NqA function

Title

NqA: an R-based algorithm for the normalization and analysis of microRNA qPCR data

Authors

Paolo Verderio, Stefano Bottelli, Chiara Maura Ciniselli, Marco Alessandro Pierotti, Manuela Gariboldi, Sara Pizzamiglio

Paper

Analytical Biochemistry: Methods in the Biological Science – Notes & Tips

(DOI: http://dx.doi.org/10.1016/j.ab.2014.05.020)

Introduction

NqA is an R function useful for identifying a small set of miRNAs for data normalization in qPCR experiments in view of subsequent validation studies.

Pre-installation

Before running NqA function, make sure you have installed the following software and R additional packages:

1. software:

- R v2.15.2 (http://www.r-project.org/) or newer
- Windows Microsoft Excel 2007 or newer

2. required additional R packages

- BiocLite (http://bioconductor.org/biocLite.R)
- affy (http://www.bioconductor.org/packages/2.12/bioc/html/affy.html)
- RColorBrewer (http://cran.r-project.org/web/packages/RColorBrewer/index.html)
- ReadqPCR (http://www.bioconductor.org/packages/2.12/bioc/html/ReadqPCR.html)
- NormqPCR (http://www.bioconductor.org/packages/2.12/bioc/html/NormqPCR.html)
- epiR (http://cran.r-project.org/web/packages/epiR/index.html)
- calibrate (http://cran.r-project.org/web/packages/calibrate/index.html)
- plotrix (http://cran.r-project.org/web/packages/plotrix/index.html)

Background

Details of the procedure have been presented elsewhere (Pizzamiglio S et at. IJC, 2013). Briefly, by considering the N miRNAs expressed in all the samples, a subset of G candidate reference miRNAs is identified according to appropriate selection criteria such as variability (coefficient of variation, CV), co-regulation (Spearman Correlation Coefficient, SCC; Artusi R et al. Int J Biol Markers, 2002) and invariance between comparison groups (Kruskal-Wallis test, KW; Hollander M and Wolfe DA. 2nd ed New York Wiley and Sons, 1999). This comparison is performed on the $\log_2$ relative quantity (RQ) of each *i-th* (i=1,2,...,N) miRNA computed according to the comparative cycle threshold (Ct) method (Livak KJ and Schmittgen TD. Methods, 2001) as $RQ_i=2^{-\Delta Ct\_N_i}$, with $\Delta Ct\_N_i=Ct_i-m\_N$, where m_N is the mean of the N miRNAs (overall mean). Subsequently the identified G miRNAs are ranked in terms of stability evaluated through both geNorm (Vandesompele J et al. Genome Biol, 2002) and NormFinder (Andersen CL et al. Cancer Res, 2004) software. The G miRNAs are then forwardly combined in S sets (S=G, where S≠1) according to their stability. Once computed for each j-th set (j=2,…S) the specific mean (m_S_j), the relative quantity of each i-th miRNA is calculated as $\log_2 RQ_{ji}=-\Delta Ct\_S_{ji}$ where $\Delta Ct\_S_{ji}=Ct_i-m\_S_j$,. Then the $\log_2 RQ_{ji}$ distribution is compared between groups by means of KW test. Finally, the smallest set of reference miRNAs showing results with the highest agreement (Fleiss JL. 2nd ed New York Wiley and Sons, 1981) with that obtained when considering the relative quantity computed by using the overall mean is identified as the best subset of reference miRNAs.

Input/Output files

1. <Input>

In order to run the NqA function different input files should be created as follows:

1.1 One .txt file for each h-th (h=1, …, H) sample processed. It is also possible to create an unique file containing all the samples processed. The file le should be formatted as follow:

- *file format*: output of the used qPCR platform (txt format). It is mandatory that the file contains the following five variables (column names):
  - ✓ Well: position of the *i-th (i=1,.., I)* miRNA on a plate
  - ✓ Sample: *h-th (h=1, …, H)* sample processed
  - ✓ Detector: name of the *i-th (i=1,.., I)* miRNA on a plate
  - ✓ Ct: Cycle Threshold (Ct) value of the *i-th (i=1,.., I)* miRNA for the *h-th (h=1, …, H)* sample
  - ✓ Ct_Avg: Average Cycle Threshold (Ct) value of the the *i-th (i=1,.., I)* miRNA for the *h-th (h=1, …, H)* sample
- *file type*: tab delimited text file (.txt)

1.2 One .txt file containing clinical information about all the processed H samples. The file should be formatted as follows:

- *file format*: it is mandatory that the file contains the following two variables (column names):
  - ✓ Sample: *h-th (h=1, …, H)* sample
  - ✓ y: dummy variable coded as 0 for Controls and 1 for Cases
- *file type*: tab delimited text file (.txt)

2. **<Output>**

The output displayed in the R console, consists in the following sections:

2.1 Detection: this section reports information regarding the evaluated miRNAs. In details:

- ➢ Number of samples in which each specific miRNA was detected according to the disease status (Controls and Cases)
- ➢ Number and percentage of the miRNAs detected in all the samples
- ➢ List of the N miRNAs detected in all the samples

2.2 Variability: this section reports information regarding the miRNAs with a Coefficient of Variation (CV) less than or equal to the 20th centile of the CV's distribution obtained by considering all the N miRNAs. In details:

- ➢ Number of miRNA with a CV ≤ 20th centile and value of the 20th centile
- ➢ List of miRNAs with a CV ≤ 20th centile and its CV value ranked according to the relative CV value

2.3 Invariance between comparison groups: this section provides the name of the miRNAs showing a statistically significant difference in the $\log_2$(RQ) distribution between Cases and Controls by using the overall mean as normalization method [Mestdagh P et al. Genome Biol, 2009] and by applying the Kruskal-Wallis Test. A specific message is displayed if there are no miRNAs differentially expressed between comparison groups. In details:

- ➢ List of miRNAs differentially expressed between Cases and Controls and relative Kruskal-Wallis p-value. (These miRNAs are excluded from the subsequent steps)

2.4 Co-regulation: this section provides the pairs of miRNAs highly correlated according to the Spearman correlation coefficient, i.e. with a Lower Limit of the Fisher Confidence Interval (Fisher R.A., 1970) greater than the value chosen by the user (suggested value: 0.80). Within each pairs the miRNA with the highest CV is excluded. A specific message is display if there are no co-regulated miRNAs. In details:

- ➢ Pairs of miRNAs highly correlated
- ➢ List of miRNA excluded within each pair

2.5 Stability: this section provides the list of the G candidate reference miRNAs sorted according to the sum of the ranks obtained by jointly considering geNorm [Vandesompele J et al. Genome Biol, 2002] and Normfinder software [Andersen CL et al. Cancer Res, 2004].

2.6 Agreement: this section reports the kappa statistic value and its 95% Confidence Interval (CI) [Fleiss JL., 1981] corresponding to the best set of reference miRNAs. The latter is the smallest set of miRNAs showing the highest agreement (i.e with the upper limit of the 95% CI ≥0.80 or the highest value of the kappa statistics when no 95% CI include the threshold of 0.80) when the results deriving by using them to normalize the data are compared with those obtained by using the overall mean as normalization method. In details:

- name and number of the best set of reference miRNAs

2.7 Summary of the results: this section provides for each of the N miRNAs detected in all samples the median values of the $\log_2(RQ)$ distribution in Cases and Controls computed according to the best set of reference miRNAs as normalization method. The miRNAs are sorted according to the p-value of the Kruskal-Wallis test (KW p-value) used for comparing Cases and Controls distributions. In addition the difference between the median value in cases and controls (Fold Change) are reported together with the –log10 of the KW p-value. These quantities are used to generate the volcano plot (see below).

3. **<Figures>**

The Graphical output, consists in following figures:

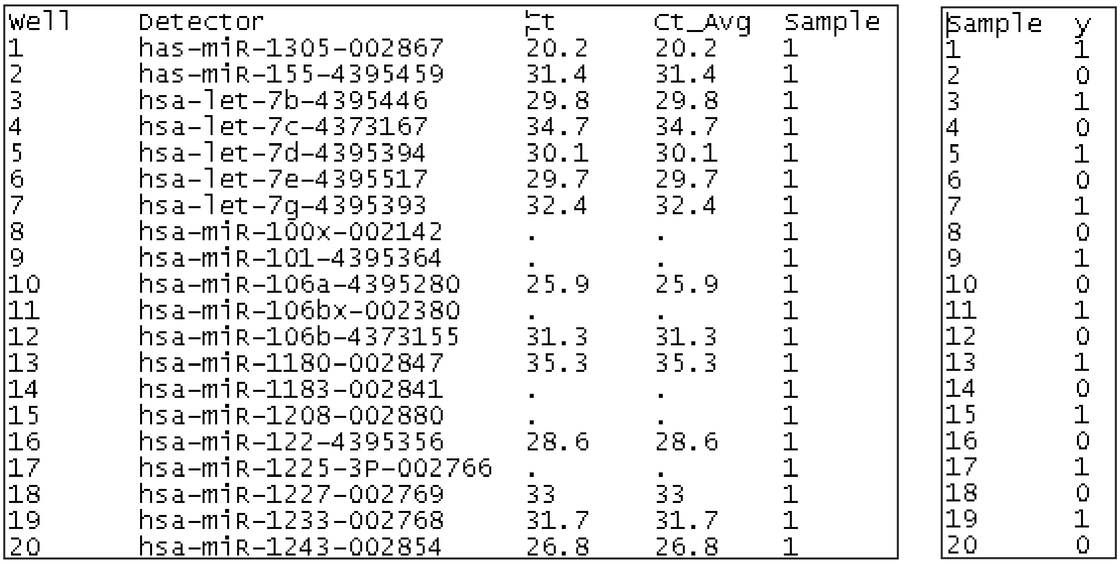
3.1 Kappa statistic: This figure reports the kappa statistic values and their 95% Confidence Intervals obtained for each of the S sets (S=G, where S≠1) of candidate reference miRNAs. The arrow indicates the best set of reference miRNAs.

3.2 Volcano Plot: This figure provides a visual identification of the N miRNAs with the large-magnitude changes between the comparison groups (Cases and Control). Specifically, the Volcano plot combines the Kruskal-Wallis test results with the magnitude of the change. The Volcano plot is generated by plotting the –log10 (KW p-value) on the y-axis and the Fold Change on the x-axis. The most relevant miRNAs are those closer to the left (down-regulated) or right (up-regulated) upper corners. The dashed line indicates the statistically significance threshold of α=5%.

3.3 Boxplots of normalized data: This figure reports the $\log_2(RQ)$ distributions in cases and controls by using different data normalization strategy: overall mean, best set of reference miRNAs by NqA and according to the reference miRNAs chosen by the user (for example those suggested by the producer's platform).The latter have to be detected in all the samples. On the contrary only the box plot with the other two data normalization strategy are displayed and a specific message is reported in the R console output.

Example of Input files:

Example of 1.1 input file (left panel) and 1.2 input file (right panel).



Well	Detector	Ct	Ct_Avg	Sample
1	has-miR-1305-002867	20.2	20.2	1
2	has-miR-155-4395459	31.4	31.4	1
3	hsa-let-7b-4395446	29.8	29.8	1
4	hsa-let-7c-4373167	34.7	34.7	1
5	hsa-let-7d-4395394	30.1	30.1	1
6	hsa-let-7e-4395517	29.7	29.7	1
7	hsa-let-7g-4395393	32.4	32.4	1
8	hsa-miR-100x-002142	.	.	1
9	hsa-miR-101-4395364	.	.	1
10	hsa-miR-106a-4395280	25.9	25.9	1
11	hsa-miR-106bx-002380	.	.	1
12	hsa-miR-106b-4373155	31.3	31.3	1
13	hsa-miR-1180-002847	35.3	35.3	1
14	hsa-miR-1183-002841	.	.	1
15	hsa-miR-1208-002880	.	.	1
16	hsa-miR-122-4395356	28.6	28.6	1
17	hsa-miR-1225-3P-002766	.	.	1
18	hsa-miR-1227-002769	33	33	1
19	hsa-miR-1233-002768	31.7	31.7	1
20	hsa-miR-1243-002854	26.8	26.8	1

Sample	y
1	1
2	0
3	1
4	0
5	1
6	0
7	1
8	0
9	1
10	0
11	1
12	0
13	1
14	0
15	1
16	0
17	1
18	0
19	1
20	0

Code to Run the NqA function (NqA_run.R):

The first part of this file allow the initialization and installation of the R packages used in NqA.

Then you have to define the work folder, call the NqA code and define its arguments:

```
path<-"path"                              # Define the work folder
setwd(path)                               # Initializing work folder specified above
source('path\\NqA_code.R')                #call the NqA code:
nomifiles<-c("file1.txt”, “file2.txt”,….., “fileH.txt")        # Insert the name(s) of the 1.1 input file(s)
nominorm<-c("mirref1"," mirref2", …)  # Insert the name(s) of reference miRNA(s) chosen by the user
q<-nqa(nomifiles,alphakw,alphacor,lwl,clinfile,ctlimit, nominorm)                    #Start NqA function
```

- *nomifiles*: string vector defined above contacting the name(s) of the 1.1 input file(s)
- *alphakw*: significance level for Kruskal-Wallis Test [suggested = 0.05]
- *alphacor*: significance level for the estimation of the Confidence Interval of the Spearman correlation coefficient [suggested =0.05]
- *lwl*: threshold for selection of miRNAs high correlated [suggested = 0.80]
- *clinfile*: name of the 1.2 input file [i.e: “clinfile.txt”]
- *ctlimit*: value of Ct to be consider as “Undetermined” [suggested= 40]
- *nominorm*: string vector defined above containing the name of the reference miRNA(s) chosen by the users (for example those suggested by the producer’s platform). If not, nominorm=NULL.

Note: the NqA function generate two working files (list_ref.txt and phenodata_norm2.txt) that should not be considered.

Step for processing the data as in Verderio P et al.:

1. First create a folder in your hard disk named NqA containing the 1.1 and 1.2 input data, the NqA_run.R and the NqA_code.R.
2. Open the NqA_run.R file in R console
3. Run the code [select all and press “Ctrl R”]