Statistical Applications in Genetics and Molecular Biology

Volume 7, *Issue* 1 2008 *Article* 22

A Comparison of Normalization Techniques for MicroRNA Microarray Data

Youlan Rao[*] Yoonkyung Lee[†] David Jarjoura[‡]
Amy S. Ruppert[**] Chang-gong Liu[††]
Jason C. Hsu[‡‡] John P. Hagan[§]

[*]The Ohio State University, rao@stat.ohio-state.edu
[†]The Ohio State University, yklee@stat.ohio-state.edu
[‡]The Ohio State University, david.jarjoura@osumc.edu
[**]The Ohio State University, amy.ruppert@osumc.edu
[††]The Ohio State University, chang-gong.liu@osumc.edu
[‡‡]The Ohio State University, jch@stat.ohio-state.edu
[§]The Ohio State University, microrna@gmail.com

A Comparison of Normalization Techniques for MicroRNA Microarray Data*

Youlan Rao, Yoonkyung Lee, David Jarjoura, Amy S. Ruppert, Chang-gong Liu, Jason C. Hsu, and John P. Hagan

Abstract

Normalization of expression levels applied to microarray data can help in reducing measurement error. Different methods, including cyclic loess, quantile normalization and median or mean normalization, have been utilized to normalize microarray data. Although there is considerable literature regarding normalization techniques for mRNA microarray data, there are no publications comparing normalization techniques for microRNA (miRNA) microarray data, which are subject to similar sources of measurement error. In this paper, we compare the performance of cyclic loess, quantile normalization, median normalization and no normalization for a single-color microRNA microarray dataset. We show that the quantile normalization method works best in reducing differences in miRNA expression values for replicate tissue samples. By showing that the total mean squared error are lowest across almost all 36 investigated tissue samples, we are assured that the bias correction provided by quantile normalization is not outweighed by additional error variance that can arise from a more complex normalization method. Furthermore, we show that quantile normalization does not achieve these results by compression of scale.

KEYWORDS: microRNA, median normalization, cyclic loess normalization, quantile normalization, robust estimates, smoothing spline, mean squared error

---

*This material is based in part upon work supported by the National Science Foundation under Agreement No. 0635561. Jason C. Hsu's research is supported by NSF Grant Number DMS-0505519

1 Introduction

In microarray experiments, variation of expression measurements among arrays can be attributed to many sources, such as differences in sample RNA preparation, cDNA labeling, image intensity and microarray hybridization/wash efficiency. Normalization of expression levels applied to microarray data can help in removing this error. Different methods, including cyclic loess, quantile normalization (Bolstad et al. 2003) and median or mean normalization (Churchill 2002, Churchill 2003, Churchill and Oliver 2001, Kerr and Churchill 2001, and Wolfinger et al. 2001), have been utilized to normalize microarray data. Briefly, cyclic loess makes the MA plot of probe intensities from every pair of arrays scatter about the $M = 0$ axis, quantile normalization makes the distributions of expression levels the same across arrays, and median or mean normalization shifts the individual log-intensities on each array so that the median or mean log-intensities, respectively, are the same across arrays. These normalization algorithms can be applied either globally to an entire data set or locally to some physical subset of the data (Quackenbush 2002). Irizarry et al. (2003) applied the quantile normalization procedure to normalize dilution data and spike-in data from Affymetrix arrays, and showed how quantile normalization removed bias as compared to no normalization. Their analysis was unique in that they knew the true expression levels and could therefore determine the degree of bias reduction from quantile normalization.

MicroRNAs (miRNAs) are noncoding RNAs of 19-24 nucleotides that are negative regulators of gene expression. Recently implicated as important in development and normal physiology, microRNAs are abnormally expressed in many human cancers (Volinia et al. 2006, Lu et al. 2005). Moreover, aberrant microRNA expression has been shown to initiate and promote carcinogenesis (reviewed in Hagan and Croce 2007). These microRNA expression signatures may reveal new oncogenetic pathways in human cancers. For systematic investigation of microRNA expression, oligonucleotide-based microarrays for microRNAs in human and mouse tissues have been developed recently (Liu et al. 2004) and several commercial platforms are now available. To date, more than a hundred published reports have used microRNA microarrays to investigate their expression profiles, where more than two-thirds have used single color versus two color hybridization systems. Although there is substantial literature regarding normalization techniques for mRNA microarray data, there are no published reports comparing normalization techniques for microRNA (miRNA) microarray data, which are subject to the similar sources of error variation.

Many statistical reports on mRNA microarrays have focused on Affymetrix

mRNA arrays, which have an exceedingly high density of probes that are *in situ* synthesized on the array. For example, in one Human Genome U133 Plus2.0 GeneChip, probe sets for each mRNA, including numerous housekeeping genes, consist of eleven oligonucleotide probes selected to maximize specificity and to have similar melting temperatures across the entire array. In contrast, microRNA microarrays are often lower density spotted arrays. Our focus is on single color microRNA microarray. This type of microarray is used predominantly in comparison to dual color arrays. Results from the Version 3.0 microRNA microarray used in this study and its earlier versions have appeared in more than 40 publications. The Version 3.0 microarray contains 3790 probes spotted in duplicate. The probes are 40 nucleotides in length, consisting of the genomic sequence that has the mature microRNA sequence and additional flanking bases. With the exception of six probes designed against *Arabidopsis thaliana* microRNAs, the rest of the probes are derived from known and predicted human and mouse microRNAs. This design allows for the detection of mature as well as precursor miRNAs and is particularly helpful in determining if computationally predicted miRNAs are real. Although U6 snRNA is frequently used as a control for microRNA experiments, this noncoding RNA has been shown to vary as much as five fold for equivalent amounts of total RNA by both microarray and Northern analysis (Hagan and Liu, unpublished observations). Hence, probes for U6 snRNA were not included in the Version 3.0 microarray. Most, if not all, commercially available microRNA microarrays do not have controls for endogenous RNAs that have been shown to be largely invariant between tissue samples.

Given the short length of miRNAs and the fact that far more mRNAs are known than miRNAs, it is important to compare normalization methods specifically for the miRNA microarray data. Although microRNA microarrays are lower density spotted arrays than mRNA microarrays, they are not "boutique" arrays. For example, microRNA arrays do not meet the following criteria: "more than half the probes might be differentially expressed between any two samples and that the differential expression might be predominately in one direction" (Oshlack et al. 2007). We also do not expect global differences across miRNA arrays. As an example, the biggest difference in miRNA expressions was expected between brain and heart tissues, we found only 15% of miRNAs were differentially expressed with a greater than 2 fold difference, when comparing these distinct tissue types. Other examples include the referenced miRNA studies in cancer (Calin et al. 2005, Volinia et al. 2006, Yanaihara et al. 2006) and tissue differentiation (Babak et al. 2004, Barad et al. 2004, Garzon et al. 2004) in Davison et al. (2006). For the three referenced cancer studies that used microRNA microarrays, the number of differentially expressed microRNAs are

$13/245$ (5.3%), $22/228-57/228$ ($9.6\%-25.0\%$, range depends on which of six tumor/normal comparisons were performed) and $43/352$ (12.2%). For the three referenced differentiation studies, the number of differentially expressed microRNA are $19/399$ (4.8%), $25/154-35/154$ ($15.2\%-22.7\%$, range depends on the specific pairwise tissue comparison) and $35/150-57/150$ ($23.3\%-38.0\%$, range depends on the specific pairwise tissue comparison). We can conclude with confidence that much less than 50% of miRNAs are differentially expressed based on our experience and assessment of the literature. In addition to our custom microRNA microarrays, there are numerous commercially available miRNA microarrays. For example, LC Sciences, Exiqon, Agilent, Invitrogen, and Ambion sell miRNA microarrays, with 1564, 4000, 15000, 3000, and 1224 miRNA probes, respectively. Hence, the probe density of our array is similar to many currently available commercial platforms. Importantly, high throughput sequencing of microRNAs is rapidly expanding the number of known microRNAs. Hence, our custom arrays soon will need to be updated with even more probes to reflect the recently identified microRNAs. The microRNA registry (Version 10.1) currently has sequences for 5395 miRNAs. Even though microRNA microarrays are not "boutique" arrays in general, a few cases exist where large numbers of microRNAs will be differentially expressed in only one direction. Knockouts of essential microRNA biogenesis proteins such as Drosha, DGCR8, or Dicer1 lead to a dramatic reduction in steady state microRNA levels by blocking production of mature microRNAs (Kumar et al. 2007). These global downregulation cases are exceptionally easy to detect by microarray as the percentage of microRNAs expressed above background is considerably different in comparison to controls. Other confirmed examples that show unidirectional microRNA regulation are quite rare. Using a novel bead-based microRNA profiling system, microRNAs were reported to be downregulated primarily in cancers (129 of 217 investigated). Almost all studies of microRNAs in cancer, including all the research referenced in this manuscript, have found roughly balanced numbers or a slight enrichment for upregulated microRNAs in cancer, casting doubt on the conclusions of Lu et al. (2005). Even research that at first glance might seem to support the conclusions of Lu and colleagues demonstrates unequivocally the opposite. For example, Chang et al. (2008) reported that Myc expression leads to widespread repression of microRNAs. As their Supplemental Table 1 shows for 313 human microRNAs investigated, 11 and 17 microRNAs are upregulated and downregulated, respectively, at least two fold upon induced Myc expression. Although vigilance must be exercised to make sure that the underlying assumptions are valid, the normalization methods that we present are compatible for the vast majority of studies using microRNA microarrays.

In this paper, we compare the performance of median, cyclic loess, quantile, and no normalization for miRNA microarray data. The data included 72 microarrays obtained from RNA from 26 human and 10 mouse tissues that were hybridized as techinal replicates. Hence, each RNA sample was hybridized to two independent microarrays. Since replicate samples should, in theory, have almost identical values for expressions, one can compare different normalization techniques in terms of the closeness of normalized measurements in the replicated samples. Moreover, there are no confounding biological effects that come from tissues from different individuals. The differences between these paired expression levels with and without normalization can be divided into a bias and variance components by expression level. Both of these miRNA-by-miRNA differences components should be reduced after applying normalization methods. We used these differences to provide direct evidence of the capability of each method of reducing these two components. It was critical to examine the effects on both quantities because the complexity of a transformation may increase the error variance over and above its bias reduction. To resemble how normalization is typically applied to samples, normalization was done globally across all 72 samples. This is an important distinction from normalizing each of 36 replicate pairs separately, where this level of normalization could produce artificially low variance and bias.

Section 2 describes the normalization methods in detail. Section 3 describes the miRNA data used in this paper. Section 4 compares normalization methods.

2 Normalization Methods

Three commonly used normalization techniques are reviewed. Suppose that we have the (log base 2 transformed) probe level expression values from p miRNAs and n arrays in a $p \times n$ matrix $\mathbf{X}$.

Median normalization shifts miRNAs expressions on each array by additive constants so that the medians of miRNAs expressions are the same across arrays by the following steps:

- Take the median of each column of $\mathbf{X}$ and generate a n-dimensional median vector M;
- Calculate the overall median of the vector M;
- Shift miRNAs expression values of each array by subtracting the difference between the median of each array and the overall median from them.

Instead of matching the median only across the arrays, **Quantile normalization** makes the distributions of expression levels the same across arrays by the following steps:

- Sort each column of $\mathbf{X}$ separately to generate a sorted $p \times n$ matrix $\mathbf{Y}$;
- Take the mean of each row of $\mathbf{Y}$ and generate a p-dimensional vector A_b, called the baseline array;
- Get the normalized miRNAs expressions for each array by rearranging the baseline array A_b to have the same ordering of the corresponding column of the matrix $\mathbf{X}$ so that empirical distributions of miRNA expressions are the same as that of the baseline array across arrays.

Cyclic loess considers the MA plot of probe intensities from every pair of arrays $(X_{ij}, X_{ij'})$, with fixed $j \neq j'$ and $i = 1, 2, ..., p$, and makes the M and A pairs scattered around the $M = 0$ axis by the following steps:

- Compute $M_i = X_{ij} - X_{ij'}$ and $A_i = \frac{1}{2}(X_{ij} + X_{ij'})$;
- Fit a loess curve by regression M on A, and denoted the fitted vector by $\hat{M}$;
- Setting the vector $D = (M - \hat{M})/2$, get the normalized miRNAs expressions for $(X_{ij}, X_{ij'})$ by modifying X_{ij} to $X_{ij} + D_i$ and $X_{ij'}$ to $X_{ij} - D_i$, $i = 1, 2, ..., p$.

3 Description of Data

Total RNA was purchased from Ambion Inc. Microarray labeling and hybridization were performed as previously described in Liu et al. (2004), except for the exceptions noted below. The Ohio State University Comprehensive Cancer Center Version 3.0 microRNA microarray was used and this array contains 3790 oligo probes derived from 578 mature miRNAs spotted in duplicate (329 Homo sapiens, and 249 Mus musculus) that are annotated in the miRNA registry http://microrna.sanger.ac.uk/ sequences/ (Accessed Nov. 2005). Of the 396 evolutionarily conserved mature microRNAs between mice and human in Version 10.1 of the microRNA registry, 68% are identical in length and sequence. Hence, many of the mouse probes serve as additional controls for their human counterparts and vice versa. In addition, 1493 human and 1137 mouse oligo

probes for miRNAs computationally predicted in human and mouse, respectively, are also spotted in duplicate. Often, more than one probe set exists for a given mature miRNA. Additionally, there are duplicate probe spots corresponding to most precursor miRNAs. Hybridization signals were ultimately detected with Streptavidin-Alexa 647, conjugate and scanned images (Axon 4000B) were quantified using the Genepix 6.0 software through a local background correction (Axon Instruments, Sunnyvale, CA).

4 Analysis

Background-corrected median signals for duplicate probes on an array were averaged. After normalization across all 72 arrays, let X_i be the log base 2 transformed expression value of the ith miRNA for a certain tissue, and let Y_i be the log base 2 transformed expression value of the ith miRNA for the replicate of the tissue.

Bias. The average $A_i = (X_i + Y_i)/2$ and the difference $M_i = X_i - Y_i$ of expression values for each miRNA can then be computed. The MA plot of the two vectors X_i and Y_i is a 45-degree rotation and axis scaling of their scatter plot. This plot is particularly useful for array data because M_i represents the log fold change and A_i represents the average log intensity for the ith miRNA. When the loess curves of the MA plot deviate from the horizontal line at $M = 0$, this demonstrates differences in the intensity levels between two arrays from the same tissue (Gentleman et al. 2005). In contrast, if the loess curves align with $M = 0$, the normalization method is considered to exhibit little bias at all levels of expression. When MA plots and loess curves were made for the replicate array data from human brain tissue using no normalization, median normalization, quantile normalization and cyclic loess, we observed that the quantile normalization method removed bias the best (Figure 1C), the loess curve closely followed the horizontal line at $M = 0$. No normalization, median normalization and cyclic loess behaved similarly in that their loess curves are not aligning with $M = 0$ closely enough (Figure 1A, 1B and 1D).

Binning. To compare the normalization methods in how much they reduced error variance in addition to reducing bias, we formally modeled the mean and variance of differences in replicate arrays as a function of their expression levels. In order to obtain reliable estimates of the expression levels, we binned duplicates according to their average expression level first and then proceeded by modeling the mean and variance based on the binned data.

We created equally-sized bins containing 34 miRNAs probes. For each bin,

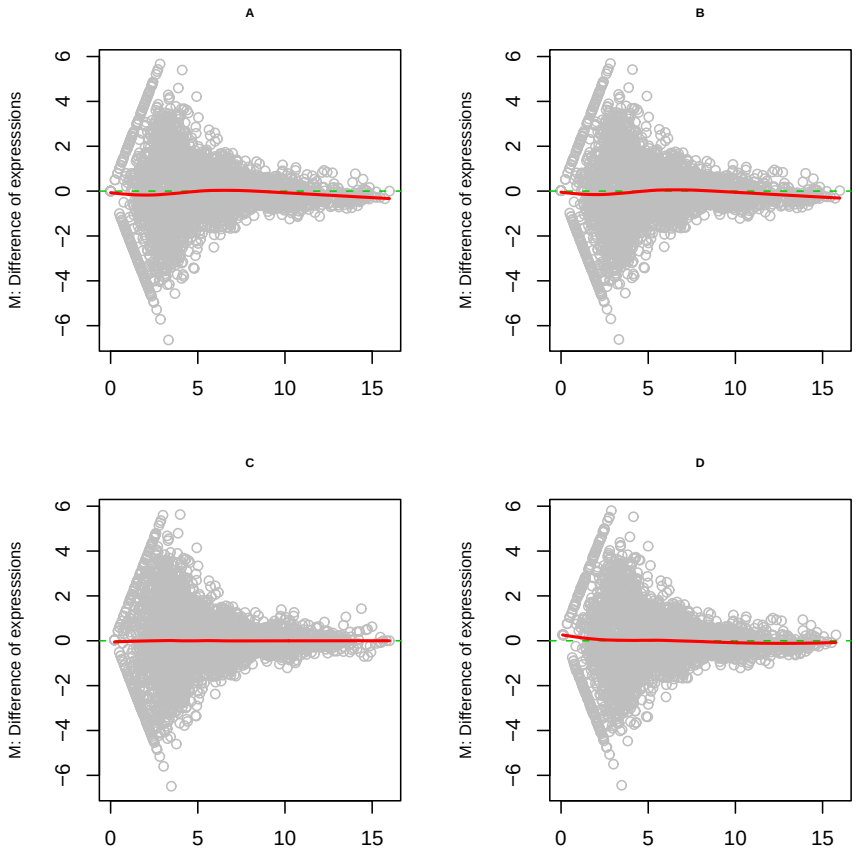

Figure 1: MA and loess plot of expression values for the human brain tissue data. A) without normalization, B) after median normalization, C) after quantile normalization and D) after cyclic loess.

we summarized the differences in the replicate arrays by median absolute deviation (MAD) of the differences and median of the differences to obtain robust estimates of variance and bias, respectively (Lin et al. 2002). The smoothed MADs and medians of the differences were used to detect systematic effects due to the different normalization methods as a function of expression levels. Lower values of smoothed MADs and smoothed medians closer to zero across average expressions correspond to a superior normalization method.

As stated above, each bin consisted of 34 miRNAs probes. For fixed k $(1 \leq k \leq K)$, let $X_{(i)k}$ $(i = 1, 2, ..., 34)$ be the expression value of the ith miRNA in the kth bin for a specific tissue, and let $Y_{(i)k}$ $(i = 1, 2, ..., 34)$ be the

expression value of the ith miRNA in the kth bin for the replicate of the tissue. The difference between the replicate arrays expression values for each miRNA in the kth bin can be denoted by $D_{(i)k} = X_{(i)k} - Y_{(i)k}$ $(i = 1, 2, ..., 34)$, and the corresponding observations by $d_{(i)k}$. We assume that for fixed k,

$$D_{(i)k} \overset{\text{i.i.d.}}{\sim} N(\mu_k, \sigma_k^2) \quad i = 1, 2, ..., 34$$

and use

$$md_k = \underset{1 \le i \le 34}{\text{median}}(d_{(i)k})$$

as a robust location (center) estimate of $\mu_k = \mathrm{E}[D_{(1)k}]$, and

$$MADd_k = \underset{1 \le i \le 34}{\text{median}}|d_{(i)k} - \underset{1 \le i \le 34}{\text{median}}(d_{(i)k})|,$$

as a robust estimate of scale (spread), which is proportional to $\sigma_k = \sqrt{\text{var}[D_{(1)k}]}$ under normality.

For the average expression values of miRNAs in the kth bin across certain tissue replicates, let $A_{(i)k} = (X_{(i)k} + Y_{(i)k})/2$ $(i = 1, 2, ..., 34)$ and $a_{(i)k}$ be the ith observation. Similarly, for estimation of the center of the average expression values in each bin, we consider

$$ma_k = \underset{1 \le i \le 34}{\text{median}}(a_{(i)k}).$$

As Figure 1A suggests, it is sensible to model μ_k and σ_k as a function of the center of the average expression values of miRNA replicates in the kth bin.

For the paired observations $(ma_1, md_1), (ma_2, md_2), ..., (ma_K, md_K)$, we modeled the median difference as a smooth function of the median average

$$md_k = \eta(ma_k) + \epsilon_k, \quad k = 1, 2, ..., K$$

with $\epsilon_k \sim N(0, \sigma_{m,k}^2)$ and with a different variance for each bin. The smoothed relationship η was obtained by the weighted smoothing spline with weights equal to the reciprocal of the squared MAD of difference. Quantile normalization gave the best results when comparing the weighted smoothed curves for the median difference in expression values using the human brain tissue data (Figure 2).

Similarly, for the paired observations $(ma_1, MADd_1)$, $(ma_2, MADd_2)$, ..., $(ma_K, MADd_K)$, we considered the following model with unequal variance

$$MADd_k = \xi(ma_k) + \epsilon_k, \quad k = 1, 2, ..., K$$

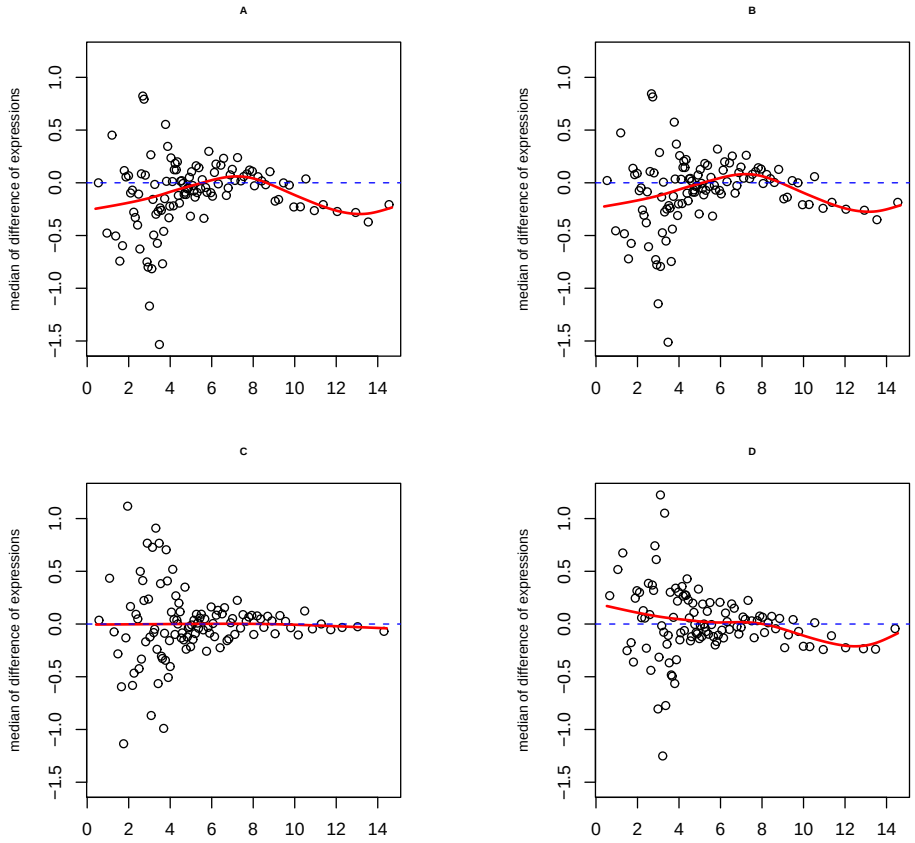

Figure 2: weighted smoothed medians of difference of expression values for the human brain tissue data. A) without normalization, B) after median normalization, C) after quantile normalization and D) after cyclic loess.

and $\epsilon_k \sim N(0, \sigma^2_{MAD})$. The smoothed MAD of differences ξ can again be obtained by smoothing splines with the smoothing parameter selected by generalized maximum likelihood (GML) (Gu 2002). It was difficult to see differences in the relationship between $MADd$ and ma among the normalization methods (Figure 3), but they became more apparent if the bias and variance were combined into a mean-squared error statistic.

Confidence intervals. The fitted medians of differences η is the smoothed estimate of bias parameter μ_k, and the fitted MAD of differences ξ is the smoothed estimate of scale parameter. We used the fitted MAD to estimate confidence intervals around bias and obtained a pointwise confidence interval for the bias by

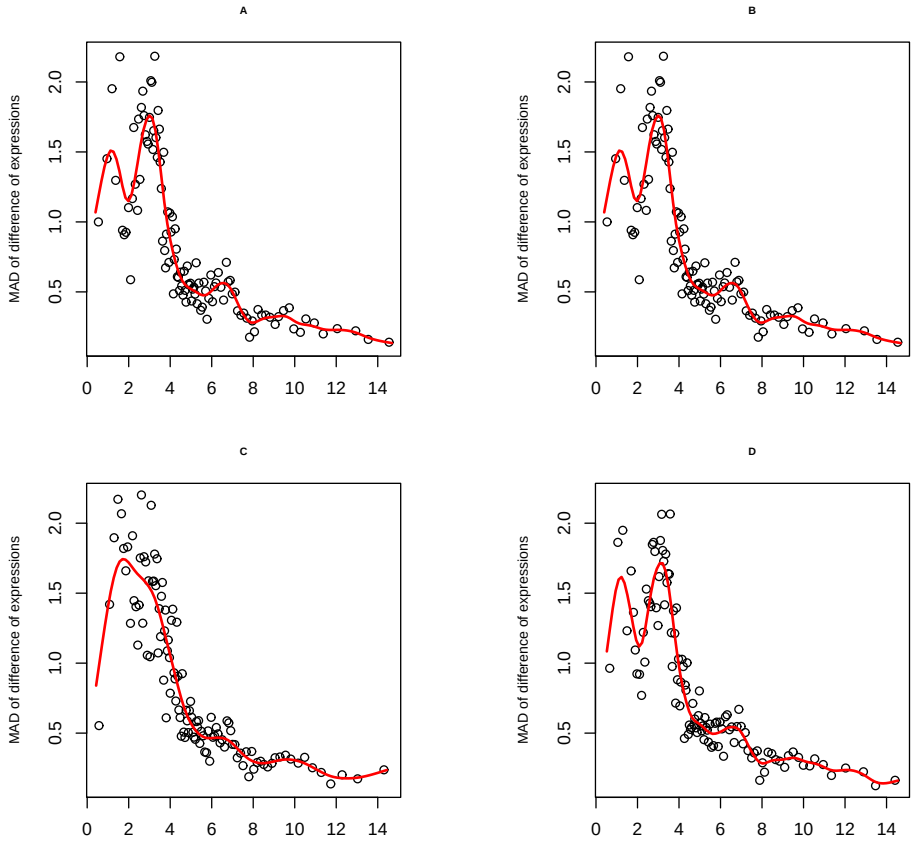

Figure 3: smoothed MADs versus median averages for the human brain tissue data. A) without normalization, B) after median normalization, C) after quantile normalization and D) after cyclic loess.

binned expression values as

$$\hat{\eta}(ma_k) \pm \frac{3.98}{\sqrt{34}}\hat{\xi}(ma_k),$$

(see Hoaglin et al. 2000). The confidence band after quantile normalization encompasses the horizontal line at $M = 0$, while those using no normalization, median normalization or cyclic loess do not include zero for larger expression values (Figure 4).

Mean Squared Error. We obtained the mean squared error (MSE) of the difference in expression values (including variance and squared bias)

$$\text{MSE}_k = \text{E}[D^2_{(1)k}] = \text{var}[D_{(1)k}] + \text{E}[D_{(1)k}]^2 = \sigma^2_k + \mu^2_k,$$

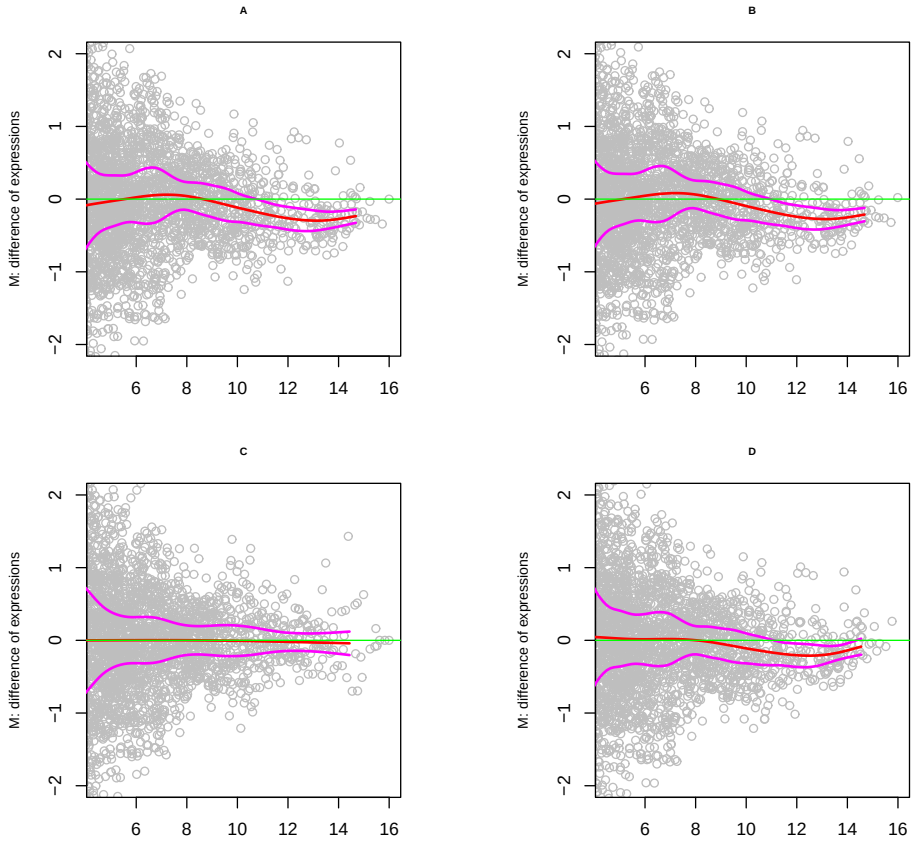

Figure 4: confidence band of the bias for the human brain tissue data. A) without normalization, B) after median normalization, C) after quantile normalization and D) after cyclic loess.

which can be estimated by the smoothed estimates

$$[\frac{\hat{\xi}(ma_k)}{0.6745}]^2 + \hat{\eta}(ma_k)^2,$$

(see Huber 2003). The estimated MSE for quantile normalization is smallest when average expression values are greater than noise levels of measurements, and the estimated MSE for cyclic loess is slightly larger than that of quantile normalization across all average expression values. Median normalization performed similarly to no normalization (Figure 5).

To evaluate the global bias and variance for each method, we averaged MSEs across expression levels greater than 4.5; the value 4.5 (log base 2 transformed)

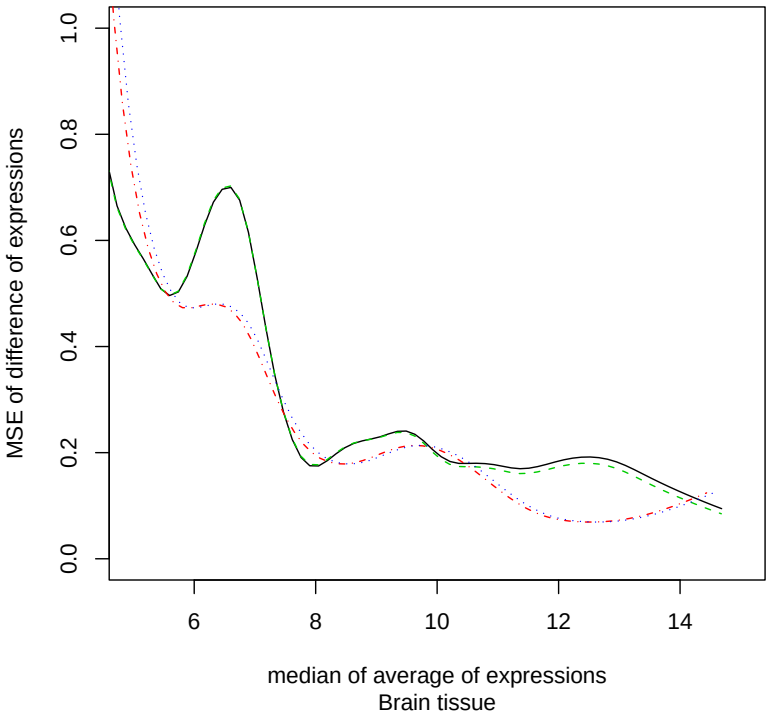

Figure 5: MSE curves without normalization (black, solid line), after median normalization (green, dashed line), and after quantile normalization (red, dot-dashed line) after cyclic loess (blue, dotted line).

was selected because 95% of the blanks (spots lacking oligonucleotide probes) gave intensities less than this value. The average MSEs for no normalization, median normalization, quantile normalization and cyclic loess using the brain tissue data were 0.278, 0.274, 0.225, 0.270 respectively. These results were found consistently across the other 35 tissue types (Figure 6), where the MSEs were lower for quantile normalization (coded 2) in almost all tissue samples compared to no normalization (coded 0), median normalization (coded 1) and cyclic loess (coded 3), except for human lung, human liver, human thymus, mouse liver and mouse lung. When the normalization methods were applied to each tissue type separately, instead of to all 72 arrays together, the results were similar.

Checking for Scale Compression. It is possible that the superior results for

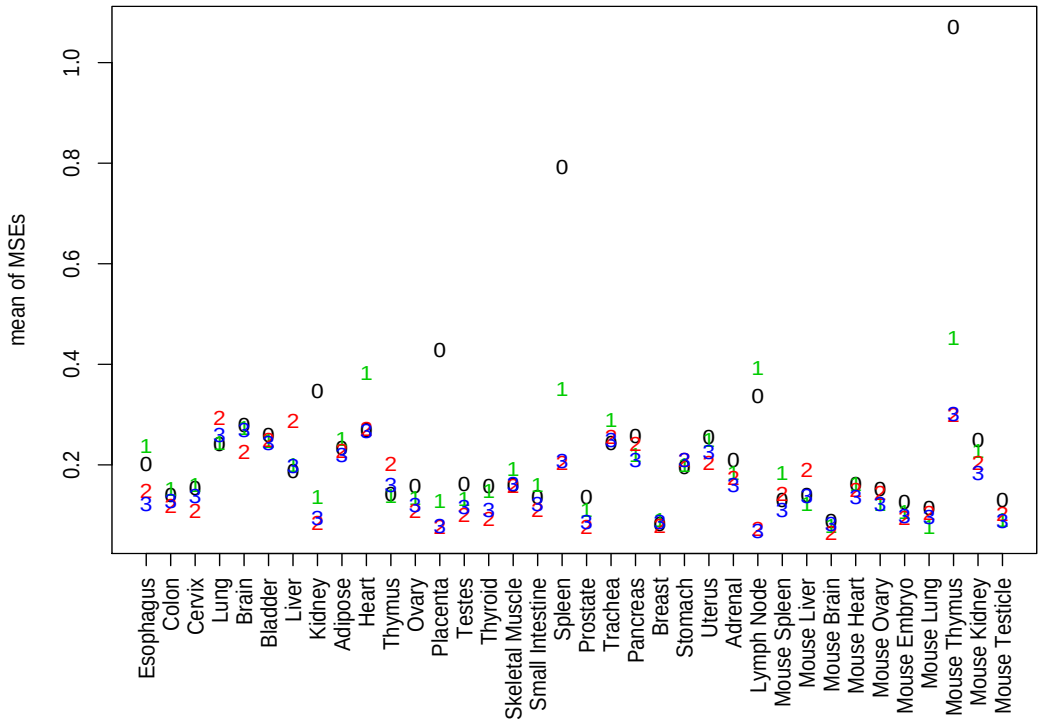

Figure 6: mean of MSEs for the difference in expression values without normalization (0 and black), after median normalization (1 and green), after quantile normalization (2 and red) and after cyclic loess (3 and blue).

quantile normalization is the result of the compression of the scale downward after transformation. To check this, we first calculated coefficients of variation (CV) as the ratio of an estimate of the standard deviation of measurement ($\sqrt{\text{MSE}}$) for each bin to the mean expression for that bin and then average the ratios across bins. We found the CVs followed the same pattern as the MSEs, that is, typically lower values for quantile normalization across tissues (Figure 7). It is also possible that the superior results for quantile normalization is the result of compressing the scale from both ends after transformation; thereby reducing spread and sensitivity of transformed measurements. To check this, we calculated the average variance of expression levels across the 36 tissues for each miRNA. This variance consists of true variance across tissues and measurement

error as obtained with the MSE. Averaging the variance across miRNAs and the MSEs across tissues, we found the ratios of signal (true) variance to noise (measurement error) variance were $12.0, 14.0, 16.3$ and 16.3 for no, median, quantile and cyclic loess normalization respectively.

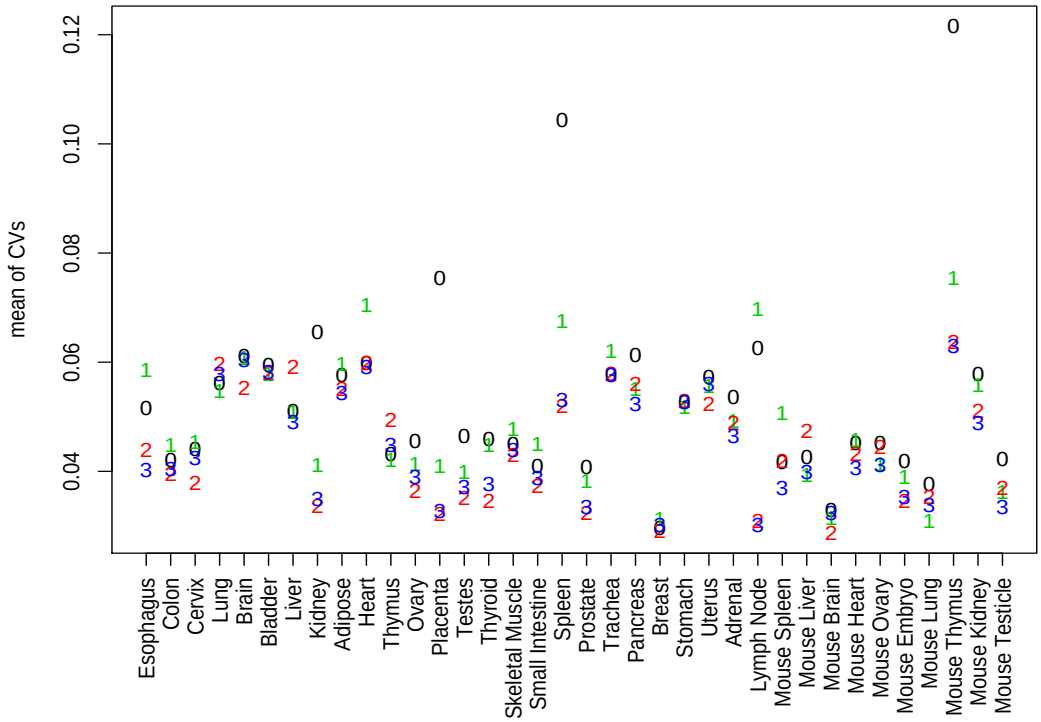

Figure 7: mean of CVs for the difference in expression values without normalization (0 and black), after median normalization (1 and green), after quantile normalization (2 and red) and after cyclic loess (3 and blue).

Comparative Study We compare real-time RT-PCR miRNA data (Lee et al. 2008) with our microarry miRNA data, since twenty-one tissues were common to both datasets. Specifically, we focused on brain and heart, since these tissues are quite biologically distinct and have substantial differences in their miRNA expression profiles. If a normalization technique was overly aggressive, then there would be an ”averaging-out” effect, leading to a significant decrease in the number of differentially expressed miRNAs. A well known difference between

microarray and RT-PCR data is that the fold changes observed by microarray tend to be compressed in comparison with fold changes observed by RT-PCR. We found 51 miRNAs were characterized by a four fold difference in expression by RT-PCR. For the microarray data on identical miRNAs, we found that 36, 35, 35, 35 miRNAs were two fold differentially expressed for no, median, cyclic loess and quantile normalization respectively. This set of miRNAs was found to have roughly an 70% overlap with the RT-PCR data. The observed values for fold changes varied little with respect to the normalization method used. In this respect, we could not conclude any superior normalization method based strictly on this analysis, but we could at least conclude that quantile normalization is not worse than other methods in terms of its sensitivity.

5 Conclusion

We showed that the quantile normalization method works best in reducing differences in miRNA expression values for duplicate tissue samples, cyclic loess works slightly worse than quantile normalization, whereas no normalization and median normalization behave similarly and seem to be inferior to quantile normalization and cyclic loess with regard to bias. This is not surprising because quantile normalization adjusted better for differential bias across the scale of expression values. By showing that the total MSE was lower across almost all 36 tissue samples, we were assured that the bias correction provided by quantile normalization was not outweighed by additional error variance that can arise from a more complex normalization method. Furthermore, we showed that quantile normalization does not achieve smaller replication error by compressing the scale downward or by compressing the scale from both ends.

References

Babak, T., Zhang, W., Morris, Q., Blencowe, B. and Hughes, T. (2004). Probing microRNAs with microarrays: Tissue specificty and functional inference, *RNA* **10**: 1813–1819.

Barad, O., Meiri, E., Avniel, A., Aharonov, R., Barzilai, A., Bentwich, I., Einav, U., Gilad, S., Hurban, P., Karov, Y., Lobenhofer, E. K., Sharon, E., Shiboleth, Y. M., Shtutman, M., Bentwich, Z. and Einat, P. (2004). MicroRNA expression detected by oligonucleotide microarrays: System establishment and expression profiling in human tissues, *Genome Research* **14**: 2486–2494.

Bolstad, B. M., Irizarry, R. A., Astrand, M. and Speed, T. P. (2003). A comparision of normalization methods for high density oligonucleotide array data based on variance and bias, *Bioinformatics* **19**: 185–193.

Calin, G., Ferracin, M., Cimmino, A., DiLeva, G., Shimizu, M., Wojcik, S., Iorio, M., Visone, R., Sever, N., Fabbri, M., Iuliano, R., Palumbo, T., Pichiorri, F., Roldo, C., Garzon, R., Sevignani, C., Rassenti, L., Alder, H., Volinia, S., Liu, C. G., Kipps, T. J., Negrini, M. and Croce, C. M. (2005). A microRNA signature associated with prognosis and progression in chronic lymphocytic leukemia, *The New England Journal of Medicine* **353**: 1793–1801.

Chang, T., Yu, D., Lee, Y., Wentzel, E., Arking, D., West, K., Dang, C. V., Thomas-Tikhonenko, A. and Mendell, J. T. (2008). Widespread microRNA repression by myc contributes to tumorigenesis, *Nature Genetics* **40(1)**: 43–50.

Churchill, G. A. (2002). Fundamentals of experimental design for cdna microarrays, *Nature Genetics* **32**: 490–495.

Churchill, G. A. (2003). Discussion to statistical challenges in functional genomics-comment, *Statistical Science* **18**: 64–69.

Churchill, G. A. and Oliver, B. (2001). Sex, flies and microarrays, *Nature Genetics* **29**: 355–356.

Davison, T., Johnson, C. and Andruss, B. (2006). Analyzing micro-RNA expression using microarrays, *Methods in Enzymology* **411**: 14–34.

Garzon, R., Pichiorri, F., Palumbo, T., Iuliano, R., Cimmino, A., Aqeilan, R., Volinia, S., Bhatt, D., Alder, H., Marcucci, G., Carlin, G., Liu, C. G., Bloomfield, C., Andreeff, M. and Croce, C. (2006). MiRNA fingerprints during human megakaryocytopoiesis, *Proceedings of the National Academy of Sciences of the United States of America* **101**: 5078–5083.

Gentleman, R., Carey, V. J., Huber, W., Irizarry, R. and Dudoit, S. (2005). *Bioinformatics and computational biology solutions using R and bioconductor*, Springer: New York.

Gu, C. (2002). *Smoothing Spline ANOVA Models*, Springer: New York.

Hagan, J. and Croce, C. (2007). MicroRNAs in carcinogenesis, *Cytogenetic and Genome Research* **118**: 252–259.

Hoaglin, D. C., Mosteller, F. and Tukey, J. W. (2000). *Understanding Robust and Exploratory Data Analysis*, John Wiley & Sons.

Huber, P. (2003). *Robust Statistics*, John Wiley & Sons.

Irizarry, R. A., Hobbs, B., Collin, F. and Speed, T. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data., *Biostatistics* **4**: 249–264.

Kerr, M. K. and Churchill, G. (2001). Experimental design for gene expression microarrays., *Biostatistics* **2**: 183–201.

Kumar, M., Lu, J., Mercer, K., Golub, T. and Jacks, T. (2007). Impaired microRNA processing enhances cellular transformation and tumorigenesis, *Nature Genetics* **39(5)**: 673–677.

Lee, E., Baek, M., Gusev, Y., Brackett, D. J., Nuovo, G. and Schmittgen, T. (2008). Systematic evaluation of microRNA processing patterns in tissues, cell lines, and tumors, *RNA* **14**: 35–42.

Lin, Y., Nadler, S. T., Lan, H., Attie, A. D. and Yandell, B. S. (2003). Adaptive gene picking with microarray data: detecting important low abundance signals, *in* G. Parmigiani, E. S. Garrett, R. A. Irizarry and S. L. Zeger (eds), *The Analysis of Gene Expression Data: Methods and Software*, Springer-Verlag.

Liu, C., Calin, G., Meloon, B., Gamliel, N., Sevignani, C., Ferracin, M., Dumitru, C., Shimizu, M., Zupo, S., Dono, M., Alder, H., Bullrich, F., Negrini, M. and Croce, C. (2004). An oligonucleotide microchip for genome-wide microRNA profiling in human and mouse tissues, *Proceedings of the National Academy of Sciences of the United States of America* **101(26)**: 9740–9744.

Lu, J., Getz, G., Miska, E., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B. L., Mark, R., Ferrando, A., R., D. J., Jacks, T., Horvitz, H. R. and Golub, T. R. (2005). MicroRNA expression profiles classify human cancers, *Nature* **435(7043)**: 843–848.

Oshlack, A., Emslie, D., Corcoran, L. and Smyth, G. (2007). Normalization of boutique two-color microarrays with a high proportion of differentially expressed probes, *Genome Biology* **8(1):R2**.

Quackenbush, J. (2002). Microarray data normalization and transformation, *Nature Genetics* **32**: 496–501.

Volinia, S., Calin, G., Liu, C., Ambs, S., Cimmino, A., Petrocca, F., Visone, R., Iorio, M., Roldo, C., Ferracin, M., Prueitt, R., Yanaihara, N., Lanza, G., Scarpa, A., Vecchione, A., Negrini, M., Harris, C. and Croce, C. (2006). A microRNA expression signature of human solid tumors defines cancer gene targets, *Proceedings of the National Academy of Sciences of the United States of America* **103(7)**: 2257–2261.

Wolfinger, R., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushe, P., Afsha, C. and Paules, R. . (2001). Assessing gene significance from cdna microarray expression data via mixed models, *Journal of Computational Biology* **8**: 625–637.

Yanaihara, N., Caplen, N., Bowman, E., Seike, M., Kumamoto, K., Yi, M., Stephens, R., Okamoto, A., Yokota, J., Tanaka, T., Carlin, G., Liu, C. G., Croce, C. and Harris, C. (2006). Unique miRNA molecular profiles in lung cancer diagnosis and prognosis, *Cancer Cell* **9(3)**: 189–198.