

METHODOLOGY ARTICLE

Open Access

Incorporation of subject-level covariates in quantile normalization of miRNA data



Anvar Suyundikov^{1,2}, John R. Stevens^{1*}, Christopher Corcoran¹, Jennifer Herrick³, Roger K. Wolff³ and Martha L. Slattery³

Abstract

Background: Most currently-used normalization methods for miRNA array data are based on methods developed for mRNA arrays despite fundamental differences between the data characteristics. The application of conventional quantile normalization can mask important expression differences by ignoring demographic and environmental factors. We present a generalization of the conventional quantile normalization method, making use of available subject-level covariates in a colorectal cancer study.

Results: In simulation, our weighted quantile normalization method is shown to increase statistical power by as much as 10 % when relevant subject-level covariates are available. In application to the colorectal cancer study, this increase in power is also observed, and previously-reported dysregulated miRNAs are rediscovered.

Conclusions: When any subject-level covariates are available, the weighted quantile normalization method should be used over the conventional quantile normalization method.

Keywords: Normalization, Differential expression, miRNA

Background

A critical data analysis step in detecting differentially expressed microRNA (miRNA) features is normalization. The normalization procedure reduces technical variation and maintains true biological changes between arrays. Various normalization techniques exist, but until recently, all were developed for messenger RNA (mRNA) arrays. The miRNA data are very different from mRNA data due to the small total number of miRNAs (a few hundred versus 10,000 to 50,000 genes in mRNA data), and the majority of miRNAs are either not expressed or are expressed at very low levels [1]. Therefore, normalization methods used for mRNA expression arrays may not be appropriate for miRNA arrays. More information about the biology of miRNAs and their role in cancer development is given in Suyundikov [2] and Suyundikov et al. [3], which also present the colorectal cancer (CRC) study motivating this paper. Briefly, the study data used here has miRNA data from paired tumor and normal samples in

over 400 subjects, with over 2,000 miRNAs. (These are the first available subjects from about 2,000 subjects in the overall study.) The study was approved by the Institutional Review Board for Human Subjects at the University of Utah (IRB_00002335 and IRB_00055877). All participants provided informed written consent prior to participating in the study.

The importance of finding an appropriate normalization method for miRNA data in our colorectal cancer (CRC) study [3] motivated us to develop a normalization method that accounts for the characteristics of data, removes any artificial variations, and keeps the crucial biological information. CRC is the third most common type of cancer and the second leading cause of cancer death in the United States [4]. Most colorectal cancers are due to demographic, lifestyle, and health-related factors, with only a small number of cases due to underlying genetic disorders [5, 6]. Cunningham et al. [7] and Watson and Collins [8] listed older age, male gender, high intake of fat, alcohol or red meat, obesity, smoking, and a lack of physical exercise as risk factors of CRC. Taken together with known and hypothesized associations of miRNA with CRC, this suggests that the expression levels

*Correspondence: john.r.stevens@usu.edu

¹Department of Mathematics and Statistics, Utah State University, 3900 Old Main Hill, Logan, UT, 84322-3900 U.S.A

Full list of author information is available at the end of the article

of miRNAs in tissues from risk group (for example, older and smoker) patients are more likely to be differentially expressed than the expression levels from non-risk group (younger and non-smoker) patients.

In our CRC study, we have collected extensive information about demographic and lifestyle variables of CRC patients along with the miRNA features from normal and tumor samples. These data may be helpful to consider not only the artificial intra- and inter-array differences, but also the differences caused by the demographic and lifestyle characteristics of patients, and to maintain only biological differences during the normalization procedure.

In this paper, we incorporate subject-level covariates (specifically the demographic and lifestyle variables) in miRNA normalization, which has not been done before. We modify the quantile normalization method from Bolstad et al. [9] that is commonly used in miRNA data analysis and was found as an efficient method to remove the artificial differences across arrays [10–14]. The quantile normalization equalizes the distributions of expression intensities across samples while ignoring any differences of characteristics of samples. In the quantile normalization method, each subject's normalized distribution of expression values depends on all other subjects' distributions equally. Such normalization of miRNA expressions while ignoring the characteristics of data results in loss of important biological information. In our modified normalization method, we assume that the normalized distribution of miRNAs from one subject should depend on the weighted distribution of miRNAs from other subjects. The weights of subjects are determined from the distance matrix generated from various distance metrics of multiple covariates. The elements of the generated distance matrix represent the pairwise distances between two subjects based on demographic and lifestyle variables. The distances (weights) among subjects are considered in the quantile normalization of miRNA.

This paper is arranged in the following manner: first, we provide an overview of normalization methods developed for miRNA data and explain our modified normalization method in detail. Then we show the application of normalization techniques using simulation and real data sets. Finally, we conclude with a discussion of the important findings presented in this paper.

Methods

Normalization methods for miRNA data

Normalization (along with background correction and summarization [14]) is one of the important steps of preprocessing of miRNA data. The procedure removes the systematic differences between arrays that do not represent true biological variation. Meyer et al. [15] state that

“normalization, often an underestimated aspect of data processing, can minimize systematic technical or experimental variation and thus has significant impact on the detection of differentially expressed miRNAs.” Bolstad et al. [9] highlight that the need for normalization arises naturally when multiple arrays are involved in experiments. There are two types of variations that can be expected: the first variation is an “interesting” variation, which represents biological differences between the expression levels of genes (or miRNAs) from tumor and normal tissues (for example, as in our CRC study), and the other is an “obscuring” technical variation, which is not interesting for the researchers and needs to be removed by a normalization procedure. Currently used normalization methods in miRNA data analysis were primarily developed for mRNA arrays, which have an exceedingly high density of probes with 10,000–50,000 genes. In comparison, miRNAs are lower density arrays with a few hundred to a couple of thousand genes. Wang and Xi [1] mentioned that the majority of miRNAs are either not expressed or are expressed at very low levels. Therefore, researchers have generally concluded that off-the-shelf normalization methods for mRNA arrays may not be appropriate for miRNA arrays [11, 13, 15].

Several studies have compared the performance of normalization methods for mRNA data to see how these methods can reduce the experimentally induced variation and maintain true biological changes between arrays in miRNA data analysis. Rao et al. [11] applied commonly used normalization methods, including cyclic loess, quantile, median or mean, and no normalization techniques to normalize miRNA expression arrays. Their analyses show that the quantile normalization method works better than other normalization techniques in removing differences across arrays in miRNA expression data. Pradervand et al. [13] also examined the impacts of mRNA array normalization procedures based on scaling, quantile, and variance stabilizing normalization (VSN) on miRNA data. They found that the quantile normalization was the most robust procedure and performed at least as well as the developed normalization method based on the set of invariants (invariants-based) among the mRNA normalization techniques that they considered (including quantile, invariants-based, scaling, VSN, and no normalization methods) over all experimental conditions tested. All normalization methods performed better than no normalization. For this reason, we chose not to perform the no normalization technique along with other normalization methods in our analysis. Over the last decade, the quantile normalization method has been commonly used in miRNA data analysis compared to the other normalization techniques developed for mRNA data [10, 12, 14]. We explain the algorithm of quantile normalization in the “Conventional quantile normalization” section.

A number of modifications to normalization methods for mRNA data have been performed to adapt to the characteristics of miRNA data. Though each modified normalization method has been shown to perform well based on the characteristics of tested miRNA data [16–20], a universal normalization method for miRNA data has not been developed yet. Meyer et al. [15] strongly suggest selecting the optimal normalization method based on the characteristics of the data set, and then examining the normalized data carefully in specific biological contexts. The choice of normalization method is one of the primary factors that affects the inference of differential expression [18].

Quantile normalization: conventional and modified to incorporate covariates

Conventional quantile normalization

Bolstad et al. [9] were among the first to apply quantile normalization to microarray data analysis. They compared its performance with the cyclic loess and contrast based normalization methods that had already been successfully used in mRNA data analysis. The purpose of the quantile normalization is to force the distribution of probe intensities for each array in a set of arrays to have the same or at least similar distribution. Bolstad et al. [9] were motivated by the idea that a quantile-quantile plot demonstrates the same distribution for two data vectors if the plot is a straight diagonal line, and not the same distribution if the plot is other than a diagonal line. They extended this concept to N dimensions of data vectors so that all data vectors have the same distribution. The quantiles of N data vectors (here, the intensities of N arrays) are plotted in such a way that (after normalization) the plot gives a straight line along the line given by the unit vector $(\frac{1}{\sqrt{N}}, \dots, \frac{1}{\sqrt{N}})$. To achieve this normalization, one can make the distribution of a set of data vectors the same if one projects the points of the N dimensional quantile onto the diagonal unit vector. More details about the projection of the quantiles onto the diagonal are provided in Bolstad et al. [9]. The numerical dependence induced by this normalization method is relatively minimal [21].

Bolstad et al. [9] provided the following algorithm to perform a quantile normalization: arrange the logarithmic transformed microarray data into a $G \times N$ matrix \tilde{X} , where G and N are total numbers of genes and arrays, respectively. Sort each column of \tilde{X} to give \tilde{X}_{sort} . Then take the means across the rows of \tilde{X}_{sort} and assign this mean to each element in the row to get \tilde{X}'_{sort} . At the end, obtain the normalized version \tilde{X}_{norm} of \tilde{X} by rearranging each column of \tilde{X}'_{sort} to have the same ordering as in the original \tilde{X} .

Another algorithm to carry out the quantile normalization was introduced by Amaratunga and Cabrera [22].

They described the algorithm as follows: “calculate the percentiles (Q_{i0}, \dots, Q_{i100}) of the i th array and the percentiles (Q_{M0}, \dots, Q_{M100}) of the median mock array. For any value X_{gi} , find the interval, $[Q_{ih}, Q_{i(h+1)}]$, to which it belongs and obtain its normalized value, X'_{gi} , by linearly interpolating between the pair points (Q_{Mh}, Q_{ih}) and ($Q_{M(h+1)}, Q_{i(h+1)}$)” [22]. In this algorithm, X_{gi} means the logarithmic transformed spot intensity measurement for gene g on array i . They define the median mock array as the array fashioned out of the medians of the arrays being normalized.

While the algorithm from Bolstad et al. [9] is more widely applied in practice than the algorithm of Amaratunga and Cabrera [22], no study has yet been published that evaluates their relative performance. In our analysis, we used the `normalize.quantiles` function from the R package `preprocessCore` [23] that is based upon the concept of the quantile normalization from Bolstad et al. [9]. We further refer to the quantile normalization based on the algorithm from Bolstad et al. [9] as the conventional quantile normalization.

Weighted quantile normalization

The conventional quantile normalization does not account for additional characteristics of samples when it normalizes the miRNA arrays. In this respect, we propose a novel approach that removes the non-biological differences across samples while incorporating the demographic and lifestyle characteristics of sample-subjects in normalization. Here, we assume that the normalized distribution of miRNA expression levels from one sample should depend on the weighted distribution from other samples. The weights of subjects are determined from a distance matrix, which is aggregated [2] from various normalized distance matrices (values are between 0 and 1 [24]; based on Euclidean, Manhattan, Minkowski, and other methods) of multiple covariates. The elements of the aggregated distance matrix (\tilde{D}) represent the pairwise distances between two subjects based on relevant demographic and lifestyle variables.

The algorithm of our proposed method is based on the quantile normalization algorithm from Bolstad et al. [9], but accounts for the weighted distance metrics of demographic and lifestyle variables. The algorithm is as follows:

1. Obtain the logarithmic (log₂-based) transformed miRNA data as a $G \times N$ matrix \tilde{X} , where G is the total number of miRNAs and N is the total number of subjects. The log-transformation of miRNA data is performed to reduce the effect of outliers on the calculation of miRNA expression levels.
2. Sort each column-subject of \tilde{X} to give \tilde{X}_{sort} .

- Obtain an aggregated distance matrix as a $N \times N$ matrix \underline{D} .
- Obtain the between-subject weight matrix as a $N \times N$ matrix \underline{W} , with elements w_{ij} . We calculate the weight of subject i in the normalized expression distribution of subject j as in (1):

$$w_{ij} = 1 - \frac{d_{ij}}{\max\{d_{i1}, \dots, d_{iN}\}} \quad (1)$$

where d_{ij} (an element of \underline{D}) is the distance between subjects i and j with $i, j = 1, \dots, N$. If $i = j$, then $w_{ij} = 1$. Note that the matrix \underline{W} is not symmetric and the elements of row i correspond to the weights of the N subjects in the normalized expression distribution of subject i .

- Normalize the between-subject weight matrix \underline{W} by the sum of the weights of row-subjects and obtain a $N \times N$ matrix \underline{W}^* . For example, the weighted means of weights of subjects in the normalized expression distribution of subject i can be found as in (2):

$$w_{ij}^* = \frac{w_{ij}}{\sum_{j'=1}^N w_{ij'}} \quad \text{s.t.} \quad \sum_{j=1}^N w_{ij}^* = 1. \quad (2)$$

- Calculate the weighted means across the rows of X_{sort} and assign this weighted mean to each element in the row to get a $G \times N$ matrix X'_{sort} . The weighted means of X_{sort} can be computed as in (3):

$$X'_{\text{sort}} = X_{\text{sort}} \left[\underline{W}^* \right]^T. \quad (3)$$

- Obtain the normalized version X_{norm} of X by rearranging each column of X'_{sort} to have the same ordering as in the original X .

The above mentioned algorithm generalizes the quantile algorithm of Bolstad et al. [9], in which all $w_{ij}^* = \frac{1}{N}$. While the conventional quantile normalization method equally weights all subjects, this weighted quantile normalization method instead weights subjects according to their similarity to each other. That is, the weighted-normalized expression values for a given subject are affected more by the expression values of similar subjects than by those of unsimilar subjects. This weighted quantile normalization algorithm is implemented (with a demonstration using simulated data) in code written for the R language [25], and is provided as Additional file 1 (see “Additional files” section).

In contrast to the normalization methods for miRNA data where disjoint clusters of miRNAs were considered

(Mestdagh et al. [26], Bargaje et al. [27]), this normalization technique will consider both clustered and overlapped subjects by accounting for each subject’s weight in the average. The clustered subjects will have heavier weights in the average than the weights of subjects who are not clustered. This normalization technique will contribute to reduction of intra- and inter-array technical variability while maintaining biological differences. We subsequently refer to this proposed method as the weighted quantile normalization.

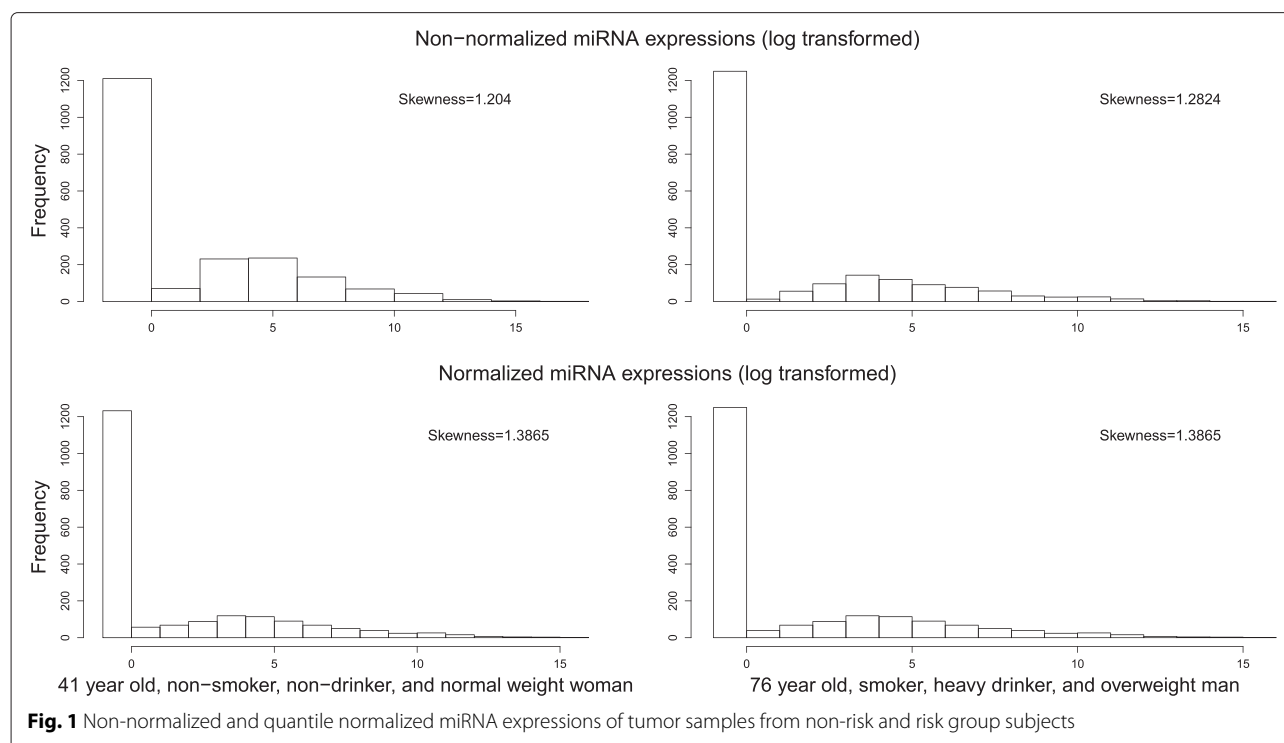
Results and discussion

In this section, we demonstrate the performance of the weighted quantile normalization method over the conventional quantile normalization method using different simulation scenarios.

Motivating example

Figure 1 illustrates the potential danger of normalizing without regard for relevant demographic or environmental factors. The miRNA distributions for two CRC subjects are displayed. For each individual (one from the “non-risk” and one from the “at-risk” sample), both the non-normalized and conventionally quantile-normalized distributions are plotted. The first (non-risk) subject is a 41-year-old, non-smoking, non-drinking, normal-weight woman, while the second (at-risk) is a 76-year-old, current-smoking, heavy drinking, overweight man. The non-normalized miRNA expression histograms, at the top of Fig. 1, show clear disparities between the two subjects. The distribution of the risk-group patient is clearly more variable and right-skewed in comparison to the miRNA distribution of the non-risk patient. (Skewness is quantified here in terms of the Fisher-Pearson coefficient of skewness, and reported in Fig. 1.) However, after performing the quantile normalization of miRNA expressions, both subjects have almost indistinguishable distributions. This suggests that their similarity is merely an artifact of the normalization, and that important differences in gene expression could be masked by ignoring demographic and environmental factors. Our repeated observation of such examples has motivated our development of the novel weighted quantile normalization method (see the “Weighted quantile normalization” section) that properly removes any technical variations, while preserving important biological information with regard to expression differences, and further allowing us to account for additional covariates.

We evaluate the performance of our proposed weighted quantile normalization method using simulated data in sections “Simulation data sets” through “Normalization accounting for unrelated covariates”. We return to the motivating example in the “Application to motivating example (real CRC data)” section.



Simulation data sets

The normalization analyses were carried out on bimodally distributed paired data matrices of $G = 2000$ miRNA expression features (rows) for each of the normal and tumor samples with sample sizes of $N = 200$ and 400 subjects (columns). We simulated expression levels of miRNAs for normal and tumor samples by controlling true differentially expressed miRNAs of tumor samples across all simulations. The simulated bimodal miRNA data sets were generated by the mixture of two normal distributions, and reflect the two modes (for non-expressed and expressed features) generally seen in miRNA data. Particularly, all miRNA features of normal samples and only non-differentially expressed miRNA features of tumor samples were simulated based on $\mu = 0.75$ and $\sigma = 0.025$ for the first distribution and $\mu = 4.0$ and $\sigma = 0.5$ for the second distribution, while the differentially expressed miRNA features of tumor samples, which consisted of 20 % of all miRNA features of tumor samples, were simulated based on $\mu = 0.75$ and $\sigma = 0.025$ for the first distribution and $\mu = 3.25$ or $\mu = 4.75$ and $\sigma = 0.5$ for the second distribution. These parameters were chosen based on observed characteristics of our CRC study data. We performed 20 simulations for each sample size.

Moreover, we simulated demographic and lifestyle variables of subjects in such a way that they could reflect the characteristics of our CRC study and also carry some useful information for the normalization procedure. In our analysis, we simulated all available 19 noncollinear

demographic and lifestyle variables (as listed in Tables 1 and 2) from the CRC study. Briefly, the covariates' values were simulated to be associated with the expression values of several randomly-selected miRNAs that were controlled as truly differentially expressed between tumor and normal; for details, see Suyundikov [2] and Suyundikov et al. [3].

We started the simulation analyses by generating miRNA expressions of normal and tumor samples and demographic and lifestyle variables based on the simulation parameters mentioned above. During simulation of miRNA expressions, we controlled arbitrarily 20 % of miRNAs from tumor samples as differentially expressed features. Euclidean distance was used for continuous covariates and Manhattan distance for discrete or binary covariates. The two Euclidean and Manhattan

Table 1 Summaries of continuous covariates in real CRC data

Covariate	Mean	SD
Age at diagnosis or selection (years)	64.1	9.8
Average num. cigarettes per day	12.5	14.7
Calories (kcal)	2504.7	1199.3
BMI	27.6	5.4
lutein + zeaxantin (mcg)	3119.3	2542.3
Vitamin D (mcg)	6.7	5.0
Lycopene (mcg)	8850.5	8195.1

Table 2 Summaries of binary or discrete covariates in real CRC data

Covariate	Summary
Gender	54 % male, 46 % female
Recent aspirin/NSAID use	64 % no, 36 % yes
Recent smoker	83 % no, 17 % yes
(among women) menopause	12 % pre, 88 % post
(among post-menopausal women)	
taking HRT within 2 years	30 % yes, 70 % no
Data collection center	79 % Kaiser, 21 % Utah
Race	81.6 % White, 8.5 % Hispanic, 7.6 % Black, 2.1 % other
Smoking status	13 % current, 45 % former, 42 % never
Long-term alcohol consumption	38 % none, 35 % moderate, 27 % high
SEER summary stage	1 % in situ, 34 % localized, 52 % regional, 12 % distant, 1 % unknown
AJCC severity stage	1 % 0 (in situ), 26 % 1, 31 % 2, 30 % 3, 12 % 4 (distant)
Colon or rectal cancer	76 % colon, 24 % rectal

between-subject distance matrices were normalized by scaling between 0 and 1 [24] and aggregated into a single between-subject distance matrix by taking their weighted average [2].

As an aside, the application of the weighted quantile normalization method is not computationally burdensome; it took less than one minute to normalize the expressions of 2000 miRNAs from 400 subjects on a machine with CPU speed of 1.86 GHz and 2 GB RAM.

Differential expression testing

We carried out the normalization methods mentioned in the “Methods” section. We conducted the differential expression analyses on tumor-normal differences (using a per-miRNA Wilcoxon signed rank test, or SRT) on the normalized data sets to check whether the weighted quantile normalization method has an equal statistical power in finding differentially expressed miRNA as the conventional quantile normalization. First, we obtained the Wilcoxon SRT statistic and *p*-value for each miRNA feature in each normalized data set and controlled the false discovery rate (FDR) at 0.05 within each simulation. Then, we calculated the true positive rate (TPR) and the false discovery rate (FDR) based on the miRNAs which were controlled as truly differentially expressed in the simulations. The TPR was defined and calculated as in Bolstad [28] and Stevens et al. [29], and the FDR was defined as in Benjamini and Hochberg [30].

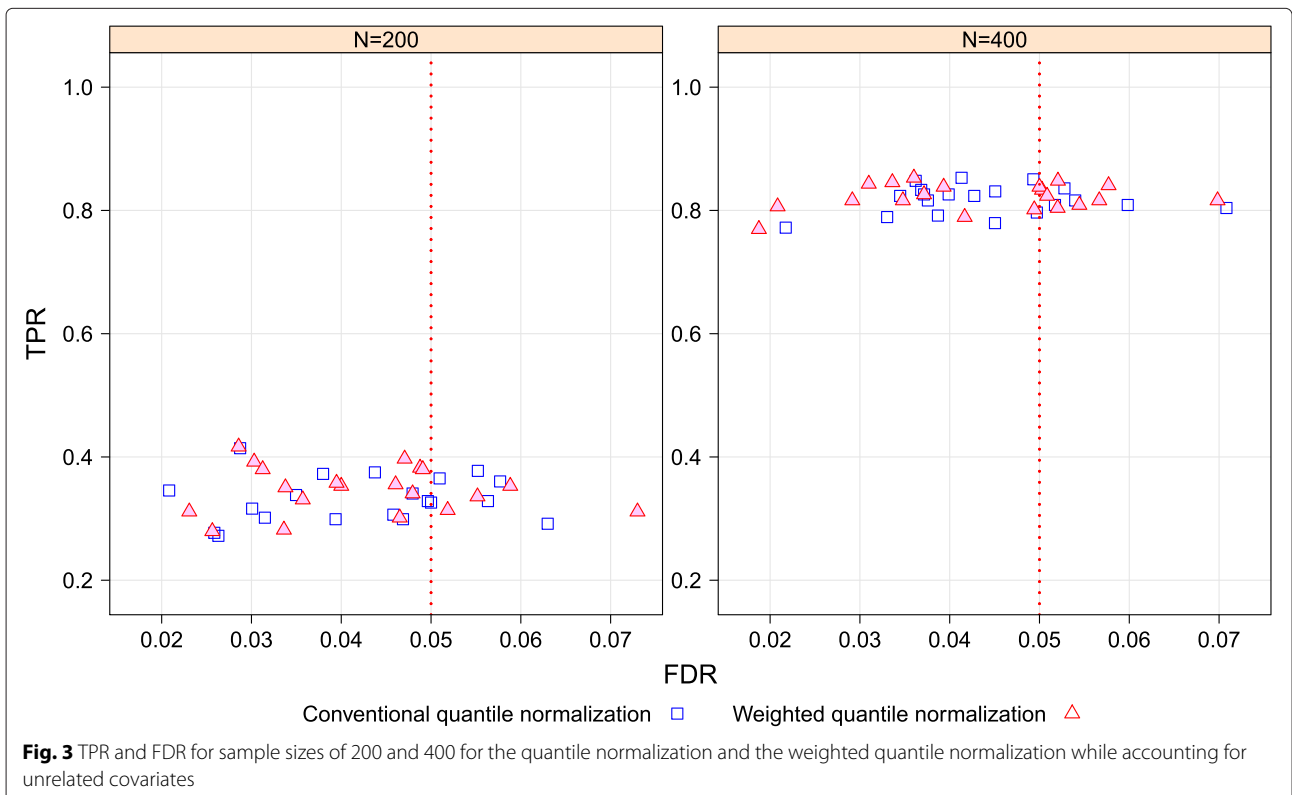
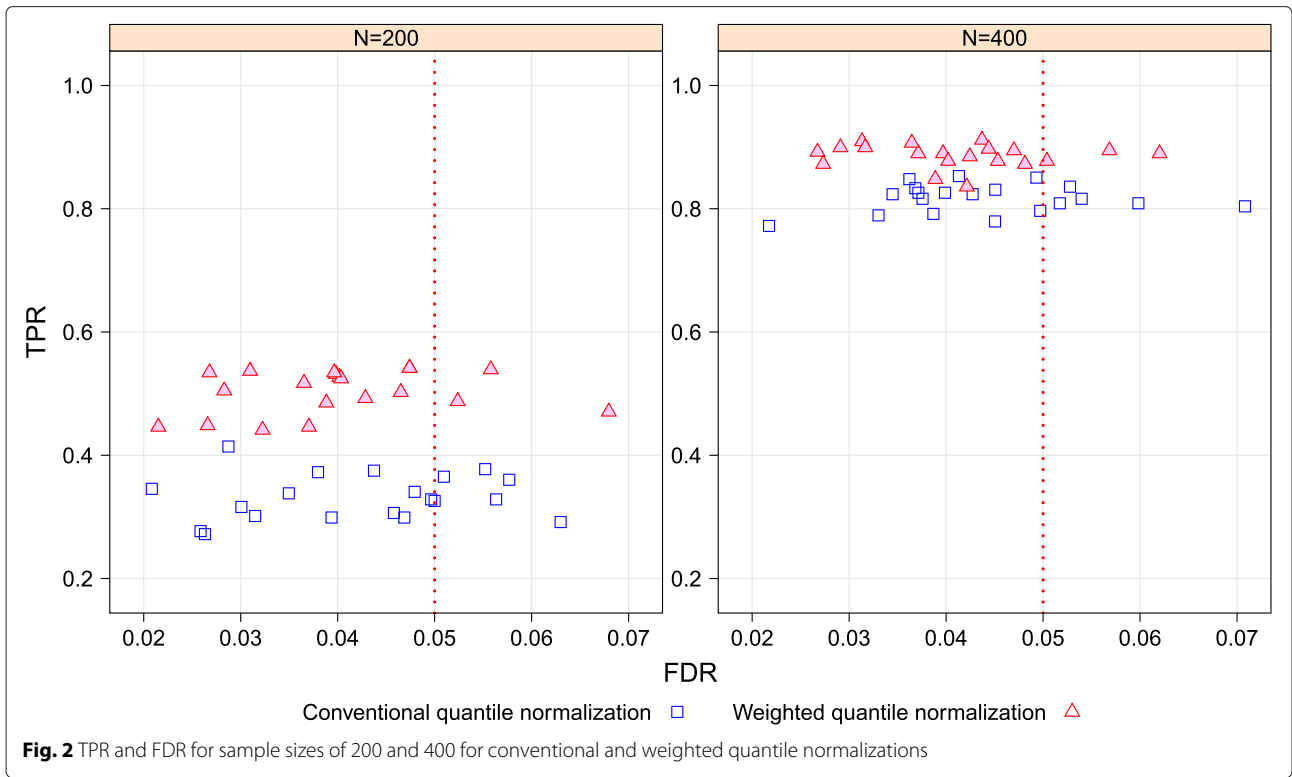
Figure 2 shows the performance (including power and FDR control) of the Wilcoxon SRT on the data sets normalized by the conventional quantile and the weighted quantile methods for the numbers of subjects of 200 and 400. As shown in this scatter plot, the power (i.e., the TPR values) increases (as would be expected) for the conventional quantile (a blue open rectangular symbol) and the weighted quantile (a red solid triangular symbol) normalization methods with larger sample sizes. The weighted quantile normalization has clearly higher power than the conventional quantile method. For 400 subjects, which is more similar to the size of our CRC study, the differences of TPR values between the two methods are up to 10 %: the power for the weighted quantile normalization is in the range of 83.6–90.5 %, while the power for conventional one is in 77.2–85.3 %. That is, one can increase power as much as 10 % by using the weighted quantile normalization method rather than the conventional quantile method. Both normalization methods generally control the FDR near 0.05 for both sample sizes. Results were generally similar when Manhattan distance was used for all covariates (see Additional file 2), rather than Euclidean for continuous covariates and Manhattan for discrete (as in Fig. 2).

Normalization accounting for unrelated covariates

The application of the weighted quantile normalization method only increases power (compared to conventional quantile normalization) when demographic and lifestyle variables (on which between-subject distance is based) are relevant to the “treatment” group comparison of interest. To demonstrate this, we again simulated the 19 covariates, but this time only as noise, without any reference to the miRNA data as in Suyundikov et al. [3]. Figure 3 shows the performance (the TPR versus the FDR) of the Wilcoxon SRT on the simulated data sets that are normalized by the conventional quantile and the weighted quantile methods while accounting for these unrelated (pure noise) covariates. The power and the FDR control are essentially the same (overlap in most simulations) for both normalization methods. Thus, the weighted quantile normalization performs at least as well as the conventional quantile normalization when demographic and lifestyle variables are not associated with the treatment group. Results were generally similar when Manhattan distance was used for all covariates (see Additional file 3), rather than Euclidean for continuous covariates and Manhattan for discrete (as in Fig. 3).

Application to motivating example (real CRC data)

We used the Wilcoxon SRT to identify differentially expressed miRNAs in the paired tumor-normal miRNA data from our CRC study [3]. The miRNA data were normalized by the conventional quantile and by the weighted



quantile methods while accounting for the demographic and lifestyle characteristics of CRC subjects. These data sets contain the first available 527 subjects with 2006 miRNA on each sample. In this analysis, we used all available 19 noncollinear demographic and lifestyle variables, as summarized in Tables 1 and 2.

Figure 4 shows a scatter plot of the FDR adjusted p -values in logarithmic scale. The green circles (there are 121) in the plot represent the miRNAs that are found significant from the quantile normalized data, but not found significant from the weighted quantile normalized data. The red circles (there are 119) show the miRNAs that are found significant only in the weighted quantile normalized data. There is no information about the truly differentially expressed miRNAs that could be helpful to evaluate the performances of both normalization methods. However, we can see from Fig. 4 that many miRNAs (in the lower right triangle of the plot) that are found significant in the quantile normalized data are found to be even more significant in the data set normalized by the weighted quantile method. The plot shows that the proposed weighted quantile normalization method has more power than the conventional quantile method.

By way of validation, our results (from the use of the weighted quantile normalization method) can be placed in the context of recent CRC miRNA studies by comparing the resulting list of candidate miRNAs with those published by Dong et al. [31] and Mazeh et al. [32].

Specifically, we focus on those miRNAs reported in Table 2 of [31] as being prognostic or predictive markers for CRC, and those miRNAs reported in Table 1 of [32] as dysregulated (by at least four studies) in colorectal tissue samples. A total of 41 miRNAs were thus considered.

Due to changes over time in miRNA naming schemes and platform capabilities, any direct comparison of results is inherently uncertain. For example, what was previously reported as miR-203 (up-regulated in tumor [32]) could appear in our CRC study as miR-203a (up-regulated in tumor), miR-203b-3p (no significant difference in tumor), or miR-203b-5p (no significant difference in tumor). Similarly, what was previously reported as miR-195 (down-regulated in tumor [32]) could appear in our CRC study as miR-195-3p (up-regulated in tumor) or miR-195-5p (down-regulated in tumor). With this in mind, and assuming that what previous studies reported for a general-named miRNA was actually what we found for more specific-named miRNA(s), we can report that of those 41 miRNAs previously reported [31, 32] as differentially expressed in tumor compared to normal colorectal tissue, we reach the same conclusion (for both statistical significance and direction of dysregulation) for 38 miRNAs. (This 93 % validation rate is summarized in Additional file 4.) Of the three remaining discrepancies, one (miR-1 down-regulated in tumor [32]) we found with marginal significance (FDR-adjusted p -value .08), and the other two involved possible naming scheme discrepancies.

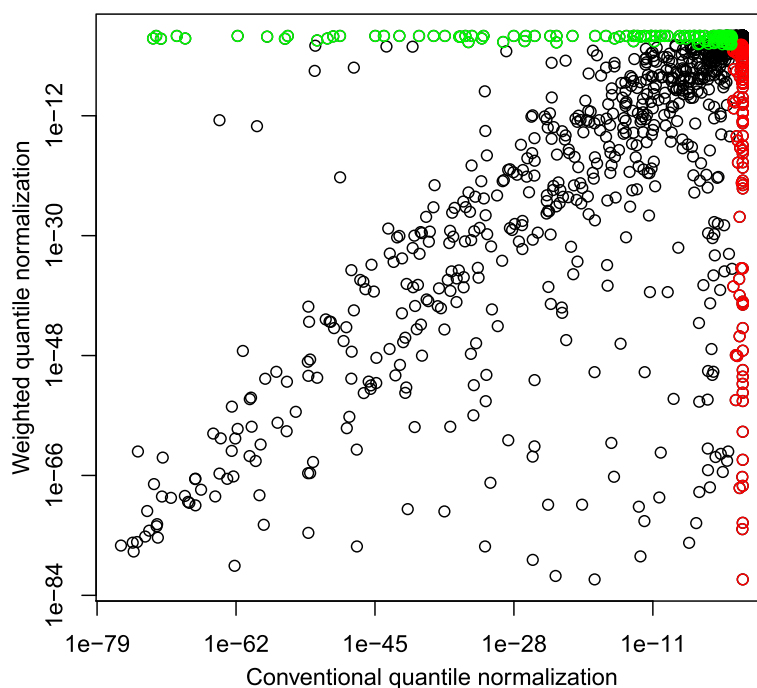


Fig. 4 Scatter plot of adjusted p -values of the CRC miRNA data, normalized by the quantile and the weighted quantile normalization methods (in log-scale). The green and red circles represent the miRNAs that are found significant only in the horizontal and vertical analyses, respectively

Specifically, miR-106a was previously reported as up-regulated in tumor [31, 32], and our CRC study found no significant difference for miR-106a-3p (FDR-adjusted p -value 1), and did not measure miR-106a-5p (which perhaps was the miRNA actually studied previously). Also, miR-30a-3p was previously reported as down-regulated in tumor [32]; while we found no significant difference in miR-30a-3p (FDR-adjusted p -value 1), we did find miR-30a-5p significantly down-regulated (FDR-adjusted p -value < 0.0001). In short, the weighted quantile normalization method allowed the rediscovery of nearly all of the dysregulated miRNAs previously reported by [31] and [32].

In the 38 dysregulated miRNAs rediscovered here, the weighted quantile normalization results tended to have lower p -values than those from the conventional quantile normalization (Additional file 4), illustrating the greater power alluded to in Fig. 4.

Conclusion

In this paper, we modified the quantile normalization method from Bolstad et al. [9] to reduce not only the artificial variations across samples, but also the variations caused by the characteristics of data, as well as to maintain true biological differences across samples. The proposed normalization method incorporates demographic and lifestyle variables by considering the distances (weights) among subjects based on their characteristics, and considers these weights in equating the quantiles of distributions of miRNA expressions. Accounting for such subject-level covariates in the normalization step helps avoid the masking of important expression differences, which can otherwise occur with conventional quantile normalization (Fig. 1).

We performed differential expression tests on simulated bimodally distributed miRNA expressions and demographic and lifestyle variables that reflect the characteristics of our CRC study. We applied the Wilcoxon SRT to identify differentially expressed miRNAs from the normalized simulated miRNA data. The scatter plots of TPR versus FDR help to evaluate the impact of normalization techniques on the results of the differential expression test (Fig. 2). The simulation study provides strong evidence that our weighted quantile approach yields up to a 10 % gain in power in comparison to the conventional normalization method. Both methods generally control the FDR near 0.05 for both sample sizes, and the weighted quantile normalization method is computationally convenient.

When irrelevant subject-level covariates are used in weighted quantile normalization, power and FDR control are essentially the same as when conventional quantile normalization is used (Fig. 3). This suggests that whenever subject-level covariates are available, weighted quantile normalization should be used because it is at least as good

as conventional quantile normalization (in terms of power and FDR control), but substantially better in the presence of relevant demographic or environmental factors.

For the differential expression tests on the paired tumor-normal miRNA data from our CRC study, the Wilcoxon SRT found many miRNAs which were called significant in the conventional quantile normalized data, even more significant in the data set normalized by the weighted quantile method (Fig. 4).

Additional files

Additional file 1: Provides an implementation of the weighted quantile normalization algorithm (as describe in the “Weighted quantile normalization” section) written for the R language [25], with a demonstration using simulated data. (R 2 kb)

Additional file 2: Reproduces Fig. 2, but with Manhattan distance used for all covariates (rather than Euclidean for continuous covariates and Manhattan for discrete, as in Fig. 2). (PDF 6 kb)

Additional file 3: Reproduces Fig. 3, but with Manhattan distance used for all covariates (rather than Euclidean for continuous covariates and Manhattan for discrete, as in Fig. 3). (PDF 5 kb)

Additional file 4: Summarizes the reproduction of significance results from previous CRC miRNA literature [31, 32] as discussed in the “Application to motivating example (real CRC data)” section, using the weighted quantile normalization method. (XLSX 16 kb)

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JH, RKW, and MLS designed the experiments. JRS and AS devised the novel statistical method in consultation with CC, JH, RKW, and MLS. AS wrote the code and conducted the statistical analysis. AS and JRS wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Erica Wolff and Michael Hoffman for miRNA assessment, Sandie Edwards, Courtney Maxfield, and Lila Mullany for tissue collection, Dr. Wade Samowitz for pathology review, and Brett Milash for miRNA bioinformatics assessment. We also thank the Division of Research Computing at USU for providing technical resources to perform numerous study simulations. This research was supported by an NIH grant, award number 1R01CA163683-01A1.

Author details

¹Department of Mathematics and Statistics, Utah State University, 3900 Old Main Hill, Logan, UT, 84322-3900 U.S.A. ²BioStat Solutions, Inc., 5280 Corporate Drive, Suite C200, Frederick, MD, 21703 U.S.A. ³Division of Epidemiology, Department of Internal Medicine, University of Utah School of Medicine, 383 Colorow Road, Salt Lake City, UT, 84108 U.S.A.

Received: 14 July 2015 Accepted: 6 November 2015

Published online: 09 December 2015

References

1. Wang B, Xi Y. Challenges for microRNA microarray data analysis. *Microarrays*. 2013;2(2):34–50.
2. Suyundikov A. Statistical dependence in imputed high-dimensional data for a colorectal cancer study. PhD thesis, Utah State University. 2015. <http://digitalcommons.usu.edu/etd/4371>.
3. Suyundikov A, Stevens JR, Corcoran C, Herrick J, Wolff RK, Slattery ML. Accounting for dependence induced by weighted KNN imputation in paired samples, motivated by a colorectal cancer study. *PLOS ONE*. 2015;10(4):0119876.

4. U.S. Preventive Services Task Force Colorectal Cancer: Screening. Final Recommendation Statement(n.d). <http://www.uspreventiveservicestaskforce.org/Page/Document/RecommendationStatementFinal/colorectal-cancer-screening>.
5. Hawk ET, Levin B. Colorectal cancer prevention. *J Clin Oncol*. 2005;23(2):378–91.
6. Weber MF, Banks E, Ward R, Sitas F. Population characteristics related to colorectal cancer testing in New South Wales, Australia: results from the 45 and up study cohort. *J Med Screen*. 2008;15(3):137–42.
7. Cunningham D, Atkin MW, Lenz HJ, Lynch HT, Minsky B, Nordlinger B, et al. Colorectal cancer. *The Lancet*. 2010;375(9719):1030–1047.
8. Watson AJ, Collins PD. Colon cancer: a civilization disorder. *Dig Dis*. 2011;29(2):222–8.
9. Bolstad BM, Irizarry RA, Åstrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinforma*. 2003;19(2):185–93.
10. Laurent LC, Chen J, Ulitsky I, Mueller FJ, Lu C, Shamir R, et al. Comprehensive microRNA profiling reveals a unique human embryonic stem cell signature dominated by a single seed sequence. *Stem Cells*. 2008;26(6):1506–1516.
11. Rao Y, Lee Y, Jarjoura D, Ruppert AS, Liu C-g, Hsu JC, et al. A comparison of normalization techniques for microRNA microarray data. *Stat Appl Genet Mol Biol*. 2008;7(1):22.
12. Sengupta S, den Boon JA, Chen IH, Newton MA, Stanhope SA, Cheng YJ, et al. MicroRNA 29c is down-regulated in nasopharyngeal carcinomas, up-regulating mRNAs encoding extracellular matrix proteins. *Proc Natl Acad Sci*. 2008;105(15):5874–878.
13. Pradervand S, Weber J, Thomas J, Bueno M, Wirapati P, Lefort K, et al. Impact of normalization on miRNA microarray expression profiling. *RNA*. 2009;15(3):493–501.
14. Lopez-Romero P, Gonzalez MA, Callejas S, Dopazo A, Irizarry RA. Processing of Agilent microRNA array data. *BMC Research Notes*. 2010;3:18.
15. Meyer SU, Pfaffl MW, Ulbrich SE. Normalization strategies for microRNA profiling experiments: a 'normal' way to a hidden layer of complexity? *Biotechnol. Lett*. 2010;32(12):1777–1788.
16. Deo A, Carlsson J, Lindlöf A. How to choose a normalization strategy for miRNA quantitative real-time (qPCR) arrays. *J Bioinforma Comput Biol*. 2011;9(06):795–812.
17. Suo C, Salim A, Chia KS, Pawitan Y, Calza S. Modified least-variant set normalization for miRNA microarray. *RNA*. 2010;16(12):2293–303.
18. Garmire LX, Subramaniam S. Evaluation of normalization methods in mammalian microRNA-Seq data. *RNA*. 2012;18(6):1279–1288.
19. Zhao Y, Wang E, Liu H, Rotunno M, Koshiol J, Marincola FM, et al. Evaluation of normalization methods for two-channel microRNA microarrays. *J Transl Med*. 2010;8(1):69.
20. Kang J, Xu EY. An integrated hierarchical Bayesian approach to normalizing left-censored microRNA microarray data. *BMC Genomics*. 2013;14(1):507.
21. Stevens JR, Nicholas G. Assessing numerical dependence in gene expression summaries with the Jackknife Expression Difference. *PLOS ONE*. 2012;7(8):.
22. Amaratunga D, Cabrera J, Vol. 446. *Exploration and Analysis of DNA Microarray and Protein Array Data*; 2004.
23. Bolstad BM. preprocessCore: a collection of pre-processing functions. R package version 1.32.0. 2015. <http://bioconductor.org/packages/release/bioc/html/preprocessCore.html>.
24. Visalakshi NK, Thangavel K. Impact of normalization in distributed k-means clustering. *Int J Soft Comput*. 2009;4(4):168–72.
25. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2015. <http://www.R-project.org/>.
26. Mestdagh P, Van Vlierberghe P, De Weer A, Muth D, Westermann F, Speleman F, et al. A novel and universal method for microRNA RT-qPCR data normalization. *Genome Biol*. 2009;10(6):64.
27. Bargaje R, Hariharan M, Scaria V, Pillai B. Consensus miRNA expression profiles derived from interplatform normalization of microarray data. *RNA*. 2010;16(1):16–25.
28. Bolstad BM. Low-level analysis of high-density oligonucleotide array data: background, normalization and summarization. Berkeley: PhD thesis University of California; 2004.
29. Stevens JR, Bell JL, Aston KI, White KL. A comparison of probe-level and probeset models for small-sample gene expression data. *BMC Bioinforma*. 2010;11(1):281.
30. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol*. 1995;57(1):289–300.
31. Dong Y, Wu WKK, Wu CW, Sung JY, Ng SSM. MicroRNA dysregulation in colorectal cancer: a clinical perspective. *Br J Cancer*. 2011;104(6):893–8.
32. Mazeh H, Mizrahi I, Ilyayev N, Halle D, Brucher BLD, Bilchik A, et al. The diagnostic and prognostic role of microRNA in colorectal cancer—a comprehensive review. *J Cancer*. 2013;4(3):281–95.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

