Analytical Biochemistry 461 (2014) 7–9

Contents lists available at ScienceDirect

Analytical Biochemistry

journal homepage: www.elsevier.com/locate/yabio

Notes & Tips

NqA: An R-based algorithm for the normalization and analysis of microRNA quantitative real-time polymerase chain reaction data

Paolo Verderio [a,*], Stefano Bottelli [a], Chiara Maura Ciniselli [a], Marco Alessandro Pierotti [a], Manuela Gariboldi [a,b], Sara Pizzamiglio [a]

[a] Unit of Medical Statistics, Biometry, and Bioinformatics, Fondazione IRCCS Istituto Nazionale dei Tumori, 20133 Milan, Italy
[b] Fondazione Istituto FIRC di Oncologia Molecolare, 20139 Milan, Italy

ARTICLE INFO

Article history:
Received 28 March 2014
Received in revised form 16 May 2014
Accepted 22 May 2014
Available online 2 June 2014

Keywords:
Software
R package
Algorithm
Normalization
qPCR
microRNA

ABSTRACT

In this note, we propose an R function named NqA (Normalization qPCR Array, where qPCR is quantitative real-time polymerase chain reaction) suitable for the identification of a set of microRNAs (miRNAs) to be used for data normalization in view of subsequent validation studies with qPCR data. NqA is available through the website of the Fondazione IRCCS Istituto Nazionale dei Tumori of Milan (http://www.istitutotumori.mi.it/modules.php?name=Content&pa=showpage&pid=812) with a dedicated user's guide. We applied our function on a qPCR dataset downloaded from the Gene Expression Omnibus (GEO) database. Results show that NqA provides a functional subset of reference miRNAs and a set of promising significantly modulated miRNAs for subsequent validation studies.

MicroRNAs (miRNAs)[1] are small noncoding RNAs involved in tumorigenesis and in the development of various cancers [1]. Quantitative real-time polymerase chain reaction (qPCR) is the most commonly used tool to investigate miRNA expression, and qPCR low-density arrays are increasingly being used for the identification of potentially relevant miRNAs. In this context, one crucial pre-processing step is data normalization, aimed at reducing nonbiological sources of variation. Data normalization strategies used with qPCR high-throughput data generally take advantage of the huge amount of data and are often based on their mean expression value [2–4]. However, data-driven methods are preferentially used during the discovery phase and are almost never applicable in subsequent validation studies that are mainly focused on a limited number of miRNAs. To overcome this issue, we developed a comprehensive procedure that, starting from high-throughput qPCR data, identifies a small set of miRNAs to use as reference for data normalization in view of subsequent validation studies. In this note, we propose an R function, named NqA (Normalization qPCR Array), developed by updating the procedure we recently published [5].

Details of the procedure are presented in Fig. 1. Briefly, by considering the N miRNAs expressed in all of the samples, a subset of G candidate reference miRNAs is identified according to appropriate selection criteria such as variability (coefficient of variation, CV), coregulation (Spearman correlation coefficient, SCC) [6], and invariance between comparison groups (Kruskal–Wallis test, KW) [7]. This comparison is performed on the $\log_2$ relative quantity (RQ) of each ith ($i = 1, 2, \ldots, N$) miRNA computed according to the comparative cycle threshold (Ct) method [8] as $RQ_i = 2^{-\Delta Ct\_Ni}$, 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 [9] and NormFinder [10] software. The G miRNAs are then forwardly combined into S sets ($S = G$, where $S \neq 1$) according to their stability. Once computed for each jth set ($j = 2, \ldots, S$) the specific mean (m_S_j), the relative quantity of each ith 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 Kruskal–Wallis test. Finally, the smallest set of reference miRNAs showing results with the highest agreement with that obtained when considering the relative quantity computed by using the overall mean is identified as the best subset of reference miRNAs.

Output printed in R console by running the NqA function consists of frequency distribution of evaluated miRNAs according to the comparison group (cases and controls), list of the N miRNAs

* Corresponding author. Fax: +39 02 23902095.
E-mail address: paolo.verderio@istitutotumori.mi.it (P. Verderio).

[1] *Abbreviations used:* miRNA, microRNA; qPCR, quantitative real-time polymerase chain reaction; NqA, Normalization qPCR Array; CV, coefficient of variation; SCC, Spearman correlation coefficient; KW, Kruskal–Wallis test; RQ, relative quantity; Ct, cycle threshold; CI, confidence interval; GEO, Gene Expression Omnibus.

http://dx.doi.org/10.1016/j.ab.2014.05.020
0003-2697/

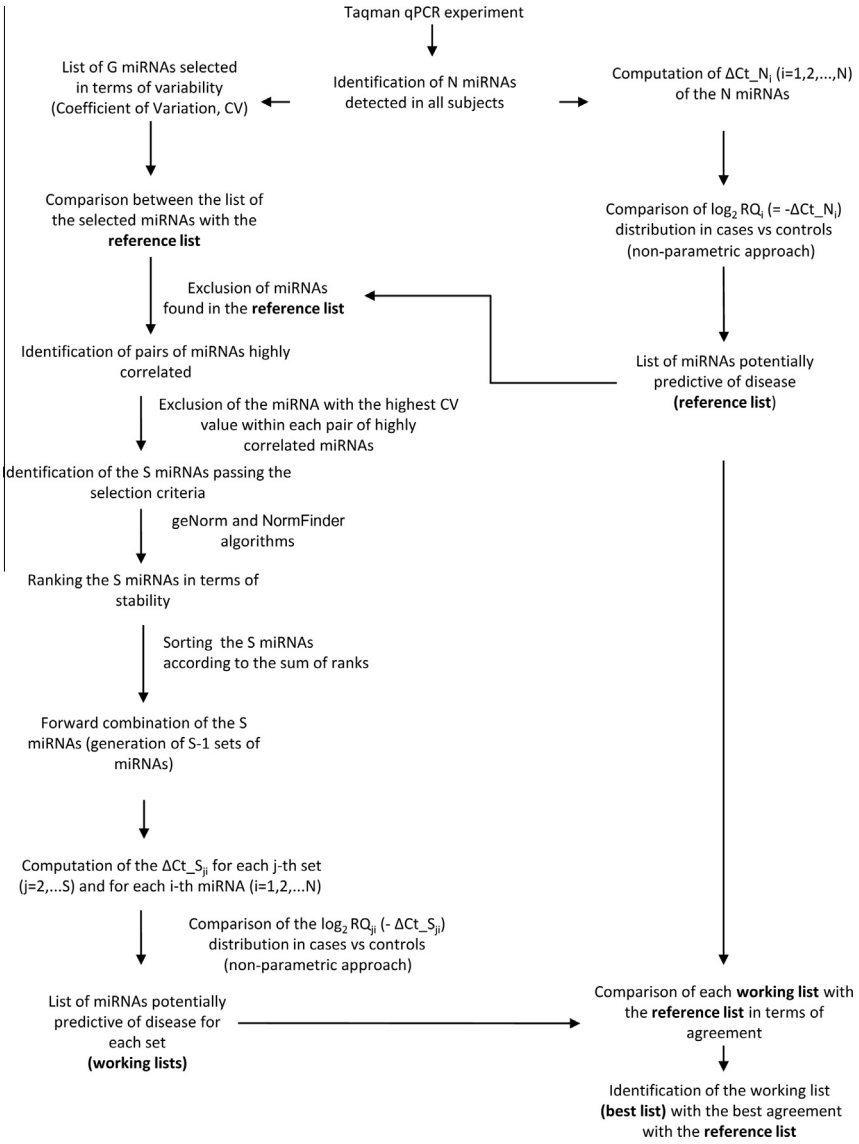

Fig.1. Procedure developed for identification of the best subset of reference miRNAs.

expressed in all of the samples and of those with a CV value lower than the pre-fixed threshold (20th centile of the distribution of the CV of the *N* miRNAs). In addition, the output reports the list of miRNAs with a statistically different relative quantity computed according to the overall mean between cases and controls (KW *P* value < 0.05). These miRNAs are excluded from the subsequent steps. Furthermore, the possible pairs of highly correlated miRNAs (lower limit of the SCC confidence interval (CI) $\geqslant$ 0.80) as well as the specific miRNAs excluded within each pair (i.e., with the higher CV) are reported. Regarding the stability evaluation, the output provides the list of the *G* candidate reference miRNAs sorted according to the sum of the ranks obtained by jointly considering geNorm and NormFinder. Subsequently, the best subset of reference miRNAs showing the highest agreement in terms of the kappa statistic [11] and its relative CI is reported. In addition, some descriptive statistics related to each miRNA are provided in the final section of the output.

As an illustration of the algorithm, we applied the NqA function to data downloaded from the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/gds) regarding the assessment of the Applied Biosystems TaqMan Array Human MicroRNA Cards (A+B Card Set, version 3) in plasma samples from hepatocellular carcinoma (HCC) patients and healthy donors (GSE50013) [12]. Starting from 257 miRNAs detected in a total of 40 plasma samples (20 from individuals with HCC [cases] and 20 from healthy donors [controls]), $N = 38$ miRNAs were detected in all samples and, among them, $G = 7$ miRNAs were selected. The best subset of reference miRNAs was found to be composed of miR-30c and miR-30b. An overview of the graphical output provided by NqA is represented in Fig. 2. Specifically, Fig. 2A reports the kappa statistic values for each set of candidate reference miRNAs. The best set of miRNAs is indicated by an arrow to immediately visualize its performance in comparison with that of the others. In Fig. 2B, the Volcano plot obtained by using miR-30c and miR-30b as reference is reported. Fig. 2C to 2E report box plots of normalized data according to different approaches, giving an immediate picture of their appropriateness in the considered scenario. In conclusion, we have presented in this note the NqA function as a suitable tool for selection of a subset of reference miRNAs and for evaluation of promising miRNAs during the discovery phase based on qPCR high-throughput data. This represents the first step of a workflow, moving from the discovery to the validation, which should produce reliable results to be applied in the clinical scenario.

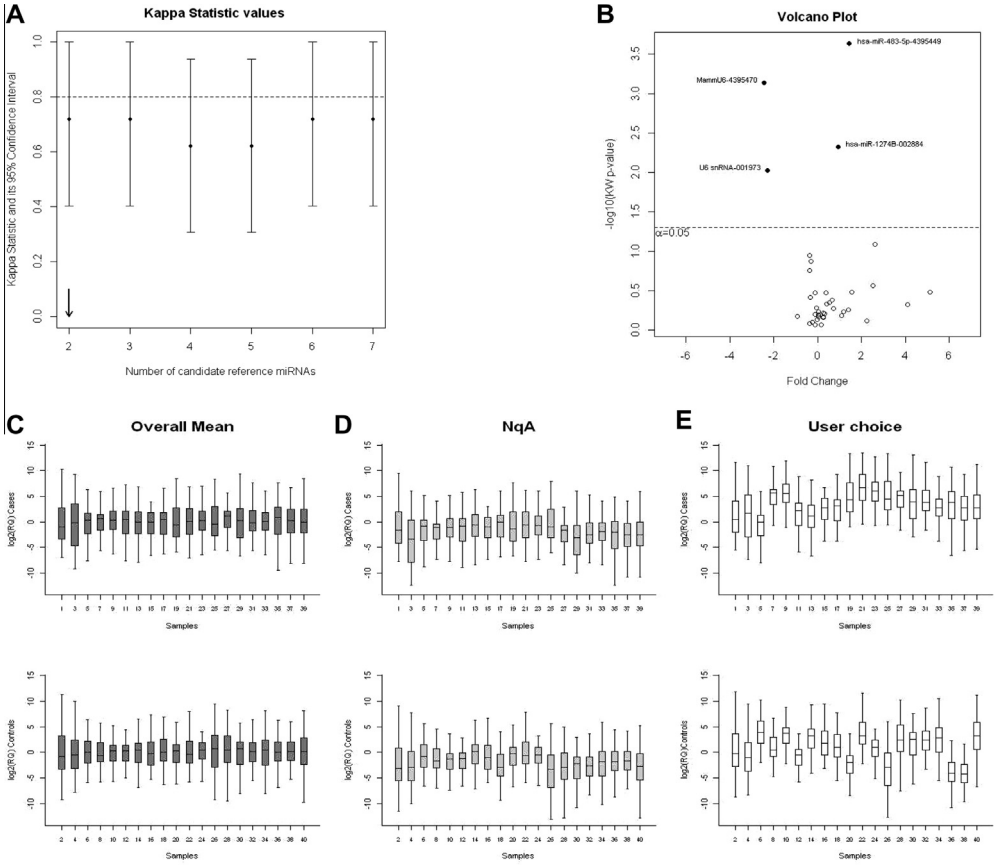

Fig.2. Graphical output provided by NqA. (A) Kappa statistic values for each set of candidate reference miRNAs. (B) Volcano plot obtained by normalizing data using the best subset of reference miRNAs identified by NqA. This plot is generated by plotting the $-\log_{10}$(KW *P* value) on the *y* axis versus the fold change [difference between the median value of the $\log_2$(RQ) in cases and controls] on the *x* axis. (C–E) Box plots reporting the $\log_2$(RQ) distributions in cases and controls computed according to the overall mean (C), the mean of the best subset of reference miRNAs provided by NqA (D), and the mean of the reference miRNAs chosen by the user (in this example, the two endogenous controls suggested by the producer's platform, U6-snRNA and MammU6, are chosen) (E).

Acknowledgments

This work was supported by Grants from Associazione Italiana per la Ricerca sul Cancro (AIRC, Grants 10529 and 12162 to M.A. Pierotti).

References

[1] J.F. Reid, V. Sokolova, E. Zoni, A. Lampis, S. Pizzamiglio, C. Bertan, S. Zanutto, F. Perrone, T. Camerini, G. Gallino, P. Verderio, E. Leo, S. Pilotti, M. Gariboldi, M.A. Pierotti, MicroRNA profiling in colorectal cancer highlights miR-1 involvement in MET-dependent proliferation, Mol. Cancer Res. 10 (2012) 504–515.

[2] A. Deo, J. Carlsson, A. Lindlöf, How to choose a normalization strategy for miRNA quantitative real-time (qPCR) arrays, J. Bioinform. Comput. Biol. 9 (2011) 795–812.

[3] P. Mestdagh, P. Van Vlierberghe, A. De Weer, D. Muth, F. Westermann, F. Speleman, J. Vandesompele, A novel and universal method for microRNA RT-qPCR data normalization, Genome Biol. 10 (2009) R64.

[4] J.C. Mar, Y. Kimura, K. Schroeder, K.M. Irvine, Y. Hayashizaki, H. Suzuki, D. Hume, J. Quackenbush, Data-driven normalization strategies for high-throughput quantitative RT-PCR, BMC Bioinformatics 10 (2009) 110.

[5] S. Pizzamiglio, S. Bottelli, C.M. Ciniselli, S. Zanutto, C. Bertan, M. Gariboldi, M.A. Pierotti, P. Verderio, A normalization strategy for the analysis of plasma microRNA qPCR data in colorectal cancer, Int. J. Cancer 134 (2014) 2016–2018.

[6] M. Hollander, D.A. Wolfe, Nonparametric Statistical Methods, 2nd ed., John Wiley, New York, 1999.

[7] R. Artusi, P. Verderio, E. Marubini, Bravais-Pearson and Spearman correlation coefficients: meaning, test of hypothesis, and confidence interval, Int. J. Biol. Markers 17 (2002) 148–151.

[8] K.J. Livak, T.D. Schmittgen, Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta Ct}$ method, Methods 25 (2001) 402–408.

[9] J. Vandesompele, K. De Preter, F. Pattyn, B. Poppe, N. Van Roy, A. De Paepe, F. Speleman, Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes, Genome Biol. 3 (2002). RESEARCH0034.

[10] C.L. Andersen, J.L. Andersen, T.F. Ørntoft, Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets, Cancer Res. 64 (2004) 5245–5250.

[11] J.L. Fleiss, Statistical Methods for Rates and Proportions, 2nd ed., John Wiley, New York, 1981.

[12] J. Shen, A. Wang, Q. Wang, I. Gurvich, A.B. Siegel, H. Remotti, R.M. Santella, Exploration of genome-wide circulating microRNA in hepatocellular carcinoma: MiR-483-5p as a potential biomarker, Cancer Epidemiol. Biomarkers Prev. 22 (2013) 2364–2373.