.2, we introduce a between-sample random effect S_s that accounts for the sample-specific effects. Consequently, model (34) is replaced by

$$\log \lambda_{ij} = \beta_{0kl} + \beta_{1kl} X_{ij} + R_{i(s)} + S_s \tag{35}$$

with $R_{i(s)} {\sim} N(0, \sigma_1^2)$, $S_s {\sim} N(0, \sigma_2^2)$ and X_{ij} defined as before. Again only the data from references k and l are used.

The rationale for this stability measure is based on two arguments. First, when given a set of candidate reference genes in a CNV setup, we expect these reference genes to have similar copy numbers. If two candidate reference genes, say k and l, are in agreement (e.g. they both show a close to diploid copy number in e.g. a human sample), the estimate $\hat{\beta}_{1kl}$ is expected to be close to zero, or, equivalently, the ratio of their estimated concentrations will be close to one (Eqs. (20) and (22)). Thus, a reference k that gives a large $\sum_{l \in S_k} \hat{\beta}_{1kl}^2$ is said to be a biased reference.

Second, a good reference gene will be stable across replicates or samples, warranting the inclusion of the $\mathrm{Var}(\hat{\beta}_{1kl})$ estimate: if copy numbers across replicates or samples are highly variable, this will be reflected in a large variability of $\hat{\beta}_{1kl}$ (i.e. a large $\mathrm{Var}(\hat{\beta}_{1kl})$). Thus, references that are highly variable are less ideal and will be penalized. A detailed discussion is presented in Section 3.2.

To allow for different quantities of reference genes in a relative expression scenario (see detailed discussion in Section 3.2), model (35) is used, but the relative expression gene stability is simplified to

$$\mathrm{Stability}_k = \frac{\sum_{l \in S_k} \mathrm{Var}\left(\hat{\beta}_{1kl}\right)}{K - 1} \tag{36}$$

i.e. only taking variability into account.

2.4. Generic formulation

Generally, the model can be written as a combination of fixed effect parameters – such as those of target and reference genes but also confounders or baseline variables, for example, age and gender – and random effect parameters such as interreplicate variation, interreference gene variation but also to account for e.g. variation between multiple laboratories. The general model can be written as

$$\log \lambda_{.} = \beta_0 + \beta_1 X_{1.} + \cdots + \beta_p X_{p.} + Z_{1.} + \cdots + Z_{q.} \tag{37}$$

for p fixed effects and q random effects, where the subscripts which are here denoted by $\cdot$, refer to the variables used for constructing the X and Z variables. The model can further be extended to allow for interactions, and random effects may be nested if implied by the study design.

One example of such an extension could be relative concentration estimation with multiple patients and with an effect of gender on the target concentration:

$$\log \lambda_{ij} = \beta_0 + X_{1ij}\beta_1 + X_{2ij}X_{3ij}\beta_2 + Z_{1i} \tag{38}$$

for partition j obtained from patient i, X_{1ij} is one if partition j of patient i is from the reference gene and zero if it is from the target gene, X_{2ij} is zero if patient i is male and one if the patient is female and X_{3ij} is zero for reference genes and one for target genes. The CNV for a male remains as in Eq. (22), but for females it becomes:

$$\widehat{\mathrm{CNV}} = \frac{\hat{c}_{\mathrm{target}}}{\hat{c}_{\mathrm{ref}}} N_b = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_2)}{\exp(\hat{\beta}_0 + \hat{\beta}_1)} N_b = \exp(\hat{\beta}_2 - \hat{\beta}_1) N_b. \tag{39}$$

Similarly, random effects can be added to account for e.g. stratification in multicenter trials. Absolute concentration estimation with an age effect, a patient effect and a center effect, could be modelled as:

$$\log \lambda_{ijc} = \beta_0 + X_{1ijc}\beta_1 + Z_{1c} + Z_{2i(c)} \tag{40}$$

where for center c, patient i and partition j, β_1 is the age effect, Z_{1c} is the random center effect (with variance σ_1^2) and $Z_{2i(c)}$ is the random patient effect (nested within the center effect, with variance ${\sigma_2}^2$). Using results similar as in Eq. (14), the mean concentration for a patient of age *age* can then be estimated as

$$\hat{c} = \hat{\lambda} / V_{\mathrm{partition}} = \frac{\exp(\hat{\beta}_0 + age \times \hat{\beta}_1 + 0.5\hat{\sigma}_1^2 + 0.5\hat{\sigma}_2^2)}{V_{\mathrm{partition}}}. \tag{41}$$

For all models, the corresponding binary GLMM with complementary log-log link function is available in statistical software (e.g. R [14,16]).

2.5. Case study data

14 genes of interest (13 target loci located on chromosomes 13, 18, 21, X and Y along with a single reference locus for normalization, *RPP30*, located on chromosome arm 10q for normalization) from 10 samples with chromosomal abnormalities and 4 control samples were analyzed. DNA was extracted from blood samples using the QIAamp DNA Blood Mini Kit (Qiagen) according to the manufacturer's instructions, after which DNA concentration was measured using UV spectrophotometry (Nanodrop). All patient samples and healthy control samples were diluted to 5 ng/μl using nuclease-free water and 10× carrier solution (Roche's tRNA from brewer's yeast, cat-no 10109517001, 50 ng/μl) making the final tRNA carrier concentration 5 ng/μl. The no template control (NTC) sample also contained 5 ng/μl carrier. Two μl of the diluted DNA sample was added into the final ddPCR reaction resulting in a 10 ng sample input. Primers and probes were diluted to a work solution of 5 μM and 2 μM, respectively, using 1× IDTE buffer, pH8 (cat-no 11-05-01-09). One ddPCR master mix was created per assay/probe pair (FAM/HEX or VIC) to perform 3 reactions per sample, control and NTC. One 20 μl reaction consisted out of following reagents: 10 μl 2× ddPCR Supermix for probes (Bio-Rad; cat-no 1863010), 250 nM of each forward and reverse primer (final concentration), 100 nM of each probe (final concentration), and 2 μl of sample DNA (5 ng/μl, 10 ng final input) or nuclease-free water as NTC. The 20 μl ddPCR reaction mix was added to the droplet generator cartridge together with 70 μl droplet generation oil (Bio-Rad, cat-no 1863005). Droplets were generated using a Bio-Rad QX100 droplet generator, followed by gentle transfer to a Twin.tec semi-skirted 96-well PCR plate. Using a Bio-Rad T100 thermal cycler, the following temperature cycling program was used for target amplification: 10 min 95 °C activation, 40 cycles of 30 s 95 °C and 1 min 59 °C, followed by 10 min 98 °C. After PCR amplification, the plate was analyzed using a QX100 droplet reader using Quantasoft software. Data was exported as a CSV file for further processing.

Primers and hydrolysis probes were designed using an in house developed primerXL assay design engine (Lefever et al., in preparation; http://www.primerxl.org), avoiding SNPs under the primer and probe annealing sites, avoiding secondary structures, and assessing primer specificity using genome wide Bowtie [17] primer alignment, avoiding primers that have fewer than three mismatches to a possible off-target homologous sequence. All primers and probes were ordered at Integrated DNA Technolgies, except for the *RPP30* probe which was ordered from Life Technologies. Sequences are available in Supplementary Table 1.

To allow comparison with digital PCR results, samples were also analyzed using arrayCGH, karyotyping and/or FISH in an ISO 15189 accredited lab at the Center for Medical Genetics, Ghent University Hospital as described before [18–20].

2.6. Data analysis

To assess the influence of ignoring interreplicate variation in the case of absolute quantification, GLM and GLMM models were fitted. The GLM model (Eq. (5)) did not account for interreplicate variation, and thus the analysis was based on pooling the negative and positive droplet counts. The GLMM (Eq. (11)), on the other hand, was fitted using a random effect to account for the interreplicate variation. Estimates of both approaches and their variances were compared.

Using the proposed framework, the copy numbers for each of the target loci was calculated by using the *RPP30* locus as a reference (model equation (20) or (22), Fig. 1, panel A or B, depending on the target locus). Copy numbers were also calculated, again by constructing GLMM models, by using all loci located on autosomal chromosomes with close to normal diploid copy number as references (model equation (30), Fig. 1, panel E), except for sample 10 where no loci with normal diploid copy number were distinguishable. Model evaluation was done by calculating the mean absolute deviation from the closest integer copy number over all non-reference locus copy numbers in a given sample.

To assess accuracy and precision with an increasing number of reference loci, target locus copy numbers were estimated by sequentially adding reference loci. To limit computational burden this was done for both a best case scenario, where the most stable reference loci were added first, and a worst case scenario, where the least stable reference loci were added first. After each step of adding a reference locus, accuracy was assessed by calculating the mean absolute deviation from integer copy numbers of the targets. Precision was assessed by calculating the mean width of all target 95% confidence intervals.

For each of the patient samples in the case study, reference locus stability was determined using the stability measure described in Section 2.3 (model (34)). Retaining the most stable reference loci only, final copy numbers were determined for each of the target loci (Eq. (30), setup as in Fig. 1, panel E).

3. Results and discussion

There is a clear need for more flexible data analysis tools for dPCR experiments. The GLMM framework is ideally suited to accommodate a wide range of dPCR experimental setups. As outlined in Section 2, the framework can be used for absolute quantification, CNV calculation and gene expression analysis (with or without replicates), but it can also accommodate other applications such as mutation quantification both in singleplex, duplex or higher multiplexing mode, making it compatible with all existing dPCR instruments. As further demonstrated in Section 2, the framework also allows adjusting for e.g. clinical baseline covariates such as age or gender, including treatments effects, or analyzing multicenter trials.

The type of the data we have collected, allows us to assess differences between classical approaches and those described in this paper. The initial setup of the experiment was to study 13 target loci

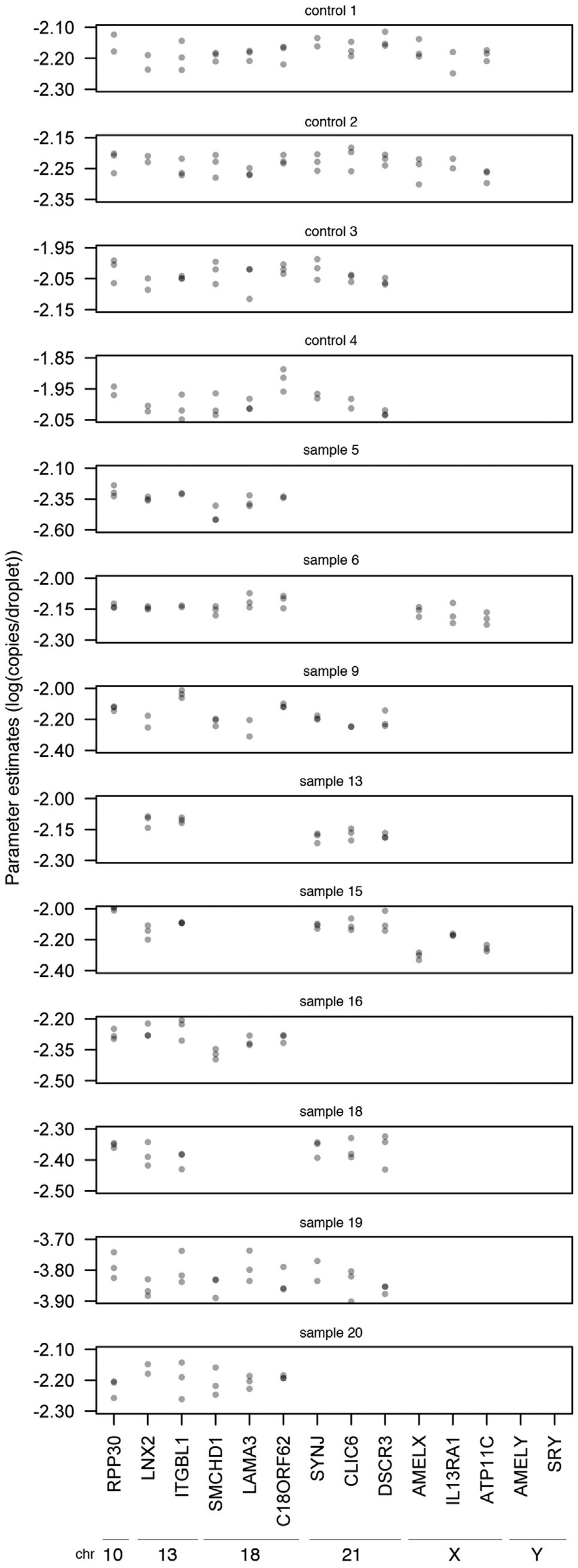

Fig. 2. Variability across replicates and reference loci. Parameterestimates of absolute quantities of the 14 candidate reference loci in 13 samples. Loci displayed are those with close to diploid copy number. Loci with a single or three copies are not displayed. For each reference locus and each sample, technical replicates are shown. The graphs demonstrate the presence of substantial variability between technical replicates, between samples and between reference loci.

along with a single reference locus for normalization. Because the patients in this study suffer from only one type of chromosomal abnormality, loci located on the non-affected chromosomes will still retain their normal copy number status. In principle, this allows the use of loci located on unaffected chromosomes as reference loci, in addition to the *RPP30* reference locus. Studying chromosomal loss or gain furthermore has the advantage that under ideal reaction conditions and assuming no mosaicism, copy numbers are expected to be integers. This is in contrast to e.g. expression levels that may take any (non-integer) value. This property of the study allows to measure accuracy as a deviance from an integer copy number.

3.1. Variance modelling

The need for modelling the different sources of variability becomes apparent from the estimates as obtained from the GLM models. As evidenced by Fig. 2 there is often substantial variability between technical replicates as well as between reference loci. Not accounting for these sources of variability may lead to overly optimistic uncertainty estimates. This is observed when comparing the uncertainty estimates of a naive pooling strategy (ignoring interreplicate variability) with those of a GLMM model accounting for interreplicate variability in an absolute quantification scenario. Fig. 3 shows a histogram of the ratio of the variance of the absolute quantification obtained with the pooling strategy relative to the variance calculated from the GLMM method. Even though the variability estimates of both methods are often close in our experimental data (in a majority of samples the fold difference was approximately one), the variability was underestimated up to a factor 10 when using a naive pooling strategy (Fig. 3). It is thus recommended to always incorporate the interreplicate variability to avoid overoptimistic variance estimates. This has also been demonstrated by Jacobs et al. [21] through simulation: pooling replicate data results in faulty confidence intervals, while accounting for interreplicate variation results in correct variance estimates and, consequently, correct confidence intervals.

The same holds true for accounting for variability between reference loci: there is often a discrepancy between different candidate reference loci (Fig. 2) and ignoring this may result in an underestimation of the uncertainty on relative quantity or CNV estimates.

Dorazio and Hunter [12] argue that accounting for additional sources of variability should only be considered if concentrations are expected or can be shown to be different. One can argue that due to the imperfectness of e.g. sample processing in a typical experiment (e.g. using replicates) such sources of variation will always be present and should always be considered [21]. Dorazio and Hunter [12] furthermore propose the use of deviance statistics to assess model goodness of fit. They argue that their model is not worse than the saturated model based on this deviance statistic, but they do not make a comparison with models accounting for additional sources of variability such as interreplicate variability. Even if such a comparison would demonstrate that a model not accounting for interreplicate variability is not significantly worse than one accounting for interreplicate variability, this may be a weak conclusion as goodness of fit tests may suffer from a lack of power [22]. For example, using our data we found that a lack of fit was detected only when the uncertainty using the GLMM method was approximately at least three times that of the uncertainty estimated using the GLM method (results not shown). When unsure about the presence or power of detecting additional sources of variation, it is thus better to be cautious and to allow for the possibility of these additional sources of variability, rather than to take the risk to obtain too optimistic results.

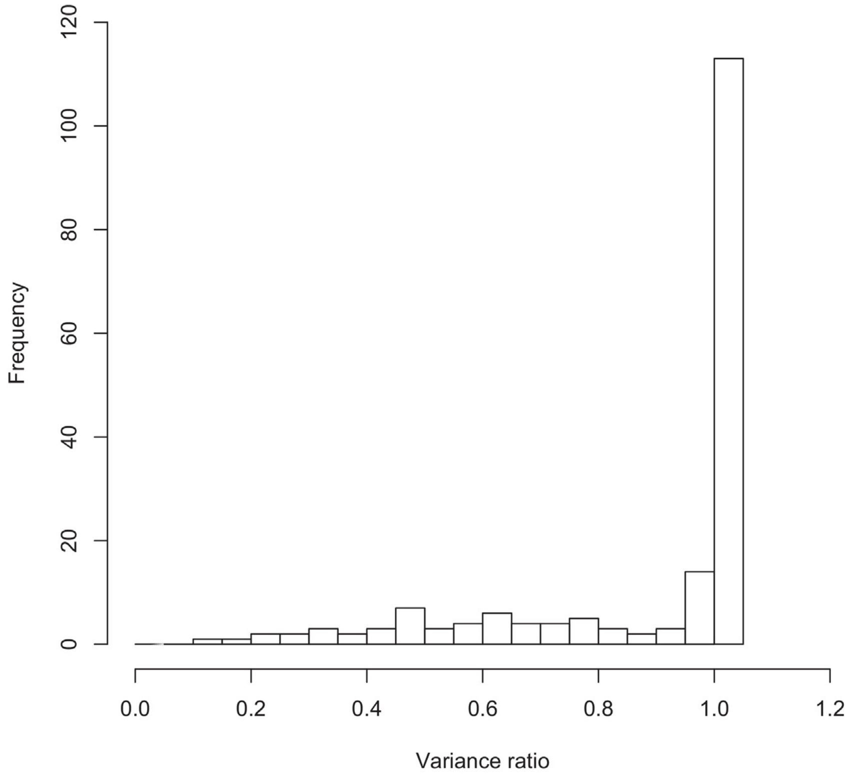

Fig. 3. Importance of modelling variability. Histogram of the ratio of variances of the parameter estimates of the naive pooling method relative to the GLMM. A ratio smaller than one indicates that the pooling method is too optimistic: the variance estimate of the pooling method is smaller than the variance estimate accounting for interreplicate variation. Ignoring this variation results in a variance estimate that is too small and consequently confidence intervals that are too narrow. It is recommended to always account for the possibility of variation between replicates to get an accurate variance estimate.

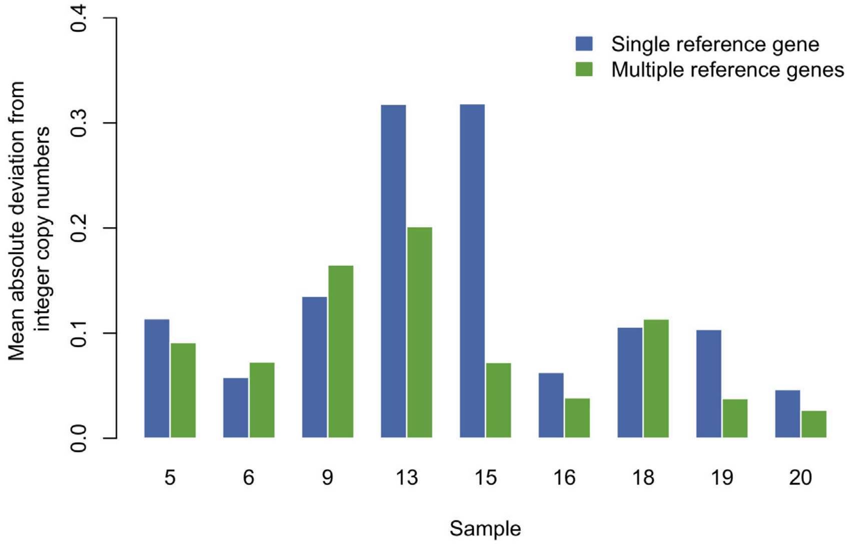

Fig. 4. Effect of normalization strategy Effect of normalization strategy on the mean absolute deviation from integer copy numbers for nine samples with chromosomal abnormalities. A low mean absolute deviation indicates that the obtained estimates are close to the expected copy numbers. Using multiple stable reference loci can result in a drastic increase in accuracy if the single reference locus was biased. If it is not beneficial, the difference with the single reference locus is small.

3.2. Multiple reference loci

It is recommended to use multiple reference loci, for relying on a single reference locus for calculating a CNV may result in reduced accuracy [7,23]. Especially in the case of e.g. erroneous amplification (e.g. due to inhibition) or copy number alteration of the candidate reference locus, estimated copy numbers may be far from accurate. This is clearly observed in the results from the case study, particularly in sample 15: when relying solely on the locus located on the *RPP30* gene, which is a popular reference gene in CNV studies [24], an underestimate of copy numbers of target loci is obtained. This has also been observed in other studies. For example, Versluis et al. [8] observed a gain of the *TERT* reference gene in two samples, resulting in erroneous results. They argue that relying on 3 to 4 reference loci would circumvent this problem.

These problems can indeed be remediated by relying on multiple reference loci, in which case e.g. erroneous amplification of one of the reference loci will only partially influence the estimated copy number. This is illustrated in Fig. 4 where relying on the predetermined reference locus on the *RPP30* gene often results in much larger mean absolute deviations from integer copy numbers when compared to relying on multiple reference loci. In those samples where the single locus normalization performs much worse, the *RPP30* locus is unsuitable for normalization (Supplementary Material 2).

The accuracy of the estimate can be further improved by using a selection of multiple stable reference loci. This can also be done using the GLMM model (possibly accounting for interreplicate variability). Our measure of stability takes both bias and variance in account. Because stable reference loci are expected to have the same copy number, the ratio of their concentrations should be close to one, or, in terms of model (34) or (35), the parameter estimates $\hat{\beta}_{1kl}$ should be close to zero. Hence, a ratio of one, or a parameter estimate of zero corresponds to no bias. By making all pairwise comparisons between a reference locus and all other reference loci, and calculating the sum of the squared deviations from zero, an estimate of the total squared bias of the reference locus is obtained: optimal reference loci will have a low bias while loci with large bias indicate that they are not in agreement with the other candidate reference loci. Variance is taken into account by using the estimate of the variance of the bias parameter estimate: loci that have a more variable concentration across replicates or samples (as propagated in the variance of the estimate) are less ideal reference loci.

The stability measure is thus penalising aberrant reference loci in two ways: loci that have a concentration deviating from other reference loci will be penalised as well as those with higher variance. Optimal reference loci are both in accordance with other reference loci and show little variation when replicated. Reference loci having similar bias will be ranked from lower to higher variance, and likewise loci with similar variance will be ranked from lower to higher bias. Fig. 5 indicates that for example for patient sample 5, the genes *LNX2*, *LAMA3* and *C18ORF62* are most stable and can be used as reference loci.

When sequentially adding reference loci from stable to less stable (best case scenario), the uncertainty of the target estimate first decreases, but increases again when more unstable reference loci are added (U shape of the curve, Fig. 5). Excluding unstable reference loci may thus lead to a more accurate estimate while inclusion of multiple stable reference loci typically also enhances the precision of the estimate. When all reference loci are stable, more reference loci means less uncertainty and exclusion of reference loci is disadvantageous (Supplementary Material 2 Figs. 7 and 8).

Sequentially adding reference loci in reverse order (worst case scenario), i.e. using one or more less stable reference loci, may result in severe bias (Fig. 5): only when the more stable reference loci are added to the pool of reference loci the bias gets close to that of the best case scenario. Even though the bias may converge to that of the best case scenario when using just a few reference loci, the uncertainty of the estimate may still remain higher than what is observed under a best case scenario.

Relative expression stability determination necessitates a different approach as the ratio of reference gene concentrations is generally not close to one, but should be stable across all samples [7]. Taking a bias parameter (deviation from a ratio of one) into account would wrongly discount good reference genes for relative expression estimation. We suggest a modified stability criterion for this scenario, taking only variability of the ratio into account. Thus, genes that have a variable concentration ratio across samples and within samples (technical replication) will be considered less ideal than genes displaying a stable ratio across and within samples. This approach is similar to the widely adopted stability measurement described by Vandesompele et al. [7].

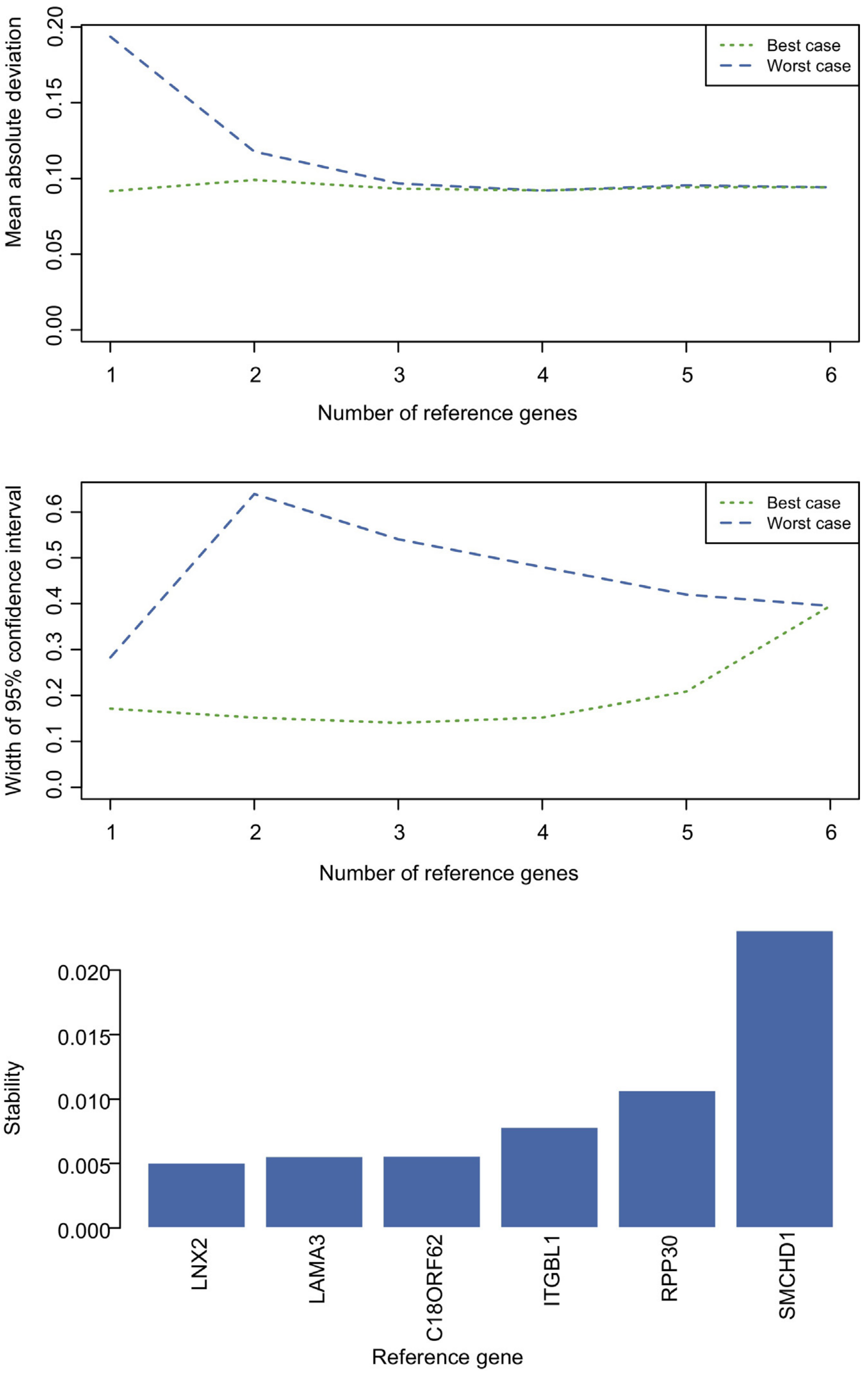

Fig. 5. Assessing reference stability. Top: Mean absolute deviation from integer copy numbers as a function of the number of reference loci in sample 5. The reference loci are added in the order of decreasing stability (best case) or increasing stability (worst case). Using an increasing number of reference loci results in a copy number estimate that is less biased. Especially when only a few reference loci are used and those reference loci are not stable, the copy number bias may be considerably larger. Middle: Uncertainty on target estimates, expressed as the width of a 95% confidence interval, as a function of the number of reference loci in sample 5. The reference loci are added in the order of decreasing stability (best case) or increasing stability (worst case). In terms of obtaining precise estimates, it is better to select the most stable reference genes and omitting the less stable reference genes. In this case, the optimal number of reference genes is 3 as the uncertainty is lowest in this case. Bottom: Stability of candidate reference loci in sample 5. Loci on *LNX2*, *LAMA3* and *C18ORF62* are more stable than the others.

3.3. Case study results

Copy numbers were measured in duplex (Fig. 1E) for 10 patient samples and analyzed using the appropriate models given in Section 2. A summary of all obtained aberrations is listed in Table 1. A full list of estimated copy numbers and confidence intervals is available as Supplementary Table 2. For 9 out of the 10 samples at least one target locus (located on the aberrant chromosome) was confirmed, i.e. the 95% confidence interval contained the integer copy number. In those 9 samples there was also a clear aberration of the non-confirmed loci located on the chromosome with the aberration, supporting the evidence for the confirmed loci (Supplementary Table 2). Sample 10 displayed an unusual profile which we were unable to identify (no loci with (close to) normal diploid copy number). The obtained aberrations using ddPCR correspond to those obtained using state-of-the-art methods (karyotyping, FISH,

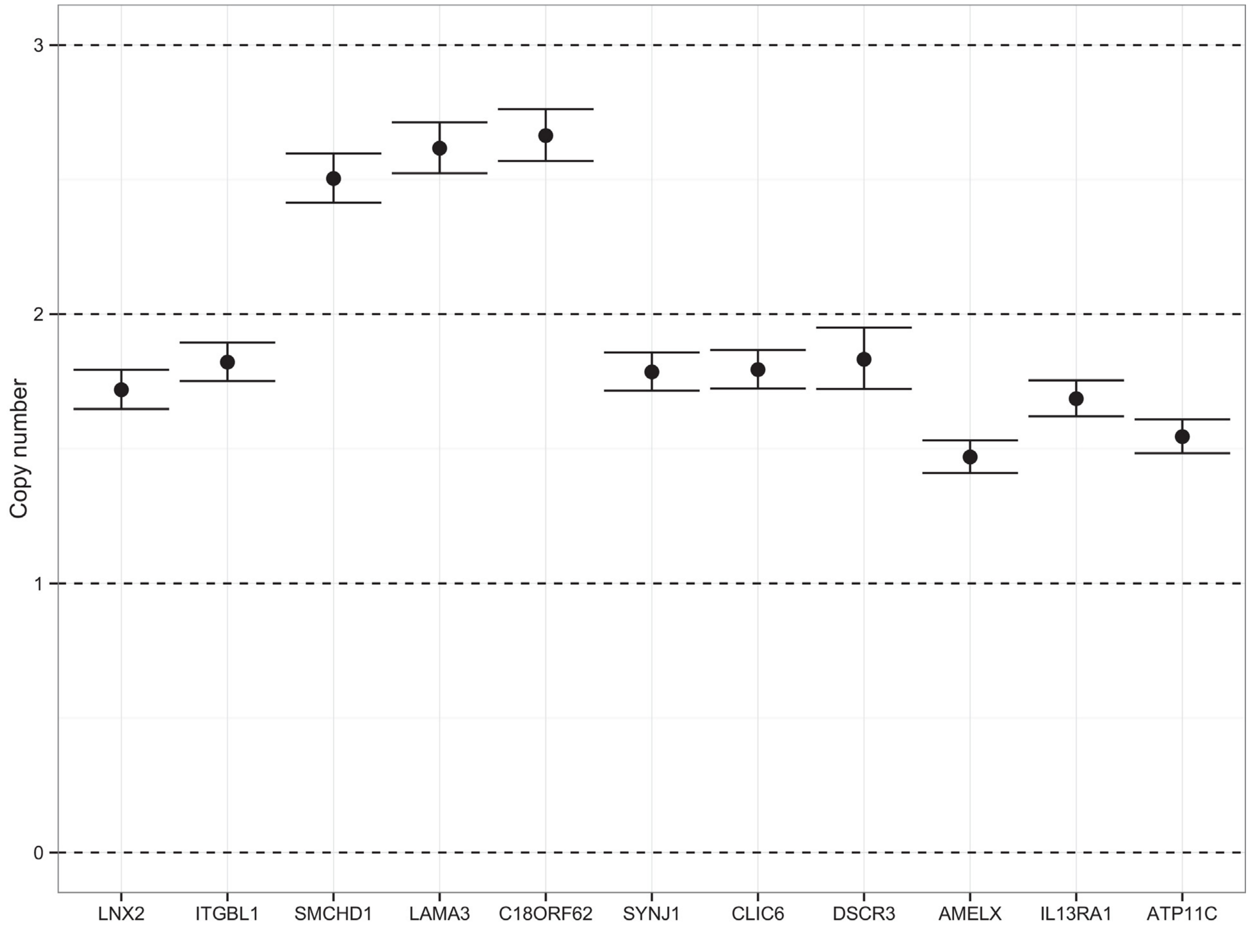

Fig. 6. Copy numbers in sample 15. Copy numbers in sample 15 after normalization using the *RPP30* locus (accounting for interreplicate variability). There appears to be a general underestimation of the copy number, which may be due to aberrant quantification of the *RPP30* locus: the measured concentration of the RPP30 locus is too high, so that the copy number of the target loci, expected to have two or three copies, are all underestimated.

Table 1
Overview of detected chromosomal abnormalities with confirmed loci.

Sample	Detected aberration	Confirmed loci
5	Chromosome 21 trisomy	*CLIC6*
6	Chromosome 21 trisomy	*DSCR3*, *SYNJ1*
9	Isochromosome Xq	*AMELX*
10	–	–
13	Chromosome 18 trisomy	*C18ORF62*, *LAMA3*
15	Chromosome 18 trisomy	*C18ORF62*, *LAMA3*, *SMCHD1*
16	Chromosome 21 trisomy	*CLIC6*, *DSCR3*, *SYNJ1*
18	Chromosome 18 trisomy	*LAMA3*
19	Isochromosome Xq	*AMELX*, *IL13RA1*, *ATP11C*
20	Chromosome 21 trisomy	*CLIC6*, *DSCR3*, *SYNJ1*

arrayCGH), except for the previously mentioned sample 10 that was not identifiable using our ddPCR results (identified to be a 45,X (Turner syndrome) sample using karyotyping, FISH and arrayCGH).

We discuss the case of patient sample 15, but the approach for the other samples is similar (Supplementary Material 3). The setup of the study was to assess chromosomal aberrations, i.e. deviations from diploid copy numbers on 13 target loci. The reference locus was located on the *RPP30* gene. The results of normalization using the *RPP30* locus (accounting for interreplicate variability) is displayed in Fig. 6. It is clear we are dealing with a sample belonging to a female (absence of positive droplets of the Y chromosome loci, Supplementary Table 3), but all copy numbers seem to be deviating: there is a general underestimation of the copy number, which may be due to aberrant quantification of the *RPP30* locus. It can furthermore be seen that the *C18ORF62*3, *LAMA3* and *SMCHD1* loci (all located on chromosome 18) seem to have a higher copy number than loci located on the other chromosomes and that a chromosome 18 trisomy is likely. We thus use the loci on chromosomes X, 13, 21 and the locus on *RPP30* as candidate reference loci and determine the stability of these loci as described in the Materials and Methods section. The stability plot (Fig. 7) indicates that the loci on *LNX2*, *IL13RA1*, *SYNJ1*, *CLIC6*, *ITGBL1* and *DSCR3* are of similar stability, with rising instability thereafter for the loci on *ATP11C*, *RPP30* and *AMELX*. The latter three are thus excluded as reference loci and the copy numbers of the three chromosome 18 loci recalculated using the six stable reference loci. The results are shown in the bottom panel of Fig. 8: the 95% confidence intervals of the multiple reference locus normalized results encompass the expected integer copy number for all three loci for a trisomy 18 case.

3.4. Comments and further research

Even though the assumption of only one chromosomal aberration turned out to be correct in our case study, it is generally to be recommended to determine candidate reference loci upfront. For CNV determination it will not always be clear whether e.g. a non-integer increase in a chromsome 18 loci combined with a non-integer decrease in chromosome 13 loci corresponds to a chromosome 18 trisomy or a chromosome 13 monosomy. This can

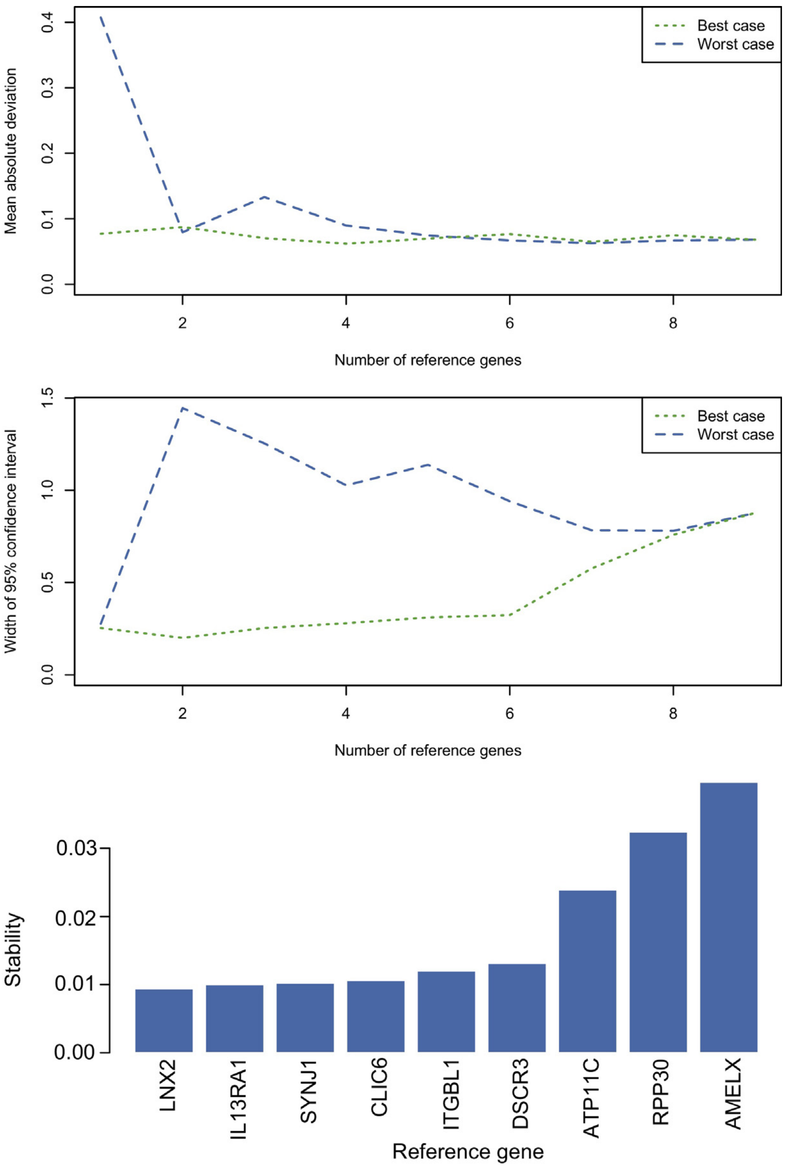

Fig. 7. Stability of candidate reference loci in sample 15. Loci on *LNX2*, *IL13RA1*, *SYNJ1*, *CLIC6*, *ITGBL1* and *DSCR3* are of similar stability, with rising instability thereafter for the loci on *ATP11C*, *RPP30* and *AMELX*.

furter be complicated due to contamination of sample material by e.g. maternal DNA. When using the framework for detection of non-integer copy numbers, the reference locus selection approach used in our case study can no longer be used and determination of stable reference loci upfront is needed.

We evaluated our method by calculating a deviation measure from integer copy numbers, assuming that the samples did not display any mosaicism. This approach is only valid under the ideal circumstances of no mosaicism, no contamination, . . .and deviation from this ideal scenario may have affected our case study findings.

We demonstrated that accounting for variation between replicates and/or reference genes is necessary to obtain correct standard errors of estimated absolute or relative quantities. These are sources of variations that can be accounted for in a majority of experiments as the use of replicate experiments and multiple reference genes are nowadays widely implemented. We want to stress that there are additional sources of variation, and depending on the experimental setup these can also be taken into account by e.g. adding additional random effects for plate effects, run effects, cartridge effects, . . .

3.5. *Software implementation*

R code and a tutorial demonstrating the implementation of these models is available as Supplementary Material 1: we give examples of how to analyze different types of data using the models proposed in this manuscript. By providing easy to use functions, the practitioner is able to determine reference gene stability and perform absolute quantification (with or without replicates) and CNV/relative expression analysis (with one or more reference genes and with or without replicates) with a few lines of code. Additionally, a Shiny web application is available at http://antonov.ugent.be:3838/dPCR/ that does not require any type of programming, but relies on a spreadsheet-like data input and point and click interface. A manual for using this web interface is also available (Supplementary Material 5).

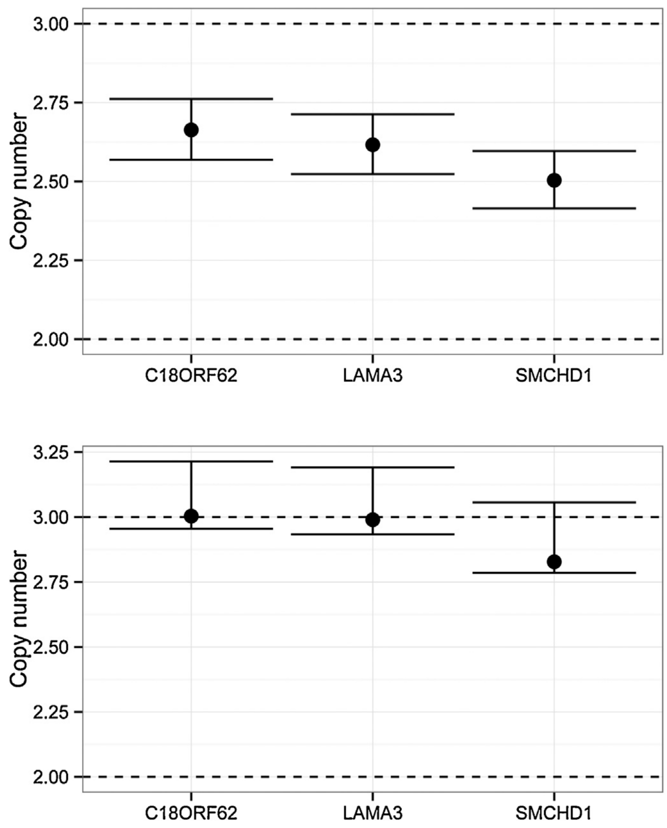

Fig. 8. Effect of normalization strategy. Effect of normalization strategy on the estimated copy numbers of 3 loci on chromosome 18 for a patient with chromosome 18 trisomy. Top: a single reference locus is used (*RPP30*), bottom: three reference loci are used simultaneously. When using multiple reference loci the copy numbers are on average closer to the expected value of three (i.e. less bias). Error bars denote 95% confidence intervals, whereby the multiple reference locus normalized results encompass the expected integer copy number for all three loci. Using an unstable (biased) single reference locus may result in biased copy number estimates. Relying on multiple reference loci typically has the advantage that the bias is reduced: if one of the reference loci is biased, this effect is partially cancelled by averaging together with two stable reference loci. See Supplementary Material 3 for more examples.

4. Conclusion

Currently available methodology and software could be used to analyze e.g. experiments with multiple reference genes, but these methods would not properly account for sources of variability introduced by these specific designs. The data analyses presented in this paper suggest that not accounting for various sources of variability can result in extremely unreliable estimates of variability of nucleic acid concentrations or copy numbers.

We introduce a general framework that covers the analysis of a wide range of experimental setups, correctly accounting for various sources of variability and allowing researchers to analyze data from experiments for which previously no appropriate methods were proposed.

A method for reference gene selection relying on this framework is suggested, allowing for an improved estimation of copy numbers or relative quantities: findings indicate that selecting stable reference genes may be beneficial in two ways: (i) bias may be reduced, (ii) uncertainty can be decreased by selecting a suitable subset. User-friendly R scripts, a Shiny web interface and tutorials are available to facilitate the use of the methodology.

Conflict of interest statement

None declared.

Acknowledgements

We are grateful to XXXfor PCR assay design using primerXL. We thank the referees for their useful comments.

Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.bdq.2016.06.001.

References

[1] A.A. Morley, Digital PCR: a brief history, Biomol. Detect. Quantif. 1 (2014) 1–2, http://dx.doi.org/10.1016/j.bdq.2014.06.001.

[2] M. Baker, Digital PCR hits its stride, Nature Methods 9 (2012) 541–544, http://dx.doi.org/10.1038/nmeth.2027.

[3] J.F. Huggett, S. Cowen, C.A. Foy, Considerations for digital PCR as an accurate molecular diagnostic tool, Clin. Chem. 61 (2015) 79–88, http://dx.doi.org/10.1373/clinchem.2014.221366.

[4] R. Sanders, J.F. Huggett, C.A. Bushell, S. Cowen, D.J. Scott, C.A. Foy, Evaluation of digital PCR for absolute DNA quantification, Anal. Chem. 83 (2011) 6474–6484, http://dx.doi.org/10.1021/ac103230c.

[5] M.C. Strain, S.M. Lada, T. Luong, S.E. Rought, S. Gianella, V.H. Terry, C.A. Spina, C.H. Woelk, D.D. Richman, Highly precise measurement of HIV DNA by Droplet digital PCR, PLOS ONE 8 (4) (2013) e55943, http://dx.doi.org/10.1371/journal.pone.0055943.

[6] A.S. Whale, J.F. Huggett, S. Cowen, V. Speirs, J. Shaw, S. Ellison, C.A. Foy, D.J. Scott, Comparison of microfluidic digital PCR and conventional quantitative PCR for measuring copy number variation, Nucleic Acids Res. 40 (2012) e82, http://dx.doi.org/10.1093/nar/gks203.

[7] J. Vandesompele, K. De Preter, F. Pattyn, B. Poppe, N. Van Roy, A. De Paepe, F. Speleman, Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes, Genome Biol. 3 (2002), research0034. doi:10.1186/gb-2002-3-7-research0034.

[8] M. Versluis, M.J. de Lange, S.I. van Pelt, C.A. Ruivenkamp, W.G. Kroes, J. Cao, M.J. Jager, G.P. Luyten, P.A. van der Velden, Digital PCR validates 8q dosage as prognostic tool in Uveal Melanoma, PLOS ONE 10 (2015) e0116371, http://dx.doi.org/10.1371/journal.pone.0116371.

[9] A. Zmienko, A. Samelak-Czajka, M. Goralski, E. Sobieszczuk-Nowicka, P. Kozlowski, M. Figlerowicz, Selection of reference genes for qPCR- and ddPCR-based analyses of gene expression in senescing Barley leaves, PLOS One 10 (2015) e0118226, doi:10.1371/journal.pone.0118226.

[10] S. Dube, J. Qin, R. Ramakrishnan, Mathematical analysis of copy number variation in a DNA sample using digital PCR on a nanofluidic device, PLoS ONE 3 (2008) e2876, http://dx.doi.org/10.1371/journal.pone.0002876.

[11] W.W. Stroup, Generalized Linear Mixed Models: Modern Concepts, Methods and Applications, CRC Press, Boca Raton, FL, 2012.

[12] R.M. Dorazio, M.E. Hunter, Statistical models for the analysis and design of digital polymerase chain reaction (dPCR) experiments, Anal. Chem. 87 (2015) 10886–10893, http://dx.doi.org/10.1021/acs.analchem.5b02429.

[13] M. Yu, K.T. Carter, K.W. Makar, K. Vickers, C.M. Ulrich, R.E. Schoen, D. Brenner, S.D. Markowitz, MethyLight droplet digital PCR for detection and absolute quantification of infrequently methylated alleles, Epigenetics 10 (9) (2015) 803–809, http://dx.doi.org/10.1080/15592294.2015.1068490.

[14] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, 2015.

[15] S. Rödiger, M. Burdukiewicz, K. Blagodatskikh, M. Jahn, P. Schierack, R as an Environment for Reproducible Analysis of DNA Amplification Experiments, The R J. 7 (1) (2015) 127–150.

[16] D. Bates, M. Machler, B. Bolker, S. Walker, lme4: Linear Mixed-Effects Models Using Eigen and S4, R package version 1 (2015) 1–10.

[17] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biology 10 (3) (2009) R25, http://dx.doi.org/10.1186/gb-2009-10-3-r25.

[18] B. Menten, N. Maas, B. Thienpont, K. Buysse, J. Vandesompele, C. Melotte, T. de Ravel, S. Van Vooren, I. Balikova, L. Backx, S. Janssens, A. De Paepe, B. De Moor, Y. Moreau, P. Marynen, J.P. Fryns, G. Mortier, K. Devriendt, F. Speleman, J.R. Vermeesch, Emerging patterns of cryptic chromosomal imbalance in patients with idiopathic mental retardation and multiple congenital anomalies: a new series of 140 patients and review of published reports, J. Med. Genet. 43 (8) (2006) 625–633, http://dx.doi.org/10.1136/jmg.2005.039453.

[19] K. Buysse, B. Delle Chiaie, R. Van Coster, B. Loeys, A. De Paepe, G. Mortier, F. Speleman, B. Menten, Challenges for CNV interpretation in clinical molecular karyotyping: lessons learned from a 1001 sample experience, Eur. J. Med. Genet. 52 (6) (2009) 398–403, http://dx.doi.org/10.1016/j.ejmg.2009.09.002.

[20] B. Menten, K. Swerts, B. Delle Chiaie, S. Janssens, K. Buysse, J. Philippe, F. Speleman, Array comparative genomic hybridization and flow cytometry analysis of spontaneous abortions and mors in utero samples, BMC Med. Genet. 10 (2009) 89, http://dx.doi.org/10.1186/1471-2350-10-89.

[21] B.K. Jacobs, E. Goetghebeur, L. Clement, Impact of variance components on reliability of absolute quantification using digital PCR, BMC Bioinform. 15 (2014) 283, http://dx.doi.org/10.1186/1471-2105-15-283.

[22] D.W. Hosmer, N.L. Hjort, Goodness-of-fit processes for logistic regression: simulation results, Stat. Med. 21 (18) (2002) 2723–2738, http://dx.doi.org/10.1002/sim.1200.

[23] B. D'haene, J. Vandesompele, J. Hellemans, Accurate and objective copy number profiling using real-time quantitative PCR, Methods 50.4 (2010) 262–270, http://dx.doi.org/10.1016/j.ymeth.2009.12.007.

[24] B.J. Hindson, K.D. Ness, D.A. Masquelier, P. Belgrader, N.J. Heredia, A.J. Makarewicz, et al., High-throughput droplet digital PCR system for absolute quantitation of DNA copy number, Anal. Chem. 83 (22) (2011) 8604–8610, doi:10.1021/ac202028g.