Cancer Informatics

ORIGINAL RESEARCH

An Empirical Evaluation of Normalization Methods for MicroRNA Arrays in a Liposarcoma Study

Li-Xuan Qin[1], Tom Tuschl[2] and Samuel Singer[3]

[1]Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY. [2]Laboratory of RNA Molecular Biology, The Rockefeller University, New York, NY. [3]Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, NY. Corresponding author email: qinl@mskcc.org

Abstract

Background: Methods for array normalization, such as median and quantile normalization, were developed for mRNA expression arrays. These methods assume few or symmetric differential expression of genes on the array. However, these assumptions are not necessarily appropriate for microRNA expression arrays because they consist of only a few hundred genes and a reasonable fraction of them are anticipated to have disease relevance.

Methods: We collected microRNA expression profiles for human tissue samples from a liposarcoma study using the Agilent microRNA arrays. For a subset of the samples, we also profiled their microRNA expression using deep sequencing. We empirically evaluated methods for normalization of microRNA arrays using deep sequencing data derived from the same tissue samples as the benchmark.

Results: In this study, we demonstrated array effects in microRNA arrays using data from a liposarcoma study. We found moderately high correlation between Agilent data and sequence data on the same tumors, with the Pearson correlation coefficients ranging from 0.6 to 0.9. Array normalization resulted in some improvement in the accuracy of the differential expression analysis. However, even with normalization, there is still a significant number of false positive and false negative microRNAs, many of which are expressed at moderate to high levels.

Conclusions: Our study demonstrated the need to develop more efficient normalization methods for microRNA arrays to further improve the detection of genes with disease relevance. Until better methods are developed, an existing normalization method such as quantile normalization should be applied when analyzing microRNA array data.

Keywords: microRNA, microarray, normalization, differential expression, cancer, sarcoma

Cancer Informatics 2013:12 83–101

doi: 10.4137/CIN.S11384

This article is available from http://www.la-press.com.

Background

MicroRNAs (miRNAs) are a class of short noncoding RNAs.[1,2] They regulate gene expression posttranscription through base pairing with mRNAs to induce their degradation and translational repression.[1,2] Humans have over 1000 miRNAs, which may target about 60% of protein-coding genes.[3] MiRNAs have been linked to a number of cancer types in several studies of individual miRNAs.[4,5] In recent years, the microarray technology has become available to study miRNAs in a high-throughput fashion, and it has been increasingly used to study the role of miRNAs in diseases such as cancer.[6–12]

Similar to mRNA arrays, miRNA arrays also exhibit array effects (that is, systematic variation due to experimental factors such as array manufacture batch, lab technician, and image scanner). We demonstrate this here with a data example from a liposarcoma study. Liposarcoma, the most common soft tissue sarcoma,[3] is subdivided into three biologic groups, one of which consists of well-differentiated liposarcoma (WD) and dedifferentiated liposarcoma (DD). In a study of miRNA expression in liposarcoma at Memorial Sloan-Kettering Cancer Center (MSKCC), we collected miRNA arrays for 56 tissue samples: 7 normal adipose, 24 WD, and 25 DD[14] using the Agilent miRNA arrays (Agilent Technologies, Santa Clara CA), which have an 8-plex format with eight arrays printed on a glass slide arranged as two rows and four columns. Arrays were generated over a period of 6 months (September 1, 2009 through February 28, 2010). Figure 1 shows a boxplot of the data (unnormalized foreground data on the $\log_2$ scale), with one box per array. Boxes are ordered by array slide and production date and colored by tissue type. It is obvious that array difference due to tissue type is dominated by differences due to array production period and array slide.

Proper normalization is essential in the analysis of array data to remove systematic variation due to experimental process and to make data from different arrays comparable.[15–18] This issue has been extensively studied in the context of mRNA expression arrays, with a number of normalization methods proposed.[19–24] A typical assumption of these normalization methods is that few genes are differentially expressed or that differential expression is symmetric.[16,17,21] These existing normalization methods have been directly applied to miRNA arrays.[25–28] However, miRNA array data have a number of distinct differences from mRNA array data. For example, the number of miRNA genes is relatively small, and differential expression is likely to be common and asymmetric as most miRNAs are expected to be expressed in a very tissue-specific manner.[29–32] It is unclear whether existing normalization

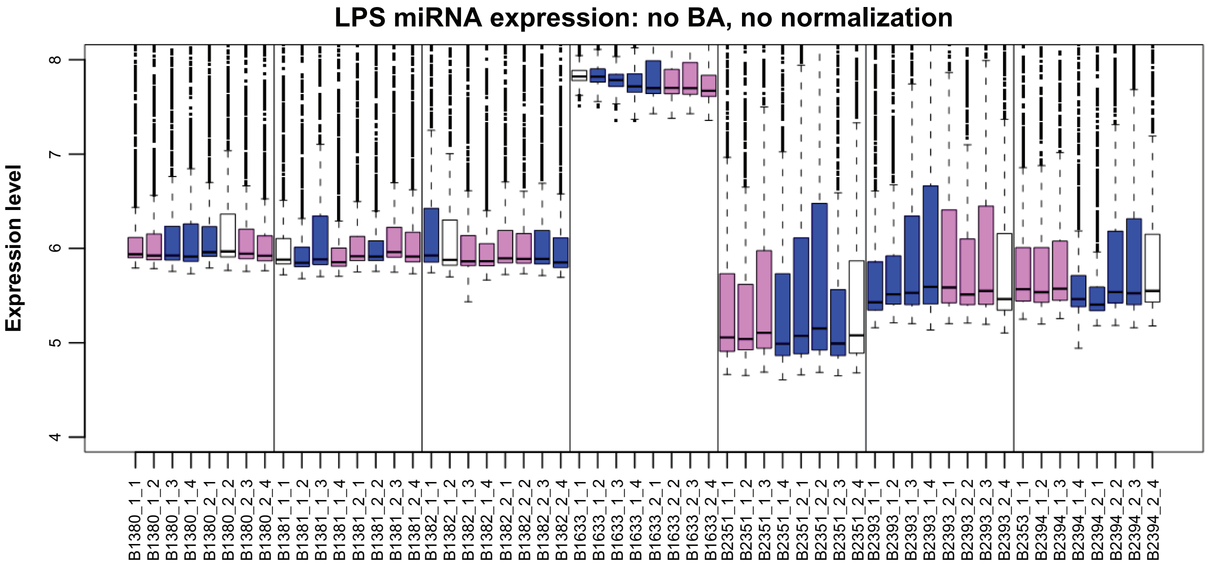

Figure 1. Boxplot of the foreground intensity on the $\log_2$ scale for the Agilent arrays (n = 56) in the liposarcoma study.
Notes: Data at the probe level with no normalization are displayed with one box per array. The arrays are ordered by array slide, with vertical lines between slides. Colors indicate the sample type (white for normal adipose tissue, violet for WD, and blue for DD).
Abbreviations: LPS, liposarcoma; BA, background adjustment.

methods, developed for mRNA arrays, are appropriate for miRNA arrays.[33,34]

Three previous reports have studied the performance of normalization methods for miRNA arrays. Rao et al[35] compared several normalization methods using an in-house array platform, with duplicate arrays for 26 human tissues and 10 mouse tissues. They used the similarity between duplicate arrays as a performance metric. Lopez-Romero et al[36] examined the performance of median and quantile normalization for Agilent arrays and used the variability between biological replicates to measure normalization performance. Pradervand et al[33] assessed normalization methods using Agilent arrays, defining true positives as a set of 59 genes that were claimed as differentially expressed by all normalization methods under study. These previous studies are important but limited by their choice of performance measure and also their definition of "true" positives.

We set out to assess normalization methods for miRNA arrays when the performance measure is to detect differentially expressed genes (a common goal in array studies) and the true positive genes are determined by an independent experimental methodology using the RNA samples derived from the same set of tissue samples. We used a set of Agilent arrays for 56 tissue samples collected at MSKCC. A subset of these samples was also profiled by Solexa sequencing of small RNA libraries at the Tuschl lab of the Rockefeller University. The small RNA libraries were prepared using a customized protocol that has been carefully developed and extensively calibrated at the Tuschl lab.[37–40] We used the sequence data on the same set of samples as the benchmark to empirically evaluate the effect of normalization on the Agilent array data and compare the performance of normalization methods.

Materials and Methods

Data collection

Our study included 56 human tissue samples. Data on these samples were collected at MSKCC and the Rockefeller University as part of a study of miRNA expression alterations in liposarcomagenesis. Details of data collection, such as sample acquisition, RNA isolation, and Solexa sequencing, were described in Ugras et al.[37,39–42] Expression arrays were processed at the MSKCC Genomic Core Lab using the Agilent Human miRNA arrays (version V2.0). Agilent array data were available for all 56 samples (7 normal adipose, 24 WD, and 25 DD). Solexa deep sequencing data were available for 28 of the samples (14 WD and 14 DD) using the same RNA samples as the Agilent arrays. Agilent arrays measure 799 miRNAs, and Solexa has reads on 597 miRNAs. These two platforms have 486 miRNAs in common, which we focused on in our analysis.

Statistical analysis of Solexa data

We looked for miRNAs differentially expressed between WD and DD by comparing the relative abundance (that is, clone count for a given miRNA divided by the total clone count in a tissue sample), using a moderated t test as implemented in the R package LIMMA.[43] Differentially expressed miRNAs were then selected using a P value cutoff $P < 0.0001$; we expected < 1 miRNA to have such small a P value by chance. In order to descriptively look at the distribution of Agilent data and how array normalization affects the data distribution, we selected three groups of miRNAs based on their abundance: (1) miRNAs ranked among the top 20 most abundant in each sample for all 28 samples, which we call " always-on" miRNAs; (2) miRNAs that had zero count in all 28 samples, which we call "always-off" miRNAs; and (3) a random subset of 20 other miRNAs, which we call "sometimes-on" miRNAs.

Statistical analysis of Agilent data and evaluation of normalization methods

We evaluated differential expression based on the Agilent data using the moderated t test implemented in the LIMMA package.[43] The P values, fold changes, and the selected top miRNAs were then compared with those from the Solexa data, which were considered to be the benchmark. For each individual array, we assessed the agreement between the Agilent and Solexa data among all miRNAs and the agreement among 'always-on' and 'sometimes-on' miRNAs using Pearson correlation.

To assess the effect of array normalization, we repeated the aforementioned analysis of Agilent data, but with normalization. We tested four normalization methods using their implementations in the R software: (1) median normalization (http://www.affymetrix.com) using R package, affy; (2) quantile normalization[20,21] using R package, processCore; (3) cyclic loess normalization[24] using R package, affy; and (4) variance stabilizing normalization[19] using R package, vsn.

The data preprocessing pipeline was as follows: (1) the probe-level foreground intensity data were $\log_2$ transformed, (2) the $\log_2$ probe-level foreground intensity data were normalized, and (3) the normalized probe-level data were converted to gene-level data by taking the average of probes corresponding to the same miRNA gene.

An exception was made for variance stabilizing normalization, where the data were first normalized and then $\log_2$ transformed.

Three of the normalization methods—quantile, cyclic loess, and variance stabilizing—depend on the entire set of samples being normalized together. In the analysis results we report here, we normalized the data for the 56 samples and then performed differential expression analysis in the 28 samples that have sequence data available. Similar analysis results were found when normalization was applied only to the 28 samples.

Results

Analysis of Solexa data

Comparing between WD (n = 14) and DD (n = 14), 25 miRNAs were differentially expressed at a *P* value cutoff of 0.0001 (Fig. 2). We refer to these 25 miRNAs as DE25 hereafter. Among these 25 miRNAs, 2 are upregulated, and 23 are downregulated in DD (a more aggressive subtype than WD). The two upregulated miRNAs are miR-9 and miR-21: miR-9 has been reported to positively regulate the malignant progression of cancer;[44] miR-21 has been shown to target a number of tumor suppressors such as PTEN and BCL2.[45,46] The 23 downregulated miRNAs include miR-143, miR-190, and miR-652, which have been shown to be relevant to cancer.[47–51] We are in the process of characterizing their roles in liposarcoma and have recently published our results from functional studies of miR-143.[41]

Seven miRNAs were always-on (Fig. 3A). They are miR-21, miR-22, miR-26a, let-7a, let-7b, let-7f, and let-7i. While the miRNAs that are on (that is, ranked as top 20 abundant in a given sample) in each individual sample accounted for about 75% of the clone count (Fig. 3B), the seven always-on miRNAs shared across the samples accounted for about 50% of the clone count (Fig. 3C). Twenty-eight miRNAs were always-off.

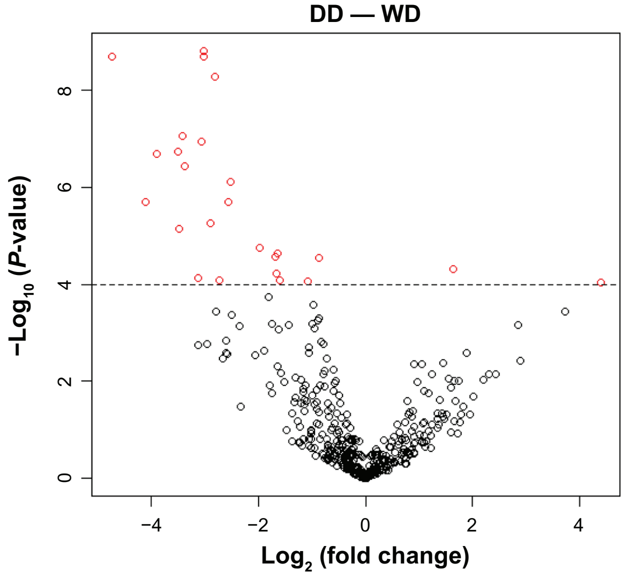

Figure 2. Volcano plot for the comparison of relative abundance (% clone count) between WD and DD.
Notes: The X axis is the $\log_2$ (fold change); the Y axis is the $-\log_{10}$ (*P* value). MiRNAs that are significant ($P < 0.0001$) are indicated in red.

Analysis of Agilent data with no normalization

A scatter plot of the Solexa data ($\log_2$ clone count) versus the Agilent data for each array is provided in Supplementary Figure 1. As expected, the Solexa data showed a wider dynamic range than the Agilent data and had better resolution for miRNAs expressed at low levels. The two data types had moderate agreement across all miRNAs, with the Pearson correlation ranging between 0.58 and 0.85 (Fig. 4). When limiting the comparison to the always-on and sometimes-on miRNAs, the Pearson correlation between the two data types increased, ranging from 0.79 to 0.95 (Supplementary Fig. 2). Distributions of the always-on, always-off, and sometimes-on miRNAs on each array are shown in Supplementary Figure 3. Based on the Agilent data with no normalization, 7 miRNAs were differentially expressed ($P < 0.0001$), which included 4 of the DE25 miRNAs (Table 1).

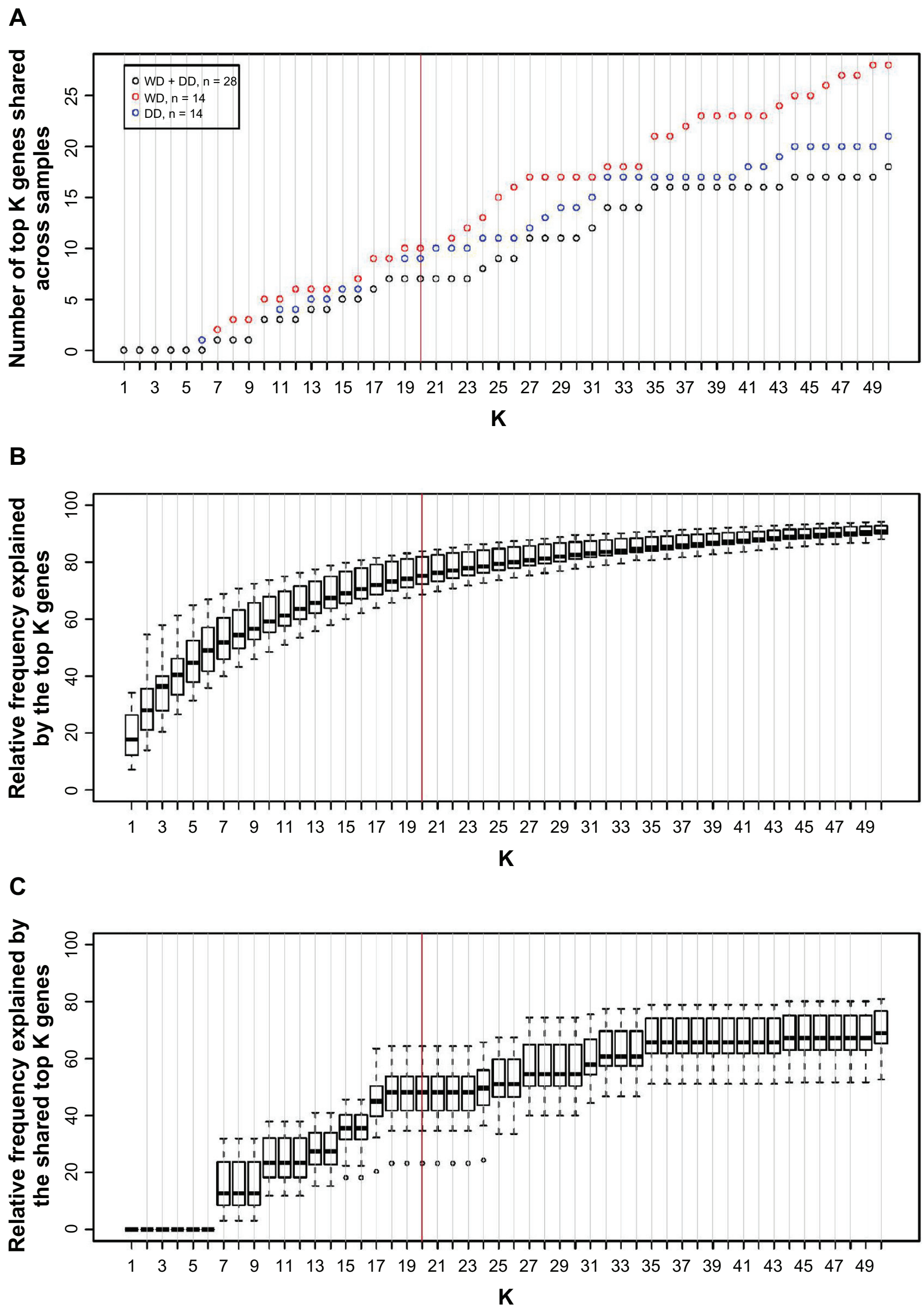

Figure 3. (**A**) The number of the top K abundant miRNAs shared among all liposarcoma samples (black), shared among all WD (red), or shared among all DD (blue). The data are derived from the Solexa sequencing. (**B**) Relative abundance explained by the top K abundant miRNAs in each sample. Its distribution among 28 liposarcoma samples is shown as a boxplot for each K. (**C**) Relative abundance explained by the top K abundant miRNAs shared among all 28 liposarcoma samples. Its distribution among 28 liposarcoma samples is shown as a boxplot for each K.

Analysis of Agilent data with normalization

Results of the Agilent data after normalization are provided in Table 1 and Figure 4.

- When median normalization was applied, 6 miRNAs were differentially expressed, including 5 of the DE25 miRNAs.
- When quantile normalization was applied, 16 miRNAs were differentially expressed, including 11 of the DE25 miRNAs. The Pearson correlation between the two data types ranged between 0.61 and 0.85 among all miRNAs (Fig. 4), and from 0.79 to 0.95 among the always-on and sometimes-on miRNAs (Supplementary Fig. 2).

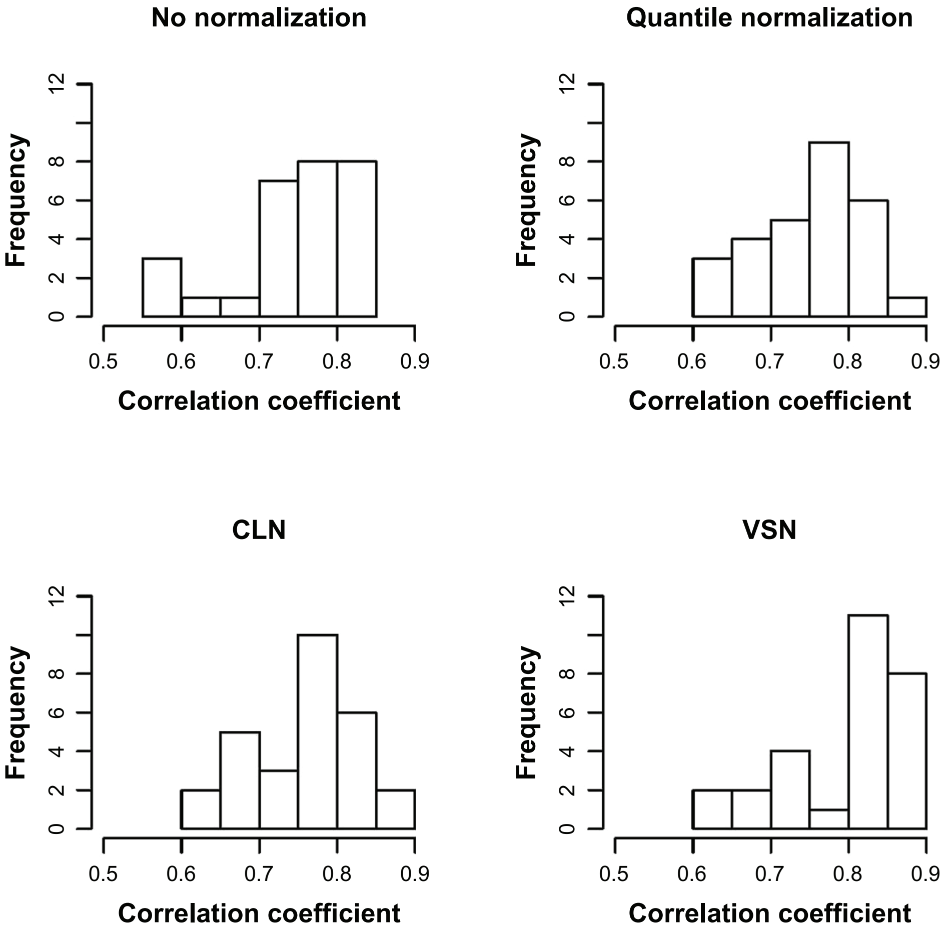

Figure 4. Histogram of correlation coefficients between Agilent data and Solexa data on each sample.
Abbreviations: CLN, cyclic loess normalization; VSN, variance stabilizing normalization.

- Similar to quantile normalization, cyclic loess normalization claimed differential expression for 15 miRNAs including 10 DE25 miRNAs and resulted in Pearson correlation between the two data types of 0.62 to 0.86 among all miRNAs and 0.80 to 0.95 among the always-on and sometimes-on miRNAs.
- Also similar were the results with variance stabilizing normalization, which claimed differential expression for 14 miRNAs including 4 DE25 miRNAs and resulted in Pearson correlation between the two data types of 0.63 to 0.89 among all miRNAs and 0.81 to 0.96 among the always-on and sometimes-on miRNAs.

To summarize, median normalization resulted in some improvement in the accuracy of the differential expression analysis, and the other three normalization methods led to a greater, yet still limited, improvement.

Effects of background adjustment

We also evaluated the effect of background adjustment method in combination with array normalization on the accuracy of detecting differentially expressed miRNAs. We assessed the effect of background adjustment on the analysis of Agilent data by subtracting background intensity from foreground intensity before $\log_2$ transformation. Background subtraction slightly improved the detection of differential expression. When combined with quantile, cyclic loess, or variance stabilizing normalization, background subtraction slightly increased the number of true positives detected while maintaining a similar number of false positives (results not shown).

Discussion

In summary, using data from a liposarcoma study, our study demonstrated that (1) array effects can significantly affect the detection of differentially

Table 1. Comparison of differentially expressed miRNAs claimed to be positive ($P < 0.0001$) or negative ($P > 0.0001$) by Agilent data versus those declared to be positive or negative by the Solexa data.

	Normalization method				
	None (+, –)	Median (+, –)	Quantile (+, –)	Cyclic loess (+, –)	Variance stablizing (+, –)
Solexa					
+	4, **21**	5, **20**	11, **14**	10, **15**	10, **15**
–	**3**, 458	**1**, 460	**5**, 456	**5**, 456	**4**, 457

Note: The numbers of false positive and false negative miRNAs identified are indicated in bold for each normalization method.

expressed miRNAs and (2) statistical methods for normalizing the arrays can improve the detection to some extent.

Array effects are not unique to our liposarcoma data, and we have observed pervasive array effects in other miRNA array datasets such as those collected by the Cancer Genome Atlas (TCGA), a multi-institutional effort led by the National Cancer Institute that aims to catalogue major cancer-causing genome alterations in human tumors through multi-dimensional genomic profiling.[52] Plots showing array effects in the TCGA ovarian miRNA data are provided in the supplementary materials.

In our study, quantile, cyclic loess, and variance stabilizing normalization performed similarly, and all three performed better than median normalization. Quantile normalization has some practical advantages over cyclic loess and variance stabilizing normalization, as cyclic loess normalization takes much more computer time and variance stabilizing normalization results in negative values that need to be arbitrarily set to 1 before $\log_2$ transformation.

More importantly, our study indicated the need for better normalization methods for miRNA data. Even with quantile normalization, the numbers of false positive and false negative miRNAs are still significant, and some of these miRNAs are expressed at moderate to high levels (Fig. 5). Hence these false miRNAs are not just a result of the low resolution of Agilent array at low expression levels, and there is still a need to develop more efficient normalization methods for miRNA arrays. Until better methods are developed, an existing normalization method such as quantile normalization should be applied when analyzing miRNA array data.

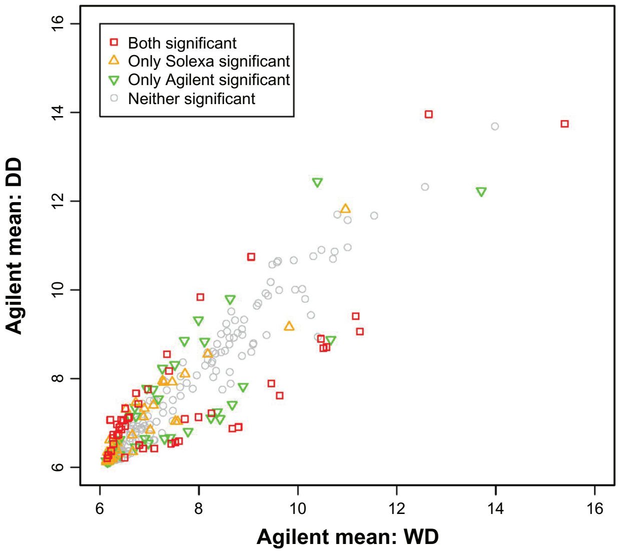

Figure 5. Scatter plot of mean expression among WD (n = 14) versus among DD (n = 14).
Notes: Red indicates miRNAs that are claimed to be significant by both the Solexa and Agilent data. Orange is for miRNAs that are claimed to be significant by the Solexa data only. Green is for miRNAs that are claimed to be significant by the Agilent data only. Grey is for miRNAs that are claimed to be significant by neither the Solexa data nor the Agilent data. Agilent data are quantile normalized.

The sequence data used as a benchmark are imperfect, but they offer the considerable advantage of being derived from an independent method. In addition to analyzing the sequence data by comparing the relative clone counts, we also analyzed the sequence data by comparing the clone count using an exact test for the negative binomial distribution and the edgeR package, which calculates an effective library size for each sample to adjust for potential sequence composition bias and allows for a dispersion parameter to account for extra variability,[53,54] and arrived at the same conclusions. Nevertheless, our study demonstrates an additional line of evidence that array normalization is useful for miRNA arrays but that better methods are needed.

Conclusions

There is a need to develop more efficient normalization methods for miRNA arrays to improve the detection of genes with disease relevance. Until better methods are developed, an existing normalization method such as quantile normalization should be applied when analyzing miRNA array data.

List of Abbreviations

miRNA, microRNA; WD, well-differentiated liposarcoma; DD, dedifferentiated liposarcoma; MSKCC, Memorial Sloan-Kettering Cancer Center; DE25, the top 25 most significant microRNAs identified in the liposarcoma sequence data.

Data Availability

Data used in this study are available upon request.

Acknowledgements

We thank Dr. Jaya Satagopan for helpful comments to the paper and Joseph Kanik for his excellent secretarial support.

Author Contributions

LXQ designed the study. SS and TT collected the data. LXQ conducted the analysis. LXQ and SS wrote the manuscript. All authors reviewed and approved of the final manuscript.

Funding

This work was supported by the National Institutes of Health (CA151947 to LXQ and SS; CA047179 to SS and LXQ; and CA140146 to SS and LXQ). The funding agents have no role in the study design, in the collection, analysis, and interpretation of data, in the writing of the manuscript, or in the decision to submit the manuscript for publication. Funding for open access charge: National Institutes of Health.

Competing Interests

Author(s) disclose no potential conflicts of interest.

Disclosures and Ethics

As a requirement of publication author(s) have provided to the publisher signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section. The external blind peer reviewers report no conflicts of interest.

References

1. Ambros V. The functions of animal microRNAs. *Nature*. 2004;431(7006):350–5.
2. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*. 2004;116(2):281–97.
3. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*. 2005;120(1):15–20.
4. He L, Thomson JM, Hemann MT, et al. A microRNA polycistron as a potential human oncogene. *Nature*. 2005;435(7043):828–33.
5. O'Donnell KA, Wentzel EA, Zeller KI, Dang CV, Mendell JT. c-Myc-regulated microRNAs modulate E2F1 expression. *Nature*. 2005;435(7043):839–43.
6. Dyrskjot L, Ostenfeld MS, Bramsen JB, et al. Genomic profiling of microRNAs in bladder cancer: miR-129 is associated with poor outcome and promotes cell death in vitro. *Cancer Res*. 2009;69(11):4851–60.
7. Jares P, Campo E. Genomic platforms for cancer research: potential diagnostic and prognostic applications in clinical oncology. *Clin Transl Oncol*. 2006;8(3):161–72.
8. Merritt WM, Lin YG, Han LY, et al. Dicer, Drosha, and outcomes in patients with ovarian cancer. *N Engl J Med*. 2008;359(25):2641–50.
9. Roldo C, Missiaglia E, Hagan JP, et al. MicroRNA expression abnormalities in pancreatic endocrine and acinar tumors are associated with distinctive pathologic features and clinical behavior. *J Clin Oncol*. 2006;24(29):4677–84.
10. Strausberg RL, Simpson AJ, Old LJ, Riggins GJ. Oncogenomics and the development of new cancer therapies. *Nature*. 2004;429(6990):469–74.
11. Yang N, Kaur S, Volinia S, et al. MicroRNA microarray identifies Let-7i as a novel biomarker and therapeutic target in human epithelial ovarian cancer. *Cancer Res*. 2008;68(24):10307–14.
12. Zhang L, Volinia S, Bonome T, et al. Genomic and epigenetic alterations deregulate microRNA expression in human epithelial ovarian cancer. *Proc Natl Acad Sci U S A*. 2008;105(19):7004–9.
13. Fletcher C, Unni K, Mertens F. *Pathology and Genetics of Tumours of Soft Tissue and Bone*. Lyon, France: IARC Press; 2002.
14. Barretina J, Taylor BS, Banerji S, et al. Subtype-specific genomic alterations define new targets for soft-tissue sarcoma therapy. *Nat Genet*. 2010;42(8):715–21.
15. Khoury MJ, McBride CM, Schully SD, et al. The Scientific Foundation for personal genomics: recommendations from a National Institutes of Health-Centers for Disease Control and Prevention multidisciplinary workshop. *Genet Med*. 2009;11(8):559–67.
16. Nguyen DV, Arpat AB, Wang N, Carroll RJ. DNA microarray experiments: biological and technological aspects. *Biometrics*. 2002;58(4):701–17.
17. Quackenbush J. Microarray data normalization and transformation. *Nat Genet*. 2002;32 Suppl:496–501.
18. Ransohoff DF. How to improve reliability and efficiency of research about molecular markers: roles of phases, guidelines, and study design. *J Clin Epidemiol*. 2007;60(12):1205–19.
19. Huber W, von Heydebreck A, Sueltmann H, Poustka A, Vingron M. Parameter estimation for the calibration and variance stabilization of microarray data. *Stat Appl Genet Mol Biol*. 2003;2:Article 3.
20. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res*. 2003;31(4):e15.
21. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003;4(2):249–64.
22. Li C, Wong WH. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A*. 2001;98(1):31–6.

23. Schadt EE, Li C, Ellis B, Wong WH. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J Cell Biochem Suppl*. 2001;Suppl 37:120–5.
24. Yang YH, Dudoit S, Luu P, et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*. 2002;30(4):e15.
25. Garzon R, Fabbri M, Cimmino A, Calin GA, Croce CM. MicroRNA expression and function in cancer. *Trends Mol Med*. 2006;12(12):580–7.
26. Pan Q, Luo X, Chegini N. Differential expression of microRNAs in myometrium and leiomyomas and regulation by ovarian steroids. *J Cell Mol Med*. 2008;12(1):227–40.
27. Perkins DO, Jeffries CD, Jarskog LF, et al. microRNA expression in the prefrontal cortex of individuals with schizophrenia and schizoaffective disorder. *Genome Biol*. 2007;8(2):R27.
28. Sengupta S, den Boon JA, Chen IH, et al. MicroRNA 29c is down-regulated in nasopharyngeal carcinomas, up-regulating mRNAs encoding extracellular matrix proteins. *Proc Natl Acad Sci U S A*. 2008;105(15): 5874–8.
29. Dennis L. MicroRNAs in early embryonic development: Dissecting the role of miR-290 through miR-295 in mouse. [dissertation]. Cambridge, MA: MIT; 2008.
30. Marson A, Levine SS, Cole MF, et al. Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell*. 2008;134(3):521–33.
31. Rosenfeld N, Aharonov R, Meiri E, et al. MicroRNAs accurately identify cancer tissue origin. *Nat Biotechnol*. 2008;26(4):462–9.
32. Sharma S, Kelly TK, Jones PA. Epigenetics in Cancer. *Carcinogenesis*. 2010;31(1):27–36.
33. Pradervand S, Weber J, Thomas J, et al. Impact of normalization on miRNA microarray expression profiling. *RNA*. 2009;15(3):493–501.
34. Tricoli JV, Jacobson JW. MicroRNA: Potential for Cancer Detection, Diagnosis, and Prognosis. *Cancer Res*. 2007;67(10):4553–5.
35. Rao Y, Lee Y, Jarjoura D, et al. A comparison of normalization techniques for microRNA microarray data. *Stat Appl Genet Mol Biol*. 2008;7(1):Article 22.
36. Lopez-Romero P, Gonzalez MA, Callejas S, Dopazo A, Irizarry RA. Processing of Agilent microRNA array data. *BMC Res Notes*. 2010;3:18.
37. Hafner M, Renwick N, Brown M, et al. RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA*. 2011;17(9):1697–712.
38. Hafner M, Renwick N, Pena J, Mihalovic A, Tuschl T. Barcoded cDNA libraries for miRNA profiling by next-generation sequencing. In: Hartmann RK, Bindereif A, Schon A, Westhof E, editors. *Handbook of RNA Biochemistry*. Weinheim, Germany: Wiley-VCH; 2005.
39. Farazi TA, Brown M, Morozov P, et al. Bioinformatic analysis of barcoded cDNA libraries for small RNA profiling by next-generation sequencing. *Methods*. 2012;58(2):171–87.
40. Hafner M, Renwick N, Farazi TA, Mihailovic A, Pena JT, Tuschl T. Barcoded cDNA library preparation for small RNA profiling by next-generation sequencing. *Methods*. 2012;58(2):164–70.
41. Ugras S, Brill E, Jacobsen A, et al. Small RNA sequencing and functional characterization reveals MicroRNA-143 tumor suppressor activity in liposarcoma. *Cancer Res*. 2011;71(17):5659–69.
42. Barcoded cDNA library preparation for small RNA profiling by next-generation sequencing. Hafner M, Renwick N, Farazi TA, Mihailović A, Pena JT, Tuschl T. *Methods*. 2012 Oct;58(2):164–70. doi: 10.1016/j.ymeth.2012.07.030. Epub 2012 Aug 7. PMID:22885844.
43. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*. 2004;3:Article 3.
44. Ma L, Young J, Prabhala H, et al. miR-9, a MYC/MYCN-activated microRNA, regulates E-cadherin and cancer metastasis. *Nat Cell Biol*. 2010;12(3):247–56.
45. Si ML, Zhu S, Wu H, Lu Z, Wu F, Mo YY. miR-21-mediated tumor growth. *Oncogene*. 2007;26(19):2799–803.
46. Meng F, Henson R, Lang M, et al. Involvement of human micro-RNA in growth and response to chemotherapy in human cholangiocarcinoma cell lines. *Gastroenterology*. 2006;130(7):2113–29.
47. Yamamoto Y, Yoshioka Y, Minoura K, et al. An integrative genomic analysis revealed the relevance of microRNA and gene expression for drug-resistance in human breast cancer cells. *Mol Cancer*. 2011;10:135.
48. Rauhala HE, Jalava SE, Isotalo J, et al. miR-193b is an epigenetically regulated putative tumor suppressor in prostate cancer. *Int J Cancer*. 2010;127(6):1363–72.
49. Li XF, Yan PJ, Shao ZM. Downregulation of miR-193b contributes to enhance urokinase-type plasminogen activator (uPA) expression and tumor progression and invasion in human breast cancer. *Oncogene*. 2009;28(44):3937–48.
50. Beezhold K, Liu J, Kan H, et al. miR-190-mediated downregulation of PHLPP contributes to arsenic-induced Akt activation and carcinogenesis. *Toxicol Sci*. 2011;123(2):411–20.
51. Calin GA, Liu CG, Sevignani C, et al. MicroRNA profiling reveals distinct signatures in B cell chronic lymphocytic leukemias. *Proc Natl Acad Sci U S A*. 2004;101(32):11755–60.
52. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*. 2008;455(7216):1061–8.
53. Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*. 2007;23:2881–7.
54. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26:139–40.

Supplementary Figures

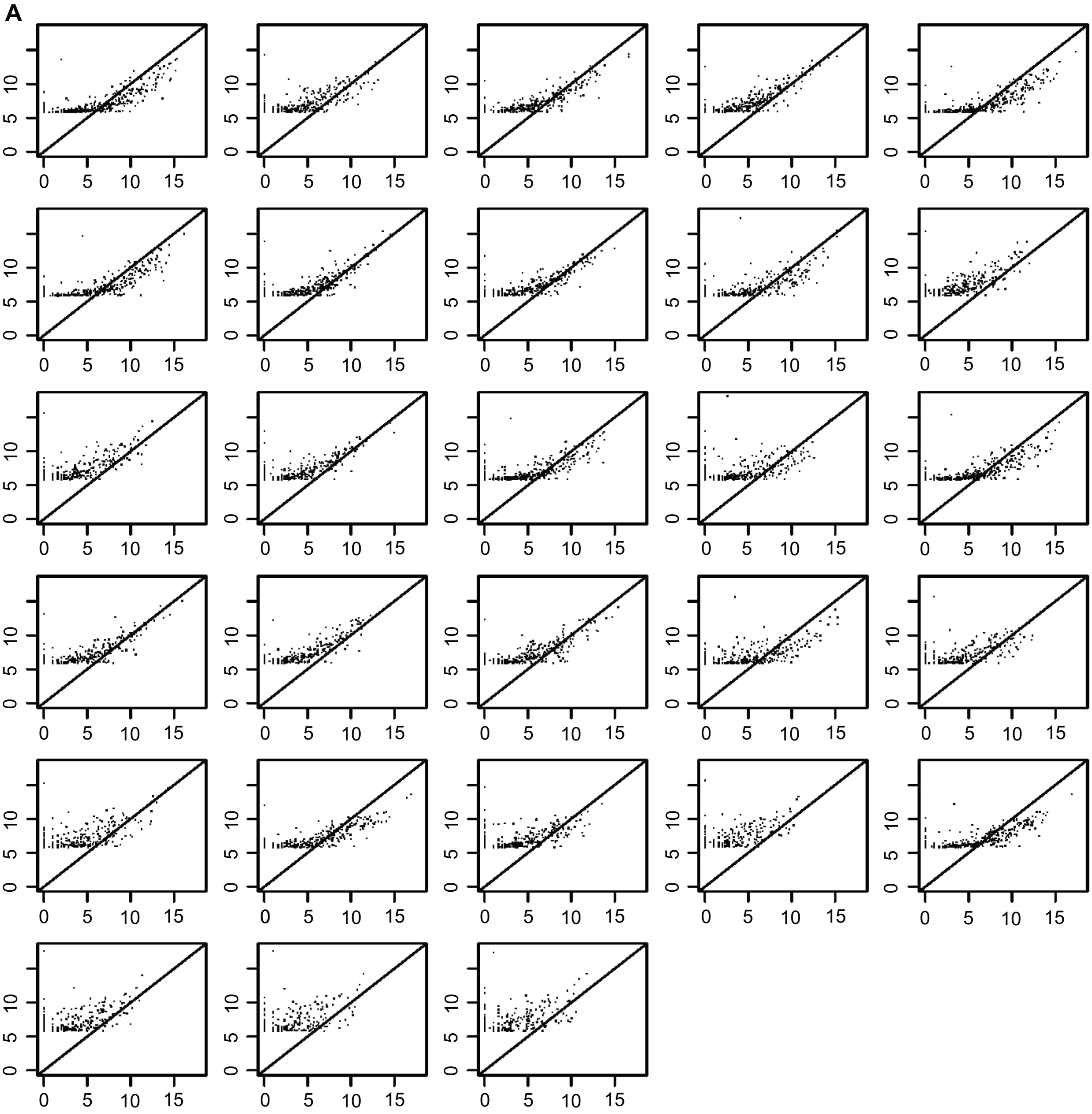

Figure S1A. Scatter plot between the deep sequencing data (x axis) and the Agilent data (y axis).
Note: No normalization.

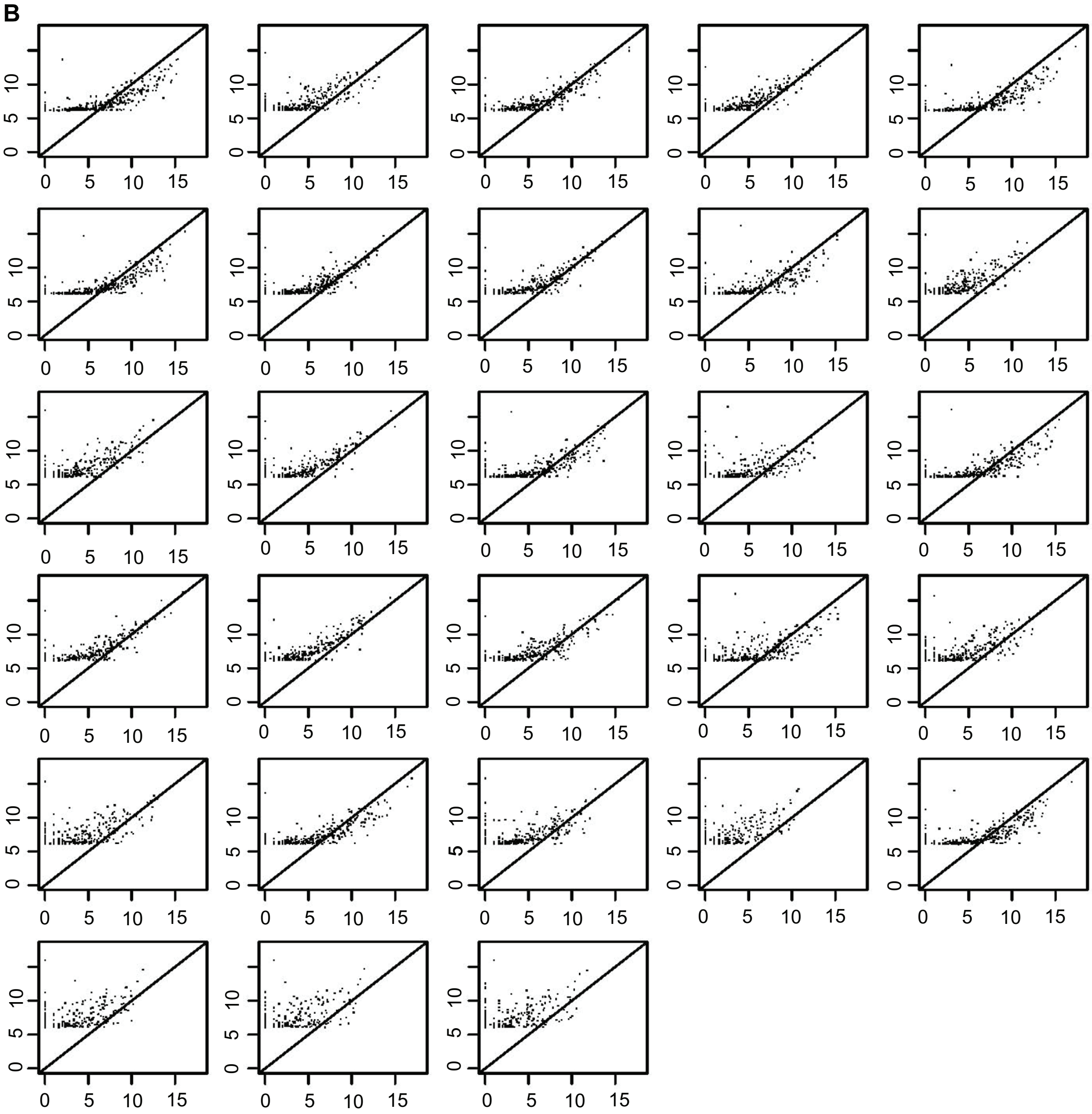

Figure S1B. Scatter plot between the deep sequencing data (x axis) and the Agilent data (y axis).
Note: Median normalization.

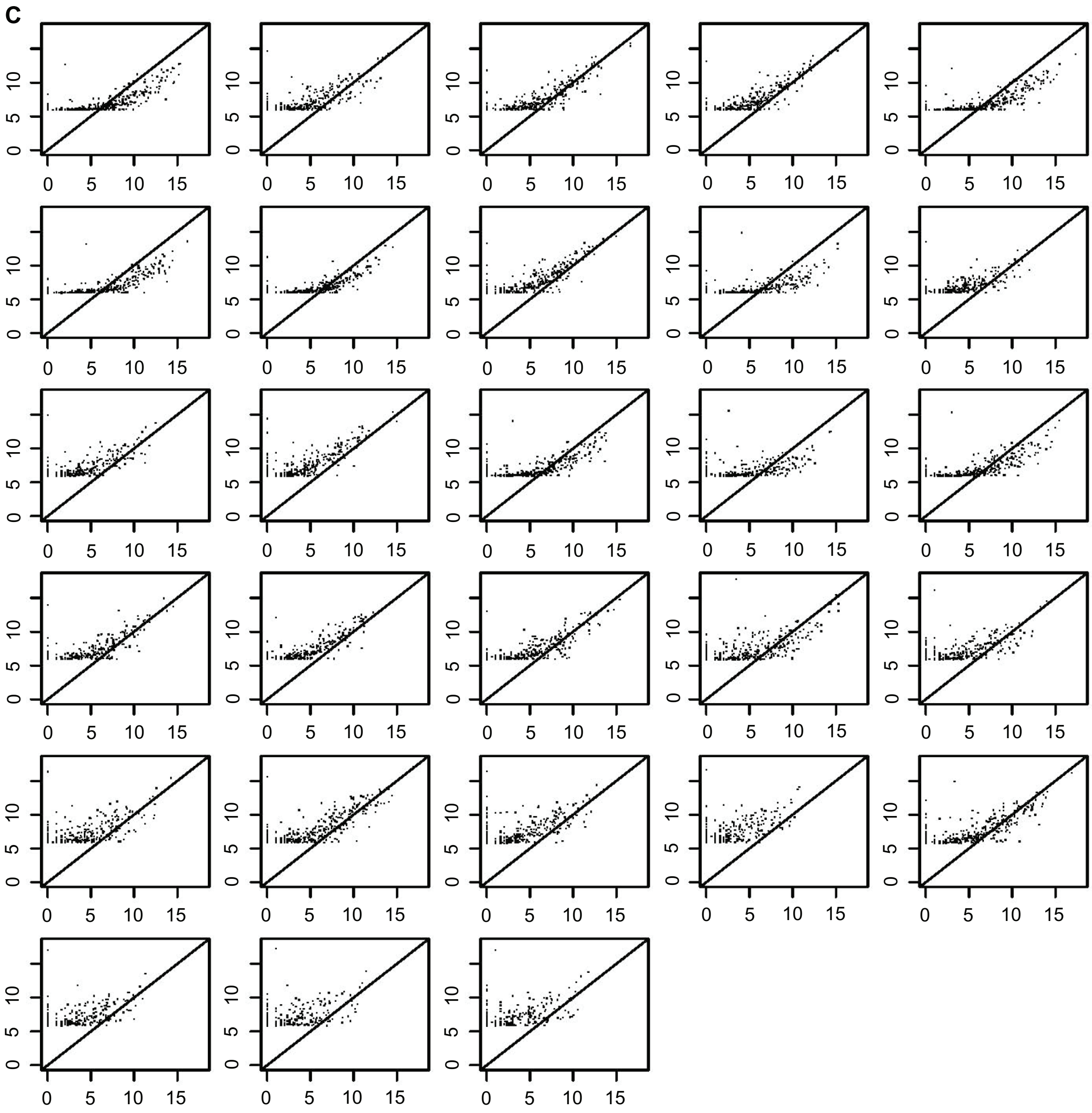

Figure S1C. Scatter plot between the deep sequencing data (x axis) and the Agilent data (y axis).
Note: Quantile normalization.

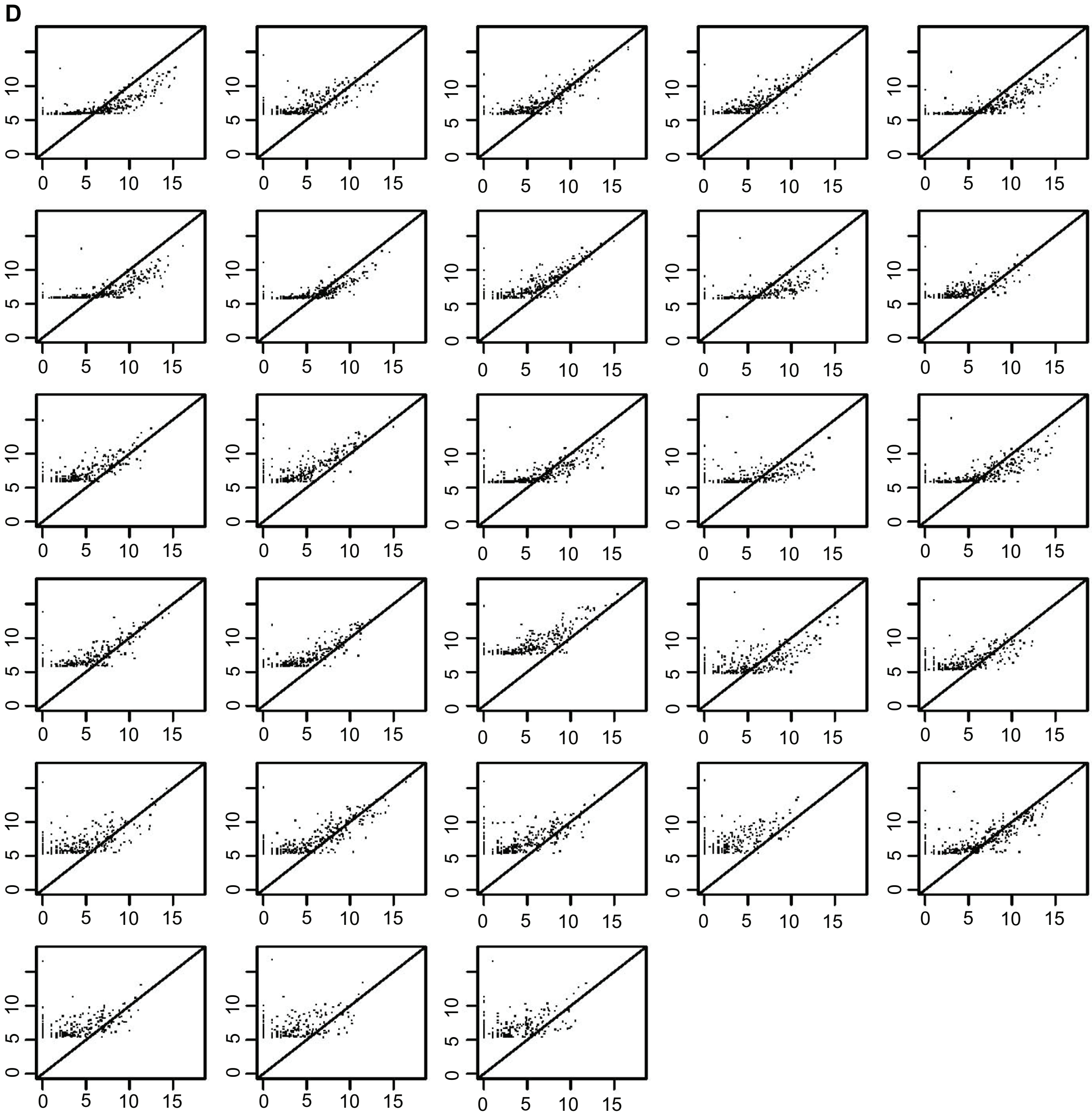

Figure S1D. Scatter plot between the deep sequencing data (x axis) and the Agilent data (y axis).
Note: Cyclic loess normalization.

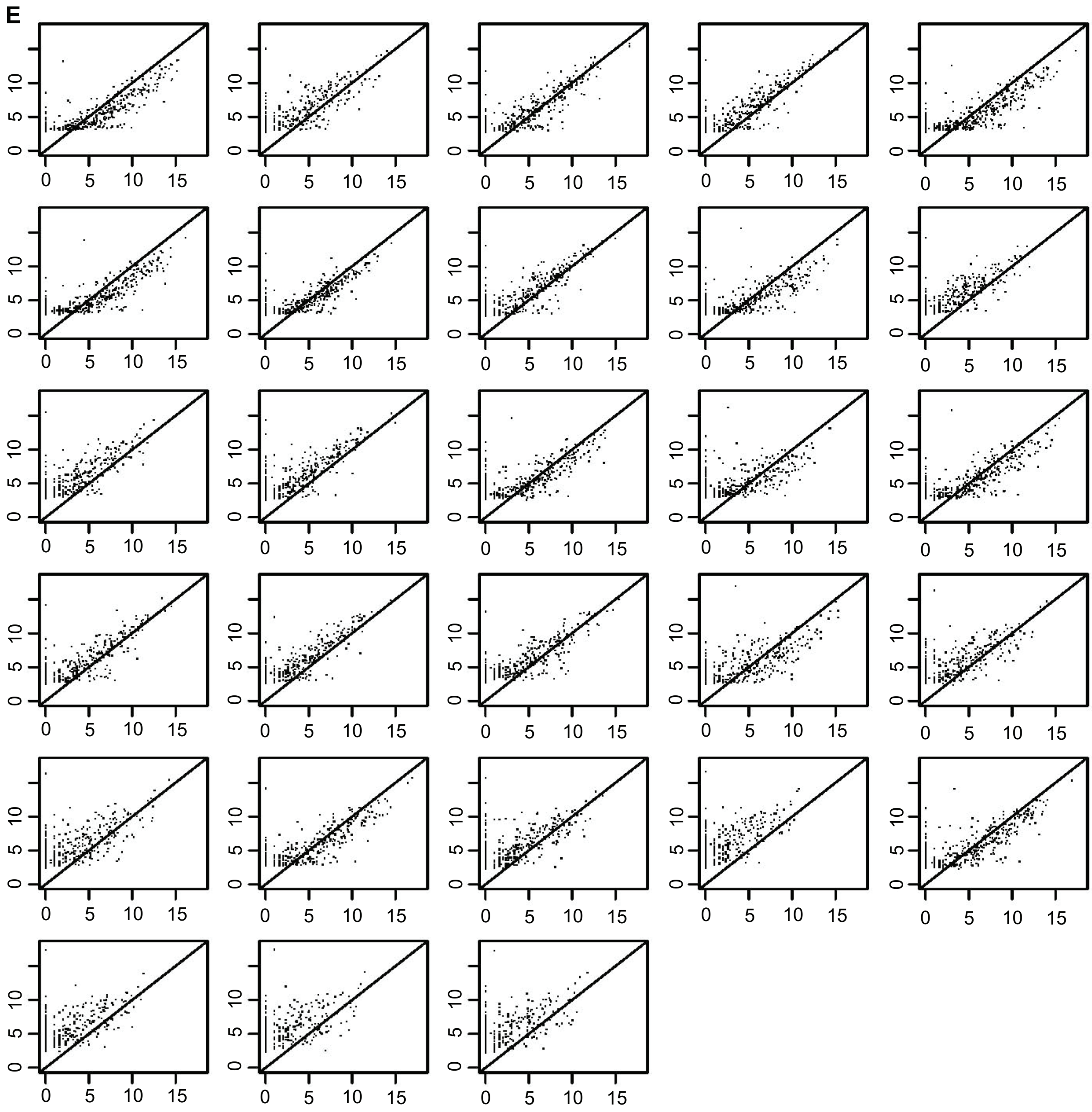

Figure S1E. Scatter plot between the deep sequencing data (x axis) and the Agilent data (y axis).
Note: Variance stabilizing normalization.

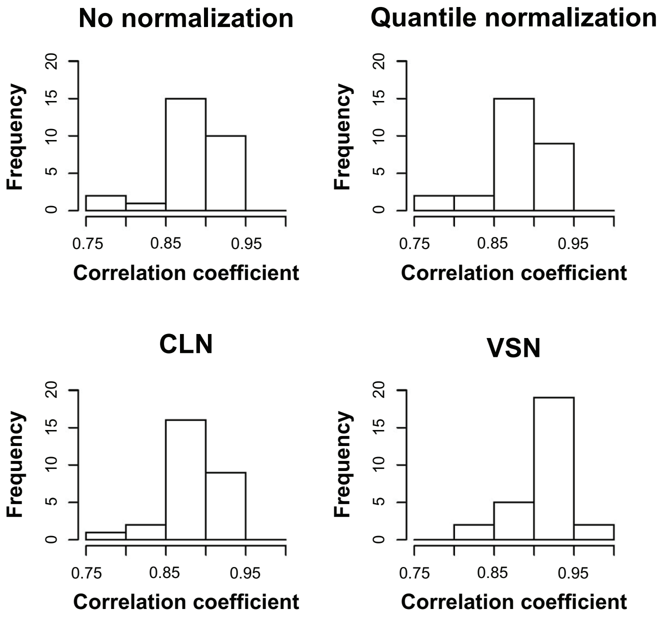

Figure S2. Histogram of the correlation coefficient between Agilent data and Solexa data among always-on genes and sometimes-on genes in each sample.

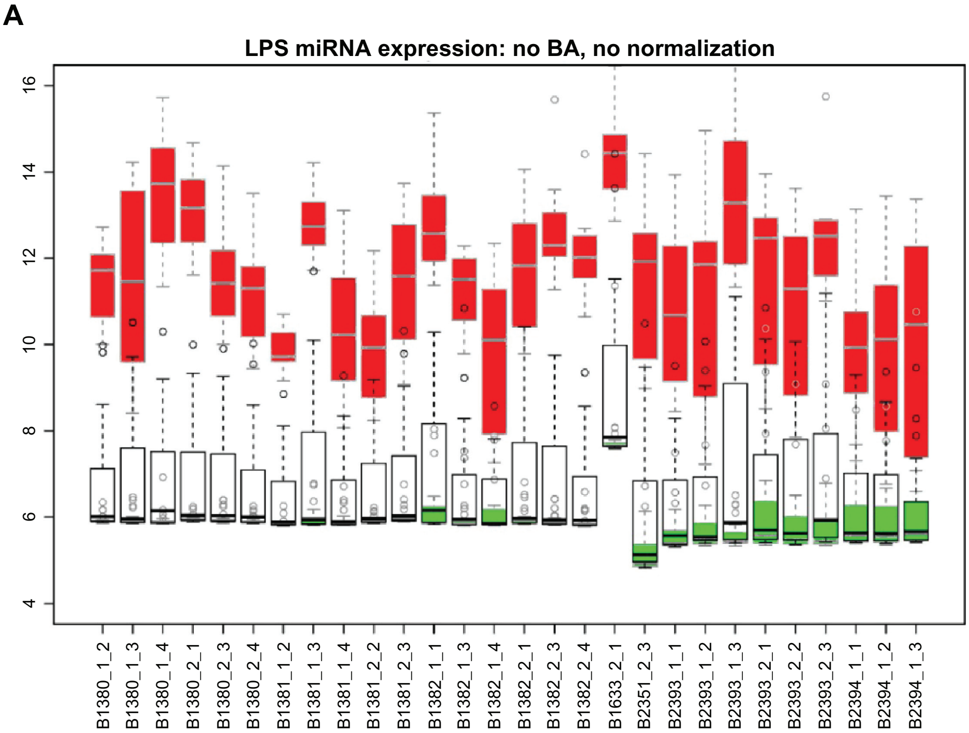

Figure S3A. Boxplot of expression levels of the 7 always-on miRNAs (red), 28 always-off miRNAs (green), and 20 sometimes-on miRNAs (white) on each array. The arrays are ordered by array slide.
Note: No normalization.

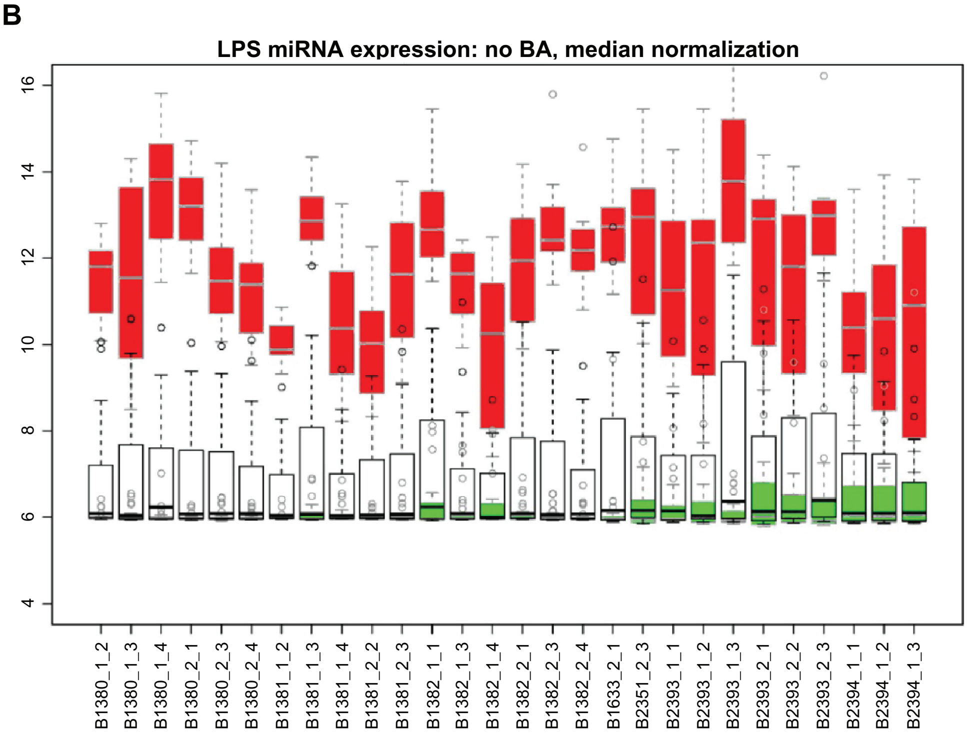

Figure S3B. Boxplot of expression levels of the 7 always-on miRNAs (red), 28 always-off miRNAs (green), and 20 sometimes-on miRNAs (white) on each array. The arrays are ordered by array slide.
Note: median normalization.

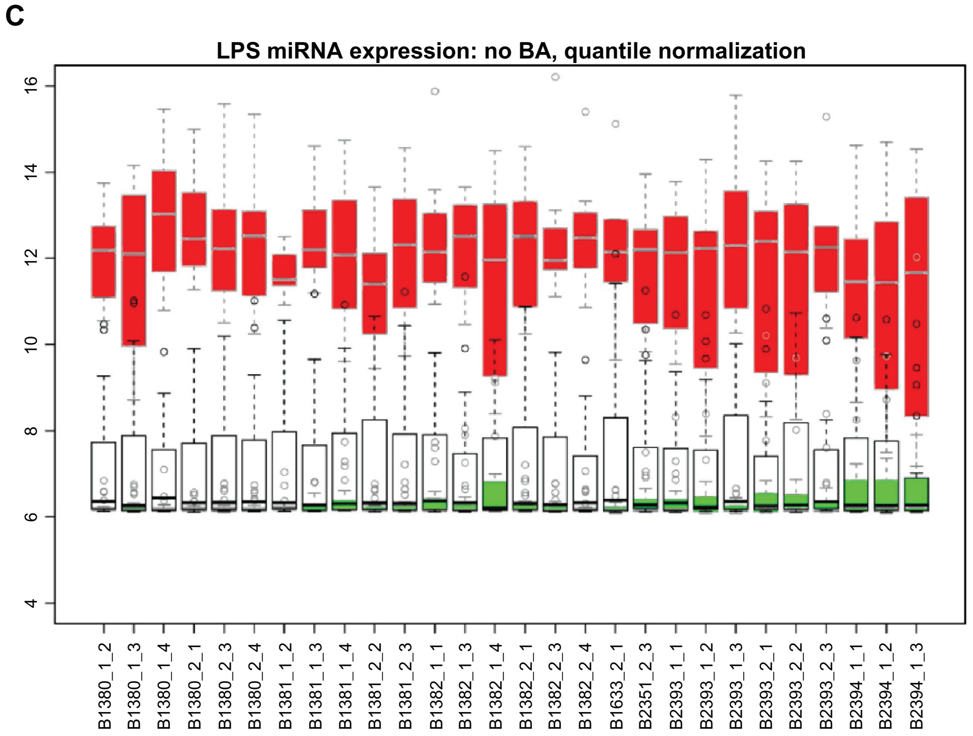

Figure S3C. Boxplot of expression levels of the 7 always-on miRNAs (red), 28 always-off miRNAs (green), and 20 sometimes-on miRNAs (white) on each array. The arrays are ordered by array slide.
Note: quantile normalization.

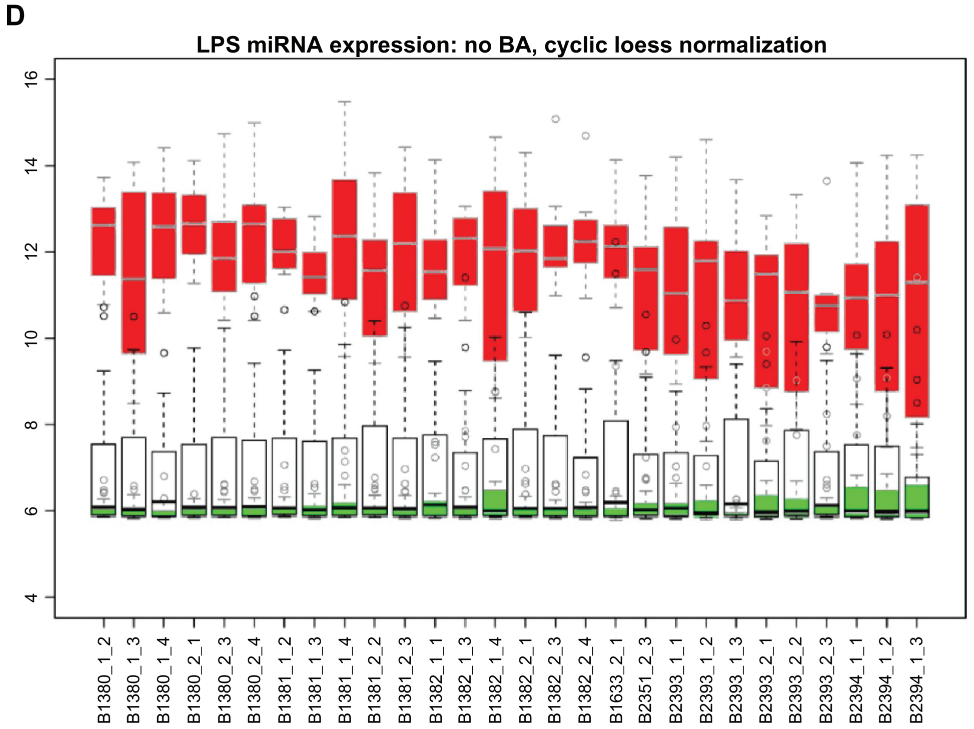

Figure S3D. Boxplot of expression levels of the 7 always-on miRNAs (red), 28 always-off miRNAs (green), and 20 sometimes-on miRNAs (white) on each array. The arrays are ordered by array slide.
Note: cyclic loess normalization.

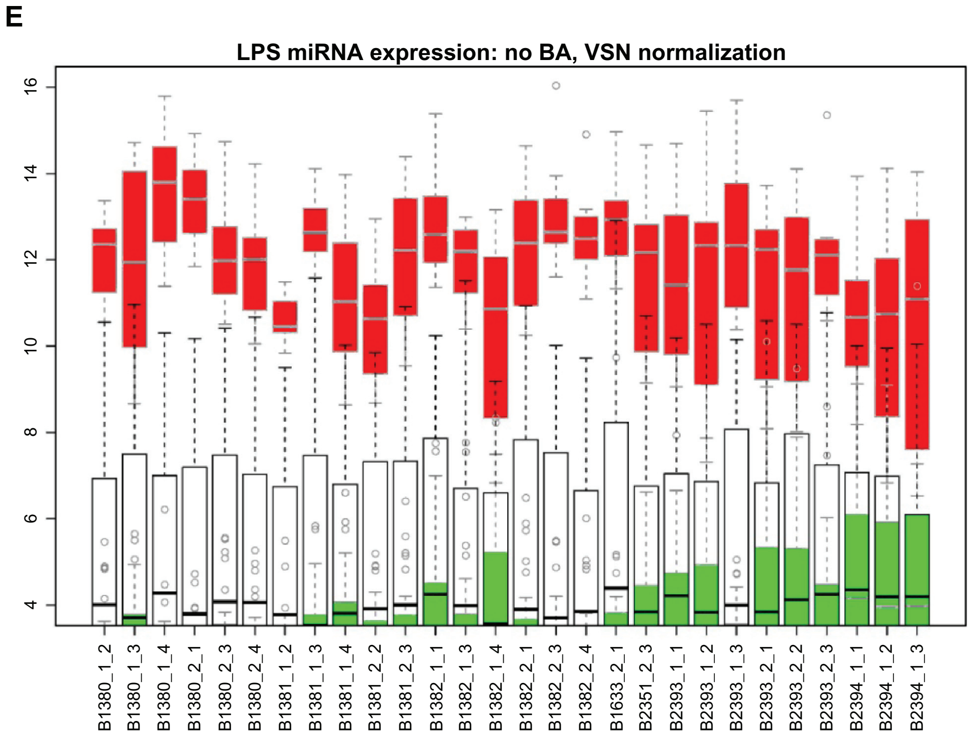

Figure S3E. Boxplot of expression levels of the 7 always-on miRNAs (red), 28 always-off miRNAs (green), and 20 sometimes-on miRNAs (white) on each array. The arrays are ordered by array slide.
Note: variance stabilizing normalization.

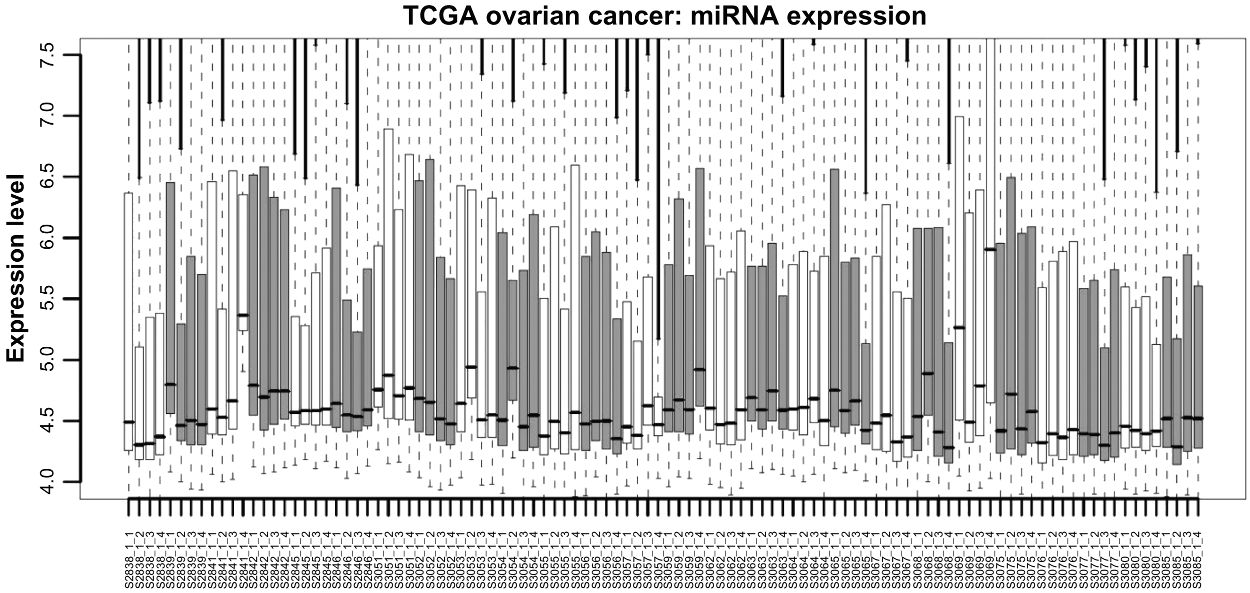

Figure S4. Boxplot of the foreground intensity on the $\log_2$ scale for a subset of the Agilent arrays (n = 104) in the TCGA ovarian study.
Notes: TCGA is a multi-institutional effort led by the National Cancer Institute that aims to catalogue major cancer-causing genome alterations in human tumors through multi-dimensional genomic profiling. One of the first three human tumor types TCGA studied was ovarian cancer. The tumor tissue samples were collected nation-wide, and the miRNA arrays were generated centrally at a Cancer Genome Characterization Center. We downloaded the first 226 ovarian arrays at the TCGA data portal. The 226 arrays belong to 26 array slides, with the number of arrays per slide ranging from 2 to 6. (Some arrays on each slide are used for the characterization center's internal controls and their data are not publicly available.) To simplify interpretation, we only looked at the row-1 arrays from slides that have all four row-1 arrays available, which consisted of 104 arrays from 26 slides. The arrays exhibit noticeable differences between array slides and arrays within a slide; these differences are not readily interpretable by tumor differences alone. Data at the probe level with no normalization are displayed with one box per array. The arrays are ordered by array slide and colored to distinguish neighboring slides.