HAYATI
Journal of Biosciences

Vol. 31 No. 3, May 2024 432-442
DOI:10.4308/hjb.31.3.432-442
ISSN: 1978-3019
EISSN: 2086-4094

Quantile Normalization for High Throughput Circulating MicroRNA Expression Study using TaqMan® Low Density Array Panels: Supporting Evidence

Azmir Ahmad[1], Syarah Syamimi Mohamed[2], Afidalina Tumian[3], Siti Marponga Tolos[4], Vijaya Mohan Sivanesan[5], Wan Ishlah Leman[6], Kahairi Abdullah[6,14], Irfan Mohamad[7], Wan Mohd. Nazri Wan Zainon[8], Luqman Rosla[9], Sharifah Nor Ezura Syed Yussof[9], Mark Paul[9,15], Kamariah Mohamed@Awang[10], Rosdi Ramli[11], Eshamsol Kamar Omar[12], Mohd. Wardah Mohd. Yassin[13], Mohd. Amin Marwan Mohamad[13], Mohd. Arifin Kaderi[13*]

[1]*Department of Basic Medical Science for Nursing, Kulliyyah of Nursing, International Islamic University Malaysia, 25200 Kuantan, Pahang, Malaysia*
[2]*Department of Surgery, School of Medical Sciences, Universiti Sains Malaysia, 16150 Kota Bharu, Kelantan, Malaysia*
[3]*PETRONAS Research Sdn. Bhd., 43000 Bandar Baru Bangi, Selangor, Malaysia*
[4]*Department of Computational and Theoretical Sciences, Kulliyyah of Science, International Islamic University Malaysia, 25200 Kuantan, Pahang, Malaysia*
[5]*Molecular Pathology Unit, Cancer Research Centre, Institute for Medical Research, National Institutes of Health, 40170 Shah Alam, Selangor, Malaysia*
[6]*Department of Otorhinolaryngology-Head and Neck Surgery, Kulliyyah of Medicine, International Islamic University Malaysia, 25200 Kuantan, Pahang, Malaysia*
[7]*Department of Otorhinolaryngology-Head and Neck Surgery, School of Medical Sciences, Universiti Sains Malaysia, 16150 Kota Bharu, Kelantan, Malaysia*
[8]*Department of Nuclear Medicine, Radiotherapy and Oncology, School of Medical Sciences, Universiti Sains Malaysia, 16150 Kota Bharu, Kelantan, Malaysia*
[9]*Department of Otorhinolaryngology, Hospital Sultan Haji Ahmad Shah, 28000 Temerloh, Pahang, Malaysia*
[10]*Department of Otorhinolaryngology, Hospital Tengku Ampuan Afzan, 25100 Kuantan, Pahang, Malaysia*
[11]*Department of Ear, Nose and Throat, Hospital Raja Perempuan Zainab II, 15586 Kota Bharu, Kelantan, Malaysia*
[12]*Ear, Nose and Throat Consultant, Kota Bharu Medical Centre, 15200 Kota Bharu, Kelantan, Malaysia*
[13]*Department of Biomedical Science, Kulliyyah of Allied Health Sciences, International Islamic University Malaysia, 25200 Kuantan, Pahang, Malaysia*
[14]*Ear, Nose and Throat Consultant, KPJ Batu Pahat Specialist Hospital, 83000 Batu Pahat, Johor, Malaysia*
[15]*Department of Otorhinolaryngology, Hospital Sultanah Aminah, 80000 Johor Bahru, Johor, Malaysia*

ARTICLE INFO

Article history:
Received August 9, 2023
Received in revised form December 4, 2023
Accepted December 20, 2023

KEYWORDS:
quantile,
normalization,
data driven,
circulating miRNAs,
high throughput

ABSTRACT

In searching for new biomarkers, high throughput technique has been widely used by researchers, including for gene expression study. However, the reliability and accuracy of results from high throughput study critically depends on appropriate data management, including normalization methods. Data driven normalization has been introduced as a normalization method for high throughput gene expression study. Thus, this study was conducted to evaluate the performance of various data driven and reference genes normalization methods using a high throughput circulating microRNA expression dataset. A quantification cycle (Cq) dataset generated from a high throughput circulating microRNA study was used to test the normalization methods using HTqPCR package in R software. The normalized Cq generated from different methods were compared descriptively using box plot analysis and coefficient of variance. The box plot analysis showed that quantile normalization produced more homogenous Cq distribution, lesser outliers and reduced coefficient of variance as compared to other normalization methods in screening and validation phases. The overview on quantile normalized Cq showed consistency in its level of expression before and after 2-ΔΔCq calculation indicating the reliability of quantile normalized Cq. Quantile normalization is suggested to be used in high throughput miRNA expression study due to its performance in homogenizing the data, reduce outliers and coefficient of variance.

1. Introduction

Reference genes for normalization is commonly used in gene expression at DNA and mRNA levels. There is a similar importance for a standard method for data normalization in miRNA expression study (Causin *et al.* 2019; Faraldi *et al.* 2019; Veryaskina *et al.* 2022). However, to date, reference genes for miRNAs that are replicable across studies have not been established, unlike in DNA and mRNA expression studies, in which reference genes such as *GAPDH* and *ACTB* genes are universally utilized as the reference genes (Veryaskina

* Corresponding Author
E-mail Address: ariffink@iium.edu.my

et al. 2022). Exogenous controls, which are miRNA genes from other organisms such as such as cel-miR-39-3p from *Caenorhabditis elegans* and ath-miR-159a from *Arabidopsis thaliana*, have been proposed as an option in the normalization technique in miRNA expression study (Kumar and Reddy 2016; Vigneron *et al.* 2016). However, the use of exogenous controls as reference genes was not recommended by Faraldi *et al.* (2019) and Dakterzada *et al.* (2020) as this controls may cause false interpretation of results. This is due to their purposive use for technical correction during the experiment but not for the other previously described intrinsic variables to which they are not biologically exposed. In contrast, endogenous miRNAs are more reflective on the cellular conditions being studied.

As advanced technologies to screen from hundreds to thousands of genes per experiment are developed these days, the use of reference genes to normalize the raw expression data for such studies is impractical (Eisenberg and Levanon 2003, 2013). Various normalization methods have been developed with scientific and mathematical evidences to increase the accuracy of a normalization technique (Vandesompele *et al.* 2002; Mestdagh *et al.* 2009; Hicks *et al.* 2018). These types of normalization methods are known as data driven normalization as they utilized all the available data that are generated by the instrument and produce a reference or control value, such as quantile, mean or median expression (Bolstad *et al.* 2003; Andersen *et al.* 2004; Deo *et al.* 2011; Hicks *et al.* 2018). Some data driven normalizations were suggested for high throughput gene expression study, such as rank invariant, mean expression and quantile normalizations (Liu *et al.* 2019).

Here, we present the comparison of some normalization methods in HTqPCR package for a quantitative polymerase chain reaction (qPCR) array, including data driven and reference genes methods. The normalization methods were tested on a qPCR dataset generated from a study on circulating miRNA expression in nasopharyngeal carcinoma (NPC) patients and control subjects. The study demonstrated the applicability of quantile normalization in high throughput gene expression study that consisted of screening phase that used representative number of samples and validation phase that use larger number of samples. The results show that for this dataset, quantile normalization performs the best in reducing variations between arrays and is better than the other data driven methods and reference genes.

2. Materials and Methods

2.1. Subject Recruitment

The protocol has been reviewed by IIUM Research Ethics Committee of International Islamic University Malaysia (IIUM) (IREC 457), Medical Research Ethic Committee of Ministry of Health Malaysia (NMRR-15-1976-27156 (IIR)) and Human Research Ethics Committee of Universiti Sains Malaysia (USM/JEPeM/16010032). Thirty six newly diagnosed and untreated NPC patients were recruited from three government and two university hospitals in Pahang and Kelantan states of Malaysia. Similar number of control subjects who have no history of NPC and blood relationship with the NPC patients in this study were recruited from the visitors of the similar hospitals and served as age-matched controls.

2.2. Circulating miRNA Extraction

Ten ml whole blood of the subjects were collected in ethylenediaminetetraacetic acid (EDTA) tube. The blood samples were centrifuged for 15 minutes at 1,200 x g at room temperature to obtain the plasma. Three hundred µL of the plasma were used for miRNA extraction using NucleoSpin® miRNA plasma kit (Macherey-Nagel, Düren, Germany) with modifications by Wozniak *et al.* (2015). The purity and RNA yield of the miRNA extracts were measured spectrophotometrically using NanoDrop™ 1,000 (Thermo Fisher Scientific, Massachusetts, United States).

2.3. Screening Phase: Circulating miRNA Screening using Taqman® Low Density Array A + B Cards

Ten plasma samples from NPC patients and eleven from controls were subjected to the circulating miRNA screening. Reverse transcription (RT) was performed using Megaplex™ Reverse Transcription kit (Applied Biosystems, California, United States) according to the manufacture's protocol. Fixed volume rather than equal quantity of miRNAs was used as starting material that consisted of one to 350 ng total RNA. Preamplification step was used after RT step using Megaplex™ PreAmp kit (Applied Biosystems, California, United States) according to the manufacture's protocol. Both RT and preamplification steps were performed using Veriti™ 96-Well Thermal Cycler (Applied Biosystems, California, United States). The preamplification products were inserted into Taqman® Low Density Array (TLDA) version

3.0 card A and B for qPCR reaction. TLDA card A and B are qPCR array that consist of more than 740 miRNAs that consistent with Sanger miRBase version 20. The qPCR reaction was performed using QuantStudio™ 12K Flex Real-Time PCR System (Applied Biosystems, California, United States) with setting of initiation stage for 10 minutes at 95°C, followed by 40 cycles amplification stage for 15 seconds at 95°C and 1 minute at 60°C. Once the qPCR completed, the qPCR data were exported to ExpressionSuite software (Applied Biosystems, California, United States) for calculation of Cq using auto-threshold and auto-baseline settings.

2.4. Validation Phase: Circulating miRNA Expression Validation using 96.96 Dynamic Array™ Integrated Fluidic Circuit Chip

The 96.96 Dynamic Array™ Integrated Fluidic Circuit (IFC) chips (Fluidigm, California, United States) with Taqman® miRNA PCR assay were used to validate the differential expression of selected miRNAs, based on method by Tan and Tan (2017). The qPCR data was exported to Fluidigm Real-Time PCR Analysis software (Fluidigm, California, United States) for calculation of Cq using auto-threshold and auto-baseline settings.

2.5. Pre-processing Data

The first pre-processing step in this study was to exclude the haemolysis-sensitive miRNAs from further analysis to increase reliability and prevent underestimation of result. This step was performed by removing the miRNAs that have been reported to be affected by haemolysis from the miRNA list in the exported .xls file (Kirschner *et al.* 2013; MacLellan 2014; Shkurnikov *et al.* 2016; Pizzamiglio *et al.* 2017). Then, the undetermined and Cq values of more than 35 have been replaced with Cq of 36, instead of Cq of 40, to minimize statistical confounding by high quantification cycle values (de Ronde *et al.* 2016; Gevaert *et al.* 2018). Further pre-processing steps were performed using the HTqPCR package from Bioconductor in RStudio software (Dvinge and Bertone 2009). The low expression miRNAs that caused analytical nuisance and confound the statistical analysis were removed using the RStudio software. As recommended by Gevaert *et al.* (2018), the miRNAs with non-informative Cq, which was Cq of 36, in more than 80% of samples were excluded from the study as they can confound the statistical analysis by increasing noise. So, for this study, the miRNAs with Cq of 36 in more than 17 out of 21 samples were excluded from further analysis.

2.6. Measures of Performance

The normalization methods offered by the HTqPCR package were quantile normalization, rank invariant normalization, delta Cq normalization and geometric mean normalization (Dvinge and Bertone 2009). The controls that were available in TLDA version 3.0 for delta Cq normalization purpose were U6 snRNA, RNU44 and RNU48 for endogenous reference genes and ath-miR-159a for exogenous reference gene. Besides, the study also used all the miRNAs available in the card A and B to search for any potential candidate reference genes by calculating summarized stability score (SSS), as recommended by Marabita *et al.* (2016). This formula summarizes the stability score from geNorm, NormFinder and coefficient of variance (CV), where the lower score from each of these algorithms shows more stable candidate reference genes. The SSS value was used to rank the miRNAs from the most stable miRNAs, indicated by the lowest SSS value, to least stable miRNAs, indicated by the highest SSS value.

The selection of normalization method for this study was based on box plot analysis to investigate the distribution patterns of Cq values for each sample and coefficient of variation (CV) to investigate the dispersion of Cq values for individual miRNAs, based on Sysi-Aho *et al.* (2007) and Deo *et al.* (2011). The criteria to select the best normalization method were determined by identifying the method that produced homogenous normalized Cq distribution, less Cq outliers and lower CV after the normalization as compared to raw Cq. The homogenous normalized Cq distribution and lower CV were important criteria to be considered as they indicated the good performance of a normalization method in adjusting the experimentally induced variations. The results of both analyses were generated using the HTqPCR package in RStudio software.

3. Results

3.1. Screening Phase: Circulating miRNA Screening using Taqman® Low Density Array A + B Cards

The exclusion of haemolysis-sensitive miRNAs from the raw Cq data had selected 352 and 378 miRNAs from 384 miRNAs in card A and card B, respectively. Next, the removal of the miRNAs with undetermined and unreliable Cq in more than 80% of the samples by HTqPCR package in RStudio software resulted in a list of remaining 182 and 122 miRNAs from card A

and card B, respectively, as reliable miRNAs for further downstream analyses. In total, 304 miRNAs were detected in screening phase for further pre-processing.

The quality assessment for each normalization methods in HTqPCR for card A and B are presented in Figure 1 and 2, respectively. RNU44 and RNU48 cannot

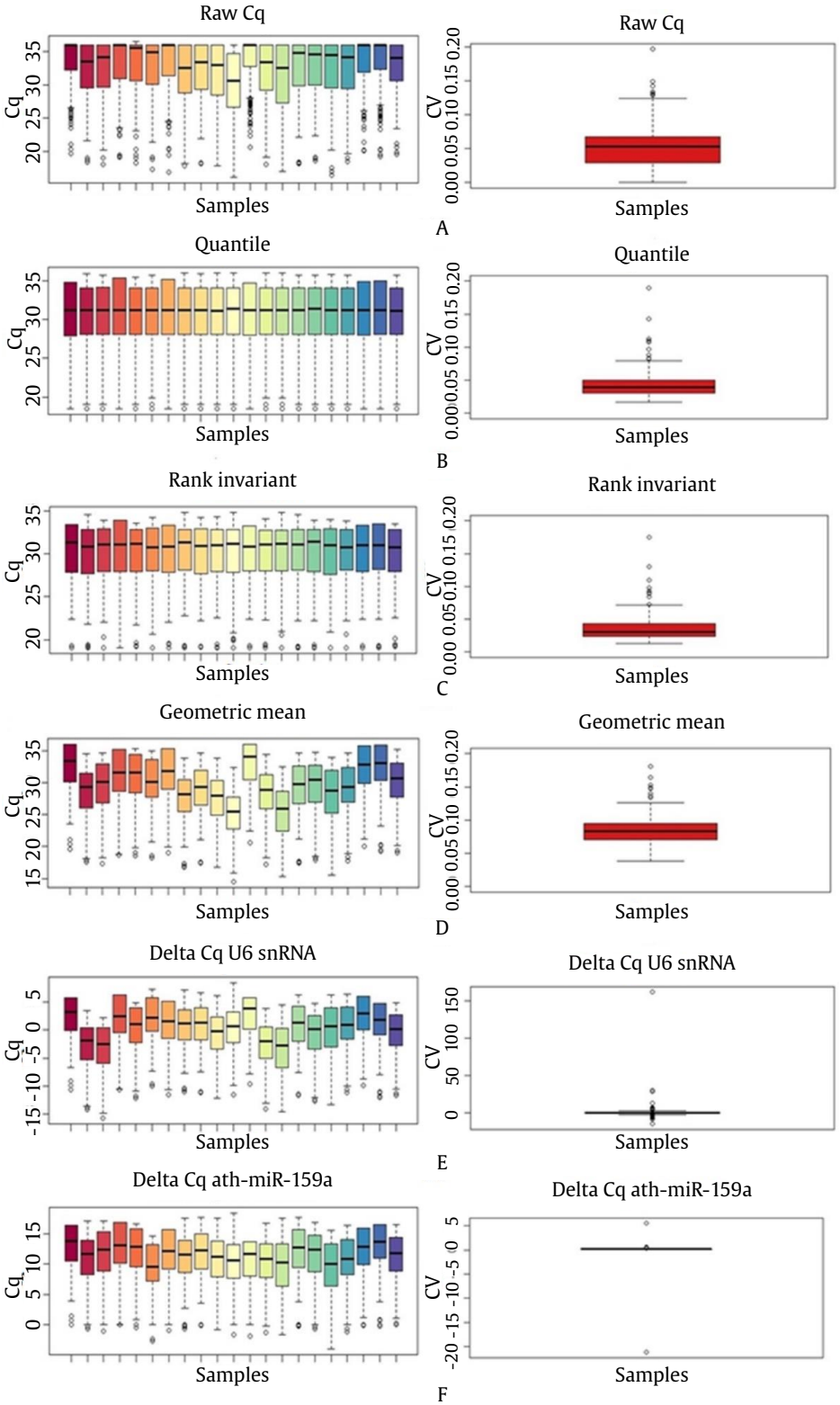

Figure 1. Cq and CV box plots for normalization methods in HTqPCR package for card A. (A) Cq and CV box plots for raw Cq. (B) Cq and CV box plots for quantile normalization. (C) Cq and CV box plots for rank invariant normalization. (D) Cq and CV box plots for geometric mean. (E) Cq and CV box plots for U6 snRNA normalization. (F) Cq and CV box plots for ath-miR-159a normalization

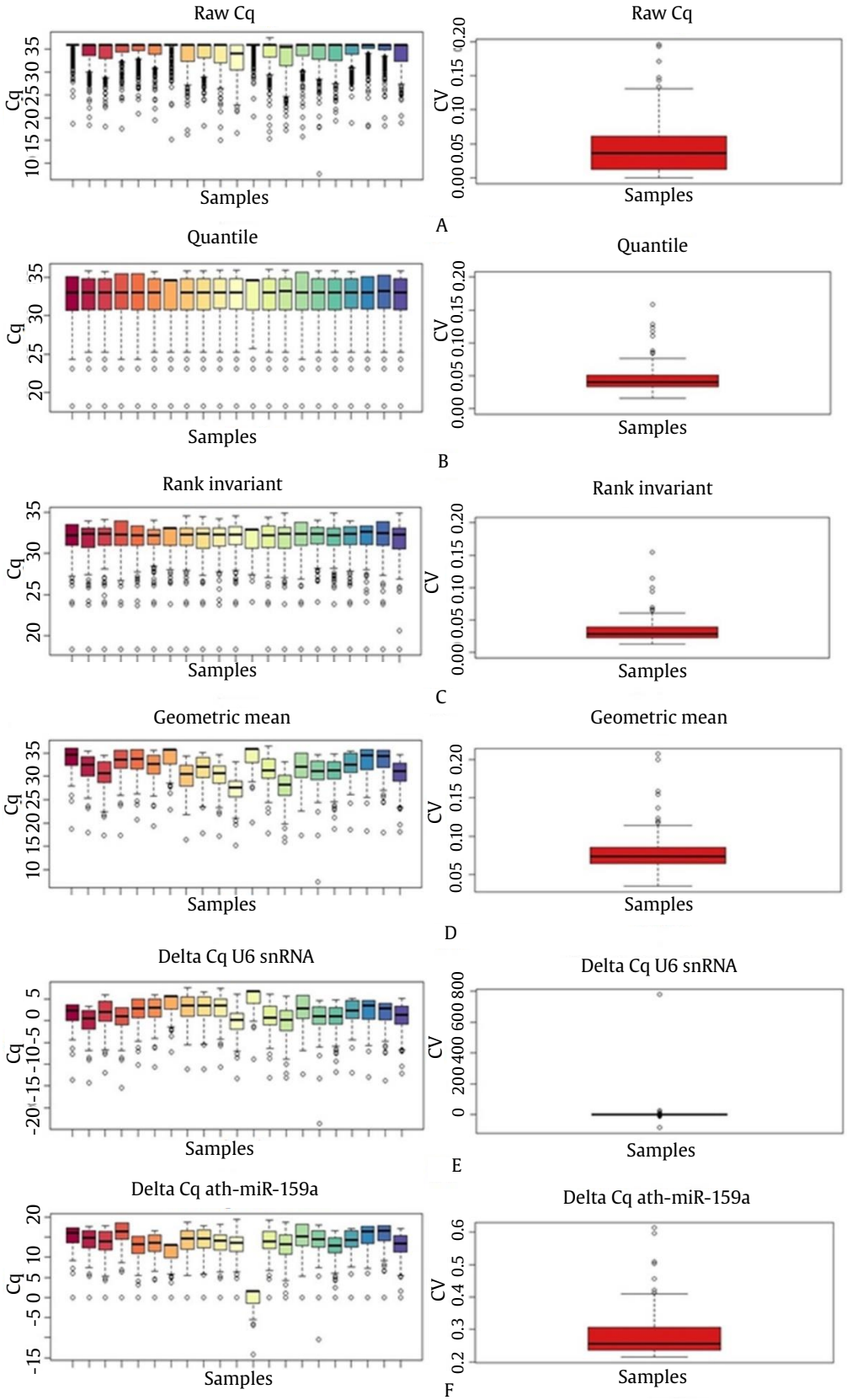

Figure 2. Cq and CV box plots for normalization methods in HTqPCR package for card B. (A) Cq and CV box plots for raw Cq. (B) Cq and CV box plots for quantile normalization. (C) Cq and CV box plots for rank invariant normalization. (D) Cq and CV box plots for geometric mean. (E) Cq and CV box plots for U6 snRNA normalization. (F) Cq and CV box plots for ath-miR-159a normalization

be assessed for their normalization performance as these reference genes have been removed from the data set during the second filtration step. The failure of RNU44 and RNU48 in passing the pre-set criteria in the pre-processing filtration indicated that these small-nucleolar RNAs are not suitable reference genes for circulating miRNAs in this study. Figure 1 shows that quantile normalization produced the best normalized Cq distribution in card A for its more homogenous mean Cq across samples and lesser Cq outliers as compared to the other methods. The quantile normalized Cq also showed a decrease in CV compared to the raw Cq. Furthermore, no obvious difference in box plots alignment has been observed in quantile normalized Cq, indicating a well-conducted experiment. Even though rank invariant normalized Cq had about the same reduction of CV as quantile normalized Cq, rank invariant normalized Cq produced less homogenous Cq distribution and had more outliers as compared to quantile normalized Cq. Geometric mean normalization produced less homogenous Cq distribution and no obvious reduction in CV was observed. Meanwhile, the normalization method using reference genes U6 snRNA and ath-miR-159a, also produced less homogenous Cq distribution after normalization. Although the CVs were reduced visually in the normalized Cq of both reference genes, the extreme outliers can be observed in their CV box plots. This indicated the inappropriateness of using reference genes as the normalizers in the high throughput study. Considering the overall assessment results on the normalization methods, the quantile normalization appeared to be the best method for normalizing the raw Cq data for card A of this study.

Quantile normalization was still the best method for normalization for card B, as illustrated in Figure 2, where it produced the best homogenous normalized Cq distribution and less Cq outliers as compared to the other methods. The reduction in CV of quantile normalized Cq was just about the same as rank invariant normalized Cq. However, the rank invariant normalized Cq had less homogenous Cq distribution as compared to quantile normalized Cq, which made quantile normalization better than rank invariant normalization. For the remaining normalization methods, namely the geometric mean, reference gene U6 snRNA and reference gene ath-miR-159a, the Cq distribution produced after normalization were apparently lesser homogenous and the CV were reduced lesser as compared to quantile normalization. Therefore, for both card A and B, quantile normalization showed the best normalization method for the screening phase of this study as it produces homogenous normalized Cq distribution across the samples, less Cq outliers and more reduced CV as compared to the other methods.

The screening of potential reference genes among candidate miRNAs from card A and B, based on the ranking by SSS is presented in Table 1. The ranking by SSS showed that hsa-miR-30b, hsa-miR-30c, hsa-miR-374, hsa-miR-301 and hsa-let-7d were among the five miRNAs with lowest SSS value. Thus, these miRNAs were selected as candidate reference genes to be tested in the validation phase.

Table 1. Ranking by SSS based on scores of candidate reference genes from geNorm, NormFinder and CV

Candidate reference genes	geNorm	NormFinder	CV	SSS
hsa-miR-30b	0.943	0.151	0.271	0.993
hsa-miR-30c	0.994	0.150	0.269	1.041
hsa-miR-374	1.139	0.163	0.293	1.187
hsa-miR-301	1.239	0.161	0.252	1.275
hsa-let-7d	1.278	0.175	0.254	1.315
hsa-miR-24	1.528	0.280	0.327	1.587
hsa-miR-331	1.561	0.143	0.277	1.592
hsa-miR-223#	1.642	0.205	0.207	1.668
hsa-miR-29a	1.756	0.154	0.282	1.785
hsa-miR-10a	2.882	0.182	0.144	2.891
hsa-miR-214	2.660	0.182	0.155	2.671
hsa-miR-323-3p	3.151	0.195	0.140	3.160
hsa-miR-642	3.068	0.188	0.148	3.077
hsa-miR-1825	3.393	0.207	0.126	3.402

CV: coefficient of variance

3.2. Validation Phase: Circulating miRNA Expression Validation Screening using 96.96 Dynamic Array™ Integrated Fluidic Circuit Chip

The pre-processing of data for validation phase was conducted in the similar way as the screening phase except for exclusion of haemolysis-sensitive miRNAs as the haemolysis-sensitive miRNAs have already been excluded in the screening phase. The second filtration, which excluded the miRNAs with undetermined and unreliable Cq in more than 80% of the samples, resulted in no exclusion of miRNAs, indicating the high percentage of acceptable Cq of miRNAs in the validation phase for good statistical analysis.

The stability of the five candidate reference genes that were selected from screening Cq data was validated first to ensure the consistency of their stability. The calculation of SSS for these candidate reference genes showed that hsa-miR-30b and hsa-miR-30c were consistent as the top two most stable candidate reference genes in plasma, as shown in Table 2. Thus, these two miRNAs was selected as reference genes for normalization on the validation Cq data.

The result on evaluation of normalization methods for Cq data in validation phase is illustrated in Figure 3. In validation phase, rank invariant normalization cannot be performed by the HTqPCR package due to inability of the methods to detect the minimum number of unique miRNAs required for the rank invariant normalization in the validation data. This may due to reduction in number of miRNAs that remained in the validation phase. The evaluation showed that quantile normalization was still the best normalization method for this study as compared to other normalization methods as it produced homogenous Cq distribution and no outliers were visualized in the box plot as compared to other methods. The CV also was decreased in quantile normalized Cq as compared to raw Cq and no outliers were visualized. Meanwhile, the other normalization methods produced the less homogenous distribution of normalized Cq compared to quantile normalization. The CV for the other normalization methods also were either increased as compared to raw Cq, which were reference gene ath-miR-159a and geometric mean, or producing outliers, which were reference genes hsa-miR-30b and hsa-miR-30c. Therefore, quantile normalization appeared to be the best normalization method for Cq data in the validation phase of this study due to its homogenous Cq distribution and lower CV as compared to raw Cq data.

4. Discussion

Normalization of Cq data is a crucial part in a qPCR analysis as it is among the steps in data pre-processing to remove experimentally induced variation and differentiating true biological changes (Steinhoff *et al.* 2006; Meyer *et al.* 2010). Thus, selection of unreliable normalization method could affect the downstream analyses and could eventually produce misleading conclusions (Dheda *et al.* 2005; Pradervand *et al.* 2009). Four normalization methods were offered in HTqPCR package, namely delta Cq normalization, quantile normalization, rank invariant normalization, and geometric mean normalization, where their performance was evaluated on suitability for Cq data of this study (Dvinge and Bertone 2009).

Delta Cq normalization is the common technique in normalizing raw Cq data using one or more reference genes that expressed stably across the samples and tissues (Vandesompele *et al.* 2002; Peltier *et al.* 2008;

Table 2. Scores of validated five candidate reference genes

Candidate reference genes	Plasma			
	geNorm	NormFinder	CV	SSS
hsa-miR-30b	0.38	0.095	0.133	0.414
hsa-miR-30c	0.38	0.055	0.112	0.399
hsa-miR-374	0.623	0.08	0.096	0.635
hsa-miR-301	0.893	0.083	0.107	0.903
hsa-let-7d	0.704	0.119	0.126	0.725

CV: coefficient of variance, SSS: summarized stability score

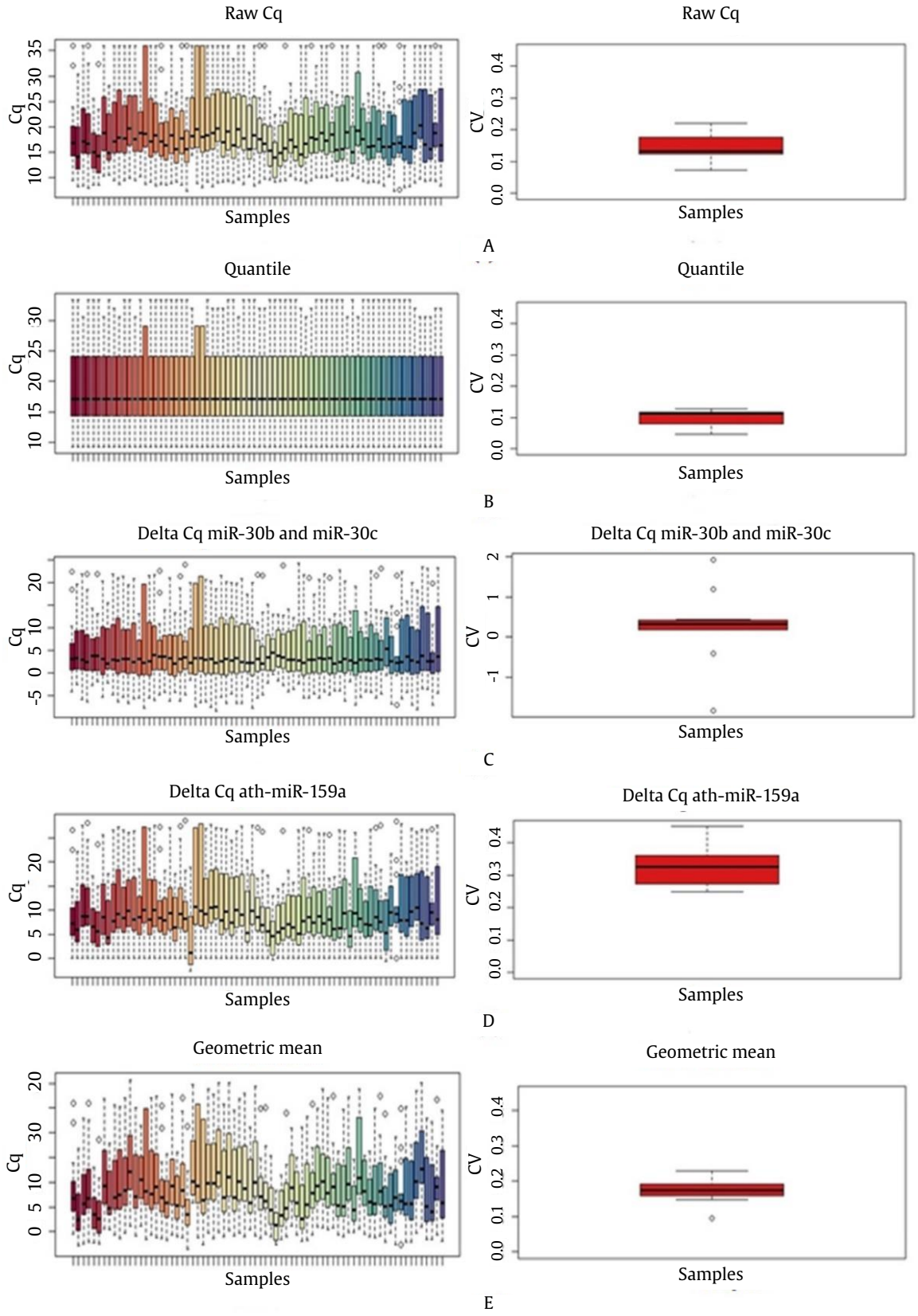

Figure 3. Evaluation of normalization methods in validation phase. (A) Cq and CV box plots for raw Cq. (B) Cq and CV box plots for quantile normalization. (C) Cq and CV box plots for delta Cq miR-30b and miR-30c normalization. (D) Cq and CV box plots for ath-miR-159a normalization. (E) Cq and CV box plots for geometric mean normalization

Meyer *et al.* 2010). In this study, three endogenous controls, namely U6 snRNA, RNU44 and RNU48, and one exogenous control, namely ath-miR-159a, were supplemented in TLDA cards as potential reference genes. RNU44 and RNU48 were excluded at early data pre-processing in the screening phase due to the presence of more than 80% of undetermined Cq in their data, indicating the unsuitability of these small RNAs for this study. The unsuitability of RNU44 and RNU48 as reference genes for circulating miRNAs has been demonstrated in previous studies. A study by Sanders *et al.* (2012) showed that RNU44 was expressed at a very low level and even undetected in some samples, while RNU48 was ranked among the least stable reference gene. Another study by McDermott *et al.* (2013) demonstrated that RNU44 and RNU48 have the least stable expression in their study, where they also used TLDA card to quantify the expression of circulating miRNAs. A recent study by Mompeón *et al.* (2020) also showed the variability expression of RNU44 and RNU48 in their plasma and serum samples, which indicated the unsuitability of these RNAs as reference genes for circulating miRNA expression analysis.

Meanwhile, the quality assessment on U6 snRNA and ath-miR-159a as reference genes showed that these small RNAs produced less homogenous Cq distribution as compared to raw Cq, as illustrated in Figure 1E, Figure 1F, Figure 2E and Figure 2F. Their CV also did not show improvement as compared to raw Cq, apart from producing some extreme outliers. Thus, these endogenous and exogenous controls demonstrated that they were not suitable to be used in this study. The use of exogenous controls as reference genes has been recommended by Vigneron *et al.* (2016). However, the current study showed that the unsuitability of exogeneous control as reference gene, namely ath-miR-159a. This is consistent with a result reported by Faraldi *et al.* (2019) who showed the less stable and unreliable of exogenous controls as reference genes that might lead to misinterpretation of results. Furthermore, Faraldi *et al.* (2019) and Dakterzada *et al.* (2020) did not agree on the use of exogenous controls as reference genes because this controls are used mainly for technical correction of variability during miRNA extraction and reverse transcription. So, the use of exogenous controls as reference genes may predispose risk to misleading interpretation of results.

Instead of the recommended candidate reference genes that were supplemented in TLDA cards, the other miRNAs in the cards were also used to search for other potential reference genes. Disagreement on the ranking of stable miRNAs was found between the results produced by geNorm, NormFinder and CV. For instance, hsa-miR-374 was the top five most stable candidate reference gene in geNorm but not in NormFinder and CV, as shown in Table 1. This may be due to different algorithms used for each method. geNorm calculates the pairwise variation of a miRNA over other candidates across the samples as the standard deviation (SD) of log-transformed expression ratios. Then, the mean variation of that miRNA with other candidates has been obtained as M-value, which is the stability value used in geNorm. NormFinder calculates the stability value based on minimal inter- and intra-group variation as the log-transformed expression ratios. The inter- and intra-group variation are estimated by considering the number of miRNAs that present in the samples and the random variation caused by biological and experimental factors. The result is reported as estimated systematic error, or rho value. Meanwhile, CV compares the variability of a miRNA across the samples after considering the total miRNA recovery for each sample. It is calculated by dividing the SD of a miRNA with its corresponding mean (Marabita *et al.* 2016). Similarly, Marabita *et al.* (2016) also reported the different rankings obtained from different algorithms in their case study analysis, which is consistent with the results in this study. Therefore, they proposed to summarize the results of different algorithms by measuring the distance from the origin in n-dimensional space, known as SSS. The SSS suggests a new ranking that summarize the results from different algorithms. In this study, two candidate miRNAs showed consistent results between the screening and validation phases based on ranking by SSS, namely the hsa-miR-30b and hsa-miR-30c. However, the evaluation on these miRNAs showed that both did not produce homogenous distribution of Cq, producing outliers and no obvious improvement of CV, as illustrated in Figure 3C. This indicates unsuitability of hsa-miR-30b and hsa-miR-30c as reference genes in this study.

The box plot analysis showed that quantile normalization was the best normalization method in this study due to the homogenous distribution and

decreased CV of its normalized Cq as compared to other methods. Principally, quantile normalization uniformizes the statistical distribution of Cq values across samples and assumes any differences observed in the distribution are induced by the technical variation (Bolstad *et al.* 2003; Dvinge and Bertone 2009). In quantile normalization, the observed distributions are forced to be the same to achieve normalization and the average distribution (average of each quantile across samples) is used as the reference (Hicks *et al.* 2018). These are shown in Figure 1B, Figure 2B and Figure 3B, where the quantile normalization corrects the distribution of raw Cq data to a uniform distribution across the samples. It also reduces the variance in gene expression data but with tolerable bias-variance trade-off, which consistent with the result of CV box plots of quantile normalized Cq in Figure 1B, Figure 2B and Figure 3B (Bolstad *et al.* 2003; Qiu *et al.* 2014). The recognition of quantile normalization as the best data driven normalizer as compared to other normalization methods was revealed in previous studies (Deo *et al.* 2011; Wozniak *et al.* 2015; Chekka *et al.* 2022).

The results from this study show that the simultaneous evaluation of several normalization methods is crucial in high throughput miRNA gene expression study to find the optimal normalizer for the data set and to prevent misleading results. For the data set in this study, quantile normalization is the recommended normalizer due to its performance in removing variation across samples with the same background, indicated by its homogenous distribution and reduced CV of normalized Cq.

Conflict of Interest

The authors declared that there is no conflict of interest in the current study.

Acknowledgements

This study was funded by Fundamental Research Grant Scheme (FRGS15-237-0478) of Ministry of Education and Research Initiative Grant Scheme (RIGS15-079-0079) of International Islamic University Malaysia. The authors would like to acknowledge the nurses from Sultan Ahmad Shah Medical Centre of IIUM, University of Science Malaysia Hospital, Tengku Ampuan Afzan Hospital, Sultan Haji Ahmad Shah Hospital and Raja Perempuan Zainab II Hospital who involved in the current study for sample collection.

References

Andersen, C.L., Jensen, J.L., Orntoft, T.F., 2004. 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, 5245–5250. https://doi.org/10.1158/0008-5472.CAN-04-0496

Bolstad, B.M., Irizarry, R.A., Astrand, M., Speed, T.P., 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics.* 19, 185–193. https://doi.org/10.1093/bioinformatics/19.2.185

Causin, R.L., Pessôa-Pereira, D., Souza, K.C.B., Evangelista, A.F., Reis, R.M.V., Fregnani, J.H.T.G., Marques, M.M.C., 2019. Identification and performance evaluation of housekeeping genes for microRNA expression by reverse transcription-quantitative PCR using liquid-based cervical cytology samples. *Oncol. Lett.* 18, 4753–4761. https://doi.org/10.3892/ol.2019.10824

Chekka, L.M.S., Langaee, T., Johnson, J.A., 2022. Comparison of data normalization strategies for array-based microRNA profiling experiments and identification and validation of circulating microRNA as endogenous control in hypertension. *Front. Genet.* 13, 836636. https://doi.org/10.3389/fgene.2022.836636

Dakterzada, F., Targa, A., Benítez, I.D., Romero-ElKhayat, L., de Gonzalo-Calvo, D., Torres, G., Moncusí-Moix, A., Huerto, R., Sánchez-de-la-Torre, M., Barbé, F., Piñol-Ripoll, G., 2020. Identification and validation of endogenous control miRNAs in plasma samples for normalization of qPCR data for Alzheimer's disease. *Alzheimers Res. Ther.* 12, 163. https://doi.org/10.1186/s13195-020-00735-x

de Ronde, M.W., Pinto, Y.M., Pinto-Sietsma, S.J., 2016. Circulating microRNA biomarkers for cardiovascular risk prediction: are we approaching clinical application? *Ann. Transl. Med.* 4, 490. https://doi.org/10.21037/atm.2016.12.05

Deo, A., Carlsson, J., Lindlöf, A., 2011. How to choose a normalization strategy for miRNA quantitative real-time (qPCR) arrays. *J. Bioinform. Comput. Biol.* 9, 795-812. https://doi.org/10.1142/S0219720011005793

Dheda, K., Huggett, J.F., Chang, J.S., Kim, L.U., Bustin, S.A., Johnson, M.A., Rook, G.A.W., Zumla, A. 2005. The implications of using an inappropriate reference gene for real-time reverse transcription PCR data normalization. *Anal. Biochem.* 344, 141–143. https://doi.org/10.1016/j.ab.2005.05.022

Dvinge, H., Bertone, P., 2009. HTqPCR: high-throughput analysis and visualization of quantitative real-time PCR data in R. *Bioinformatics.* 25, 3325–3326. https://doi.org/10.1093/bioinformatics/btp578

Eisenberg, E., Levanon, E.Y., 2003. Human housekeeping genes are compact. *Trends Genet.* 19, 362–365. https://doi.org/10.1016/S0168-9525(03)00140-9

Eisenberg, E., Levanon, E.Y., 2013. Human housekeeping genes, revisited. *Trends Genet.* 29, 569–574. https://doi.org/10.1016/j.tig.2013.05.010

Faraldi, M., Gomarasca, M., Sansoni, V., Perego, S., Banfi, G., Lombardi, G., 2019. Normalization strategies differently affect circulating microRNA profile associated with the training status. *Sci. Rep.* 9, 1584. https://doi.org/10.1038/s41598-019-38505-x

Gevaert, A.B., Witvrouwen, I., Vrints, C.J., Heidbuchel, H., Van Craenenbroeck, E.M., Van Laere, S.J., Van Craenenbroeck, A.H., 2018. MicroRNA profiling in plasma samples using qPCR arrays: recommendations for correct analysis and interpretation. *PLoS One.* 13, e0193173. https://doi.org/10.1371/journal.pone.0193173

Hicks, S.C., Okrah, K., Paulson, J.N., Quackenbush, J., Irizarry, R.A., Bravo, H.C., 2018. Smooth quantile normalization. *Biostatistics.* 19, 185–198. https://doi.org/10.1093/biostatistics/kxx028

Kirschner, M.B., Edelman, J.J., Kao, S.C., Vallely, M.P., van Zandwijk, N., Reid, G., 2013. The impact of haemolysis on cell-free microRNA biomarkers. *Front. Genet.* 4, 94. https://doi.org/10.3389/fgene.2013.00094

Kumar, S., Reddy, P.H., 2016. Are circulating microRNAs peripheral biomarkers for Alzheimer's disease? *Biochim. Biophys. Acta.* 1862, 1617–1627. https://doi.org/10.1016/j.bbadis.2016.06.001

Liu, X., Li, N., Liu, S., Wang, J., Zhang, N., Zheng, X., Leung, K.S., Cheng, L., 2019. Normalization method for the analysis of unbalanced transcriptome data: a review. *Front. Bioeng. Biotechnol.* 7, 358. https://doi.org/10.3389/fbioe.2019.00358

MacLellan, S.A., MacAulay, C., Lam, S., Garnis, C., 2014. Pre-profiling factors influencing serum microRNA levels. *BMC Clin. Pathol.* 14, 27. https://doi.org/10.1186/1472-6890-14-27

Marabita, F., de Candia, P., Torri, A., Tegnér, J., Abrignani, S., Rossi, R.L., 2016. Normalization of circulating microRNA expression data obtained by quantitative real-time RT-PCR. *Brief. Bioinform.* 17, 204–212. https://doi.org/10.1093/bib/bbv056

McDermott, A.M., Kerin, M.J., Miller, N., 2013. Identification and validation of miRNAs as endogenous controls for RQ-PCR in blood specimens for breast cancer studies. *Plos One.* 8, e83718. https://doi.org/10.1371/journal.pone.0083718

Mestdagh, P., Van Vlierberghe, P., De Weer, A., Muth, D., Westermann, F., Speleman, F., Vandesompele, J., 2009. A novel and universal method for microRNA RT-qPCR data normalization. *Genome Biol.* 10, R64. https://doi.org/10.1186/gb-2009-10-6-r64

Meyer, S.U., Pfaffl, M.W., Ulbrich, S.E. 2010. Normalization strategies for microRNA profiling experiments: A 'normal' way to a hidden layer of complexity? *Biotechnol. Lett.* 32, 1777–1788. https://doi.org/10.1007/s10529-010-0380-z

Mompeón, A., Ortega-Paz, L., Vidal-Gómez, X., Costa, T.J., Pérez-Cremades, D., Garcia-Blas, S., Brugaletta, S., Sanchis, J., Sabate, M., Novella, S., Dantas, A.P., Hermenegildo, C., 2020. Disparate miRNA expression in serum and plasma of patients with acute myocardial infarction: a systematic and paired comparative analysis. *Sci. Rep.* 10, 5373. https://doi.org/10.1038/s41598-020-61507-z

Peltier, H.J., Latham, G.J., 2008. Normalization of microRNA expression levels in quantitative RT-PCR assays: Identification of suitable reference RNA targets in normal and cancerous human solid tissues. *RNA.* 14, 844–852. http://www.rnajournal.org/cgi/doi/10.1261/rna.939908

Pizzamiglio, S., Zanutto, S., Ciniselli, C.M., Belfiore, A., Bottelli, S., Gariboldi, M., Verderio, P., 2017. A methodological procedure for evaluating the impact of haemolysis on circulating microRNAs. *Oncol. Lett.* 13, 315–320. https://doi.org/10.3892/ol.2016.5452

Pradervand, S., Weber, J., Thomas, J., Bueno, M., Wirapati, P., Lefort, K., Dotto, G.P., Harshman, K., 2009. Impact of normalization on miRNA microarray expression profiling. *RNA.* 15, 493–501. http://www.rnajournal.org/cgi/doi/10.1261/rna.1295509

Qiu, X., Hu, R., Wu, Z., 2014. Evaluation of bias-variance trade-off for commonly used post-summarizing normalization procedures in large-scale gene expression studies. *PLoS One.* 9, e99380. https://doi.org/10.1371/journal.pone.0099380

Sanders, I., Holdenrieder, S., Walgenbach-Brünagel, G., Ruecker, A.V., Kristiansen, G., Müller, S.C., Ellinger, J., 2012. Evaluation of reference genes for the analysis of serum microRNA in patients with prostate cancer, bladder cancer and renal cell carcinoma. *Int. J. Urol.* 19, 1017–1025. https://doi.org/10.1111/j.1442-2042.2012.03082.x

Shkurnikov, M.Y., Knyazev, E.N., Fomicheva, K.A., Mikhailenko, D.S., Nyushko, K.M., Saribekyan, E.K., Samatov, T.R., Alekseev, B.Y., 2016. Analysis of plasma microRNA associated haemolysis. *Bull. Exp. Biol. Med.* 160, 748–750. https://doi.org/10.1007/s10517-016-3300-y

Steinhoff, C., Vingron, M., 2006. Normalization and quantification of differential expression in gene expression microarrays. *Brief. Bioinformatics.* 7, 166–177. https://doi.org/10.1093/bib/bbl002

Sysi-Aho, M., Katajamaa, M., Yetukuri, L., Oresic, M., 2007. Normalization method for metabolomics data using optimal selection of multiple internal standards. *BMC Bioinform.* 8, 93. https://doi.org/10.1186/1471-2105-8-93

Tan, G.W., Tan, L.P., 2017. High-throughput RT-qPCR for the analysis of circulating microRNAs. *Methods Mol. Biol.* 1580, 7–19. https://doi.org/10.1007/978-1-4939-6866-4_2

Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., Speleman, F., 2002. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* 3, RESEARCH0034. https://doi.org/10.1186/gb-2002-3-7-research0034

Veryaskina, Y.A., Titov, S.E., Ivanov, M.K., Ruzankin, P.S., Tarasenko, A.S., Shevchenko, S.P., Kovynev, I.B., Stupak, E.V., Pospelova, T.I., Zhimulev, I.F., 2022. Selection of reference genes for quantitative analysis of microRNA expression in three different type of cancers. *PLoS One.* 17, e0254304. https://doi.org/10.1371/journal.pone.0254304

Vigneron, N., Meryet-Figuière, M., Guttin, A., Issartel, J.P., Lambert, B., Briand, M., Louis, M.H., Vernon, M., Lebailly, P., Lecluse, Y., Joly, F., Krieger, S., Lheureux, S., Clarisse, B., Leconte, A., Gauduchon, P., Poulain, L., Denoyelle, C., 2016. Towards a new standardized method for circulating miRNAs profiling in clincal studies: interest of the exogenous normalization to improve miRNA signature accuracy. *Mol. Oncol.* 10, 981–992. https://doi.org/10.1016/j.molonc.2016.03.005

Wozniak, M.B., Scelo, G., Muller, D.C., Mukeria, A., Zaridze, D., Brennan, P., 2015. Circulating microRNAs as non-invasive biomarkers for early detection for non-small-cell lung cancer. *PLoS One.* 10, e0125026. https://doi.org/10.1371/journal.pone.0125026