BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 2 2010, pages 228–234
doi:10.1093/bioinformatics/btp655

Gene expression

A personalized microRNA microarray normalization method using a logistic regression model

Bin Wang[1], Xiao-Feng Wang[2], Paul Howell[3], Xuemin Qian[3], Kun Huang[1], Adam I. Riker[4], Jingfang Ju[5] and Yaguang Xi[3,*]

[1]Department of Mathematics and Statistics, University of South Alabama, Mobile, AL 36688, [2]Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland, OH 44195, [3]Mitchell Cancer Institute, University of South Alabama, Mobile, AL 36604, [4]Department of Surgery, Ochsner Health System, Ochsner Cancer Institute, New Orleans, LA 70121 and [5]Department of Pathology, School of Medicine, Stony Brook University, Stony Brook, NY 11794, USA

Received on August 7, 2009; revised on November 18, 2009; accepted on November 19, 2009

Advance Access publication November 23, 2009

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: MicroRNA (miRNA) is a set of newly discovered non-coding small RNA molecules. Its significant effects have contributed to a number of critical biological events including cell proliferation, apoptosis development, as well as tumorigenesis. High-dimensional genomic discovery platforms (e.g. microarray) have been employed to evaluate the important roles of miRNAs by analyzing their expression profiling. However, because of the small total number of miRNAs and the absence of well-known endogenous controls, the traditional normalization methods for messenger RNA (mRNA) profiling analysis could not offer a suitable solution for miRNA analysis. The need for the establishment of new adaptive methods has come to the forefront.

Results: Locked nucleic acid (LNA)-based miRNA array was employed to profile miRNAs using colorectal cancer cell lines under different treatments. The expression pattern of overall miRNA profiling was pre-evaluated by a panel of miRNAs using Taqman-based quantitative real-time polymerase chain reaction (qRT-PCR) miRNA assays. A logistic regression model was built based on qRT-PCR results and then applied to the normalization of miRNA array data. The expression levels of 20 additional miRNAs selected from the normalized list were post-validated. Compared with other popularly used normalization methods, the logistic regression model efficiently calibrates the variance across arrays and improves miRNA microarray discovery accuracy.

Availability: Datasets and R package are available at http://gauss.usouthal.edu/publ/logit/

Contact: xi@usouthal.edu

1 INTRODUCTION

MicroRNAs (miRNAs) are naturally occurring small single-stranded non-coding RNAs (ncRNAs) that mediate gene expression at the post-transcriptional and translational level in both plants and animals. The first miRNA, *lin-4*, was initially discovered over a decade ago in *Caenorhabditis elegans* and controls the timing and progression of the nematode life cycle (Feinbaum and Ambros, 1999; Lee *et al*., 1993; Reinhart *et al*., 2000). However, the importance of miRNA research has not been appreciated until recently with the discoveries of hundreds of miRNAs in worm, fly and mammalian genomes (Berezikov *et al*., 2005; Lagos-Quintana *et al*., 2003). Many miRNAs are evolutionarily conserved, indicating that these miRNAs are involved with essential biological processes such as development, cell growth, differentiation, apoptosis and tumorigenesis (Baskerville and Bartel, 2005; Carmell *et al*., 2002; Esquela-Kerscher and Slack, 2006; Karube *et al*., 2005; Lee *et al*., 2005; Sempere *et al*., 2003; Takamizawa *et al*., 2004).

The miRNA research has come to the forefront thanks to their unique signatures. In contrast to messenger RNA (mRNA), miRNAs are regulatory molecules that come in small numbers (<1000). Their small size translates into the stable analysis of clinically archived samples (Xi *et al*., 2007). Moreover, miRNA regulates >30% of all human genes at the post-transcriptional and translational levels. The substantial value of miRNAs for diagnostic and prognostic determination as well as for eventual therapeutic intervention has been demonstrated (Nakajima *et al*., 2006; Xi *et al*., 2006a). Along with increasing interest in miRNAs, most well-established molecular and biological technologies have been successfully transferred into miRNA research, such as microarray and qRT-PCR.

Microarray is a high-dimensional discovery tool for genomic research. Probe-target hybridization is the central concept to determine relative abundance of nucleic acid sequences through fluorescence-based detection (D'Auria *et al*., 2003). Therefore, in microarray experiments, variations of expression measurements among arrays can be attributed to many different sources, including sample preparation, dying, image intensity and microarray hybridization, scanning and equipment errors, etc. Normalization is an essential step to reduce non-biological errors and convert raw data to valid results. For mRNA, we can assume that (i) the total number of mRNA transcripts is abundant; (ii) the expression level of a majority of genes is constant. These assumptions are valid when a large transcriptome chip with thousands of genes is applied. The miRNA arrays are usually low density spotted arrays due to the fact

*To whom correspondence should be addressed.

that the total number of miRNAs (∼1000) is much smaller. As a result, normalization methods based on the above two assumptions might not be feasible for miRNAs. Also, there are few validated housekeeping miRNAs that can be perfectly used as specific control spots for normalization.

In this study, we propose to calibrate the probe-to-probe variations on miRNA array data by introducing external information from the qRT-PCR results. qRT-PCR is a prevalent molecular analysis technique for amplification and simultaneous quantification of a target molecule. Over the past few years, the development of novel chemistries and instrumentation platforms enabling detection of PCR products on a real-time basis have led to widespread adoption of qRT-PCR. It is becoming the preferred method and the gold standard for validating results obtained from array analyses and other techniques that evaluate global gene expression changes (Schmittgen *et al*., 2008). Here, we built a logistic regression model using qRT-PCR results together with some auxiliary information in order to normalize the array data. The performance of the proposed method is also compared with some existing normalization methods.

2 MATERIALS AND METHODS

2.1 Cell lines and reagents

The HCT-116 (wt-p53) and HCT-116 (null-p53) cell lines were a gift from Dr Bert Vogelstein at The Johns Hopkins University (Baltimore, MD, USA) and were described in detail previously (Bunz *et al*., 1998, 1999). Both cell lines were maintained in McCoy's medium supplemented with 10% fetal bovine serum, 1 mM/l sodium pyruvate, 2 mM/l L-Glutamine and antibiotics. All cell lines were grown at 37°C in a humidified incubator with 5% CO_2. 5-Fluorouracil (5-FU), oxaliplatin (OX) and Irinotecan (CPT-11) were purchased from Sigma Inc. (St. Loius, MO). Two cell lines were treated by these three drugs individually at the concentrations of 5 μM, 0.5 μM and 5 μM for 24 h prior to being harvested. The non-treated cell lines served as controls. Throughout this article, we regard each cell line under any of the above three drugs or no treatment as a 'sample'. Thus, a total of eight (8) 'samples' were involved in this study.

2.2 miRNA expression analysis using locked nucleic acid miRNA array

Total RNA was isolated from treated samples and controls using the standard protocol (Xi *et al*., 2006b). One microgram of total RNA was labeled with the miRCURY LNA miRNA Array labeling kit (Exiqon Inc., Vedbaek, Denmark) following the manufacturer's instructions. The labeled samples were loaded on the miRCURY LNA miRNA Array v7.5.0 (Exiqon Inc.) and hybridized for 16 h at 60°C. The slides were scanned by an Axon GenePix Professional 4200A microarray scanner (Molecular Devices Corp., Sunnyvale, CA). ImaGene 7.0 (BioDiscovery Inc., El Segundo, CA) gridded the images and generated the digital raw data. Each sample was duplicated for the array experiment, and each array includes four replicated probes for every miRNA. As a result, a total of eight measurements were captured for each miRNA.

2.3 qRT-PCR analysis for miRNA expression

Total RNA isolated from each sample was profiled using Taqman-based qRT-PCR on an ABI 7500HT instrument (Applied Biosystems Inc., Foster City, CA). All qRT-PCR reagents were purchased from ABI, unless otherwise mentioned. The experiments were conducted by following strictly the manufacturer's instructions. Each sample was amplified in triplicate. Profiling data based on 37 randomly selected miRNA assays were initially employed to build a logistic regression model for normalization. The additional 20 miRNAs selected from the normalized data were post-tested using qRT-PCR again. The results were used to validate predictions made by the logistic regression model. Gene expression, ΔC_T, values of the selected miRNAs from each sample were calculated by normalizing with the internal control RUN6B, and relative quantitation (RQ) values were calculated according to standard formulas (Livak and Schmittgen, 2001).

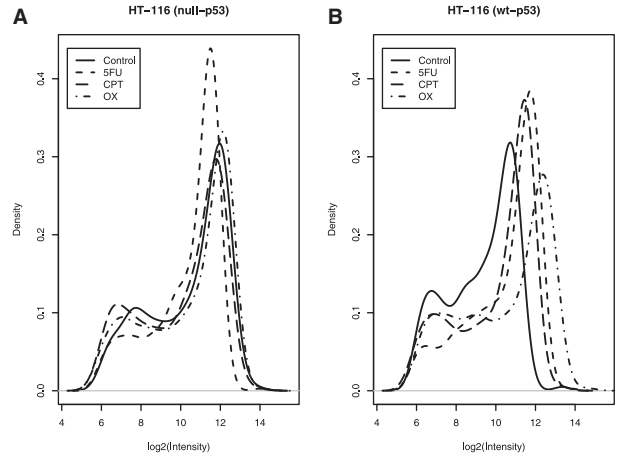

Fig. 1. Distributions of the $\log_2$-transformed intensity measures for HCT-116 (null-p53) (**A**) and HCT-116 (wt-p53) (**B**). The density curves were generated with function *density* in CRAN R with default parameters.

2.4 Data

A total of 359 miRNAs were profiled for all eight samples using locked nucleic acid (LNA) miRNA microarray. Due to the fact that the intensity measures of many miRNAs are close to the background, the measurements close to zero are dense. To gain a better view of the distributions of the data from all arrays, we took the base-2 logarithms of the averaged intensities of the replicates after background noise subtraction. The estimated density curves are shown in Figure 1. For both cell lines, the distributions under different treatments demonstrate similar shapes.

2.5 Normalizing miRNA data by fitting a logistic regression model

Let x_{ijk} be the intensity of miRNA k under treatment j for sample i, where $k=1,\ldots 359$, $j=0,1,2,3$ for control, OX, 5-FU and CPT-11, respectively, and $i=1,2$ for HCT-116 (wt-p53) and HCT-116 (null-p53), respectively. The proposed normalization method consists of the following two stages:

Stage 1: normalize arrays from the three treatments for different samples by assuming that the majority of the miRNAs do not change significantly. To achieve this, we first compute the fold changes (FCs) by

$$\mathrm{FC}_{ijk}=\frac{x_{ijk}}{x_{i0k}} \quad \text{for } i=1,2,\ j=1,2,3 \text{ and } k=1,\ldots,359. \tag{1}$$

Second, we sort the FC values and compute a p%-trimmed mean, $\overline{\mathrm{FC}}_{ij}$, by dropping p% of the FC values at each of the two ends. The $(100-2p)$% of the miRNAs with FC values in the middle are not believed to be differentially expressed. The choice of the value of p depends on the distribution of FC values. In our study, $p=25$ is sufficient. Third, normalize the arrays by

$$x'_{ijk}=\frac{x_{ijk}}{\overline{\mathrm{FC}}_{ij}} \quad \text{for } i=1,2,\ j=1,2,3 \text{ and } k=1,\ldots,359. \tag{2}$$

Stage 2: fit a logistic regression model and justify the expression patterns by external information from qRT-PCR and auxiliary variables (sample and treatment types). Throughout this study, we assume that the qRT-PCR results reflect the true expression patterns and they will be used as gold standards to

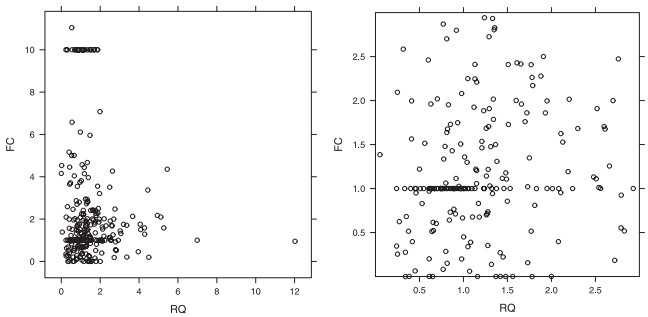

Fig. 2. Scatter plot of the array intensity fold changes (labeled as FC on Y-axis) against the qRT-PCR RQ values (labeled as RQ on X-axis).

calibrate the array results (Schmittgen *et al*., 2008). A set of 37 miRNAs, of which we had Taqman assays available in our lab, were selected and tested by qRT-PCR. Figure 2 shows the scatter plot of the qRT-PCR RQ values against the corresponding array intensity FCs after a scaling normalization by the median. We find that there is no significant linear relationship between the qRT-PCR RQ values and the intensity FC values. The Pearson correlation coefficient between the RQ and FC is -0.072 with P-value 0.485, which indicates that the relationship between the RQ and FC might be non-linear, or there might exists gene- and platform-specific effects. As a result, a linear normalization method can hardly work well. To improve the consistency between the two platforms and hence improve the normalization, a non-linear normalization method should be considered. In this study, we relax the assumption by assuming that the miRNA expression patterns (regulation trends such as up- or downregulated or no significant change) from the array and qRT-PCR results are consistent. This also relaxed our assumption of using the qRT-PCR results as a 'gold standard', and led to a less strict assumption that for the same sample the two platforms (array and qRT-PCR) shall have similar predicted regulation trends . We built a logistic regression model as follows using the RQ values:

$$\text{logit}(z_{ijk}) = \beta_0 + \beta_1 \log(x'_{ijk}) + \beta_2 \log(\text{FC}_{ijk}) + T_j * S_i + \epsilon_{ijk}, \qquad (3)$$

where $i=1,2$, $j=1,2,3$, $k=1,\ldots,359$ and $\text{logit}(x)=\log(x/(1-x))$, $z_{ijk}=Pr(Y_{ijk}=1)$ is the probability of $Y_{ijk}=1$. Y_{ijk} is an indicator showing whether the k-th miRNA is up- or downregulated according to the RQ values. The reason that we consider taking the logarithm of FC_{ijk} is that the FC values are asymmetric: downregulated miRNAs occupy the scale from 0 to 1, whereas the upregulated miRNAs occupy the scale from 1 to ∞. After taking the logarithm, the scales for both up- and down-regulated genes become symmetric. In addition to associating the expression patterns from the qRT-PCR and arrays, we also considered the effects of sample type S_i and treatment type T_j. We checked and found there was no significant interaction effect between the sample and treatment types. The logarithm of the normalized intensity x'_{ijk} was included in the model due to the fact that the significance of the observed FCs depends on the absolute expression levels of the miRNAs. The biological variance of a miRNA with lower intensity measurements is more likely to be masked by the non-biological variances and the inclusion of $\log(x'_{ijk})$ can reflect this effect to some extent. An error term ϵ_{ijk} is adopted to explain the non-biological errors from different experiments.

The up- and downregulation trends were predicted separately. Specifically, to predict the upregulation trends, we first determine whether a miRNA is upregulated ($Y_{ijk}=1$) or not ($Y_{ijk}=0$) by the qRT-PCR results with a preselected cutoff. Throughout this study, we choose 2.0 as a cutoff for the qRT-PCR RQ values: a miRNA is classified to be upregulated if $\text{RQ}_{ijk} > 2.0$. Second, we fit the model in (3) based on the 37 selected miRNAs. Third, we use the fitted model to predict that the upregulation trends of the miRNAs are not selected. To predict the downregulation trends, we set $Y_{ijk}=1$ if $\text{RQ}_{ijk} < 1/2.0$, and $Y_{ijk}=0$ otherwise. The rest of the steps are similar to those in predicting the upregulation trends.

By discretizing the RQ values into up/downregulation trends, attention needs be paid to the overdispersion in the logistic regression model fitting. In this study, we adjust for overdispersion with the quasilikelihood approach (Agresti, 1996).

3 RESULTS

3.1 Measures of the performance of the normalization methods

We assume the microarray results can reveal the actual changes of miRNAs after appropriate normalization, and the intensity-based FCs are supposed to show consistency with the qRT-PCR results. To evaluate the performance of the proposed normalization method, another set of 20 miRNAs were selected from the normalized list to be post-tested for validation. To measure the consistency, we adopted the following four criteria:

(1) Pearson's correlation coefficient, which measures the linear association between two random variables. After normalization, the Pearson's correlation coefficient between RQ and FC will be computed.

(2) Weighted kappa test, which can be used to measure the consistency between two raters (Cohen, 1960; Fleiss, 1981; Fleiss and Cohen, 1973; Fleiss *et al*., 1969). There are only three possible outcomes for each miRNA, upregualted, downregulated or no significant change, based on either the qRT-PCR RQ values or FCs. Here, we considered a weighted Kappa statistic to assess the consistency between the predicted and qRT-PCR results. The qRT-PCR and LNA array results will be treated as two raters who rate the miRNAs with scores: -1 (downregulation), 0 (no significant change) and 1 (upregulation). Let $w(z_1,z_2)$ be the weight for the qRT-PCR result z_1 and the predicted result z_2. The weights were selected as follows: $w(-1,0)=w(0,-1)=w(1,0)=w(0,1)=1/2$, $w(-1,-1)=w(0,0)=w(1,1)=1$ and $w(-1,1)=w(1,-1)=0$. Landis and Koch (1977) interpreted the kappa values as following: (a) <0: no agreement; (b) 0.0–0.2: slight agreement; (c) 0.21–0.40: fair agreement; (d) 0.41–0.60: moderate agreement; (e) 0.61–0.80: substantial agreement; (f) 0.81–1.00: almost perfect agreement.

(3) False acceptance rate (FAR), which is defined as the total probability of either a non-upregulated miRNA being classified as upregulated or a non-downregulated miRNA being classified as downregulated.

(4) False rejection rate (FRR), which is defined as the total probability of either an upregulated miRNA being rejected as an upregulated miRNA or a downregulated miRNA being rejected as a downregulated miRNA.

3.2 Existing normalization methods

To evaluate the performance of the proposed method, the following six existing normalization methods will be applied and their performances will be compared with the proposed method.

(1) No normalization: as suggested by some literature, results will be obtained by simply performing background signal subtraction on the miRNA microarray (Baskerville and Bartel, 2005; Liang *et al*., 2005; Schmittgen *et al*., 2004).

(2) Median normalization: all the miRNA microarrays are assumed to have a common median and each of the miRNA microarrays is normalized based on its median (Calin *et al*., 2004a, b; Liu *et al*., 2004).

(3) Scaling method by U6 probes: U6 is a ncRNA that is a component of the spliceosome, which is involved in RNA splicing of the pre-mRNA. The RNA sequence of U6 is the most highly conserved across organisms of all five of the small nuclear RNAs (snRNAs) (Brow and Guthrie, 1988), suggesting that the function of the U6 snRNA is both crucial and unchanged through evolution. The Exiqon miRNA array has included two U6 probes (U6-snRNA-1 and U6-snRNA-2). Due to their stability, the two U6 probes can be used as normalizers.

(4) Invariants: efforts have been made to normalize miRNA microarray data by finding one or more probes that do not change across arrays (Davison *et al*., 2006; Garzon *et al*., 2008; Hua *et al*., 2008; Pan *et al*., 2008; Perkins *et al*., 2007; Pradervand *et al*., 2009; Rao *et al*., 2008) In the same spirit of Pradervand *et al*. (2009), we find a set of miRNAs that do not change significantly according to their qRT-PCR RQ values. The miRNAs with RQ values that are not significantly different from 1.00 across arrays were selected, and the data from different arrays were then normalized by the mean value of the selected miRNAs.

(5) Cyclic loess: this approach was first presented by Dudoit *et al*. (2002) and Mascellani *et al*. (2008). For each pair of arrays $(X_{ijk}, X_{ij'k})$ with fixed $j \neq j'$, $i = 1, 2$ and $k = 1, \ldots, 359$, we first consider the $M = \log(X_{ij}/X_{ij'})$ versus $A = (\log(X_{ij} + \log(X_{ij'})/2$ plot. Second, we fit a loess curve by regressing M on A, and denote the fitted values by $\widehat{M}$. Third, we set $D = \exp((M - \widehat{M})/2)$ and justify X_{ijk} and $X_{ij'k}$ by $X'_{ijk} = X_{ijk} \times D_k$ and $X'_{ij'k} = X_{ij'k}/D_k$.

(6) Quantile method: the quantile normalization method is first proposed under the assumption that there is an underlying common distribution of intensities across arrays (Bolstad *et al*., 2003; Garzon *et al*., 2008; Northcott *et al*., 2009). It is based upon the concept of quantile–quantile plot extended to *n*-dimensions. Figure 1 shows that although the total number of miRNAs is small, the density curves of the intensity measurements from different arrays have similar shapes. Thus, the 'common distribution' assumption by the quantile normalization method is not severely violated and hence it can be applied in this analysis. The quantile normalization method has been implemented in R package *affy* and freely available from the *The Comprehensive R Archive Network* servers over the internet.

Methods 2–4 are scaling normalization methods and methods 5 and 6 and the proposed new method are non-linear. Rao *et al*. (2008) and Pradervand *et al*. (2009) performed a fairly comprehensive analysis to compare the performances of the existing one-color miRNA microarray normalization techniques. The results from both papers showed that the quantile normalization method performed best, while Pradervand *et al*. (2009) showed that both the invariant method and quantile method achieved satisfying performances.

Table 1. Performance comparisons among different normalization methods

Method	PCC	*P*-value	WKappa	FAR (%)	FRR (%)
None	−0.055	0.60	0.083	67.74	51.22
Median	−0.103	0.32	0.066	70.91	60.98
U6	−0.097	0.35	0.071	69.09	58.54
Invariant	0.028	0.79	0.072	80.00	58.54
Loess	0.021	0.84	0.123	64.41	48.78
Quantile	−0.005	0.96	0.220	63.16	48.78
Logit	0.198	0.055	0.224	61.29	41.46
Logit*	−0.072	0.485	0.069	69.23	60.98
Logit**	0.060	0.611	0.041	70.83	65.85

Note: PCC refers to the Pearson's correlation coefficient. Here, PCC is used to measure the linearity between the two vectors of average intensities from the control and treated samples after background subtraction; WKappa is the values of the weighted kappa statistic. FAR refers to the overall rate of miRNAs that are falsely classified as up- or downregulated miRNAs; FRR refers to the overall rate of miRNAs that are falsely rejected as up- or downregulated miRNAs.

3.3 Performance comparisons

The validation results for the different methods are summarized in Table 1. In Table 1, the second column shows the correlation coefficients between RQ and FC, and the third column shows the corresponding *P*-value. The correlation coefficient for the logistic regression method is computed based on the FCs after normalization from Stage 1. We see that none of the normalization methods being compared show significant linear relationship between RQ and FC.

The values of the weighted kappa statistic are shown in column 4. Based on the results of the weighted kappa tests, we see that only the quantile normalization and logistic regression methods demonstrate fair agreement (both have a *P*-value of 0.003), which shows that both the logistic regression and quantile normalization methods work well. Fair consistency can be observed between the two platforms of qRT-PCR and miRCURY LNA array after employment of either the logistic regression or the quantile normalization method. All the other methods have *P*-values >0.05, which indicates that the weight kappa statistics are not significantly different from zero.

The information of FAR and FRR are shown in columns 5 and 6. In terms of FAR and FRR, the logistic regression method has better performance than the other methods: the logistic regression method achieves the smallest FAR and FRR values. The performances of the loess method and the quantile normalization method are similar.

To demonstrate the role of Stage 1 preprocess, we analyzed the data with and without Stage 1. The results by Stage 1 alone are shown in the line labeled 'Logit*'; and the results without Stage 1 are shown at the bottom line labeled 'Logit**'. We found that Stage 1 is important to improve the overall performance of the proposed method. It can roughly align the FCs of samples under different treatments and enhance the prediction accuracy. More details are provided in Table 2 about the logistic regression and quantile normalization methods. From Table 2, we see that the quantile normalization method is somewhat better than the logistic regression method in identifying the downregulated miRNAs, while the logistic regression method works better than the quantile normalization in identifying upregulated miRNAs.

Figure 1 shows that the distributions of the intensity measurements from different experiments have similar patterns.

Table 2. Comparisons of the predicted results by the logistic regression and quantile normalization methods

	Logistic regression			Quantile normalization		
	rt.down	rt.nsc	rt.up	rt.down	rt.nsc	rt.up
fc.down	4	7	5	5	14	5
fc.nsc	4	25	4	4	26	8
fc.up	4	22	20	3	14	16
FRR (%)	66.7	53.7	31.0	58.3	51.8	44.8

Note: a prefix 'rt' indicates that the results are obtained from the qRT-PCR and a prefix 'fc' indicates results from array under a specific normalization method. 'down' indicates downregulation; 'nsc' indicates no significant change; and 'up' indicates upregulation. FRR refers to the overall rate of miRNAs that are falsely rejected as up- or downregulated miRNAs.

Table 3. Choice of the number of miRNAs

n	Avg. WKappa	Standard Error (s.e)	Avg.FAR (%)	Standard Error (s.e)	Avg.FRR (%)	Standard Error (s.e)
29	0.224	–	61.29	–	41.46	–
25	0.210	0.019	61.99	1.07	43.73	4.4
20	0.192	0.028	62.84	1.64	47.56	6.73
15	0.175	0.042	63.72	3.06	52.66	9.94
10	0.134	0.057	66.24	4.35	61.17	12.3
8	0.114	0.053	67.53	4.64	66.02	11.93

Thus, the 'common distribution' assumption by the quantile normalization method is not severely violated and hence the quantile normalization works well in this analysis.

3.4 Investigation of miRNA number for building the logistic regression model

Among the 37 miRNAs, 8 miRNAs have no valid qRT-PCR measures and the information of 29 miRNAs were used in fitting the logistic regression model. To investigate the possibility of using a smaller amount of miRNAs to fit a logistic regression model for normalization, we randomly selected a subset of size n, fitted a logistic regression model, and checked the performance of the fitted model. This procedure was repeated 100 times, and the results are shown in Table 3.

In Table 3, the first row shows the results of the fitted model by using the information of all available miRNAs. For $n=25, 20, 15, 10$ and 8, the average weighted kappa statistics, FAR and FRR were computed based on the results from 100 iterations, and the corresponding standard errors were also listed. We found that when the number of miRNAs decreases, the average weighted kappa decreases, meanwhile the average FAR, average FRR and standard errors increase. Overall, the larger the number of miRNAs, the better the fitted model and its performance. Though, when n becomes too small, say $n=6$, we had trouble fitting the logistic regression model based on the selected subset of miRNAs, and thus failed to obtain robust prediction results. From Tables 2 and 3, we see that when $n=29$, the logistic regression method has slight improvement over quantile normalization. When $n=20$ and $n=25$, although the logistic regression method has smaller weighted kappas than quantile normalization, its FAR and FRR are smaller than those of quantile normalization. This also demonstrates the potential of using the proposed logistic regression model to improve the array data normalization with increased miRNA numbers.

4 DISCUSSION

Normalization is an essential matter for discovery experiments using microarray. It has a profound impact on accuracy, precision and overfitting (Argyropoulos *et al.*, 2006). The fundamental assumption of most established normalization methods suitable for high-density arrays is that relatively few genes will be dramatically up- or downregulated compared with the total number genes, and the overall distribution for each slide will have a similar pattern. However, miRNA has broken this assumption because of its small total number. Consequently, current miRNA microarray platforms possibly do not include enough miRNAs with stable expression (Davison *et al.*, 2006).

In order to overcome this obstacle, specific controls have been recommended for miRNA normalization. Careful selection of appropriate controls is extremely critical since variation has been observed for the most commonly used housekeeping markers in mRNA analyses: for example, β-Actin and GAPDH (de Kok *et al.*, 2005). The ideal controls should be consistently stable and highly abundant despite tissue types or treatments. Also, they should have characteristics similar to miRNAs, including size, biogenesis and stability. According to this criteria, mRNA and synthesized spike-in controls are not perfectly suitable for the normalization purpose. ncRNAs are the other class of small RNA molecules including transfer RNA (tRNA), snRNA and small nucleolar RNA (snoRNA) (Kiss, 2002). Their sizes range from 45 to 200 nt and their characters are closer to miRNA. Some array platforms have adapted to utilize ncRNAs as normalization controls included on slides, such as the new version of Exiqon miRCURY LNA miRNA Array and Luminex FlexMIR panels. Nevertheless, in one of our ongoing projects, we found some ncRNAs used for normalization controls contained in these arrays could be influenced by chemo drug treatments, such as 5-FU, Cisplatin or Doxorubicin (data not shown). As a result, we need to be aware of the stability of normalization controls across a relatively wide variety of tissues, cell lines and conditions. A number of normalization controls is thus recommended to be included and their stability validation is suggested to be performed before normalization.

Given the change of a relatively high portion of total miRNAs under varied situations and the difficulty of finding appropriate specific normalization controls, how could we efficiently take advantage of the benefit brought by a high-dimensional microarray platform for miRNA research? In order to address this question, we attempted to introduce the personalized concept into miRNA array normalization. Our assumption is that the expression pattern of an entire miRNA population under certain conditions should remain consistent within any applied discovery and validation platforms, such as microarray and qRT-PCR. qRT-PCR is becoming the gold standard for relative gene expression analysis and one of the major validation methods after high-dimensional discovery including miRNA research (Schmittgen *et al.*, 2008). In this study,

we estimated the overall expression pattern of the entire miRNA profile using a panel of representative miRNAs through qRT-PCR validation. The results were led to fitting a logistic regression model which was used for array data normalization. The additional 20 normalized miRNAs were post-validated by qRT-PCR. After comparison with other normalization methods, we found that the new model improves the normalization of miRNA array-based profiling analysis by lowering FAR and FRR.

Several concerns will surely come to the forefront. Will implementation be cost-effective? How many miRNAs need to be included to build such a model? In our study, no obvious increase of cost was applicable because we took advantage of our on-hand Taqman qRT-PCR miRNA assays. They have been continually stocked for previous and ongoing projects for years. Each Taqman assay can perform hundreds of reactions which leads the average cost for an individual project relatively minimal. Core laboratory and multi-lab collaboration can also practically reduce the cost because the same panel could be applied repeatedly to different projects involving miRNA array. Compared with the expensive array costs, especially in large-scale biomarker discovery using multiple samples, the cost to perform qRT-PCR and accomplish critical normalization will not be a huge burden. Also, the cost will surely be decreased along with technology development. For example, SYBR Green assays have been well developed for miRNA research with a much lower cost. This study also showed a potential for developing an affordable low cost assay for miRNA profiling data normalization by using the proposed logistic regression model.

We have used 37 randomly selected miRNAs to generate such a logistic regression model for miRNA array normalization purposes. Due to the difference of individual projects, it is not realistic to assume a certain number as a universal standard to fit a normalization model. Optimization is strongly recommended. In our results (Table 3), we investigated the efficiency of models composed of a different number of miRNAs. In our experimental model, the size could be as low as 6. However, the small size will cause higher FAR, FRR and standard errors.

Studies have confirmed that >30% of miRNAs exhibited differential expression in varied conditions (Davison *et al.*, 2006). This percentage indicates the potential of a small number of miRNA candidates to reflect the overall expression pattern. In our study, 8 out of 37 miRNAs had no measurements. We recommend choosing miRNAs with higher expression levels (higher than the overall mean or median expression value) to build the logistic regression model. Also, we purposely did not include any confirmed miRNAs associated with our experimental model in this study. However, logically, we believe that previous accomplishments or confirmed markers could improve the efficiency of optimization. Likely, it is executable to pick up a few interesting or known miRNAs in the target experimental model to design a panel for building the normalization model and further data mining. In this study, the logistic regression model was built based on one-colored array data. The application of the proposed model to two-colored arrays will be investigated in our further studies. As mentioned earlier, the current prevalent normalization methods for mRNA analysis are not well adapted to miRNA profiling studies. We are proposing a concept of personalized normalization in allusion to miRNA analysis, ultimately pushing for the development of more feasible and adaptive normalization methods for miRNA research.

ACKNOWLEDGEMENTS

We also thank Dr Cynthia Schneider for her help in revising this article.

Funding: AACR-Colorectal Cancer Coalition Fellows Grant, in memory of Lisa Dubow (to Y.X.).

Conflict of Interest: none declared.

REFERENCES

Agresti,A. (1996) *Categorical Data Analysis*. Wiley, New York.

Argyropoulos,C. *et al.* (2006) Operational criteria for selecting a cDNA microarray data normalization algorithm. *Oncol. Rep.*, **15**, 983–996.

Baskerville,S. and Bartel,D.P. (2005) Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA*, **11**, 241–247.

Berezikov,E. *et al.* (2005) Phylogenetic shadowing and computational identification of human microRNA genes. *Cell*, **120**, 21–24.

Bolstad,B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Brow,D.A. and Guthrie,C. (1988) Spliceosomal RNA U6 is remarkably conserved from yeast to mammals. *Nature*, **334**, 213–218.

Bunz,F. *et al.* (1998) Requirement for p53 and p21 to sustain G2 arrest after DNA damage. *Science*, **282**, 1497–1501.

Bunz,F. *et al.* (1999) Disruption of p53 in human cancer cells alters the responses to therapeutic agents. *J. Clin. Invest.*, **104**, 263–269.

Calin,G.A. *et al.* (2004a) MicroRNA profiling reveals distinct signatures in B cell chronic lymphocytic leukemias. *Proc. Natl Acad. Sci. USA*, **101**, 11755–11760.

Calin,G.A. *et al.* (2004b) Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proc. Natl Acad. Sci. USA*, **101**, 2999–3004.

Carmell,M.A. *et al.* (2002) The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. *Genes Dev.*, **16**, 2733–2742.

Cohen,J. (1960) A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.*, **20**, 37–46.

D'Auria,S. *et al.* (2003) DNA arrays for genetic analyses and medical diagnosis. In Lakowicz,J.R. (ed) *Topics in Fluorescence Spectroscopy*. Kluwer Academic/Plenum Publishers, New York, pp. 213–237.

Davison,T.S. *et al.* (2006) Analyzing micro-RNA expression using microarrays. *Methods Enzymol.*, **411**, 14–34.

de Kok,J.B. *et al.* (2005) Normalization of gene expression measurements in tumor tissues: comparison of 13 endogenous control genes. *Lab. Invest.*, **85**, 154–159.

Dudoit,S. *et al.* (2002) Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments. *Stat. Sin.*, **12**, 111–139.

Esquela-Kerscher,A. and Slack,F.J. (2006) Oncomirs - microRNAs with a role in cancer. *Nat. Rev.*, **6**, 259–269.

Feinbaum,R. and Ambros,V. (1999) The timing of lin-4 RNA accumulation controls the timing of postembryonic developmental events in Caenorhabditis elegans. *Dev. Biol.*, **210**, 87–95.

Fleiss,J.L. (1981) *Statistical Methods for Rates and Proportions*, 2nd edn. John Wiley, New York, pp. 38–46.

Fleiss,J.L. and Cohen,J. (1973) The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability, *Educ. Psychol. Meas.*, **33**, 613–619.

Fleiss,J.L. *et al.* (1969) Large sample standard errors of kappa and weighted kappa. *Psychol. Bull.*, **72**, 323–327.

Garzon,R. *et al.* (2008) Distinctive microRNA signature of acute myeloid leukemia bearing cytoplasmic mutated nucleophosmin. *Proc. Natl Acad. Sci. USA*, **105**, 3945–3950.

Hua,Y.J. *et al.* (2008) Comparison of normalization methods with microRNA microarray. *Genomics*, **92**, 122–128.

Karube,Y. *et al.* (2005) Reduced expression of Dicer associated with poor prognosis in lung cancer patients. *Cancer Sci.*, **96**, 111–115.

Kiss,T. (2002) Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. *Cell*, **109**, 145–148.

Lagos-Quintana,M. *et al.* (2003) New microRNAs from mouse and human. *RNA*, **9**, 175–179.

Landis,J.R. and Koch,G.G. (1977) The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.

Lee, R.C., Feinbaum, R.L. and Ambros, V. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843–854.

Lee,Y.S. *et al.* (2005) Depletion of human micro-RNA miR-125b reveals that it is critical for the proliferation of differentiated cells but not for the down-regulation of putative targets during differentiation. *J. Biol. Chem.*, **280**, 16635–16641.

Liang,R.Q. *et al.* (2005) An oligonucleotide microarray for microRNA expression analysis based on labeling RNA with quantum dot and nanogold probe. *Nucleic Acids Res.*, **33**, e17.

Liu,C.G. *et al.* (2004) An oligonucleotide microchip for genome-wide microRNA profiling in human and mouse tissues. *Proc. Natl Acad. Sci. USA*, **101**, 9740–9744.

Livak,K.J. and Schmittgen,T.D. (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods*, **25**, 402–408.

Mascellani,N. *et al.* (2008) Using miRNA expression data for the study of human cancer. *MINERVA BIOTEC.*, **20**, 23–30.

Nakajima,G. *et al.* (2006) Non-coding microRNAs hsa-let-7g and hsa-miR-181b are associated with chemoresponse to S-1 in colon cancer. *Cancer Genomics Proteomics*, **3**, 317–324.

Northcott,P.A. *et al.* (2009) The miR-17/92 polycistron is up-regulated in sonic hedgehog-driven medulloblastomas and induced by N-myc in sonic hedgehog-treated cerebellar neural precursors. *Cancer Res.*, **69**, 3249–3255.

Pan,Q. *et al.* (2008) Differential expression of microRNAs in myometrium and leiomyomas and regulation by ovarian steroids. *J. Cell Mol. Med.*, **12**, 227–240.

Perkins,D.O. *et al.* (2007) microRNA expression in the prefrontal cortex of individuals with schizophrenia and schizoaffective disorder. *Genome Biol.*, **8**, R27.

Pradervand,S. *et al.* (2009) Impact of normalization on miRNA microarray expression profiling. *RNA*, **15**, 493–501.

Rao,Y. *et al.* (2008) A comparison of normalization techniques for microRNA microarray data. *Stat. Appl. Genet. Mol. Biol.*, **7**, Article 22.

Reinhart,B.J. *et al.* (2000) The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, **403**, 901–906.

Schmittgen,T.D. *et al.* (2004) A high-throughput method to monitor the expression of microRNA precursors. *Nucleic Acids Res.*, **32**, e43.

Schmittgen,T.D. *et al.* (2008) Real-time PCR quantification of precursor and mature microRNA. *Methods*, **44**, 31–38.

Sempere,L.F. *et al.* (2003) Temporal regulation of microRNA expression in Drosophila melanogaster mediated by hormonal signals and broad-Complex gene activity. *Dev. Biol.*, **259**, 9–18.

Takamizawa,J. *et al.* (2004) Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival. *Cancer Res.*, **64**, 3753–3756.

Xi,Y. *et al.* (2006a) Prognostic values of microRNAs in colorectal cancer. *Biomark. Insights*, **2**, 113–121.

Xi,Y. *et al.* (2006b) Differentially regulated micro-RNAs and actively translated messenger RNA transcripts by tumor suppressor p53 in colon cancer. *Clin. Cancer Res.*, **12**, 2014–2024.

Xi,Y. *et al.* (2007) Systematic analysis of microRNA expression of RNA extracted from fresh frozen and formalin-fixed paraffin-embedded samples. *RNA*, **13**, 1668–1674.