Harbin: A quantitation PCR analysis tool

Rachelle Bester[1], Pieter T Pepler[2], Dirk J Aldrich[1] and Hans J Maree[1,3*]

[1]Department of Genetics, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa.

[2]Institute of Biodiversity Animal Health and Comparative Medicine, University of Glasgow, G12 8QQ, Glasgow.

[3]Agricultural Research Council, Infruitec-Nietvoorbij: Institute for Deciduous Fruit, Vines and Wine, Private Bag X5026, Stellenbosch, 7599, South Africa

*Corresponding author: Hans J Maree

email: hjmaree@sun.ac.za

Telephone number: 021 808 9579

Fax number: 021 808 5833

Abstract

Objectives

To enable analysis and comparisons of different relative quantitation experiments, a web-browser application called Harbin was created that uses a quantile-based scoring system for the comparison of samples at different time points and between experiments.

Results

Harbin uses the standard curve method for relative quantitation to calculate concentration ratios (CRs). To evaluate if different datasets can be combined the Harbin quantile bootstrap test is proposed. This test is more sensitive in detecting distributional differences between

data sets than the Kolmogorov-Smirnov test. The utility of the test is demonstrated in a comparison of three grapevine leafroll associated virus 3 (GLRaV-3) RT-qPCR data sets.

Conclusions

The quantile-based scoring system of CRs will enable the monitoring of virus titre or gene expression over different time points and be useful in other genomic applications where the combining of data sets are required.

Keywords

Bootstrap; genetic variants; GLRaV-3; Location-scale-shape problem; Quantile-based scoring; RT-qPCR

Article section

Plant Cell Technology

Introduction

Quantitative polymerase chain reaction (qPCR) is a widely used technique to measure expression levels of nucleic acids. Absolute quantitation uses a fixed calibration curve that makes comparing different experiments easier, however relative quantitation compensates for differences in tissue types, environmental conditions, integrity of RNA, loading error and reaction efficiency. A concentration ratio (CR) can be obtained to compare the concentration of a gene of interest relative to stable reference genes. A relative quantitation model with an efficiency correction is recommended since a small difference in target assay efficiency and reference gene assay efficiency can result in a false expression ratio (Pfaffl 2001; Bester et al. 2014).

One of the most important viral diseases of grapevine worldwide is grapevine leafroll disease (GLD) with grapevine leafroll-associated virus 3 (GLRaV-3) considered as the main etiological agent contributing to the disease (Maree et al. 2013). Currently, the complete genomes of 13 distinct GLRaV-3 isolates representing five of the eight major genetic variant groups are available (Maree et al. 2015). Little is known about the biological characteristics of the different GLRaV-3 genetic variants and it is therefore important to investigate whether there is significant variation between the variant groups beyond the genome. One parameter to investigate would be the CR of the different groups over time.

The comparison of replicate experiments over time is complicated by differences in the location, scale and shape of the population distributions, i.e. data with differences in these parameters are not directly comparable. The most commonly used method to determine the compatibility of data is to test for shifts in shape (distribution), location (mean) and scale (variance). To address this, we propose a new bootstrap test for hypothesis against the location-scale-shape alternative, based on quantiles of the empirical distributions of two data

sets. This test is not based on any assumptions about the shapes of the distributions, and is powerful in detecting differences in location, scale and/or shape simultaneously. The accuracy of this novel test was compared in a Monte Carlo simulation study to the well-known Kolmogorov-Smirnov test (Kolmogorov 1933).

A software tool called Harbin is presented, for the analysis of real-time qPCR data using a relative quantitation strategy. It allows for the combining of different qPCR data sets/experiments to enable comparisons of different relative quantitation experiments. Harbin runs within the R statistical computing environment (R Core Team, 2013) on all major platforms. It is also freely available as a graphical user interface (GUI) utilizing the Shiny web-based package that requires no additional software installations. The utility of Harbin was demonstrated using three GLRaV-3 RT-qPCR data sets to investigate if the data sets can be combined to study variation in virus variant concentrations.

Methods

Plant material

Three independent sample groups were selected for this study, all consisting of *Vitis vinifera* cv. Cabernet Sauvignon plants. The first data set included 30 samples of which 15 samples were infected with grapevine leafroll associated virus 3 (GLRaV-3) variant group II and 15 samples infected with GLRaV-3 variant group VI. The second data set included 12 plants singly infected with either variant group I, II, III or VI (three plants each). The third data set included 37 plants of which seven plants were infected with variant groups I, eight plants infected with variant group II, eight plants infected with variant group III, eight plants infected with variant group VI and six plants infected with variant group VII.

Due to GLRaV-3 being a phloem-limited virus, phloem material from each plant shoot was collected and stored at −80 °C. Total RNA was extracted from 2 grams of phloem material using a modified CTAB extraction protocol (Carra et al. 2009; Bester et al. 2014). All plants were confirmed to be infected with only GLRaV-3 after testing negative for frequently occurring grapevine viruses using RT-PCRs (Jooste et al. 2015). GLRaV-3 variant group status of all plants was confirmed using the previously designed real-time RT-PCR high-resolution melting curve analysis assay (Bester et al. 2012).

RT-qPCRs

In order to calculate the virus CR in each plant, RT-qPCRs were performed using previously designed assays targeting ORF1a of GLRaV-3 and three *V. vinifera* reference genes targeting actin, alpha-tubulin and glyceraldehyde 3-phosphate dehydrogenase (GAPDH) (Bester et al. 2014). The stability of the reference genes was assessed using BestKeeper (Pfaffl et al. 2004).

Data analysis

The Rotor-gene Q software version 2.3.1 (Qiagen) was used to calculate primer efficiencies, Cq values and gene quantitation values for all targets. For further analysis of the three data sets, an R based application called Harbin was developed to ease the data handling and computational aspects. Harbin runs within the R statistical computing environment (R Core Team, http://www.R-project.org/) on all major platforms, and is available under an open source licence. Harbin is dependent on base R and additional packages (psych, car, beeswarm) available from the Comprehensive R Archive Network (CRAN). Harbin is also available as a graphical user interface (GUI) utilizing the Shiny web-based package. The GUI

can be used in most web browsers and requires only an Internet connection and no installation. The Harbin user manual is available for download from within the application or at https://github.com/Rbester18/Harbin. Harbin has a direct input option for the quantitation files (.csv) generated by the Rotor-Gene Q software (version 2.3.11 and above). The application also allows for the upload of Cq values from any other qPCR platform, provided that a standard curve equation for each gene is available. An example template is available for download from within the application. Normalisation of the gene of interest concentrations are performed with a reference gene index, calculated using the geometric mean of up to ten reference genes. The calculation of fold changes between genes often entails only limited comparisons of values across two conditions, however the Harbin application allows for significance testing of two or more groups using either parametric or non-parametric tests by selecting and classifying individual data points to the number of groups specified. The non-parametric Wilcoxon rank-sum test can be used to assess statistically significant differences between samples infected with different variant groups. The Harbin application and additional information can be used and downloaded at https://rbester.shinyapps.io/Harbin/ and https://github.com/Rbester18/Harbin.

Harbin quantile-based bootstrap test

The Harbin application was used to perform the quantile-based bootstrap test (Harbin-test) to determine if the three data sets are compatible to be combined. For each data set, the 20th, 40th, 60th and 80th percentiles of the CRs distribution are calculated and assigned a score (1–5). A CR in the lowest quantile (0–20%) is assigned a “1”, and a CR in the highest quantile (80–100%) is assigned a “5”. If data from a previous experiment is available and the option to use it as a reference data set is selected, the application will compare the test data to the

reference data set. The Harbin-test adds the data set to the reference dataset and calculates the number of CRs in the reference data set for which the "scores" (1–5) have changed. This test statistic is compared to the distribution of the same statistic calculated from 1000 bootstrap samples (each of the same size as the test data) drawn from the reference data set.

The purpose of the Harbin-test function is to determine whether the samples in a new data set are compatible with those in a well-defined reference data set. The combining of different data sets is performed under the assumption that the samples originate from populations that can be described by the same probability distribution function. Suppose that $\boldsymbol{x}' = [x_1, \ldots, x_n]$ and $\boldsymbol{y}' = [y_1, \ldots, y_m]$ are representative data sets from two continuous univariate populations, G_{ref} and F, respectively. It is of interest to determine whether the two population distributions are homogeneous, or in particular, whether the new data set $\boldsymbol{y}$ is compatible with the reference data set, $\boldsymbol{x}$. The hypothesis of interest is

$$H_0 : F(x) = G_{ref}(x), \text{ for all } x \in (-\infty, \infty), \quad (1)$$

against the general location-scale-shape alternative,

$$H_1 : F(x) \neq G_{ref}(x), \text{ for some } x \in (-\infty, \infty), \quad (2)$$

where F and G_{ref} are continuous univariate probability distribution functions describing the two populations. Hypothesis (2) implies a difference at any point on the two distributions: The medians, variances and/or shapes of the two distributions differ.

The Harbin-test is a quantile-based bootstrap test for hypothesis (1) against the general alternative in (2). The test works as follows: Calculate the 20th, 40th, 60th and 80th percentiles of $\boldsymbol{x}$, indicating these percentiles with Q_{20}, Q_{40}, Q_{60} and Q_{80}, respectively. Let $g_i, i = 1, \ldots, n$ be a variable taking the values,

$$g_i = \begin{cases} 1 & \text{if } x_i \leq Q_{20}, \\ 2 & \text{if } Q_{20} < x_i \leq Q_{40}, \\ 3 & \text{if } Q_{40} < x_i \leq Q_{60}, \\ 4 & \text{if } Q_{60} < x_i \leq Q_{80}, \\ 5 & \text{if } x_i > Q_{80}. \end{cases} \quad (3)$$

Combine the reference and new data sets in a vector, $\boldsymbol{z}' = [\boldsymbol{x}'\boldsymbol{y}']$ and construct a variable, h_i $i = 1, \ldots, n$, taking the values,

$$h_i = \begin{cases} 1 & \text{if } x_i \leq Q^*_{20}, \\ 2 & \text{if } Q^*_{20} < x_i \leq Q^*_{40}, \\ 3 & \text{if } Q^*_{40} < x_i \leq Q^*_{60}, \\ 4 & \text{if } Q^*_{60} < x_i \leq Q^*_{80}, \\ 5 & \text{if } x_i > Q^*_{80}. \end{cases} \quad (4)$$

where Q^*_p indicates the p^{th} percentile of $\boldsymbol{z}$. Let

$$c_i = \begin{cases} 0 \text{ if } g_i = h_i, \\ 1 \text{ if } g_i \neq h_i. \end{cases} \quad (5)$$

The quantity $\sum_{i=1}^{n} c_i$ is thus the number of elements in $\boldsymbol{x}$ for which the "scores" (1–5) have changed in the combined data set, $\boldsymbol{z}$. The test statistic for hypothesis (1) is

$$u = \frac{1}{n}\sum_{i=1}^{n} c_i, \quad (6)$$

which is the proportion of the elements in x for which the scores have changed in the combined data set. To find the distribution of u under the null hypothesis, $r = 1000$ bootstrap samples (Efron and Tibshirani 1994) of size m are drawn from x. Let

$$z_0^{(j)} = \left[\begin{matrix} \boldsymbol{x} \\ \boldsymbol{y_0^{(j)}} \end{matrix}\right], \; j = 1, \ldots, r, \quad (7)$$

where $\boldsymbol{y}_0^{(j)}$ indicates the j^{th} bootstrap sample. Using $\boldsymbol{x}$ and $\boldsymbol{z}_0^{(j)}$, the j^{th} bootstrap replication of the test statistic, $u_0^{(j)}$, is calculated as in (6). The null hypothesis in (1) is rejected at a significance level of α if the test statistic in (6) exceeds the $100(1-\alpha)^{th}$ percentile of $\boldsymbol{u}_0' = \left[u_0^{(1)}, \ldots., u_0^{(r)}\right]$.

Two example data sets are available on github (https://github.com/Rbester18/Harbin) and will be able to serve as independent reference data sets if the same qPCR protocol and reagents are used as described in this study.

Monte Carlo simulation study

A Monte Carlo simulation study was performed to compare the size and power of the Harbin-test to the Kolmogorov-Smirnov test (Hollander et al. 2013). Compared to the number of available tests for common location and/or homogeneity of variances for two groups, relatively few tests have been proposed to test for equality of the population distributions. A

well-known non-parametric test for the two-sample hypothesis in (1) against the location-scale-shape alternative in (2) is the Kolmogorov-Smirnov test.

The Kolmogorov-Smirnov test compares the empirical distribution functions of two data sets. If differences in the locations, scales or shapes of the empirical distribution functions are sufficiently large, the conclusion is made that the two population distribution functions differ.

For the first ("reference database") group, data sets of sizes $n_1 = 10, 30\ or\ 50$ were simulated from populations with one of the following four distributions:

- 1a. Normal: $N(3, 1)$;
- 1b. Chi-squared with three degrees of freedom: x_3^2;
- 1c. Uniform distribution on the $[0, 6]$ interval;
- 1d. Bimodal: Half of observations from a $N(1.5, 0.75^2)$ distribution, with the other half from a $N(4.5, 0.75^2)$ distribution.

For all four of the distribution types, the majority of the observations will thus lie on the $[0, 6]$ interval, as can be seen in Fig. 1. For the second ("new data") group, data sets of size n_2 for ratios $\frac{n_2}{n_1} = 0.5, 1\ or\ 2$, were simulated from populations with one of the following four distribution types:

- 2a. Normal: $N(3 + \delta, 1\gamma)$;
- 2b. Chi-squared with 3γ degrees of freedom, shifted to the right by addition of the value δ; i.e. $x_{3\gamma}^2 + \delta$
- 2c. Uniform distribution on the $[0, 6\gamma]$ interval, shifted by addition of the quantity $(-3\gamma + 3 + \delta)$;
- 2d. Bimodal: Half of observations from a $N(1.5 + \delta, (0.75\gamma)^2)$ distribution, with the other half from a $N(1.5 + 3\gamma + \delta, (0.75\gamma)^2)$ distribution.

The mean shift values, $\delta = 0, 0.2, 0.5, 1, 1.5, 2$, and standard deviation shift values, $\gamma = 1, 1.5, 2$, were varied to determine the power of the two tests to detect shifts in location and scale, respectively. For each ($n_1 : n_2 :$ *Distribution 1 type*: *Distribution 2 type* : $\delta : \gamma$) factorial treatment combination, a total of $r = 1000$ simulation runs were performed. The simulation study was performed on the Rhasatsha high-performance computer (HPC) at Stellenbosch University (http://www.sun.ac.za/hpc), using R (R Core Team, 2013). For each test per simulation run, a significance level of 5 % was used to decide whether to reject the null hypothesis.

The Harbin application has the option to apply either the Harbin-test or the Kolmogorov-Smirnov test to test the two-sample hypothesis. If the hypothesis that the two data sets originated from populations with the same probability distribution function seems plausible, the Harbin application allows for the option to add the new data set to the reference data set. The quantile scores of the data in the reference data set will be adjusted according to the new combined data distribution.

Results and discussion

RT-qPCRs

The utility of the Harbin application is demonstrated in a comparison of three GLRaV-3 RT-qPCR data sets. The requisite control reactions were included in all data sets, and as expected no virus CRs were generated for GLRaV-3 negative plant samples. The statistics of the standard curves generated for each assay per data set can be seen in Table 1. The PCR efficiencies and linearity calculated from all assays' standard curves were high and no evidence of inhibition was seen from the Cq values of the dilution series. These assays

complied with the Minimum Information for publication of Quantitative real-time PCR Experiments (MIQE) guidelines to ensure the integrity of the experiments and facilitate reproducibility (Bustin et al. 2009).

Monte Carlo simulation study

The overall performance of the Harbin-test and the Kolmogorov-Smirnov test was assessed by the percentage of simulation runs for which the null hypothesis was correctly rejected (or not rejected) for the specific test. The Kolmogorov-Smirnov test had the smaller size (2.4%) and power (54%), indicating that it is conservative compared to the Harbin-test, failing to reject an incorrect null hypothesis in a larger proportion of cases. The Harbin-test was found to be consistently more accurate and powerful than the Kolmogorov-Smirnov test, but had a higher false positive rate (10.6%). Therefore the Harbin-test offers a good alternative in situations where the purpose is to avoid considering samples from two different distributions as originating from populations with the same distribution. The power of both tests increases with an increase in the size of the sample from the first ("reference data set") population. For the smallest sample size considered (n_1 = 10), the Harbin test outperformed the Kolmogorov-Smirnov test. This advantage disappeared in the larger sample size scenarios (n_1 = 30, 50), where both tests have nearly equal power.

For populations with the same distribution types (for example, 1a vs. 2a, 1b vs. 2b, etc.), it is of interest to compare the power of the two tests to detect differences in location and/or scale only. The Harbin test showed greater power (68.8%) on the simulated data compared to the Kolmogorov-Smirnov test (56.7%). The Harbin-test was the most powerful in detecting location shifts.

One important purpose of the Harbin test is to detect differences in the shapes of two population distributions, when the locations and scales of the populations are approximately equal. To assess the performance of the two tests in this regard, the power of the tests for detecting only differences in distribution types were calculated. The Harbin test is more powerful (11.4%) than the Kolmogorov-Smirnov test (2.2%) in this regard.

Considering the detection of location and/or scale shifts for two populations with different distribution types, the Harbin test is also more powerful (65.4%) than the Kolmogorov-Smirnov test (53.2%). Location shifts, scale shifts and changes in sample size from two populations with different distribution types showed the same effects on the tests as was observed overall.

Both the Kolmogorov-Smirnov test and the Harbin-test are able to compare data sets irrespective of the relationship between the data sets.

Harbin-test

When comparing the three qPCR data sets generated from the greenhouse samples, it seemed possible that the data sets originated from populations which can be described by the same probability distribution function. The Harbin-test shows that this assumption is likely, as only 6.67 % of the scores assigned to values in the first data set changed when the second data set was added (p-value = 0.906). Only 13.04 % of scores assigned to values in the newly combined data set changed with the addition of the third data set (p-value = 0.187). Therefore, it was concluded that the three data sets are compatible and can be combined for further analyses. The decision to combine data sets remains the user's responsibility. It is important to ensure that all qPCR data were generated using the same protocol and reagents. Neither the Harbin-test nor the Kolmogorov-Smirnov test takes into account any biological

factors and therefore careful consideration should be given to the experimental setup before combining data sets. The combining of data sets is beneficial when sample numbers are large, experiments need to be extended over a long period of time or when different time points need to be compared. The cumulative addition of subsequent samples to a reference data set will ensure an increase in the confidence with which each quantile score represents a true distribution of CRs unique to the specific quantile.

The distribution of the quantile scores and the change in quantile score distribution with the addition of data sets can be seen in Fig. 2. The addition of the second data set lowered one and raised one of the quantile scores of the first data set by one score. With the addition of the third data set to the combined data of data set 1 and 2, six quantiles scores were raised with one score of which four were in data set 1 and two in data set 2.

Conclusions

Harbin simplifies the analysis of high-density qPCR assays, either for individual experiments or across sets of replicates and biological conditions. The Harbin-test for the combining of data sets was shown to be less conservative than the Kolmogorov-Smirnov test, and therefore more sensitive in detecting distributional differences between data sets. Both tests are able to compare data sets irrespective of the relationship between the data sets. The quantile-based scoring system of CRs will allow for comparison of samples between experiments and different time points, aiding the monitoring of virus titre or gene expression over a season or longer period of time. The Harbin application and the Harbin-test will ease the data analysis associated with virus quantitation to monitor disease spread in vineyards. In this study a quantile score was assigned to each virus concentration ratio of GLRaV-3 single variant infections in three independent data sets. The addition of more data to the reference database

will increase the confidence of the quantile boundaries as they will eventually stabilise and provide a scoring system for virus concentrations. This enables the simplified comparison of virus concentrations between different variants of GLRaV-3. The addition of mixed variant infections and more time points to study variation over time will aid the investigation into the biological characteristics of the different variant groups and their individual contribution to GLD. It is envisioned that the Harbin-test will also be useful in other genomic applications where the combining of data sets can be beneficial. The application runs in any web-browser, and requires no programming experience from the user. This increases the accessibility of the Harbin quantitation framework for analysis of qPCR data.

Acknowledgements

The authors would like to thank Prof. Johan Burger for critical reading of the manuscript. The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the authors and are not necessarily to be attributed to the NRF.

Conflict of Interest: The authors declare that they have no conflict of interest.

References

Bester R, Jooste AE, Maree HJ, Burger JT (2012) Real-time RT-PCR high-resolution melting curve analysis and multiplex RT-PCR to detect and differentiate grapevine leafroll-associated associated virus 3 variant groups I, II, III and VI. Virol J 9:1–12. doi: 10.1186/1743-422X-9-219.

Bester R, Pepler PT, Burger JT, Maree HJ (2014) Relative quantitation goes viral: An RT-qPCR assay for a grapevine virus. J Virol Methods 210:67–75. doi: 10.1016/j.jviromet.2014.09.022

Bustin SA, Benes V, Garson JA, et al (2009) The MIQE quidelines: Minimum information for publication of quantitative real-time PCR experiments. Clin Chem 55:611–622. doi: 10.1373/clinchem.2008.112797

Carra A, Mica E, Gambino G, et al (2009) Cloning and characterization of small non-coding RNAs from grape. Plant J 59:750–763. doi: 10.1111/j.1365-313X.2009.03906.x

Efron B, Tibshirani RJ (1994) An Introduction to the Bootstrap. Taylor & Francis

Hollander M, Wolfe DA, Chicken E (2013) Nonparametric Statistical Methods. Wiley

Jooste AEC, Molenaar N, Maree HJ, et al (2015) Identification and distribution of multiple virus infections in Grapevine leafroll diseased vineyards. Eur J Plant Pathol. 142(2): 363–375. doi: 10.1007/s10658-015-0620-0

Kolmogorov AN (1933) Sulla determinazione empirica di una legge di distribuzione. G DellIstituto Ital Degli Attuari 4:83–91.

Maree HJ, Almeida RPP, Bester R, et al (2013) Grapevine leafroll-associated virus 3. Front Microbiol. 4: 82. doi: 10.3389/fmicb.2013.00082

Maree HJ, Pirie MD, Oosthuizen K, et al (2015) Phylogenomic analysis reveals deep divergence and recombination in an economically important grapevine virus. PLOS ONE 10:e0126819. doi: 10.1371/journal.pone.0126819

Pfaffl MW (2001) A new mathematical model for relative quantification in real-time RT–PCR. Nucleic Acids Res 29:e45–e45. doi: 10.1093/nar/29.9.e45

Pfaffl M, Tichopad A, Prgomet C, Neuvians T (2004) Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper – Excel-based tool using pair-wise correlations. Biotechnol Lett 26:509–515. doi: 10.1023/B:BILE.0000019559.84305.47

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

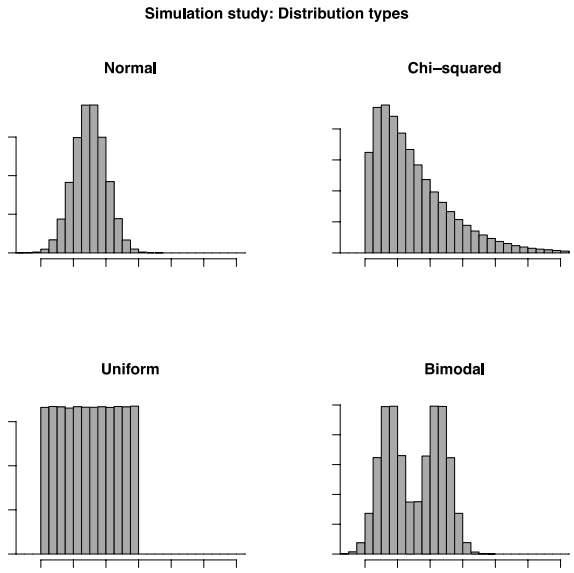
Table 1 RT-qPCR standard curve statistics per data set

Assay	Efficiency	r^2	Slope	y-intercept (b)
Data set 1				
GLRaV-3 ORF1a	1.02	0.996	-3.286	20.623
actin	0.99	0.997	-3.349	17.746
GAPDH	1.01	0.997	-3.305	20.615
alpha-Tubulin	1.00	0.996	-3.317	20.090
Data set 2				
GLRaV-3 ORF1a	0.96	0.996	-3.413	27.193
actin	1.06	0.998	-3.180	25.579
GAPDH	0.97	0.995	-3.393	24.546
alpha-Tubulin	0.94	0.995	-3.469	23.998
Data set 3				
GLRaV-3 ORF1a	0.87	0.993	-3.667	17.29
actin	0.91	0.991	-3.559	18.309
GAPDH	0.72	0.99	-4.243	19.503
alpha-Tubulin	0.92	0.995	-3.524	21.153

Fig. 1 Empirical examples of the four distribution types for the first group (1a, 1b, 1c and 1d) used in the simulation study

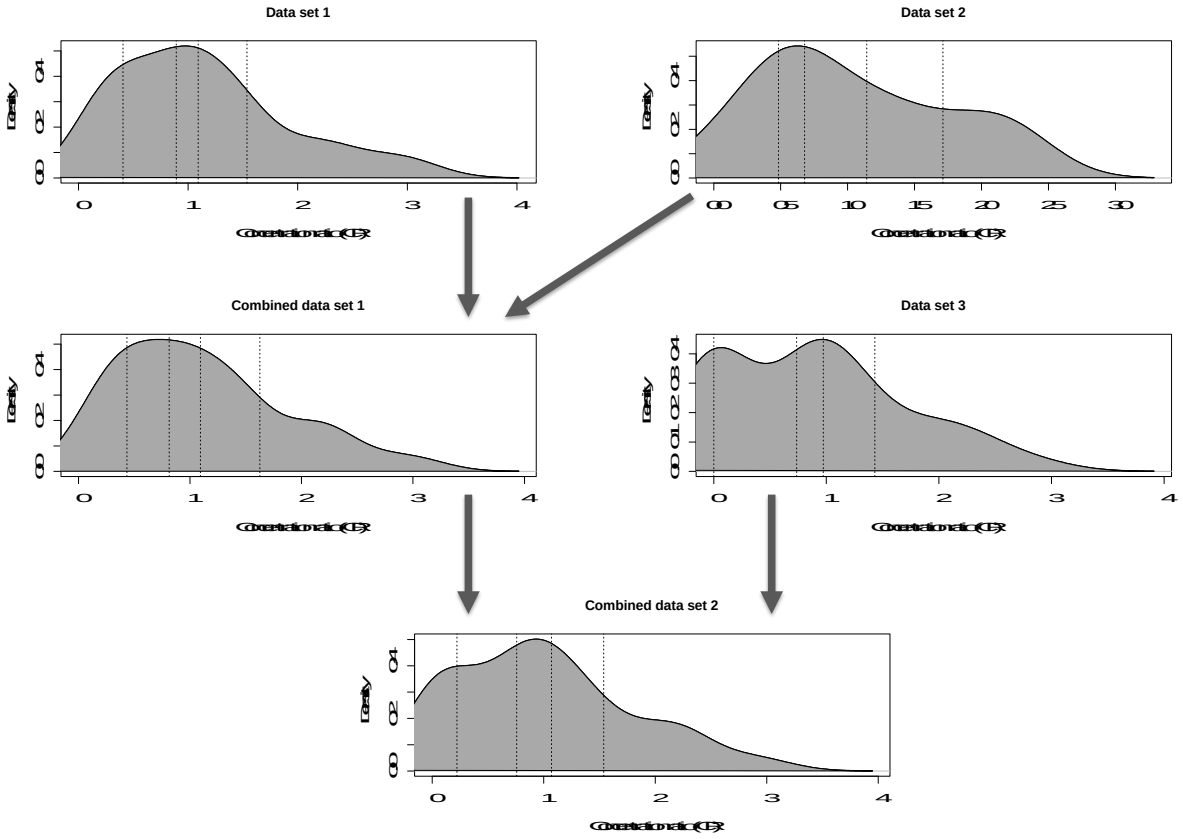
Fig. 2 Concentration ratio (CR) distribution per data set. Dotted lines indicate the quantile boundaries. The change in distribution and the quantile boundary shifts can be seen in the combined data sets.

Fig. 1


Simulation study: Distribution types
Normal
Chi–squared
Uniform
Bimodal

440 **Fig. 2**

441


Data set 1
Data set 2
Combined data set 1
Data set 3
Combined data set 2