GHENT
UNIVERSITY

Digital PCR data analyis:

A flexible framework for absolute and relative quantification and CNV

Olivier Thas

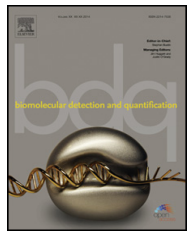
Ghent University, Belgium

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]

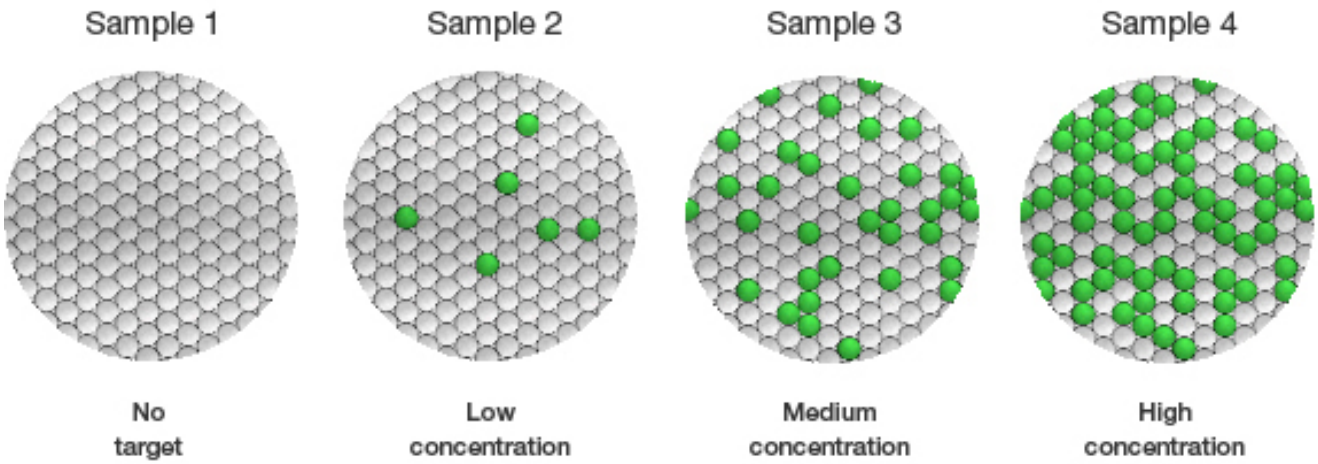
Quantification: classical approach

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 ($j = 1, \ldots, N$).

What we observe is the digital outcome

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



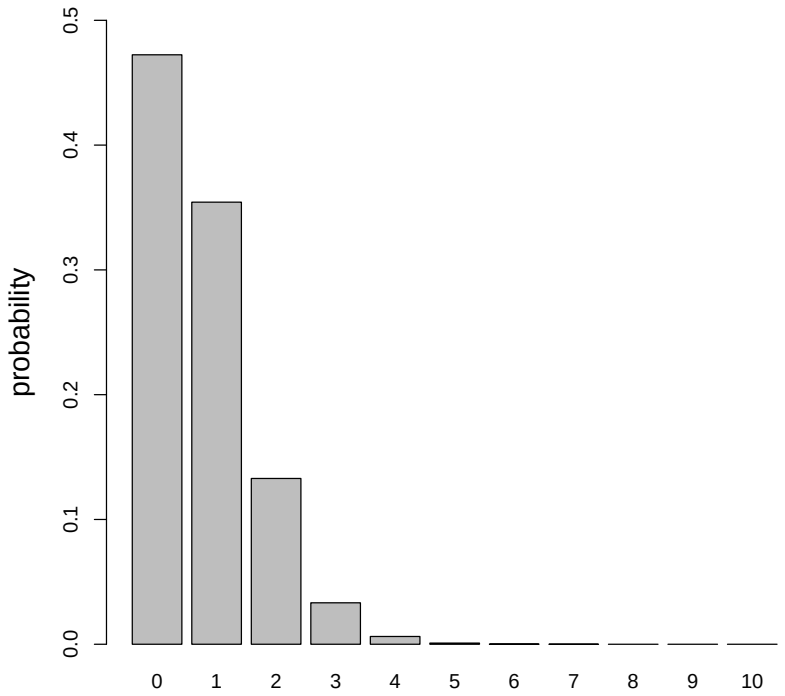
Positive/Negative digital observations (Bio-Rad Laboratories)

Quantification: classical approach

dPCR data analysis methods rely on the Poisson assumption,

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

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



Quantification: classical approach

ddPCR data analysis methods rely on the Poisson assumption,

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

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

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

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

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

The classical approach consists in utilising the Poisson distribution:

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

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:

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

Hence,

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

$\longrightarrow$ GLM software provides estimates and standard errors

Quantification: GLM

The mean number of copies per droplet is still

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

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.

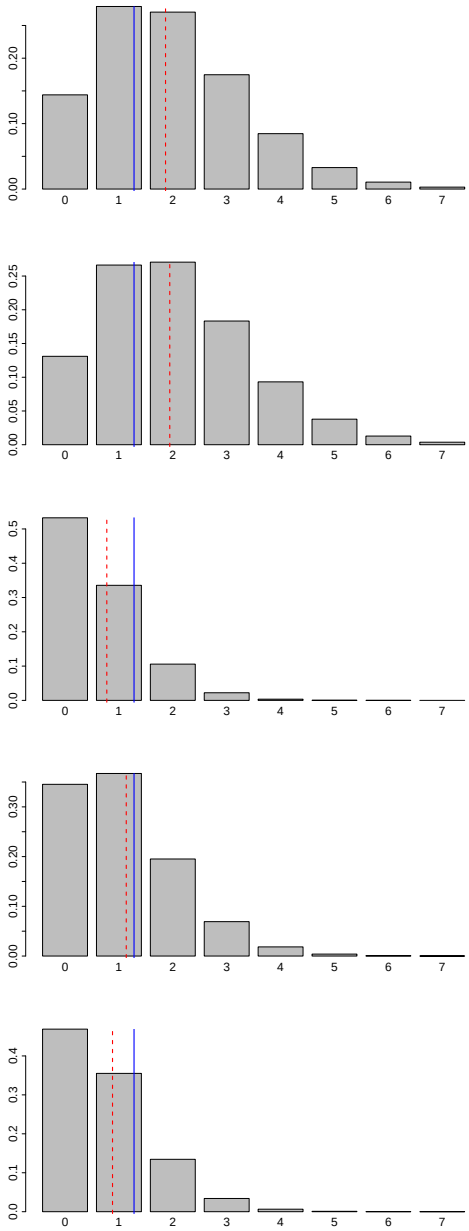
A typical design:

- 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

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

Quantification: GLMM

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

Generalised Linear Mixed Model (GLMM)

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

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

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

Once β and σ^2 are estimated, the target concentration is estimated as

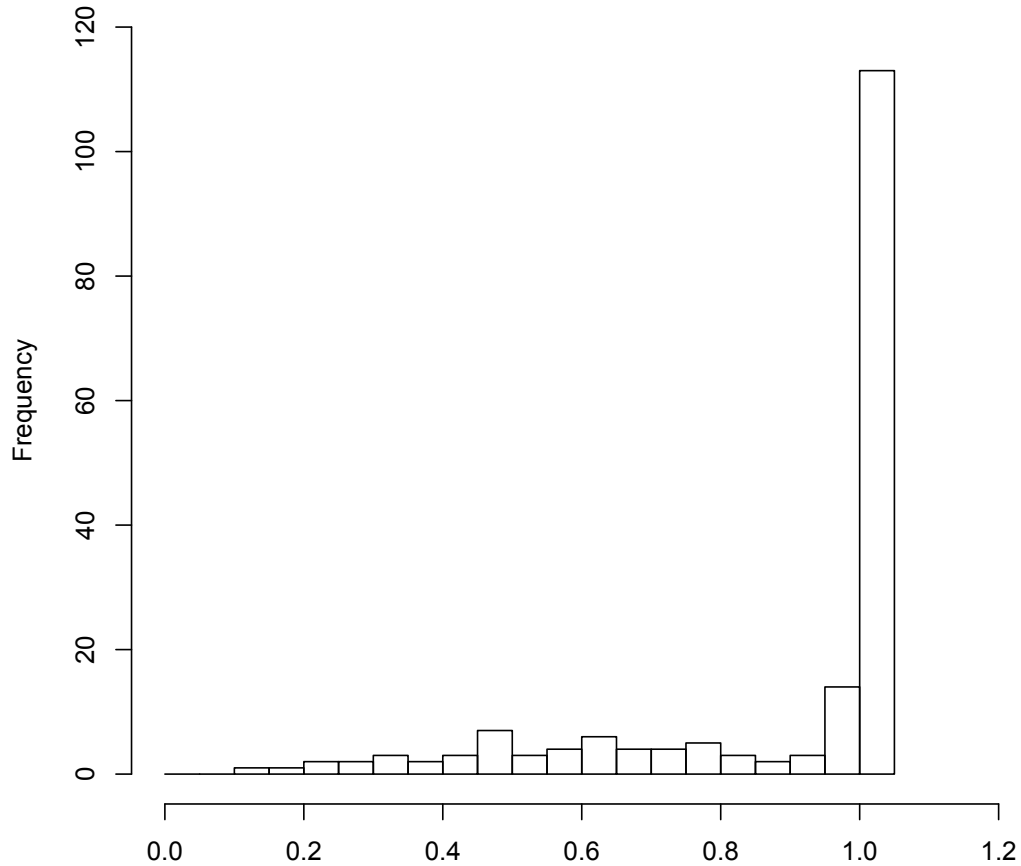
$$\hat{c} = \hat{\lambda}/V_{\text{droplet}} = \exp(\hat{\beta} + 0.5\hat{\sigma}^2)/V_{\text{droplet}}.$$

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

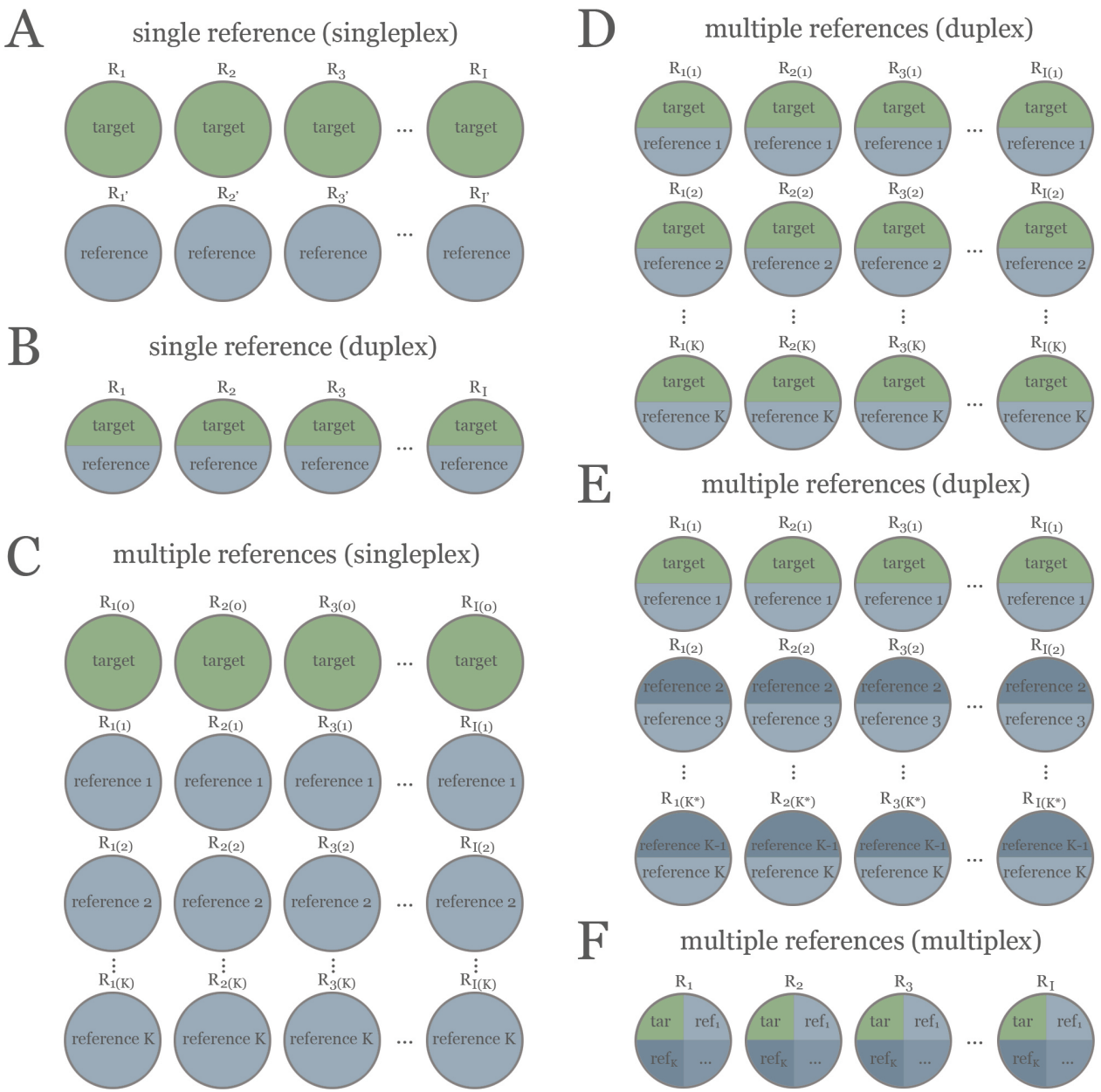
$$CNV = \frac{c_{\text{target}}}{c_{\text{ref}}} N_b,$$

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

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

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

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

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 $k = 1, \ldots, K$, 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

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

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:

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

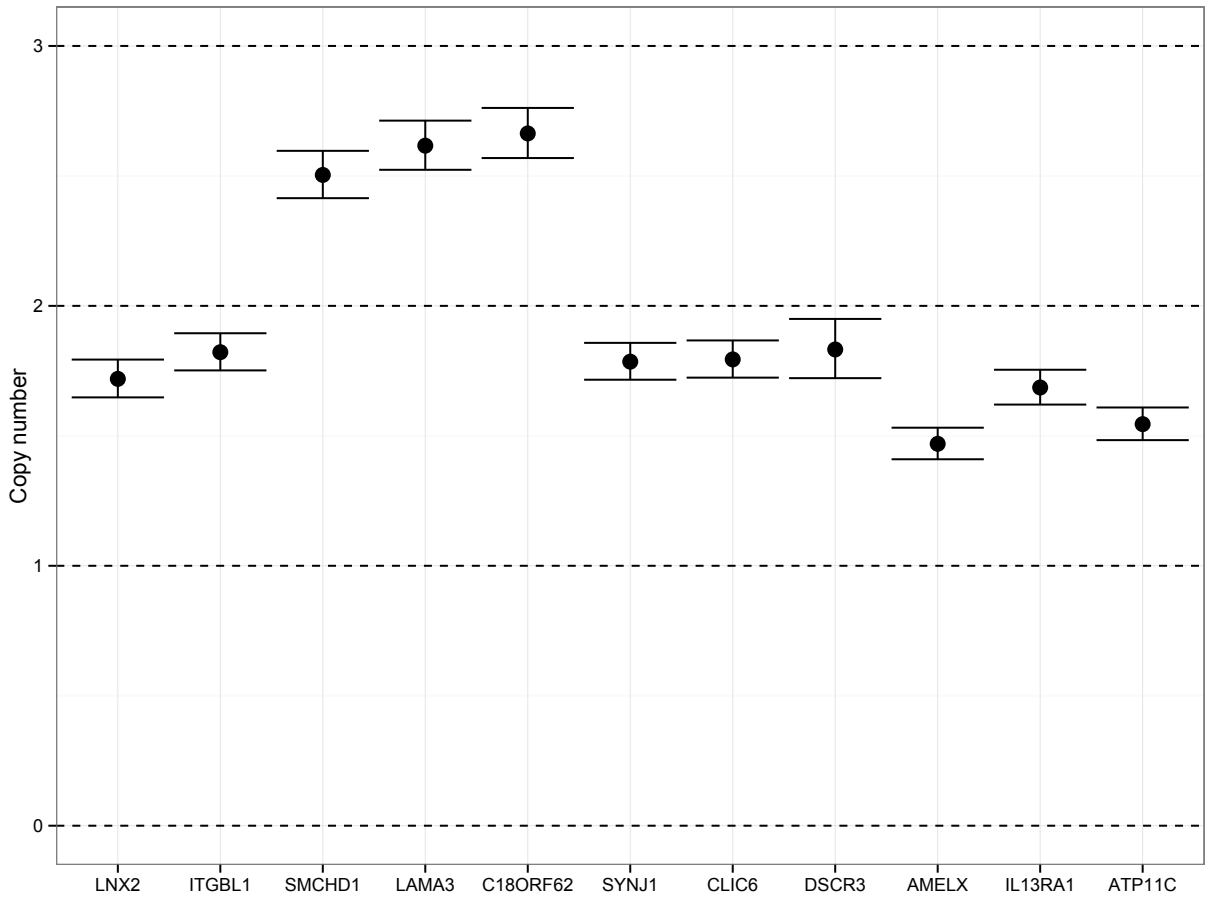
For design C the CNV is first given for target versus a single reference k, based on replicates i and i':

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

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

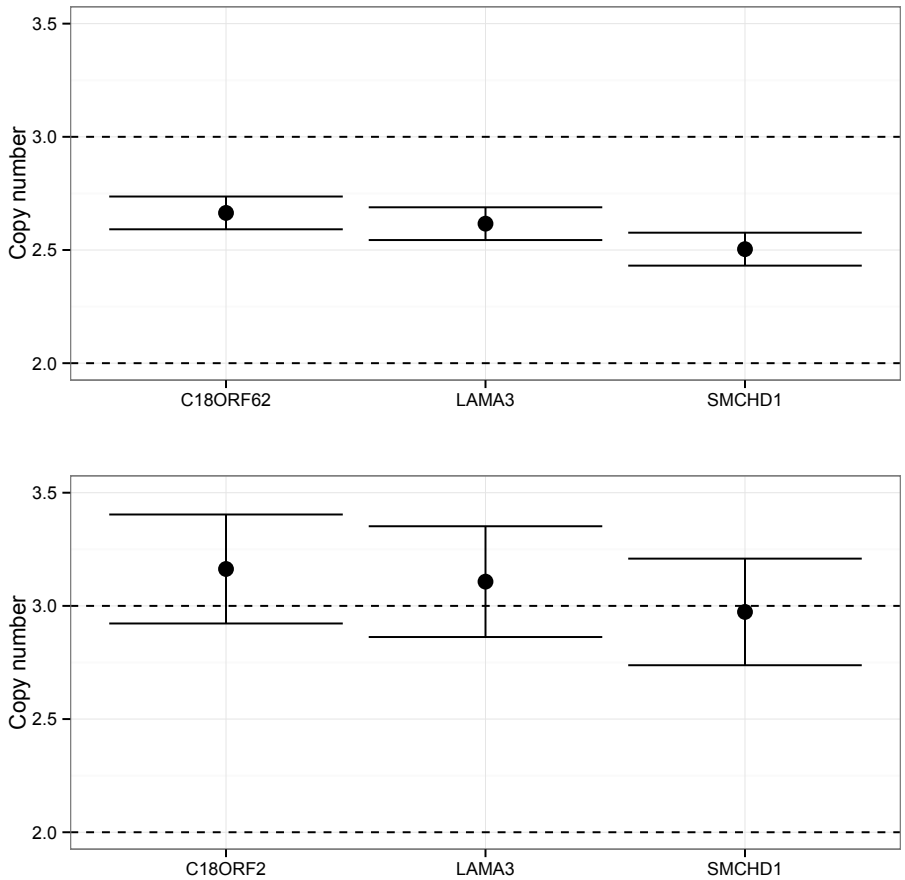
$$\mathrm{CNV} = \mathrm{E}\{\mathrm{CNV}_{i,i';k}\} = \exp\left(-\beta_1 + \frac{1}{2}\sigma_1^2 + \sigma_2^2\right) N_b. \tag{27}$$

Copy number variation: single versus multiple ref. genes



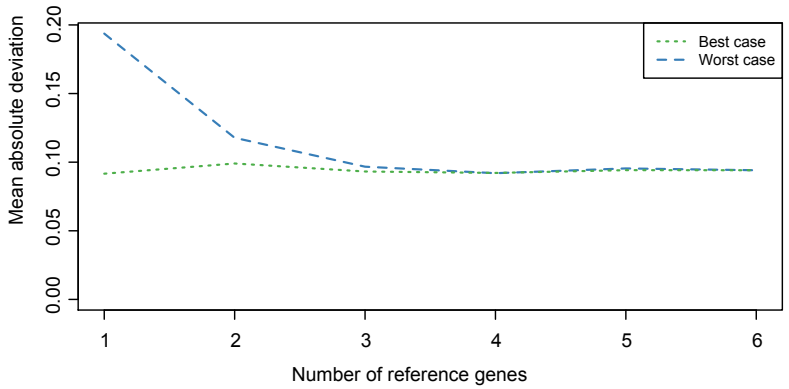
$\longrightarrow$ bias caused by single ref. gene

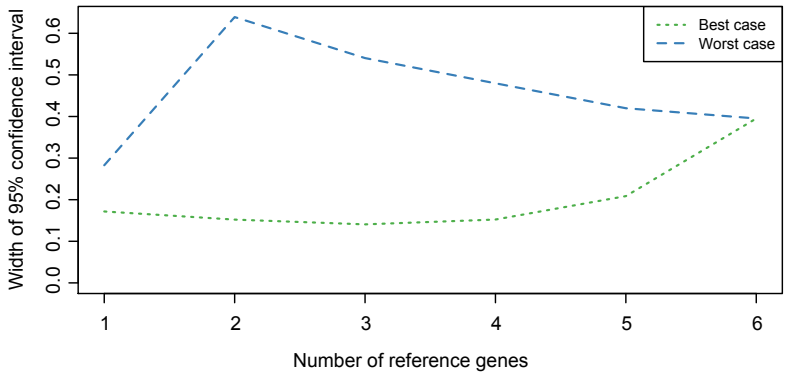
Copy number variation: single versus multiple ref. genes

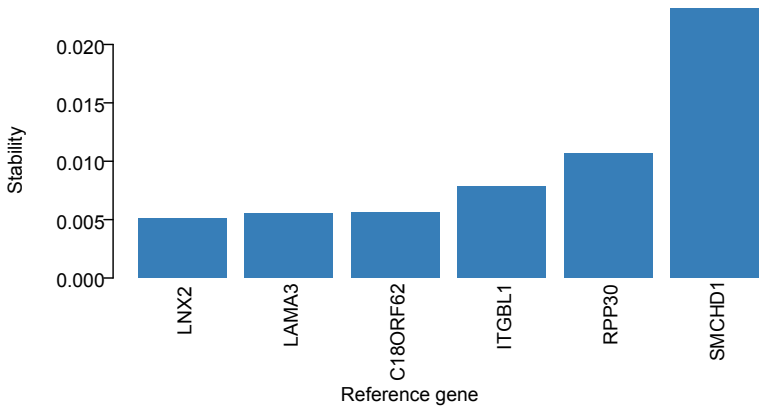


$\longrightarrow$ bias resolved by using multiple ref. genes

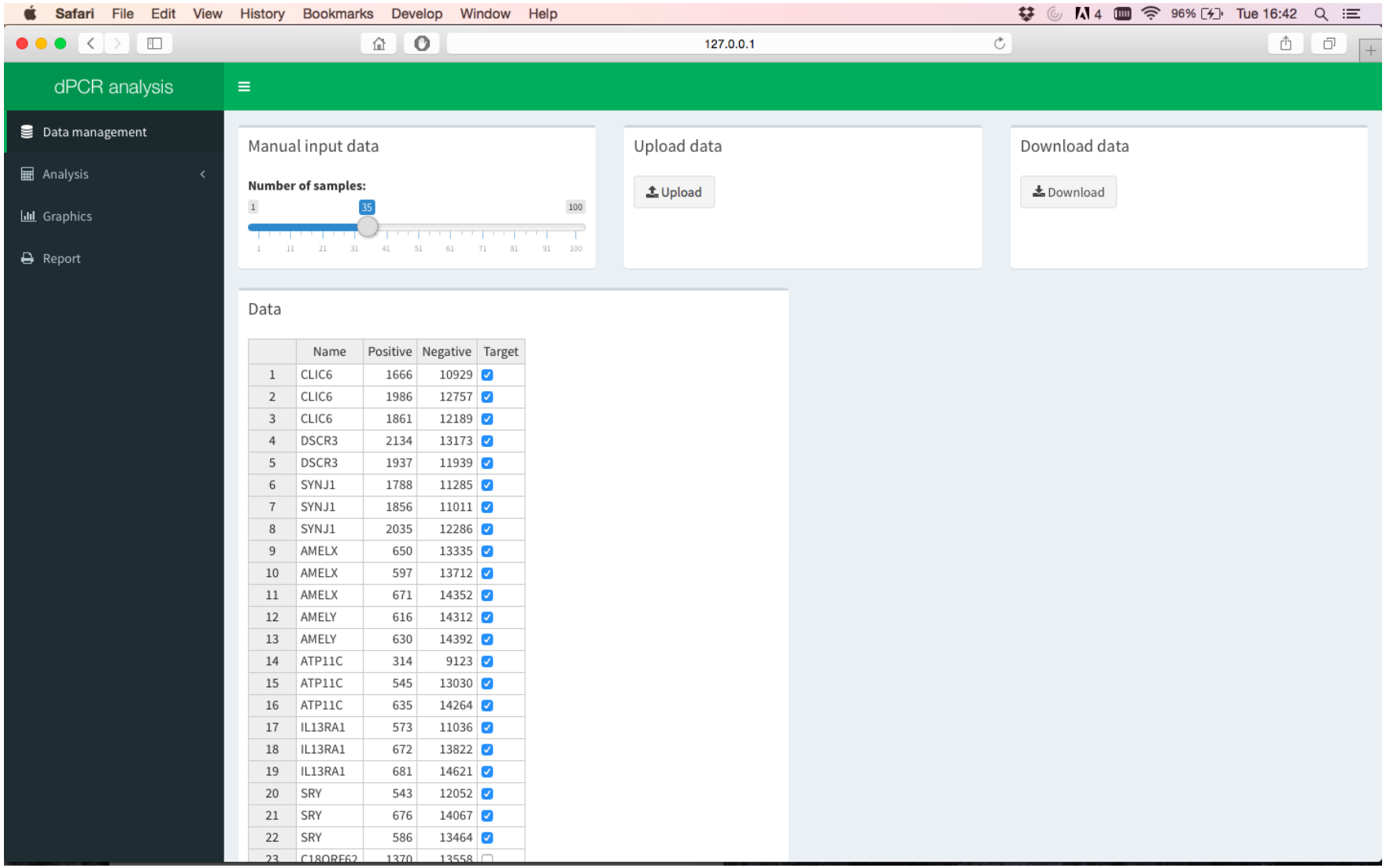
Copy number variation: selection of ref. genes



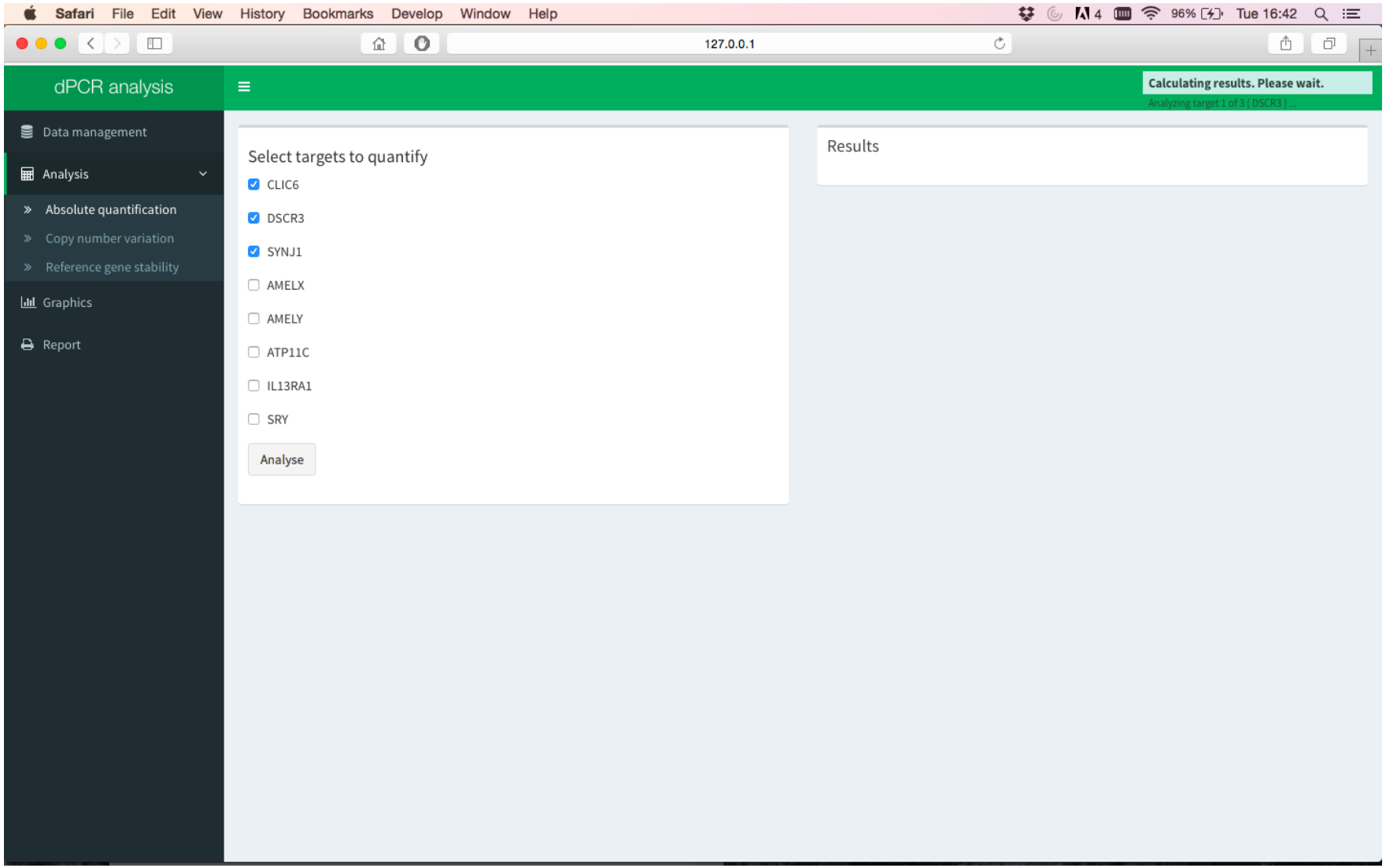




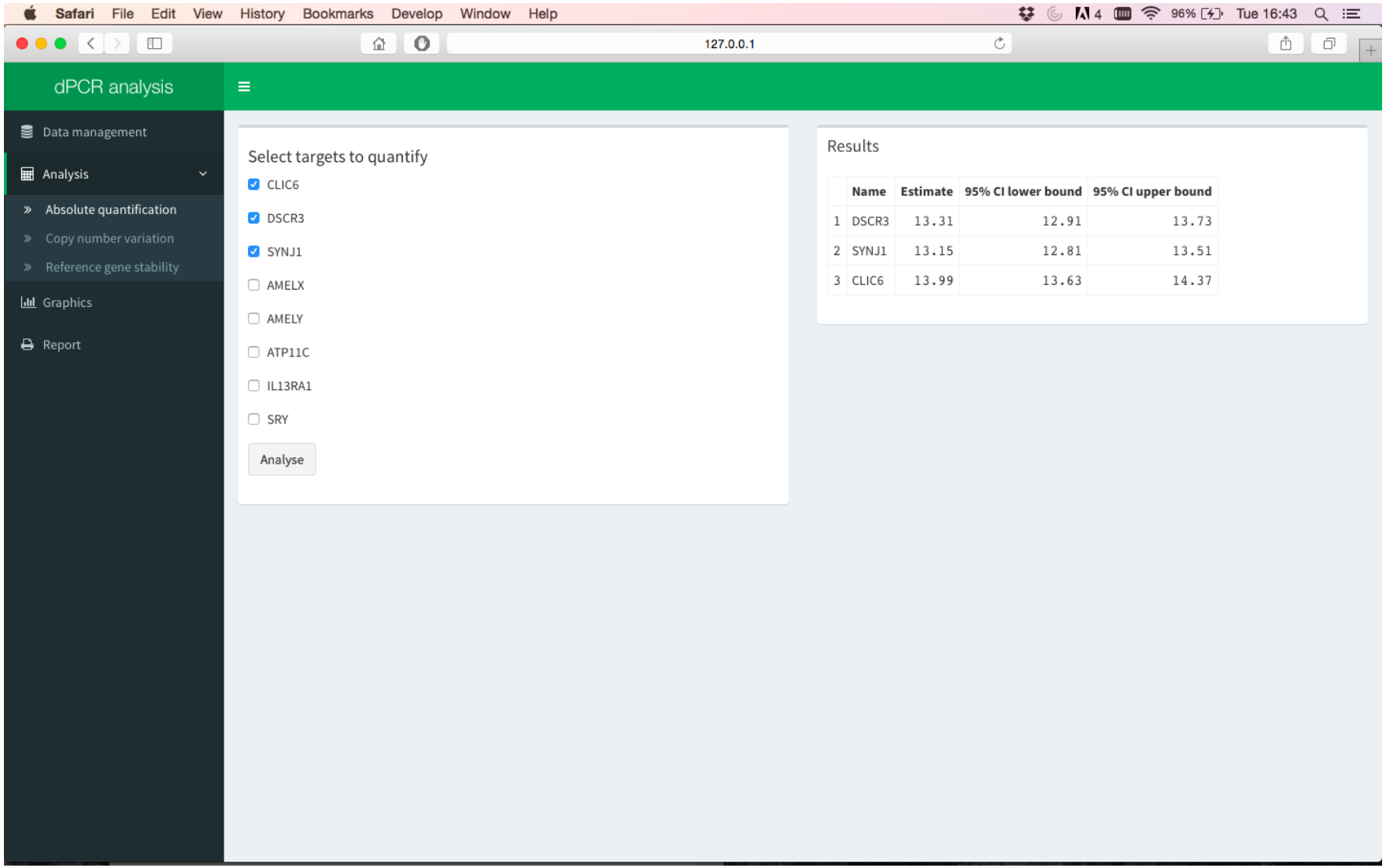
Shiny Application GLMM: read data



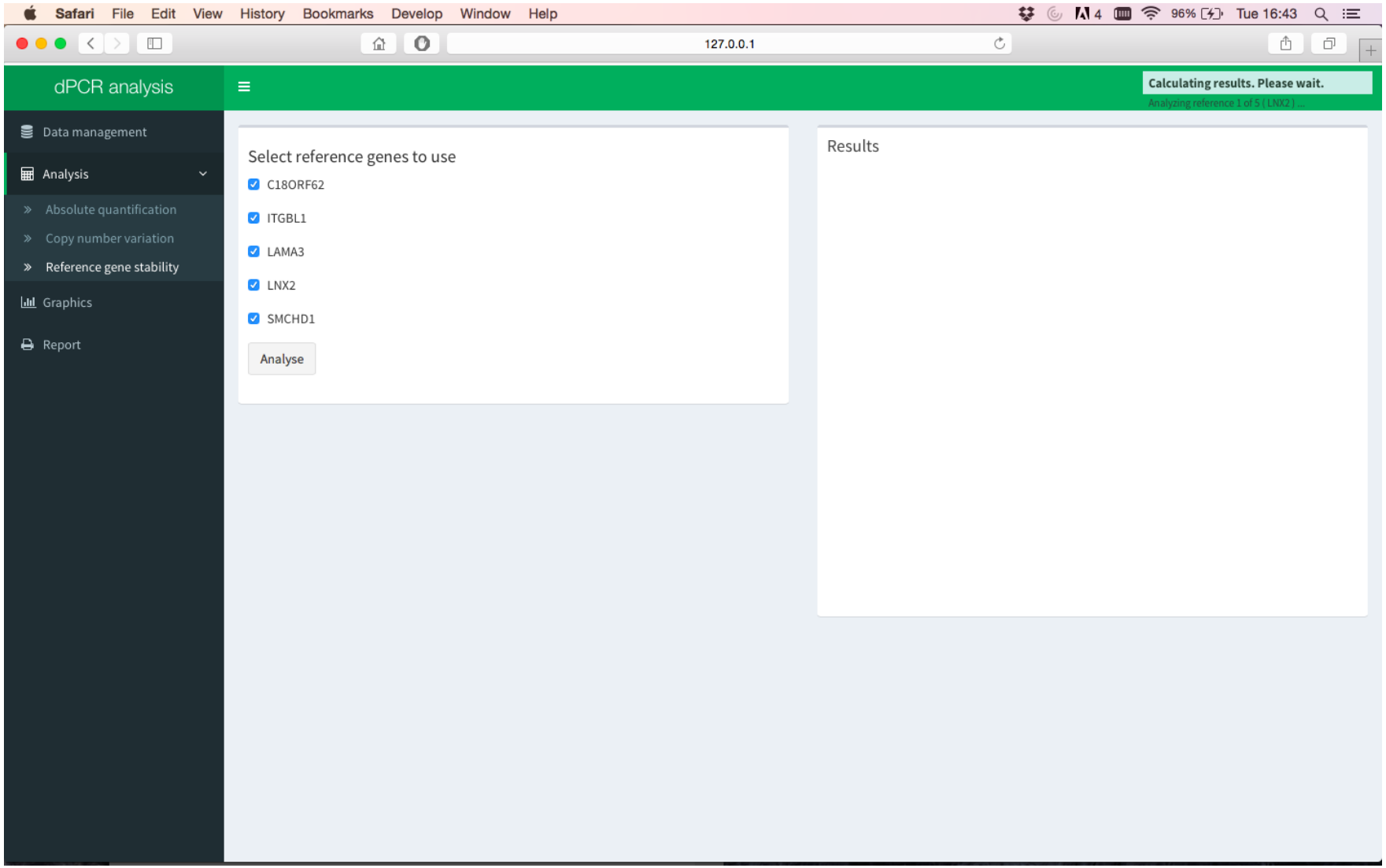
Shiny Application GLMM: absolute quantification



Shiny Application GLMM: absolute quantification



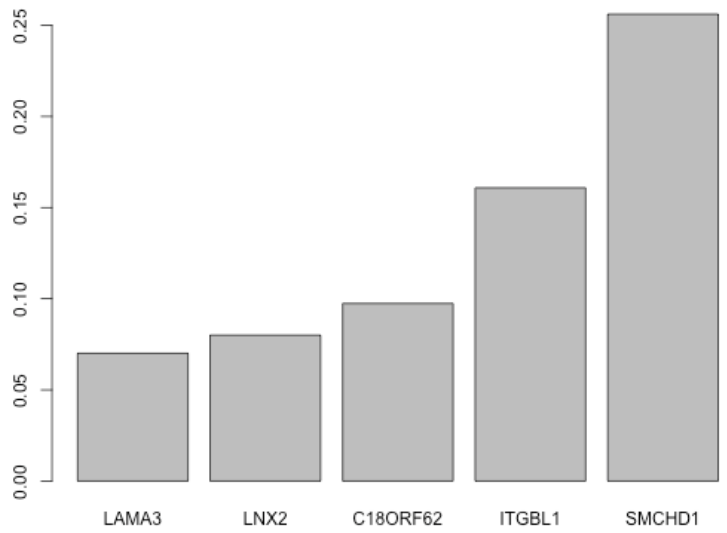
Shiny Application GLMM: gene stability



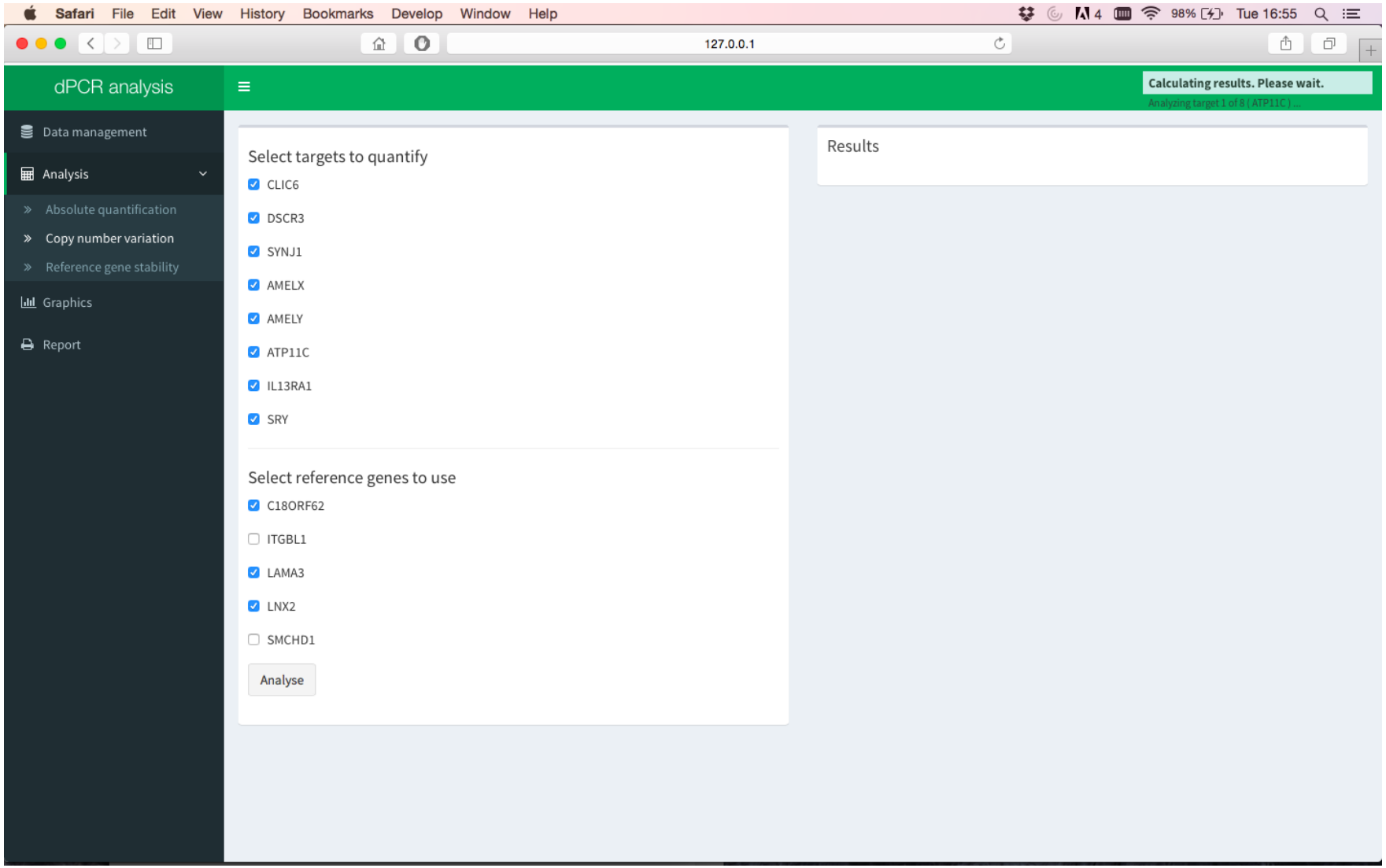
Shiny Application GLMM: gene stability

Results

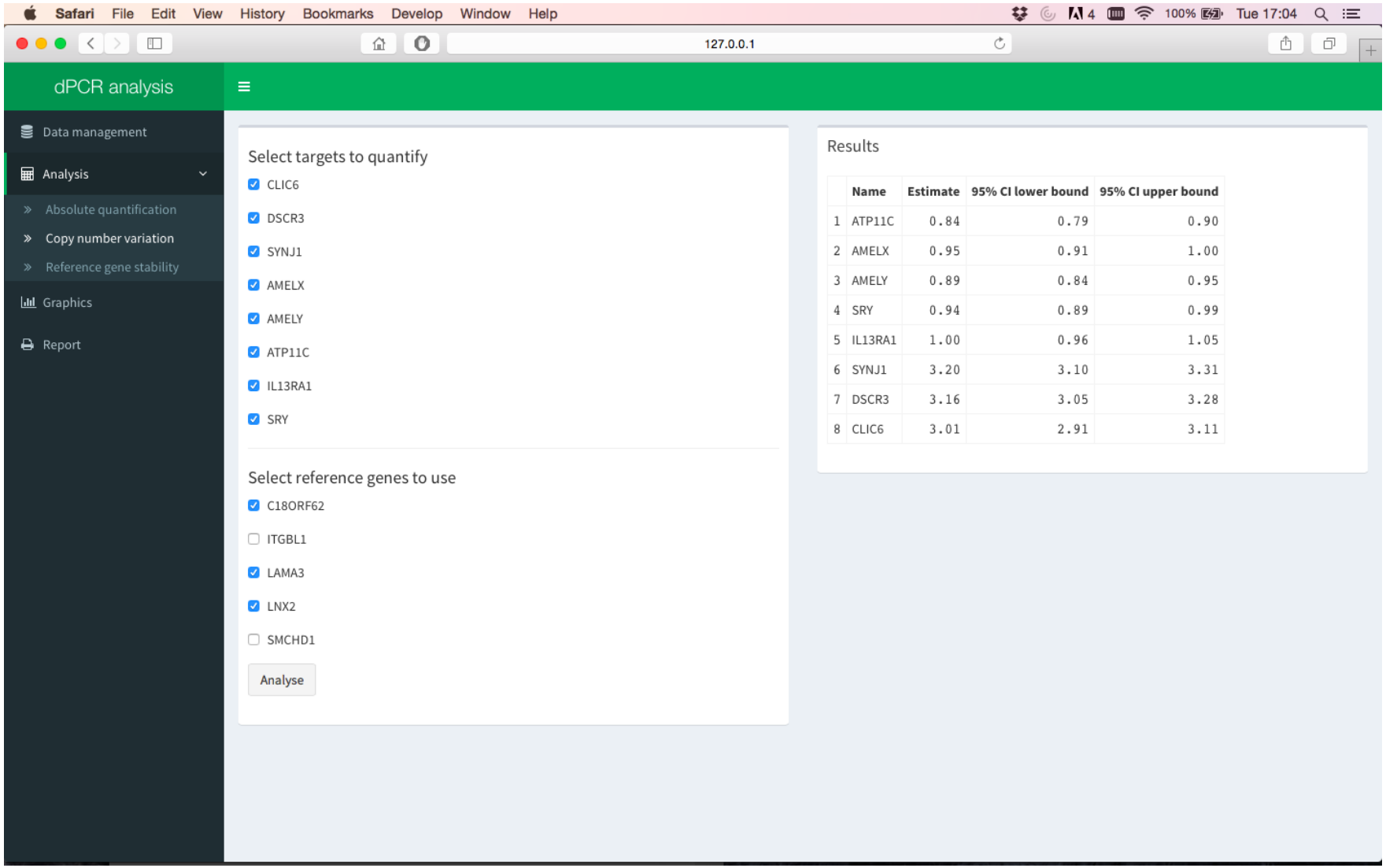
	Gene	Mean	Deviation
1	LNX2	2.05	0.08
2	SMCHD1	1.75	0.26
3	ITGBL1	2.16	0.16
4	LAMA3	2.00	0.07
5	C18ORF62	2.08	0.10



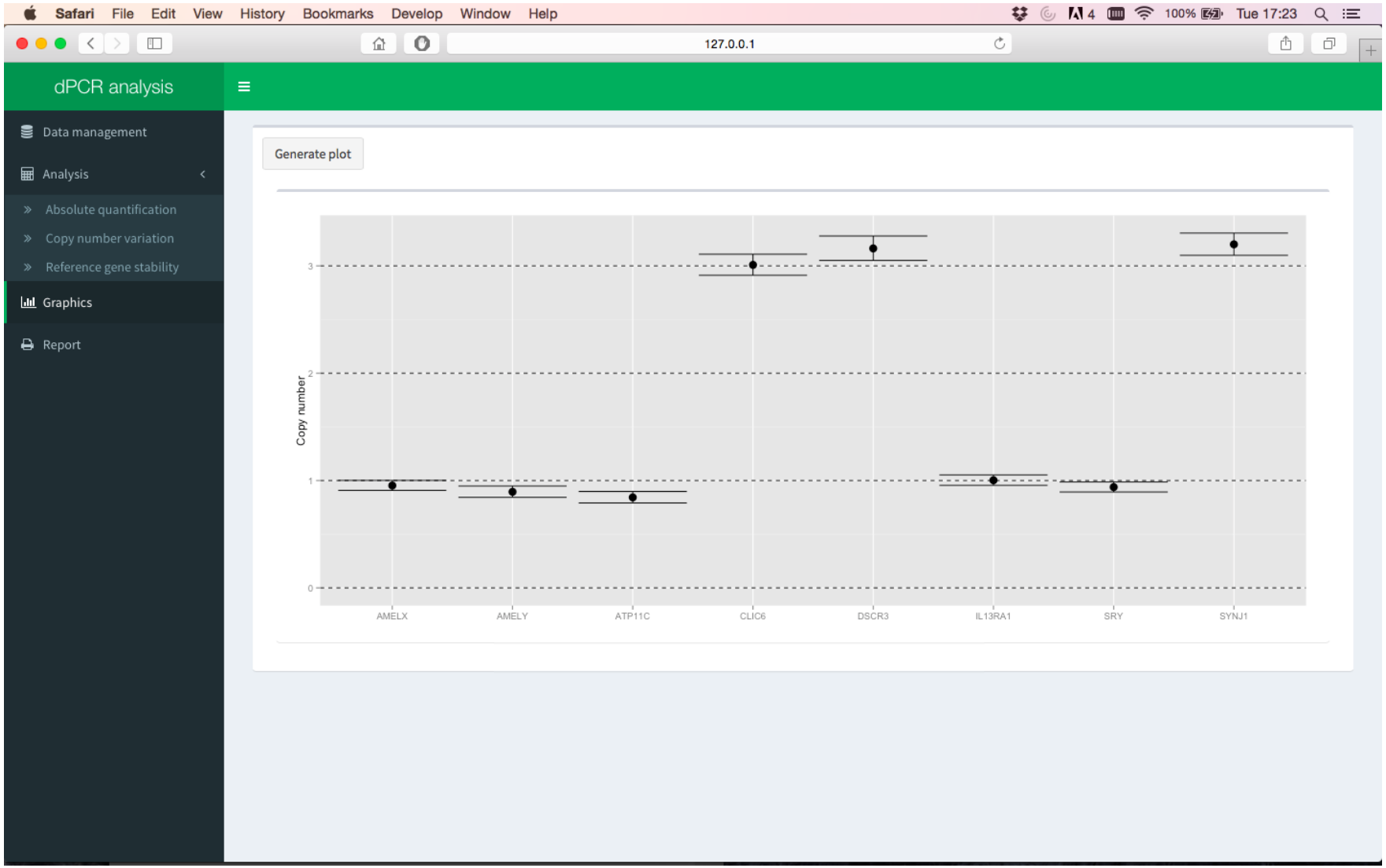
Shiny Application GLMM: CNV



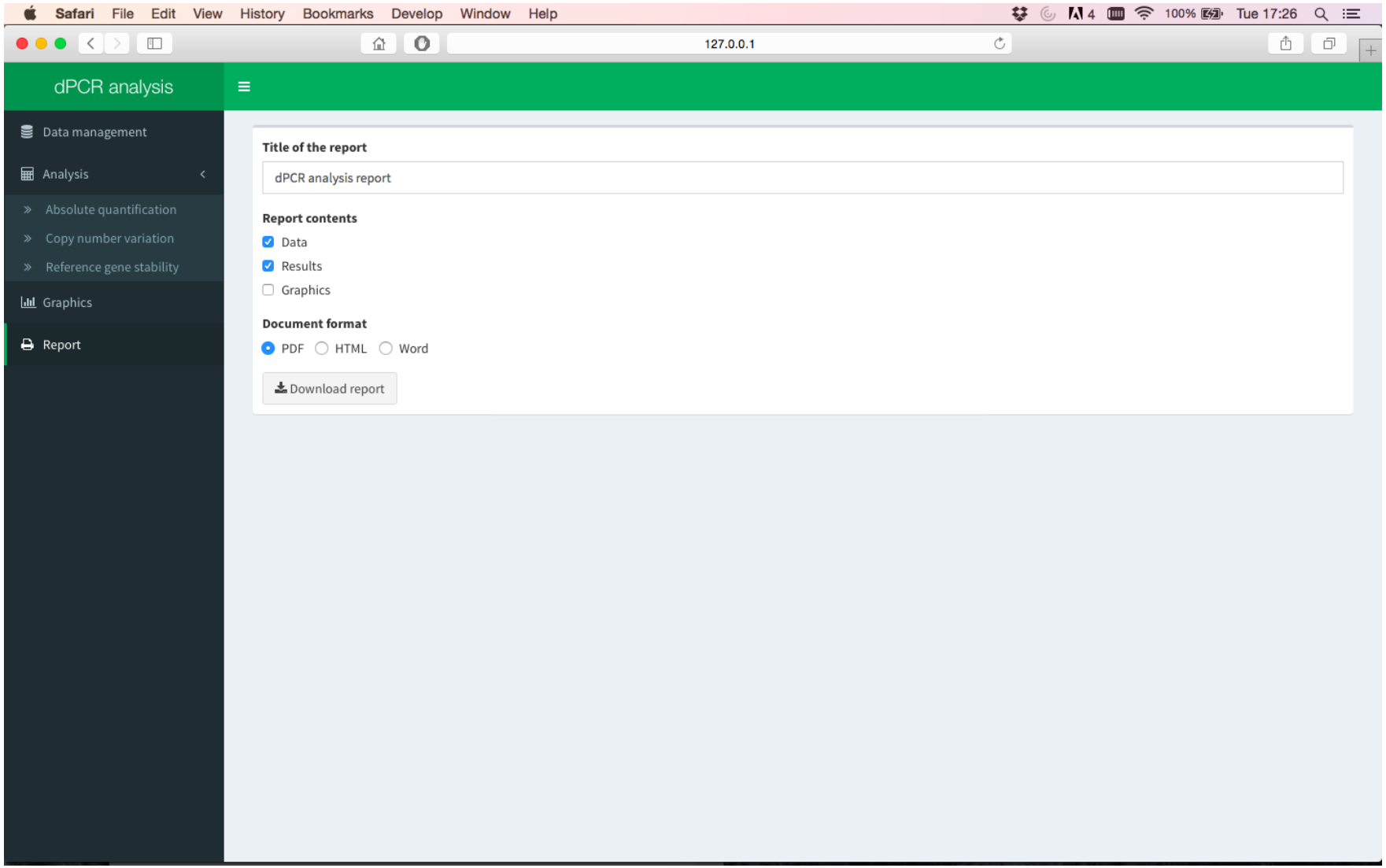
Shiny Application GLMM: CNV



Shiny Application GLMM: CNV



Shiny Application GLMM: CNV



dPCR Website UGent

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

dPCR Website UGent

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**

dpcr.ugent.be

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