HARBIN
Relative quantitation data analysis tool for real-time qPCR data

LICENSE

SYNOPSIS

Harbin is a tool for interactive evaluation of real-time qPCR data. Gene expression analysis can be performed with a relative quantitation strategy using the standard curve method and normalisation with a reference gene index. Harbin also allows for the pooling of different qPCR datasets/experiments for further analysis by assigning a score to each concentration ratio and subsequently testing if the datasets are sufficiently compatible before combining them. Differential expression analysis between biological conditions/groups is also possible.

CITATION

If you use Harbin in your work, please cite:

Bester, R., Pepler, P.T., Burger, J.T. and Maree, H.J. (2016) Harbin: An analysis tool for relative quantitation of real-time qPCR data and a quantile-based bootstrap test for data pooling.

DEPENDENCIES

Harbin was developed for the R statistical computing environment and will run on all major platforms (Windows, Mac OS, and Linux distributions).
Harbin is dependent on base R and additional packages (psych, car, beeswarm) available from the Comprehensive R Archive Network (CRAN). However, Harbin can also be used via the shiny web application, without a local installation of R. A web browser and an Internet connection are the only requirements.

INSTALL

For the web application:

1. Check to see if you have a working Internet connection.

2. Open a web browser and go to the following site:

    https://rbester.shinyapps.io/Harbin/

3. Start data analysis (see USAGE)

If you are familiar with R and shiny, Harbin can be run directly from R:

1. Download R or R studio from:

    https://cran.rstudio.com/ or
    https://www.rstudio.com/products/rstudio/download/

2. After installation load R in terminal or open the R studio app.

3. Use the following code in R/RStudio to check if packages are installed and install them if they are not:

```
pkg <- c("shiny", "psych", "car", "beeswarm")
new.pkg <- pkg[!(pkg %in% installed.packages())]
if (length(new.pkg)) {
install.packages(new.pkg)
}
```

4. Download the zip Harbin directory from GitHub:

    http://rbester18.github.com/harbin/

5. After download, UNZIP the Harbin directory and set your working path in R or RStudio to the Harbin directory:

```
setwd("Path_to_the_unzipped_harbin_directory")
```

6. Load the shiny library in R/RStudio:

```
library(shiny)
```

7. Run the Harbin app:

```
runApp("Harbin_app_new_RG")
```

8. A second window will open with the Harbin application. Click on the “Open in Browser” button in the left upper corner to open the application in your default web browser for better visualisation.

9. Start data analysis (see USAGE)

USAGE

The Harbin web application is organised into five different panels:

A: Data upload

This is the default active tab when the application starts.
Two options are available, either direct input of your gene of interest files and reference genes files from the Rotor-Gene Q software or manual import of Cq values from another platform.

For the Rotor-Gene option, at least one gene of interest file and one reference gene file need to be provided.
If data for one gene is split over multiple .csv files, multiple files can be selected for upload.
For each gene, standard curve data and sample data need to be available in one of the files selected for upload. Standard curve samples need to be tagged as "Standard" and samples to be included in the analysis need to be tagged as "Unknown" in the "Type" column (Automatic tagging system of the Rotor-Gene Q software). The values for the "Standard" tagged samples will be used to set the minimum and maximum valid Cq value to prevent extrapolation of concentrations from the standard curve beyond the standard curve range. Please note that the Rotor-Gene Q format changed in version 2.3.11 and additional rows in the header of the file and a "Color" column was added to the .csv file.

For generation of the .csv file from the Rotor-Gene Q software please visit:

https://www.qiagen.com/za/resources/resourcedetail?id=58d4a7d9-287f-4b01-85c3-5cb83db2228b&lang=en

or see the Rotor-Gene Q manual included in the GitHub directory at:

http://rbester18.github.com/harbin/

An example .csv file is available for download in the application or in the GitHub directory.

After data upload, row numbers, sample names and concentration values will be checked for consistency.
To perform normalisation, every sample name in the gene of interest file(s) need to be present in all the reference gene file(s) and every sample need to have a concentration value in the "Rep. Calc. Conc." column. If inconsistencies are detected, a warning and/or error will be shown in the "Data upload" panel.

For manual importing of Cq values from a different qPCR platform than Rotor-Gene Q, Cq values need to be provided in a single comma-separated file (.csv). Column one should contain the names of the samples, column two the gene of interest Cq values and then column three onwards the Cq values per reference gene used. Every sample in the gene of interest column will need to have a value in each of the reference gene columns. An example file is available for download in the application or in the GitHub directory.

After upload of files are complete, normalisation will be done in the background by dividing the gene of interest value per sample by the reference gene index (geometric mean of the reference gene

concentrations for each sample). The uploaded files and the normalised values can be viewed by selecting the “Rotor-Gene data output” or “Manual import data output” panels.

B: Harbin intervals

After normalisation, each concentration ratio (CR) is assigned a score based on the distribution of the data. 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”.
The normalised values and the interval scores can be downloaded at the bottom of the page.

If a previous experiment (reference data set) is available and you want to add the new data to the previous experiment, the two data sets can be compared to see if the data is compatible based on the distribution functions of the two data sets. A reference data set example file is available for download in the app or in the GitHub directory. Either the Kolmogorov-Smirnov test or the Harbin test can be performed to compare data sets.

The Kolmogorov-Smirnov test is a well-known test to assess the location, scale or shape of the empirical distribution functions of data sets and is the default option to compare data sets in the Harbin application. The Harbin test is proposed for a more conservative approach to avoid considering samples from two different distributions as originating from populations with the same distribution. The Harbin test is also applicable for scenarios with a larger reference data set than the test data set.

Both tests will produce a p value to assess the null hypothesis that both data sets have the same distribution function. If the p value is smaller than a chosen significance level (e.g. 0.05), the statistical evidence is considered sufficient to reject the null hypothesis and conclude that the data distribution functions differ from each other. For more details on the Harbin test, please see the end of this document.
The percentage of the elements in the reference data set for which the “labels” (1–5) have changed are also calculated for both tests.

Even though the analysis was performed using the combined data set, the new data has not been added to the reference data set until the option to add it has been selected. The data in the reference data set will be updated according to the new combined data distribution. After selecting this option, the new reference data set can be downloaded with the updated interval scores. The application will also check the names of the samples present the reference data set. If the new data set contains samples with the same names of samples in the reference data set, a warning will be shown to help avoid accidental duplication in the reference data set.

This panel also has a view option for the data intervals. By selecting “View intervals”, plots will be displayed for the new

data, unchanged reference data set (if option to compare to reference data set was selected) and the new reference data set (if option to add to reference data set was selected). In these plots the different data distributions can be viewed and the influence of the distribution of each data set on the interval boundaries can be seen (indicated with dotted lines).

C: Group selection

If applicable, the normalised data, the new reference data set or a different file (formatted as reference data set file) can be loaded and grouped into biological conditions/groups to perform statistical analysis. In this panel the user can select the number of groups to be compared and subsequently the same number of tables (sample name and sample value) will show up for selection of the individual samples to be classified into each group.

The parametric statistical tests included within Harbin are based on the assumption that the variables are normally distributed and group variances equal. A violation of the normality assumption can affect the nominal probability of a Type I or Type II error. Appropriate transformation of the data points can lessen the degree to which the assumptions are violated. These transformed values can then be used in the statistical tests.
The Harbin application allows for data transformation using the natural log or log base 10.

D: Data distribution

In this panel the basic statistics of each group selected in the previous panel will be displayed. A bar plot showing the data range and the mean of each group is also plotted, together with a box and whisker plot for better visualisation of each group's distribution.

Two tests are performed for the hypothesis that the sample data comes from a population with a normal distribution. As a guideline, the normality assumption is not rejected if the p value for each test is greater than 0.05.
A test for homoscedasticity (Levene's test) is performed to assess the variances of the groups. Group population variances are assumed to be equal if the p value for this test is greater than 0.05.

Even if a statistical test has been performed on a transformed variable, it is not recommended to report the basic statistics (means, standard deviation etc.) in transformed units. These statistics should be re-calculated using the untransformed data set by selecting the "do not transform option" in the "Group selection" panel. The "Data distribution" panel will refresh automatically..

E: Statistical tests

In this panel the statistical significance testing results between the concentration ratios across biological conditions can be viewed.

If the normality assumption seems justified, the parametric test results can be used. For two independent sample groups, a t test is

available. If population group variances are considered to be equal, the ordinary t test results can be used. If the group variances are not equal, the Welsh t test is available. For three or more independent sample groups a single factor analysis of variance (ANOVA) test will be performed.

For data not from normally distributed populations, the non-parametric tests are available. For two independent sample groups the Wilcoxon rank sum test is available and for three or more independent groups, the Kruskal Wallis test can be used to determine whether there are location differences between the groups.

Additional notes

Even if results from the Harbin and/or Kolmogorov-Smirnov tests indicate that the pooling of qPCR data sets seem justified, it is strongly recommended to only pool data sets that have been generated using the same RT-qPCR protocol.

Details of the Harbin test

The pooling of different data sets is performed under the assumption that the samples originate from populations, which can be described by the same probability distribution function.
A primary concern is therefore to determine whether these samples are compatible with each other. Suppose that $x' = [x_1, \ldots, x_n]$ and $y' = [y_1, \ldots, y_n]$ are representative samples from two continuous univariate populations, G and F, respectively. In many applications, it is of interest to determine whether the two population distributions are homogeneous. The hypothesis of interest is

$$H_0 : G(x) = F(x), \quad \text{for all } x, \qquad (1)$$

where $G(x)$ and $F(x)$ are continuous univariate probability distribution functions describing the two populations. Compared to the number of parametric and non- parametric tests available for testing either equality of medians or homogeneity of variances for two groups, relatively few tests have been proposed to test for equality of the population distributions.

The more general location-scale-shape alternative hypothesis is

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

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 20^{th}, 40^{th}, 60^{th} and 80^{th} percentiles of 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,

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

Let h_i, $i = 1,\ldots,n$ (corresponding to the elements of x) be a variable 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 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 x for which the "labels" (1–5) have changed in the combined data set, 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 labels have changed in the combined data set. To find the distribution of u under the null hypothesis, $r = 1000$ bootstrap samples of size m are drawn from x. Let

$$z_0^{(j)} = \begin{bmatrix} x \\ y_0^{(j)} \end{bmatrix}, \qquad j = 1,\ldots,r, \quad (7)$$

where $y_0^{(j)}$ indicates the j^{th} bootstrap sample. Using x and $z_0^{(j)}$, 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 $u_0' = \left[u_0^{(1)}, \ldots., u_0^{(r)}\right]$.

The choice of four percentiles in (3) was motivated by the application for which the test was developed. Changes in the number of percentiles in order to find the optimal number to maximise the power of the test was not studied for the purpose of this application. It is surmised that for large data sets, an increase in the number of quantiles will make the test more sensitive to detect differences between two population distributions, at the cost of a slight increase in computation time.

Contact details for authors

Dr H.J Maree
Department of Genetics
Stellenbosch University
Private Bag X1
Matieland
7602
South Africa
hjmaree@sun.ac.za