Barcoding bias in high-throughput multiplex sequencing of miRNA

Shahar Alon, Francois Vigneault, Seda Eminaga, et al.

Genome Res. 2011 21: 1506-1511 originally published online July 12, 2011
Access the most recent version at doi:10.1101/gr.121715.111

Method

Barcoding bias in high-throughput multiplex sequencing of miRNA

Shahar Alon,[1,6] Francois Vigneault,[2,3,4,6] Seda Eminaga,[2] Danos C. Christodoulou,[2] Jonathan G. Seidman,[2] George M. Church,[2,3] and Eli Eisenberg[5,7]

[1]*Department of Neurobiology, George S. Wise Faculty of Life Sciences, Tel-Aviv University, Tel-Aviv 69978, Israel;* [2]*Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA;* [3]*Wyss Institute for Biologically Inspired Engineering, Boston, Massachusetts 02115, USA;* [4]*Ragon Institute of MGH, MIT, and Harvard, Boston, Massachusetts 02129, USA;* [5]*Raymond and Beverly Sackler School of Physics and Astronomy, Tel-Aviv University, Tel-Aviv 69978, Israel*

Second-generation sequencing is gradually becoming the method of choice for miRNA detection and expression profiling. Given the relatively small number of miRNAs and improvements in DNA sequencing technology, studying miRNA expression profiles of multiple samples in a single flow cell lane becomes feasible. Multiplexing strategies require marking each miRNA library with a DNA barcode. Here we report that barcodes introduced through adapter ligation confer significant bias on miRNA expression profiles. This bias is much higher than the expected Poisson noise and masks significant expression differences between miRNA libraries. This bias can be eliminated by adding barcodes during PCR amplification of libraries. The accuracy of miRNA expression measurement in multiplexed experiments becomes a function of sample number.

[Supplemental material is available for this article.]

The discovery of microRNAs (miRNAs) has revealed the existence of a previously unrecognized layer of complexity in gene regulation (Bartel 2004). MiRNAs regulate protein expression and are involved in many cellular and physiological processes, including numerous pathological conditions (Lu et al. 2005). Therefore, detecting new miRNAs and measuring the expression profiles of known miRNAs are important tasks required for a complete understanding of various biological conditions. The relatively low number of miRNAs (about 1000 human miRNAs in miRBase version 15) (Griffiths-Jones et al. 2008) and the small size of mature miRNAs (19–25 nucleotides [nt]) allow current second-generation sequencing platforms to achieve both mentioned tasks (Creighton et al. 2009). For example, one lane of Illumina's flow cell with a sequencing depth of ~200 M bases was adequate for the identification of novel miRNAs and for the quantification of known miRNAs using library of miRNAs derived from human tissues (Morin et al. 2008). However, the cost of next-generation sequencing is still considerable, limiting the number of biological conditions to be tested. A popular solution for this hurdle is multiplexing, where many samples are being marked by some specific tag sequence (barcode) and sequenced in a single lane.

One approach to constructing multiplex libraries for miRNA expression analyses consists in introducing a DNA barcode in the 5′ oligonucleotide adapter required for miRNA ligation (Uziel et al. 2009; Tarasov et al. 2007; Zhu et al. 2008, 2009). Another possible option is to introduce barcodes during the PCR amplification of the libraries. Here, we compared miRNA expression profiles obtained from DNA libraries constructed by these two methods.

[6]These authors contributed equally to this work.
[7]Corresponding author.
E-mail elieis@post.tau.ac.il.
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.121715.111.

Results

First, biases in miRNA expression levels introduced by 5′ ligation barcodes were assessed. "Control" miRNA samples were collected from normal and diseased mouse cardiac left ventricle (Teekakirikul et al. 2010). The two samples were prepared in parallel; each mouse miRNA sample was split into 10 equal aliquots and marked by the same set of 10 different 5′ ligation barcodes during the library preparation. The two sets of 10 libraries were sequenced by Illumina's GAIIx instrument in a single flow cell, each set in a single lane (see Methods).

Significant differences were observed between the miRNA expression profiles for the same RNA sample that differed only in the barcode used to construct the library, suggesting a barcode bias (Fig. 1A; Methods). The barcode-dependent bias was easily observed using hierarchical clustering of all the miRNA expression profiles. The clustering process paired identical barcodes rather than clustering the profiles belonging to identical biological conditions (Fig. 1C). Moreover, reliable identification of differentially expressed miRNAs was not feasible (see Methods). For example, when comparing the exact same tissue with two different barcodes, one observes erroneously that 27% of the miRNAs are differentially expressed. Looking for differentially expressed miRNAs between normal and diseased mouse hearts using different barcodes for the two tissues and a stringent cutoff of twofold change results in dramatically different sets of miRNAs depending on the barcodes used. One finds <5% overlap between the lists of presumed differentially expressed miRNAs for eight different barcode pairs. Apparent variation in miRNA expression was well modeled, assuming different barcode-dependent capturing efficiency for each miRNA. Variance analysis revealed that barcode-specific capture bias could be as much as twofold, which often exceeded the biological variation in miRNA levels (Fig. 1B; Methods).

While using different ligation-based barcodes does introduce a large amount of variability, measuring the fold-change across different biological conditions with the same ligation-based barcode

21:1506–1511 © 2011 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/11; www.genome.org

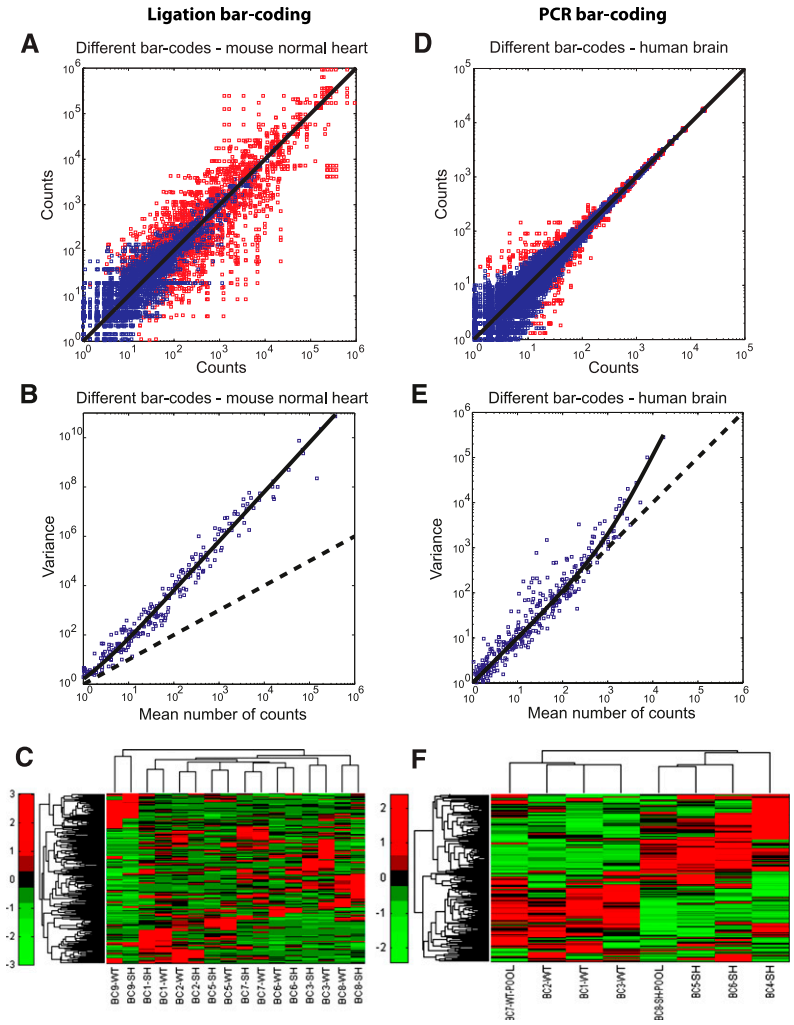

Figure 1. Barcoding bias analysis. (*A,D*) Total number of miRNA counts in each barcode compared with all the other barcodes (all the possible comparisons are plotted). The blue boxes represent points within the 99% region of Poisson noise, and the red boxes represent points outside this region. (*A*) When using ligation barcoding and normal mouse heart data, only 73% of all points fall inside this region, attesting for a barcode bias. (*D*) When using PCR barcoding and human brain data, 97% of all points fall inside the Poisson noise region. (*B,E*) The variance in counts number for a specific miRNA among the different barcodes as a function of the mean, plotted for all miRNAs. The black dotted line is the expected Poisson distribution with no barcode bias. The black full line is a fit to the general form expected for biased barcodes (see Methods). (*B*) When using ligation barcoding and normal mouse heart data, the variance due to barcodes diversity is much larger than the Poisson noise. (*E*) When using PCR barcoding and human brain data, only Poisson noise is evident for most of the experimentally relevant regime. (*C,F*) Hierarchical clustering of the miRNA expression profiles across different barcodes and biological conditions. (*C*) When using ligation-based barcodes, miRNA expression profiles cluster according to their barcodes, although they were derived from two different experimental conditions (normal and diseased mouse hearts, marked with WT and SH, respectively). (*F*) When using PCR-based barcodes, miRNA expression profiles cluster according to the experimental condition.

does provide meaningful, barcode-independent results. We confirmed this by comparing the results obtained from the different barcodes for the two biological conditions (Supplemental Fig. S1). Although two different biological conditions are compared, 79% of the miRNAs expression differences are within the Poisson noise region. As two different biological conditions are compared, some miRNAs are expected to be differentially expressed. Nevertheless, the agreement between the different miRNA profiles when comparing the same barcodes in different conditions is higher than the agreement between different barcodes in the same tissue (Fig. 1A vs. Supplemental Fig. S1). Furthermore, when comparing different barcodes in the same tissue (Fig. 1A), one finds that 26% (17%) of the points represent counts that changed by at least twofold (threefold). In comparison, comparing different biological conditions using the same barcode (Supplemental Fig. S1) results in only 10% (5%) of the points representing counts that changed by at least twofold (threefold). Lastly, looking at the eight lists of the differentially expressed miRNAs derived by comparing each barcode between the two biological conditions, one finds almost a 50% overlap between the lists. We therefore conclude that one may use ligation-based barcoding, as long as the reads-count numbers are always compared between different measurements with the same barcode only. That is, one may use two (or more) flow cell lanes with the same set of barcodes and compare the expression profile, for each barcode separately, between lanes.

Since all other steps in the process are identical and the library construction protocol used universal (i.e., barcode-independent) primers for the PCR step, we hypothesized that barcode bias was introduced during the ligation stage and not during the PCR amplification. At the time that this work was conducted, the Illumina PCR-based multiplexing approach was not compatible with miRNA (Methods and Fig. S2); we therefore developed a protocol where the barcodes are introduced during the PCR steps (see Methods). A sample of total RNA from human brain tissue was subdivided into 12 equal independent aliquots before the ligation of the adapters, marked by 12 different barcodes during the PCR amplification step and sequenced on one lane of Illumina GAIIx instrument using a 75-bp single pass sequencing read (our design also allows independent sequencing of indexing read if desired). The same variance analysis revealed that the PCR-based barcodes almost completely suppressed the barcode bias (Fig. 1D,E), bringing the typical barcode-dependent error down to ±3%. We then used the PCR-based barcodes to sequence samples of miRNAs from normal and diseased mouse heart tissues on one lane of Illumina GAIIx instrument. Again, the barcode-dependent error was estimated to be ±3%. The low level of error allowed a reliable detection of differentially expressed miRNAs (Supplemental Table S5), an important task that is not feasible with barcodes that introduce large bias (Fig. 1F; Supplemental Fig. S3; Methods).

Given the feasibility of practically bias-free multiplexing protocol, a question of major practical importance arises regarding the cost–benefit balance in barcoding. More barcodes allow for more

tissues or biological conditions to be tested in a given lane of a flow cell but allow less reads per miRNA (and thus lower detection rate and higher Poisson noise level). We note that the distribution of expression levels among the various miRNAs follows a Zipf's law with an exponential cutoff (Fig. 2A). We use this to estimate the decrease in the number of detections of differentially expressed miRNAs as a function of the total number of reads per barcode (see Methods), and we find a sublinear decrease due to the Zipf's law behavior (Fig. 2B,C). For example, using 5 million reads per barcode causes a less than 20% decrease in the number of detections of differentially expressed miRNAs compared with using 10 million reads per barcode. Therefore, within current sequencing read-output, the usage of many barcodes could lead to an overall larger number of detections, at the expense of not being sensitive to the low-level miRNAs. A similar behavior (but with higher detection values) is observed when one is interested in the detection of expressed miRNAs rather than in differential expression between two biological conditions (Fig. 2B).

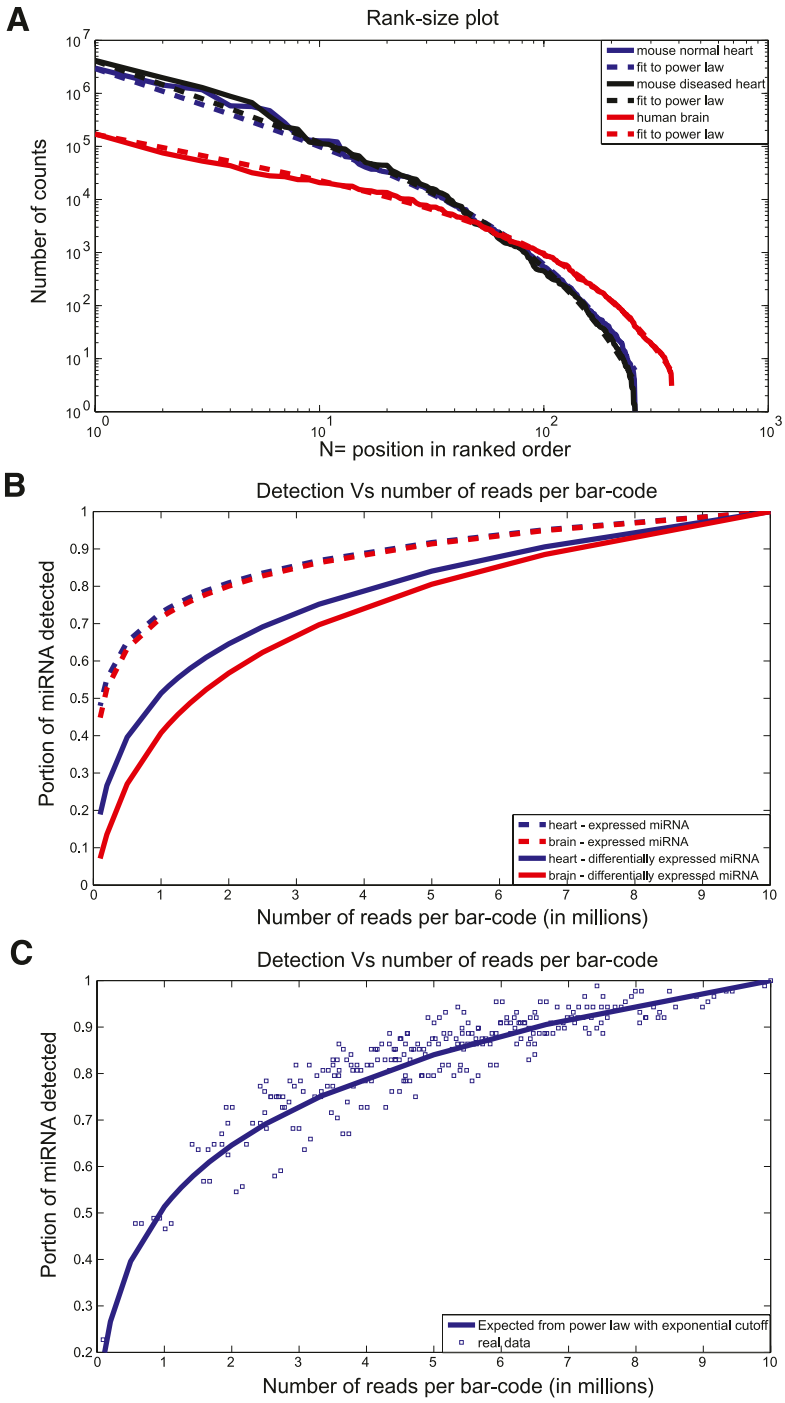

Figure 2. Modeling the detection efficiency as a function of the number of multiplexed samples. (*A*) Rank-size plot. Mouse normal heart, mouse diseased heart, and human brain data are plotted in blue, black, and red, respectively. The dashed lines are fits to power law with exponential cutoff. The fit has the form N^(−1.4) × exp(−N/47), N^(−1.4) × exp(−N/43), and N^(−0.8) × exp(−N/68) for mouse normal hearts, mouse diseased hearts, and human brain, respectively. (*B*) Expected portion of expressed miRNA (dashed lines) and differentially expressed miRNA (full lines) detected as a function of the number of reads per barcode. Human brain data are plotted in red and mouse normal heart data in blue. (*C*) Portion of differentially expressed miRNA detected as a function of the number of reads per barcode (see Methods). The blue boxes represent real data, and the blue line is the same as in *B*. Only reads uniquely aligned to miRNAs were used.

Discussion

Second-generation sequencing have revolutionized modern genomics in general and have increased our understanding about miRNAs in particular (Creighton et al. 2009). However, this still-growing field inevitably creates some biases in the vast amount of data generated. Indeed, recent reports show biases at multiple levels, from the effect of the library preparation in miRNA sequencing (Linsen et al. 2009) to preferences to specific sequence mutations (Dohm et al. 2008) up to problems caused by the wrong alignment of miRNA sequences (de Hoon et al. 2010). We have demonstrated that multiplexing of miRNA by ligation-based barcodes can create additional bias. By mapping and quantifying these biases, it will be possible to take advantage of the possibilities that second-generation sequencing has to offer without compromising the quality of the data.

Although it was possible to narrow the cause for the bias in ligation-based barcoding to the ligation stage, we did not succeed in pinpointing the exact problem caused by ligation. It is reasonable to look for some kind of sequence preference between the ligated sequence (containing the barcode) and the miRNA sequences. This sequence preference can be explained, for example, by the secondary structure formed between the miRNA and the barcode sequence. We therefore looked for any short (two to five bases) sequence in the beginning or the end of the miRNA that is over- or underrepresented in only some of the barcode libraries

from the same biological sample. This search did not result in any statistically significant sequence preference. Therefore, future studies are needed to identify the particular cause of the ligation bias.

The question of which kind of multiplexing approach to take, that is, ligation-based or PCR-based barcoding, is relevant to other applications of second-generation sequencing. As the sequencing depth will increase further, the possibility to multiplex mRNA will also be reasonable. It was not in the scope of this current work to determine if the ligation-bias will be also apparent in mRNA multiplexing. However, we believe that the framework presented here will be relevant for such future efforts.

In summary, we show here that the current ligation-based barcodes technique (Uziel et al. 2009; Tarasov et al. 2007; Zhu et al. 2008, 2009) introduces a large bias to the miRNA expression profiles. To overcome this problem, we have established a protocol for barcoding by PCR overlap that does not create bias. Recently, Illumina has introduced the TruSeq line of products that allows PCR multiplexing of miRNA. As it avoids the problematic ligation step, we believe it is likely that the TruSeq PCR-based multiplexing solution will behave like our PCR solution. One should remember that the increase in the throughput of the experiments by the use of barcodes inevitably causes a decrease in detection due to the lower number of reads per barcode. We show that this decrease can be modeled, allowing for a rational design of miRNA expression profiling with barcodes.

Methods

Multiplexing of miRNAs using PCR- and ligation-based barcoding

For a detailed description and a list of oligonucleotides, see the Supplemental Protocol.

PCR–based barcoding

Ligation of the 3′ adapter was conducted by incubating 1 μg of total RNA from the desired samples with 10 pmol of 3′ adenylated oligonucleotide, 10% DMSO, 20 U of RNaseInhibitor (Enzymatics Y924L), and 300 U of T4 RNA ligase 2 truncated (Enzymatics), for 1 h at 22°C. Following incubation, 10 pmol of 5′ adapter was added alongside 8 μM ATP (Enzymatics N207-10-L) and 20 U of T4 RNA Ligase 1 (Enzymatics L605L) and was incubated for 1 h at 20°C. A third of the reaction product was used for reverse transcription (RT) of the adapter-miRNA-adapter fragments using 25 pmol of 3′adapter compatible reverse transcription primer and 200 U of Superscript III (Invitrogen 18080-044) as previously described (Vigneault et al. 2008) followed by incubating for 30 min at 48°C. Following reverse transcription, PCR components were added directly to the RT reaction mixture by adding 1× HF Phusion buffer, 25 pmol of each PCR primers pairs (PCR1 and PCR2 barcoding primers), 250 μM dNTPs, and 1 U of Phusion hotstart DNA polymerase (NEB F-540L). The reaction was thermal cycled as follows: 30 sec at 98°C; 12 cycles of 10 sec at 98°C, 20 sec at 60°C, and 20 sec at 72°C; a final incubation of 5 min at 72°C; and pause at 4°C. The PCR products were purified on denaturing PAGE twice, as detailed in the Supplemental Protocol using crush and soak extraction (Vigneault et al. 2008). Each individually barcoded miRNA library was quality controlled on an Agilent Bioanalyzer and on a Nanodrop spectrophotometer and was combined at an equimolar concentration in one unique library containing all the barcoded samples prior to sequencing on a single Illumina lane.

Ligation–based barcoding

All steps of the protocol for ligation-based miRNA capture were conducted as described for the PCR-based barcoding protocol above with the only difference that the barcodes were introduced in the 5′ oligonucleotides during the ligation step to the miRNA.

Sequencing data filtering

The following RNA samples were used: (1) wild type mouse heart tissue RNA and (2) cardiac disease mouse tissue RNA (Teekakirikul et al. 2010), both extracted using the AMbion mirVana miRNA Isolation Kit (Ambion AM1561), and (3) FirstChoice Human Brain Reference RNA (Ambion AM6050). These three experimental conditions were sequenced on three lanes using two different flowcells (one for the human data and one for the two mouse tissues) of Illumina GAIIx instrument following the manufacturer's protocol. The total number of reads was ~18 M, ~20 M, and ~10 M reads for mouse normal hearts, mouse diseased hearts, and human brain, respectively. All the reads were filtered, demanding that (1) each read will have full barcode and (2) the quality of each read will not be below some threshold value (chosen to be 20) in more than three positions. In addition, sequences identified as 5′ or 3′ adaptors were removed. After adaptors trimming, reads with a length longer (more than 28 bases) or shorter (less than 15 bases) than the typical length of a mature miRNA were also removed: ~8 M, ~8 M, and ~5 M reads passed this filtering process for the mouse normal hearts, mouse diseased hearts, and human brain, respectively. The total number of reads per barcode after the filtering step is given in Supplemental Tables S1 through S3. Even after the removal of libraries with relatively a low number of counts, large differences (up to 18-fold) between the barcodes are observed in the mouse heart data. The differences were consistent between mouse normal hearts and mouse diseased hearts. Such large differences were not observed in the human brain barcode data that was constructed using the new, PCR-based, barcode protocol.

Constructing miRNAs profiles

The filtered reads were aligned using Bowtie (Langmead et al. 2009) against the mouse or human known pre-miRNAs taken from miRBase, allowing a total number of two mismatches. We demanded unique best hits (i.e., reads that cannot be aligned to other miRNAs with the same number of mismatches); ~5 M, ~5 M, and ~1.2 M reads were successfully aligned for the mouse normal hearts, mouse diseased hearts, and human brain, respectively. Only reads that were aligned against regions in the pre-miRNA that were annotated as mature miRNA by miRBase were further used to construct miRNA counts profile per barcode.

The miRNA profiles were normalized to allow comparison between them. We tested three types of normalizations: (1) scaling each miRNA profile by the total number of counts (Supplemental Tables S2, S3), (2) scaling each miRNA profile by the number of counts after trimming the higher and lower quartiles of changed miRNAs (as defined by their log-folds), and (3) scaling each miRNA profile such that the log-folds are distributed around zero after trimming the higher and lower quartiles of changed miRNAs (again, as defined by their log-folds). The third method is a variation of the one recently published (Robinson and Oshlach 2010); here we have weighted the log-folds using the standard deviation of log-folds from Poisson distribution of counts. All three methods of normalization gave almost the same results. Method number three is used in the following.

Measuring barcode bias

miRNA profiles from the same tissue but with different barcodes should be identical up to the statistical Poisson noise due to the

finite count numbers. Deviations between the profiles that are larger than expected for Poisson noise indicate a barcode bias. A direct way to measure the potential bias is to study the variance in the number of reads obtained in the different profiles. This variance is expected to be the sum of the Poisson noise variance, which is equal to the mean number of counts, and the variance of the barcode-specific efficiency, which is proportional to the square of the mean number of counts (Cameron and Trivedi 1998). Hence, the variance should follow the general form: Variance = mean + A $\times$ mean-squared, where A is related to the distribution P(λ) of the barcode-specific efficiencies:

$$A = (\langle\lambda^2\rangle - \langle\lambda\rangle^2)/\langle\lambda\rangle^2.$$

That is, A is the square of the relative standard deviation of λ. The smaller it is, the smaller is the relative spread of efficiencies among the different barcodes. By calculating the variance for all the miRNAs among the different barcodes as a function of the mean number of counts and by fitting the result to the above general form, one is able to estimate the constant A. This constant is equal to 0.63 and 0.66 using mouse normal hearts and mouse diseased hearts data, respectively, attesting for a barcode bias (Fig. 1B). Assuming a log-normal distribution for P(λ), A is given by A = $\exp(\sigma^2) - 1$. Thus A $\sim$ 0.65 corresponds to $\sigma \sim 0.7$, or a typical multiplicative factor-2 bias.

In contrast, for the new PCR-based barcodes, the constant A equals 0.0010 (human brain data), corresponding to $\sigma \sim 0.03$, or a typical multiplicative bias factor of 1.03. This bias is masked by the Poisson noise for all but the most highly expressed miRNAs, as seen in Figure 1E.

Comparing counts among the different miRNA profiles

In Figure 1, A and D, and Supplemental Figures S1 and S3, the miRNA counts are compared between the different miRNA profiles. We ask whether the discrepancies from the naïve y = x line in these figures originate from the expected Poisson noise. That is, for each point, which represents the counts of miRNA in two different conditions (biological or technical), can we reject the null hypothesis that these two numbers come from a Poisson distribution with means that differ only by a global normalization factor. The calculation steps were as follows: (1) the means were estimated using the actual counts (before normalization) and the global normalization factors; (2) the sum of squares of differences between the actual counts and the estimated means was computed; (3) the estimated means were used to generate 1000 realizations of counts from Poisson distribution; for each realization, the sum of squares of differences between the pseudo-counts and the estimated means was computed; and (4) points for which the sum of squares is in the top 1% of all the randomized 1000 realizations are marked as being outside the 99% region of the Poisson noise.

On average, only 73% of all points fall inside the Poisson noise region when we compare different barcodes in the same biological condition (Fig. 1A, all possible comparisons are plotted), again pointing toward barcode bias. When we compare a single barcode to a different barcode in the same biological condition, 10%–40% of the miRNAs are falsely detected as differentially expressed (i.e., they fall outside the Poisson noise region). Moreover, detecting differentially expressed miRNAs between normal and diseased mouse hearts using twofold change criteria gives dramatically different sets of genes depending on the barcodes used; choosing eight different couples of barcodes revealed <5% overlap between the sets of detected genes.

Finally, comparing different barcodes in the human brain tissue (PCR-based protocol) revealed that almost all points (97%) fall inside the Poisson noise region, as expected for no-bias barcodes (Fig. 1D). If we take into account the low residual barcode bias, modeled by a log-normal distribution with $\sigma \sim 0.03$ (see above), exactly 99% of the points fall inside the Poisson 99% noise region.

Analyzing mouse heart samples sequenced with PCR-based barcodes

One lane of Illumina's flow-cell (GAIIx instrument) was used to sequence eight libraries derived from normal and diseased mouse hearts (Teekakirikul et al. 2010); each library was marked by a different barcode using the PCR-based protocol. Three libraries were of three different normal samples; another three libraries were of three different diseased samples; one library was a pool of the three normal samples; and the last library was a pool of the three diseased samples. The sequencing data were filtered; expression profiles were constructed; and the data was normalized as described in the sections above (see Supplemental Tables S1, S4). If there is no barcode bias, one would expect that (1) the average of the three different samples for each miRNA will be the same as the pool, up to Poisson noise, and (2) the list of differentially expressed miRNAs will be almost the same if the data compared is either the three normal samples against the three diseased samples or the normal pool against the diseased pool. Indeed, by comparing the average of the three different samples for each miRNA to the pool, we see that exactly 99% of the points fall inside the Poisson noise region after we take into account the added noise with $\sigma \sim 0.03$ (see above) to the realizations of counts (Supplemental Fig. S3). These data demonstrate that the quality of the PCR-based barcodes is reproducible. Two lists of differentially expressed miRNAs were created as described above, that is, one using the three individual samples added together and the other for the pooled samples, demanding a stringent cutoff of twofold change and minimum number of 50 reads in each condition in order to avoid Poisson noise. Comparing these two lists of differentially expressed miRNAs resulted with a high overlap of 80% (for the list of overlapping differentially expressed miRNAs, see Supplemental Table S5). Thus, one can use libraries marked with these barcodes for detecting differentially expressed genes. Detecting differentially expressed miRNAs was almost impossible using the ligation-based barcodes due to the large bias that they introduced ($\sigma \sim 0.7$, i.e., twofold change).

Estimating detection efficiency as a function of the number of barcodes

The use of barcodes in multiplexing strategy results in expected decrease in the number of counts per library. This decrease can cause (1) a reduced number of miRNAs to be detected as expressed in a given library and (2) a reduced number of miRNAs to be detected as differentially expressed when one compares two libraries. These two effects can be modeled if the distribution of expression levels among the various miRNAs is known. We observed that this distribution follows Zipf's with an exponential cutoff for the two different tissues: the human brain and the mouse heart (Fig. 2A). Given the distribution of counts, estimating the decrease in detection of expressed miRNAs is straightforward: On average, a miRNA with number of counts equal or higher than the number of barcodes used will be detected. It turns out that the decrease in detection of expressed miRNAs as a function of the barcodes used depends only weakly on the investigated tissue; both human brain and mouse heart gave similar results (Fig. 2B). We estimated the decrease in the detection of differentially expressed miRNAs by identifying miRNAs with changes in expression that are significantly higher than the standard deviation of

the miRNA counts before and after the reduction in counts caused by the use of barcodes (Fig. 2B). The standard deviation in the miRNA counts is known given that the variation in the miRNAs counts follows a Poisson distribution. As the human brain tissue displays a wider distribution of miRNAs compared with mouse heart tissue (Fig. 2A), it is expected that the decrease in the number of differentially expressed miRNAs by using barcodes will be stiffer compared with mouse heart tissue. Indeed, this trend is observed in Figure 2B, but the behavior of the two tissues is similar and the difference between the two curves is within 10%. We confirmed the estimated decrease in the number of differentially expressed miRNAs by comparing with experimental data. The experimental data were derived by comparing the expression of all the miRNAs in one experimental condition (mouse normal heart) to the second experimental condition (mouse diseased heart) using libraries generated with the same barcodes. All the miRNAs that had changes in expression that cannot be explained by Poisson noise were recorded. The detection procedure is as described in the section Comparing Counts Among the Different miRNA Profiles above but with Bonferroni correction. We performed all the possible comparisons between libraries with the same barcode along the different tissues (Fig. 2C).

Data access

The sequence data from this study have been submitted to the NCBI Sequence Read Archive under accession no. SRA029326.

Acknowledgments

This work was supported by the Center for Excellence in Genome Sciences grant from the National Human Genome Research Institute. F.V. is supported by a Canadian Institutes of Health Research and Ragon Institute Fellowship. J.G.S is supported by grants from the NIH, NHLBI, the SysCODE Consortium (NIH) and the Fondation Leducq. This work was partially supported by a grant from the United States-Israel Binational Science Foundation (grant no. 2009290), Jerusalem, Israel.

Note added in proof

Ligation biases in miRNA sequencing were also independently described in a recent publication (Hafner et al. 2011), supporting the findings reported herein.

References

Bartel DP. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116:** 281–297.

Cameron AC, Trivedi PK. 1998. *Regression analysis of count data*. Cambridge University Press, Cambridge, UK.

Creighton CJ, Reid JG, Gunaratne PH. 2009. Expression profiling of microRNAs by deep sequencing. *Brief Bioinform* **10:** 490–497.

de Hoon MJ, Taft RJ, Hashimoto T, Kanamori-Katayama M, Kawaji H, Kawano M, Kishima M, Lassmann T, Faulkner GJ, Mattick JS, et al. 2010. Cross-mapping and the identification of editing sites in mature microRNAs in high-throughput sequencing libraries. *Genome Res* **20:** 257–264.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* **36:** e105. doi: 10.1093/nar/gkn425.

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. 2008. miRBase: tools for microRNA genomics. *Nucleic Acids Res* **36:** D154–D158.

Hafner M, Renwick N, Brown M, Mihailovic A, Holoch D, Lin C, Pena JT, Nusbaum JD, Morozov P, Ludwig J, et al. 2011. RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA* **17**. doi: 10.1261/rna.2799511.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** R25. doi: 10.1186/gb-2009-10-3-r25.

Linsen SE, de Wit E, Janssens G, Heater S, Chapman L, Parkin RK, Fritz B, Wyman SK, de Bruijn E, Voest EE, et al. 2009. Limitations and possibilities of small RNA digital gene expression profiling. *Nat Methods* **6:** 474–476.

Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, et al. 2005. MicroRNA expression profiles classify human cancers. *Nature* **435:** 834–838.

Morin RD, O'Connor MD, Griffith M, Kuchenbauer F, Delaney A, Prabhu AL, Zhao Y, McDonald H, Zeng T, Hirst M, et al. 2008. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res* **18:** 610–621.

Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11:** R25. doi: 10.1186/gb-2010-11-3-r25.

Tarasov V, Jung P, Verdoodt B, Lodygin D, Epanchintsev A, Menssen A, Meister G, Hermeking H. 2007. Differential regulation of microRNAs by p53 revealed by massively parallel sequencing: miR-34a is a p53 target that induces apoptosis and G1-arrest. *Cell Cycle* **6:** 1586–1593.

Teekakirikul P, Eminaga S, Toka O, Alcalai R, Wang L, Wakimoto H, Nayor M, Konno T, Gorham JM, Wolf CM, et al. 2010. Cardiac fibrosis in mice with hypertrophic cardiomyopathy is mediated by non-myocyte proliferation and requires Tgf-beta. *J Clin Invest* **120:** 3520–3529.

Uziel T, Karginov FV, Xie S, Parker JS, Wang YD, Gajjar A, He L, Ellison D, Gilbertson RJ, Hannon G, et al. 2009. The miR-17~92 cluster collaborates with the Sonic Hedgehog pathway in medulloblastoma. *Proc Natl Acad Sci* **106:** 2812–2817.

Vigneault F, Sismour AM, Church GM. 2008. Efficient microRNA capture and bar-coding via enzymatic oligonucleotide adenylation. *Nat Methods* **5:** 777–779.

Zhu QH, Spriggs A, Matthew L, Fan L, Kennedy G, Gubler F, Helliwell C. 2008. A diverse set of microRNAs and microRNA-like small RNAs in developing rice grains. *Genome Res* **18:** 1456–1465.

Zhu JY, Pfuhl T, Motsch N, Barth S, Nicholls J, Grasser F, Meister G. 2009. Identification of novel Epstein-Barr virus microRNA genes from nasopharyngeal carcinomas. *J Virol* **83:** 3333–3341.

Received February 3, 2011; accepted in revised form July 5, 2011.