

## Digital PCR data analysis:

A flexible framework for absolute  
and relative quantification and CNV

Olivier Thas

Ghent University, Belgium



ELSEVIER

Contents lists available at [ScienceDirect](#)

## Biomolecular Detection and Quantification

journal homepage: [www.elsevier.com/locate/bdq](http://www.elsevier.com/locate/bdq)



Research Paper

### Flexible analysis of digital PCR experiments using generalized linear mixed models

Matthijs Vynck<sup>a,\*</sup>, Jo Vandesompele<sup>b,c,d</sup>, Nele Nijs<sup>d</sup>, Björn Menten<sup>b,c</sup>, Ariane De Ganck<sup>d</sup>, Olivier Thas<sup>a,e</sup>



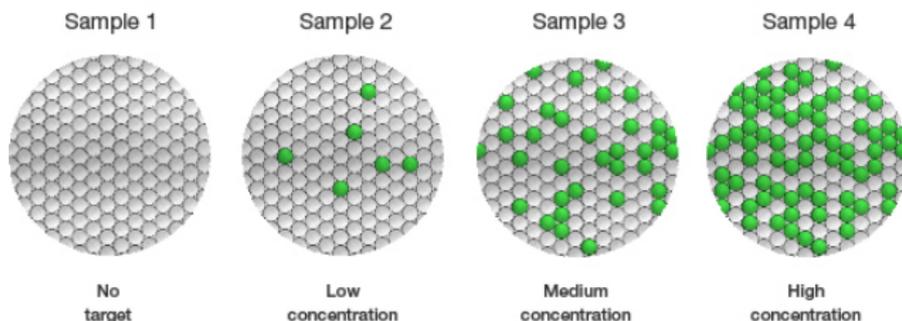
## Quantification: classical approach

Once the presence/absence of the target in each droplet has been determined, the original concentration can be determined assuming a Poisson distribution of the number of target molecules in the droplets.

Let  $Y_j^*$  denote the number of target copies in droplet  $j$  ( $j = 1, \dots, N$ ).

What we observe is the digital outcome

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



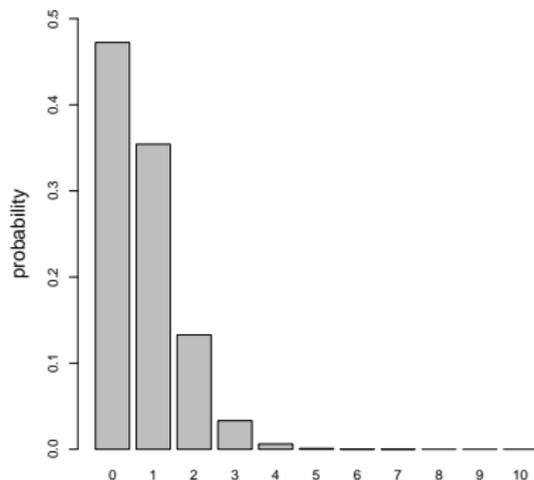
Positive/Negative digital observations (Bio-Rad Laboratories)

## Quantification: classical approach

dPCR data analysis methods rely on the **Poisson assumption**,

$$Y_j^* \sim \text{Poisson}(\lambda) \quad j = 1, \dots, N,$$

where  $\lambda = E \{ Y_j^* \}$  is the average number of target copies in a droplet.



## Quantification: classical approach

ddPCR data analysis methods rely on the **Poisson assumption**,

$$Y_j^* \sim \text{Poisson}(\lambda) \quad j = 1, \dots, N,$$

where  $\lambda = E \{ Y_j^* \}$  is the average number of target copies in a droplet.

If  $\lambda$  can be estimated, say  $\hat{\lambda}$ , the **target concentration** can be estimated as

$$\hat{c} = \frac{\hat{\lambda}}{V_{\text{droplet}}}$$

where  $V_{\text{droplet}}$  is the average droplet volume (e.g. 0.85 nL).

The classical approach consists in utilising the Poisson distribution:

$$\hat{\lambda} = -\ln P \{ \widehat{Y^*} = 0 \} = -\ln P \{ \widehat{Y} = 0 \} = -\ln \frac{\text{number of negative droplets}}{\text{total number of droplets}}.$$

## Quantification: classical approach

Good scientific practice requires reporting of standard error (or confidence interval) of  $\hat{c}$

A classical solution relies on estimating the standard error of  $\hat{c}$ , say  $s_{\hat{c}}$ .  
An approximate 95% confidence interval (CI) is then given by

$$[\hat{c} - 1.96 s_{\hat{c}}, \hat{c} + 1.96 s_{\hat{c}}]$$

For the calculation of the standard error  $s_{\hat{c}}$ :

- Approximation methods exist
- For more complicated designs, ad-hoc calculation methods are applied.

## Quantification: GLM

It has been shown that the estimation can be done with GLM software:

- $Y_j \sim \text{Bernoulli}(p)$
- $p = P \{ \text{droplet } j \text{ is negative} \}$
- $\ln(-\ln(p)) = \beta$

Hence,

$$\beta = \ln(-\ln(p)) = \ln \lambda.$$

→ GLM software provides estimates and standard errors

## Quantification: GLM

The mean number of copies per droplet is still

$$E\{Y^*\} = \lambda = \exp(\beta)$$

which can be estimated as

$$\hat{\lambda} = \exp(\hat{\beta}),$$

and from the confidence interval on  $\hat{\beta}$  the confidence interval on  $\hat{\lambda}$  can be obtained directly.

Using GLMs in this simple setting seems an overkill, but in the next part the power of statistical models will become clear.

# Quantification: GLMM

Suppose that **data from  $r$  replicate runs** are available.

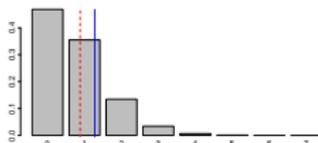
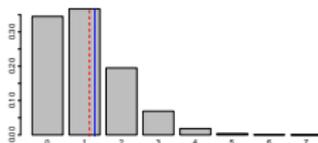
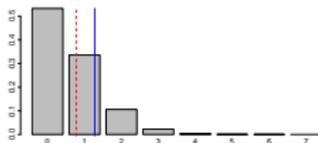
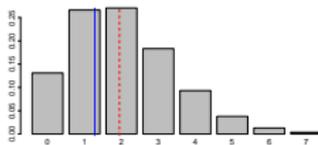
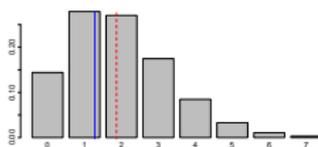
A typical design:

- **one biological sample** for which a target concentration has to be estimated
- **$r$  technical replicates**  
(e.g. separate runs and/or separate sample preparations)
- for each technical replicate  $i$ ,  $n_i$  droplets are measured.

Solutions:

- pool digital outcomes / replicates
- use GLMM

# Quantification: GLMM



# Quantification: GLMM

We introduce an **unobservable latent variable**  $Y_{ij}^*$  denoting the number of target copies

- in droplet  $j$
- of replicate  $i$ .

What we observe is the **digital** outcome

$$Y_{ij} = \min(Y_{ij}^*, 1) = \begin{cases} 0 & \text{if } Y_{ij}^* = 0 \text{ (negative droplet)} \\ 1 & \text{otherwise (positive droplet)} \end{cases}$$

## Quantification: GLMM

- Within a replicate ( $i$ ) the target copy numbers may be described by a Poisson distribution (as before).
- Each replicate comes with a **replicate effect**, say  $R_i$ .

We write for the **within-replicate droplets**

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

with **replicate-specific** mean

$$\lambda_i = \exp(\beta) \times \exp(R_i) \quad \text{or} \quad \ln(\lambda_i) = \beta + R_i$$

and

$$R_i \sim N(0, \sigma^2).$$

This model gives the **mean target copy number**

$$E \{ Y_{ij}^* \} = \exp(\beta + 0.5\sigma^2).$$

## Quantification: GLMM

The parameters will have to be estimated from the observed digital data.

### Generalised Linear Mixed Model (GLMM)

$$p_i = P \{ \text{negative droplet in replicate } i \}$$

$$Y_{ij} \mid R_i \sim \text{Bernoulli}(p_i)$$

$$\ln(-\ln(p_i)) = \beta + R_i.$$

Once  $\beta$  and  $\sigma^2$  are estimated, the **target concentration is estimated as**

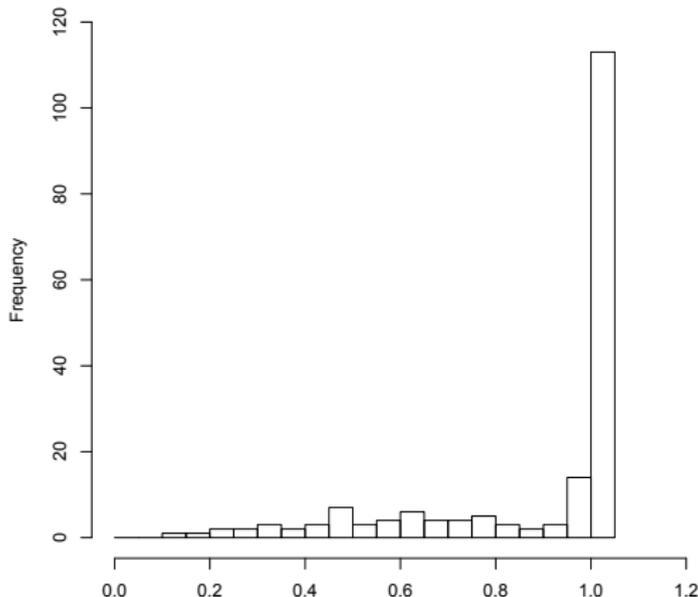
$$\hat{c} = \hat{\lambda} / V_{\text{droplet}} = \exp(\hat{\beta} + 0.5\hat{\sigma}^2) / V_{\text{droplet}}.$$

The GLMM software will give  $\hat{\beta}$  and  $\hat{\sigma}^2$  and their estimated standard errors.

## Quantification: GLMM versus pooling

Pooling: merge digital outcomes over replicates and hence ignore inter-replicate variance.

Figure shows variance ratio of  $\hat{\beta}$  for pooled analysis versus GLMM.



# Copy number variation

Copy number variation (CNV) is defined as

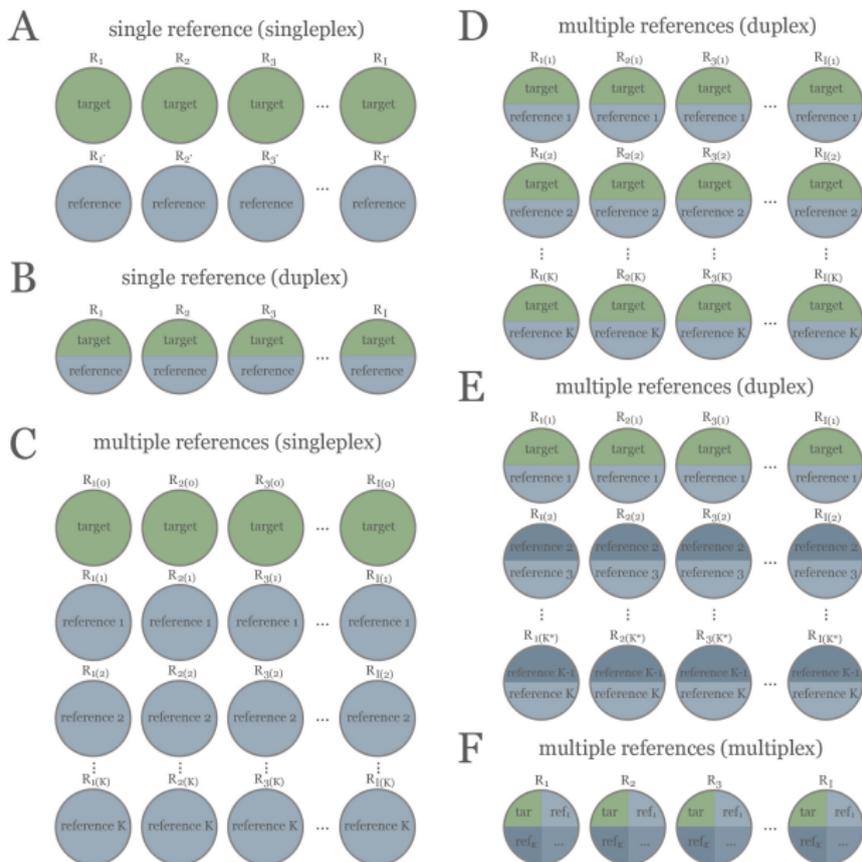
$$CNV = \frac{C_{\text{target}}}{C_{\text{ref}}} N_b,$$

where  $N_b$  is ploidy of the genome (for humans  $N_b = 2$ ).

The GLMM can be used for

- one or more target genes
- one or multiple reference genes
- single channel or multiplex measurements

# Copy number variation: designs



# Copy number variation: models

## 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

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

where

$$\log \lambda_{ij} = \beta_0 + X_{ij}\beta_1 + R_i \quad (17)$$

and

$$R_i \sim N(0, \sigma^2). \quad (18)$$

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  $\exp(\beta_0 + 1 \times \beta_1 + R_i)$ .

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

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

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

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

# Copy number variation: models

## 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  $k = 1, \dots, K$ , 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  $k$ th 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

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

with  $S_k$  the effect of reference  $k$  on the log-mean, and  $R_{i(k)}$  the effect of the  $i$ th 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:

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

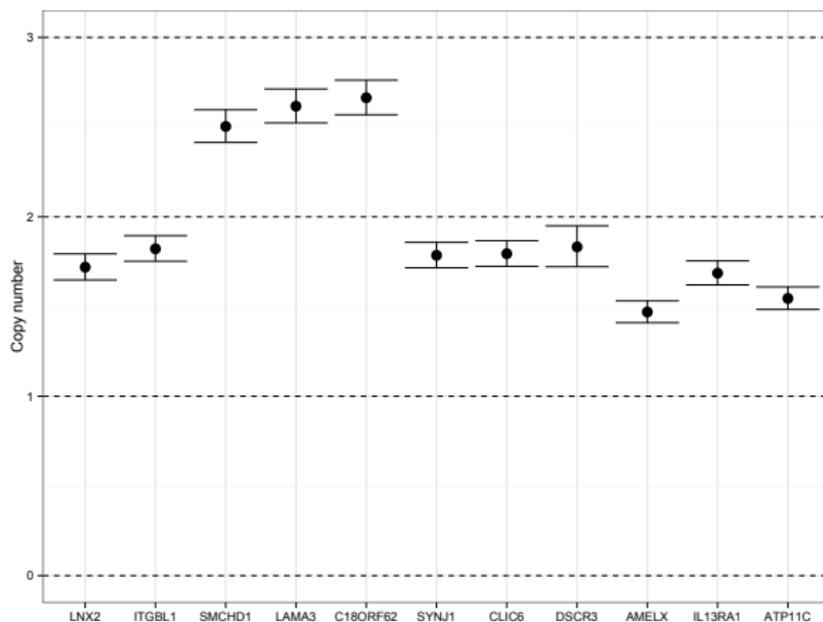
For design C the CNV is first given for target versus a single reference  $k$ , based on replicates  $i$  and  $i'$ :

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

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

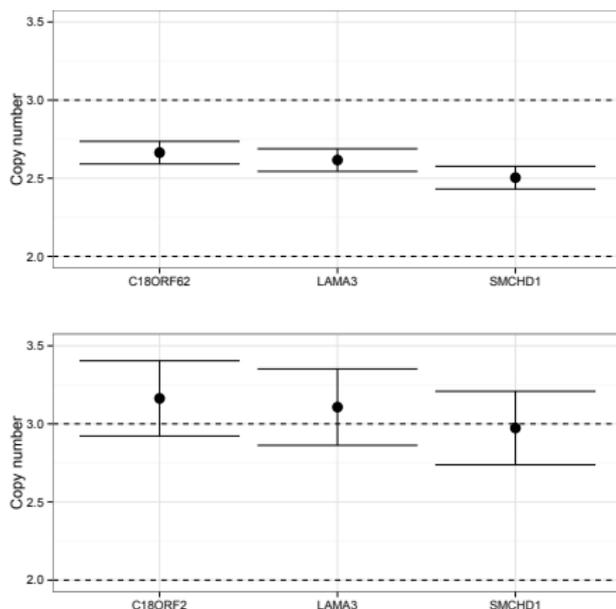
$$\text{CNV} = E(\text{CNV}_{i,i';k}) = \exp\left(-\beta_1 + \frac{1}{2}\sigma_1^2 + \sigma_2^2\right) N_b. \quad (27)$$

# Copy number variation: single versus multiple ref. genes



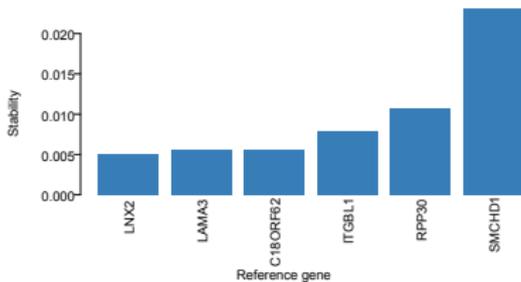
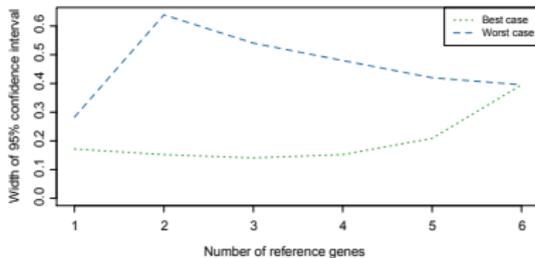
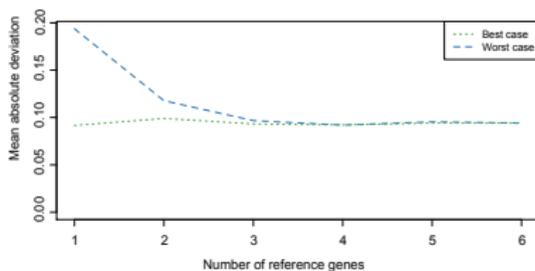
→ bias caused by single ref. gene

# Copy number variation: single versus multiple ref. genes



→ bias resolved by using multiple ref. genes

# Copy number variation: selection of ref. genes



# Shiny Application GLMM: read data

dPCR analysis

Data management

Analysis

Graphics

Report

Manual input data

Number of samples:

1 35 100

1 21 31 41 51 61 71 81 91 100

Upload data

Upload

Download data

Download

Data

	Name	Positive	Negative	Target
1	CLIC6	1666	10929	<input checked="" type="checkbox"/>
2	CLIC6	1986	12757	<input checked="" type="checkbox"/>
3	CLIC6	1861	12189	<input checked="" type="checkbox"/>
4	DSCR3	2134	13173	<input checked="" type="checkbox"/>
5	DSCR3	1937	11939	<input checked="" type="checkbox"/>
6	SYNJ1	1788	11285	<input checked="" type="checkbox"/>
7	SYNJ1	1856	11011	<input checked="" type="checkbox"/>
8	SYNJ1	2035	12286	<input checked="" type="checkbox"/>
9	AMELX	650	13335	<input checked="" type="checkbox"/>
10	AMELX	597	13712	<input checked="" type="checkbox"/>
11	AMELX	671	14352	<input checked="" type="checkbox"/>
12	AMELY	616	14312	<input checked="" type="checkbox"/>
13	AMELY	630	14392	<input checked="" type="checkbox"/>
14	ATP11C	314	9123	<input checked="" type="checkbox"/>
15	ATP11C	545	13030	<input checked="" type="checkbox"/>
16	ATP11C	635	14264	<input checked="" type="checkbox"/>
17	IL13RA1	573	11036	<input checked="" type="checkbox"/>
18	IL13RA1	672	13822	<input checked="" type="checkbox"/>
19	IL13RA1	681	14621	<input checked="" type="checkbox"/>
20	SRY	543	12052	<input checked="" type="checkbox"/>
21	SRY	676	14067	<input checked="" type="checkbox"/>
22	SRY	586	13464	<input checked="" type="checkbox"/>
23	C18ORF62	1370	13558	<input type="checkbox"/>

# Shiny Application GLMM: absolute quantification

The screenshot shows a web browser window displaying a Shiny application titled "dPCR analysis". The browser's address bar shows the URL "127.0.0.1". The application interface has a green header bar with the text "Calculating results. Please wait." on the right. On the left, there is a dark sidebar with a navigation menu containing the following items: "Data management", "Analysis" (with a dropdown arrow), "Absolute quantification" (with a right-pointing arrow), "Copy number variation" (with a right-pointing arrow), "Reference gene stability" (with a right-pointing arrow), "Graphics", and "Report". The main content area is divided into two panels. The left panel, titled "Select targets to quantify", contains a list of target genes with checkboxes: CLIC6 (checked), DSCR3 (checked), SYNJ1 (checked), AMELX (unchecked), AMELY (unchecked), ATP11C (unchecked), IL13RA1 (unchecked), and SRY (unchecked). Below this list is an "Analyse" button. The right panel, titled "Results", is currently empty.

# Shiny Application GLMM: absolute quantification

dPCR analysis

Select targets to quantify

- CLIC6
- DSCR3
- SYNJ1
- AMELX
- AMELY
- ATP11C
- IL13RA1
- SRY

Analyse

Results

	Name	Estimate	95% CI lower bound	95% CI upper bound
1	DSCR3	13.31	12.91	13.73
2	SYNJ1	13.15	12.81	13.51
3	CLIC6	13.99	13.63	14.37

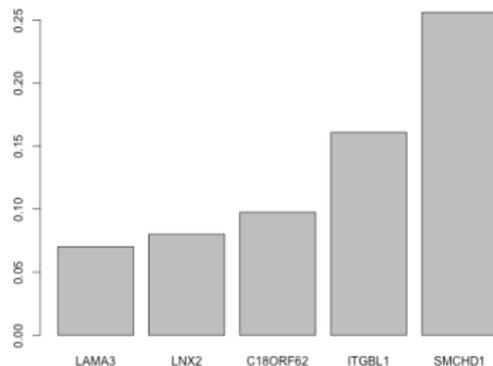
# Shiny Application GLMM: gene stability

The screenshot shows a web browser window displaying a Shiny application titled "dPCR analysis". The browser's address bar shows the URL "127.0.0.1". The application interface has a green header bar with the text "Calculating results. Please wait." on the right. On the left, there is a dark sidebar with a menu containing "Data management", "Analysis", "Absolute quantification", "Copy number variation", "Reference gene stability", "Graphics", and "Report". The "Analysis" menu is expanded, showing "Reference gene stability" as the selected option. The main content area is divided into two panels. The left panel, titled "Select reference genes to use", contains a list of five genes with checkboxes: C18ORF62, ITGBL1, LAMA3, LNX2, and SMCHD1. All checkboxes are checked. Below the list is an "Analyse" button. The right panel, titled "Results", is currently empty.

# Shiny Application GLMM: gene stability

Results

	Gene	Mean	Deviation
1	LNX2	2.05	0.08
2	SMCHD1	1.75	0.26
3	ITGBL1	2.16	0.16
4	LAMA3	2.00	0.07
5	C18ORF62	2.08	0.10



# Shiny Application GLMM: CNV

The screenshot shows a web browser window displaying a Shiny application titled "dPCR analysis". The browser's address bar shows the URL "127.0.0.1". The application interface has a green header bar with the text "Calculating results. Please wait." on the right. On the left, there is a dark sidebar menu with the following items: "Data management", "Analysis" (expanded), "Absolute quantification", "Copy number variation", "Reference gene stability", "Graphics", and "Report". The main content area is divided into two sections. The top section is titled "Select targets to quantify" and contains a list of genes with checkboxes: CLIC6, DSCR3, SYNJ1, AMELX, AMELY, ATP11C, IL13RA1, and SRY. The bottom section is titled "Select reference genes to use" and contains a list of genes with checkboxes: C18ORF62, ITGBL1, LAMA3, LNX2, and SMCHD1. Below this list is an "Analyse" button. To the right of the main content area is a box labeled "Results", which is currently empty.

# Shiny Application GLMM: CNV

dPCR analysis

Data management

Analysis

- Absolute quantification
- Copy number variation
- Reference gene stability

Graphics

Report

Select targets to quantify

- CLIC6
- DSCR3
- SYNJ1
- AMELX
- AMELY
- ATP11C
- IL13RA1
- SRY

Select reference genes to use

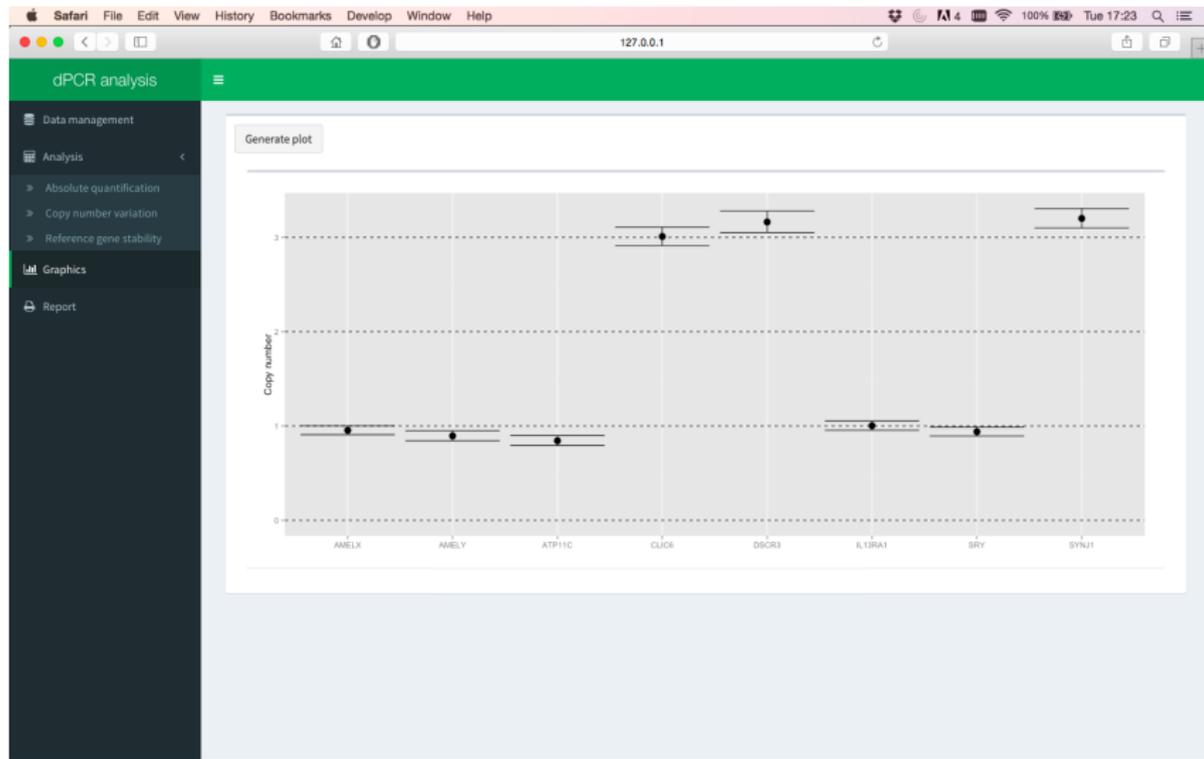
- C18ORF62
- ITGBL1
- LAMA3
- LNX2
- SMCHD1

Analyse

Results

Name	Estimate	95% CI lower bound	95% CI upper bound
1 ATP11C	0.84	0.79	0.90
2 AMELX	0.95	0.91	1.00
3 AMELY	0.89	0.84	0.95
4 SRY	0.94	0.89	0.99
5 IL13RA1	1.00	0.96	1.05
6 SYNJ1	3.20	3.10	3.31
7 DSCR3	3.16	3.05	3.28
8 CLIC6	3.01	2.91	3.11

# Shiny Application GLMM: CNV



# Shiny Application GLMM: CNV

The screenshot shows a web browser window displaying a Shiny application titled "dPCR analysis". The browser's address bar shows the URL "127.0.0.1". The application interface features a dark grey sidebar on the left with a menu containing the following items: "Data management", "Analysis", "Absolute quantification", "Copy number variation", "Reference gene stability", "Graphics", and "Report". The "Report" item is currently selected. The main content area is white and contains a form for generating a report. The form includes a text input field for the "Title of the report" with the value "dPCR analysis report". Below this, there are two sections: "Report contents" with checkboxes for "Data" (checked), "Results" (checked), and "Graphics" (unchecked); and "Document format" with radio buttons for "PDF" (selected), "HTML", and "Word". At the bottom of the form is a "Download report" button with a download icon.



## Digital PCR @ Ghent University

[HOME](#)[ORGANIZATION](#)[RESEARCH ▾](#)[EVENTS ▾](#)

## Home

Digital PCR @ Ghent University is a group of researchers (professors, postdocs, PhD students) combining their expertise from biomedical science and statistics to advance the development and dissemination of novel applications and methodology for digital PCR platforms.

Our currently used platforms include the Bio-Rad QX100, Bio-Rad QX200 and Stilla Naica platforms. In most cases, our methods are, however, applicable to other dPCR platforms.

dpcr.ugent.be

## Digital PCR @ Ghent University

HOME

ORGANIZATION

RESEARCH ▾

EVENTS ▾

## Software

- Web tools
  - ddpcrquant (Shiny web application for threshold setting)  
Access: <http://antonov.ugent.be:3838/ddpcrquant/>
  - dPCalibRate (Shiny web application for quality control)  
Access: <http://antonov.ugent.be:3838/dPCalibRate/>
  - dPCR (Shiny web application for GLMM analysis)  
Access: <http://antonov.ugent.be:3838/dPCR/>
  - Power (Shiny web application for dPCR power analysis)  
Access: <http://vandesompelelab.ugent.be/power/>
  - PrimerXL (website for primer design)  
Access: <http://www.primerxl.org>
  - Variance components

# Summary

- We propose to use GLMMs for analysis of dPCR data:
  - binomial distribution with a complementary log-log link
  - **flexible framework** that is adjustable to many settings
  - based on widely available methodology (e.g. in R)
  - R functions and **easy-to-use Shiny applications** are available