Assessment of microRNA differential expression and detection in multiplexed small RNA sequencing data

JOSHUA D. CAMPBELL,^{1,2} GANG LIU,² LINGQI LUO,² JI XIAO,² JOSEPH GERREIN,^{1,2} BRENDA JUAN-GUARDELA,³ JOHN TEDROW,³ YURIY O. ALEKSEYEV,⁴ IVANA V. YANG,⁵ MICK CORRELL,⁶ MARK GERACI,⁷ JOHN QUACKENBUSH,⁶ FRANK SCIURBA,³ DAVID A. SCHWARTZ,⁵ NAFTALI KAMINSKI,³ W. EVAN JOHNSON,^{1,2} STEFANO MONTI,^{1,2} AVRUM SPIRA,^{1,2} JENNIFER BEANE,² and MARC E. LENBURG^{1,2,4}

¹Bioinformatics Graduate Program, Boston University, Boston, Massachusetts 02215, USA

²Section of Computational Biomedicine, Boston University Medical Center, Boston, Massachusetts 02118, USA

³Simmons Center for Interstitial Lung Disease and Department of Medicine, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania 15213, USA

⁴Department of Pathology and Laboratory Medicine, Boston University School of Medicine, Boston, Massachusetts 02118, USA

⁵Center for Genes, Environment and Health and Department of Medicine, National Jewish Health, Denver, Colorado 80206, USA

⁶Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, Massachusetts 02215, USA

⁷Department of Medicine, University of Colorado School of Medicine, Aurora, Colorado 80045, USA

ABSTRACT

Small RNA sequencing can be used to gain an unprecedented amount of detail into the microRNA transcriptome. The relatively high cost and low throughput of sequencing bases technologies can potentially be offset by the use of multiplexing. However, multiplexing involves a trade-off between increased number of sequenced samples and reduced number of reads per sample (i.e., lower depth of coverage). To assess the effect of different sequencing depths owing to multiplexing on microRNA differential expression and detection, we sequenced the small RNA of lung tissue samples collected in a clinical setting by multiplexing one, three, six, nine, or 12 samples per lane using the Illumina HiSeq 2000. As expected, the numbers of reads obtained per sample decreased as the number of samples in a multiplex increased. Furthermore, after normalization, replicate samples included in distinct multiplexes were highly correlated ($R > 0.97$). When detecting differential microRNA expression between groups of samples, microRNAs with average expression >1 reads per million (RPM) had reproducible fold change estimates (signal to noise) independent of the degree of multiplexing. The number of microRNAs detected was strongly correlated with the $\log_2$ number of reads aligning to microRNA loci ($R = 0.96$). However, most additional microRNAs detected in samples with greater sequencing depth were in the range of expression which had lower fold change reproducibility. These findings elucidate the trade-off between increasing the number of samples in a multiplex with decreasing sequencing depth and will aid in the design of large-scale clinical studies exploring microRNA expression and its role in disease.

Keywords: microRNA; differential expression; detection; sequencing; multiplexing; coverage

INTRODUCTION

MicroRNAs are short, noncoding RNAs that can down-regulate a target mRNA by binding to the 3' UTR (Winter et al. 2009; Thomas et al. 2010). Advances in sequencing technology have produced new methods to interrogate the entire microRNA transcriptome at single base resolution. High-throughput sequencing can be used to quantify and compare the levels of microRNA expression between biological conditions or disease states. Additionally, as sequencing does not require a priori knowledge about the microRNA sequences, it can be used to find novel sequences such as novel microRNA and microRNA isoforms (Friedländer et al. 2008; Cloonan et al. 2011; Perdomo et al. 2013).

Cost and throughput are critical factors in the design of large-scale clinical studies that examine microRNA expression and its relationship with disease. While sequencing methods have the capacity to characterize the transcriptome at an unprecedented level of detail, they can be costly and have lower throughput than array-based methods. Sequencers such as the Illumina HiSeq 2000 have flow cells with a fixed number of lanes (e.g., eight), limiting the total number of samples that can be assayed per run and increasing the cost per sample. Multiplexing allows multiple samples to be run per lane and potentially mitigates these disadvantages.

© 2015 Campbell et al. This article is distributed exclusively by the RNA Society for the first 12 months after the full-issue publication date (see <http://rnajournal.cshlp.org/site/misc/terms.xhtml>). After 12 months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

Corresponding author: mlenburg@bu.edu

Article published online ahead of print. Article and publication date are at <http://www.rnajournal.org/cgi/doi/10.1261/rna.046060.114>.

Multiplexing is often accomplished by attaching adapters with a unique sequence index (also known as a barcode) to the ends of the RNA or cDNA molecules in each sample (Alon et al. 2011; Hafner et al. 2012; Vigneault et al. 2012). Samples with different indices can then be pooled and sequenced in the same lane. As both the small RNA molecule and index are sequenced, each read obtained from a lane can be assigned back to the original biological sample from which it was derived.

The trade-off between increasing the number of samples sequenced per lane and decreasing the coverage per sample has not been fully explored in the context of small RNA sequencing. In this study, we assess the effect of varying levels of multiplexing on microRNA differential expression and detection using lung tissue samples collected in a clinical setting.

RESULTS

Sequencing of multiplexed small RNA samples

We sequenced the small RNA of lung tissue samples from the Lung Genome Research Consortium ($n = 15$). Each sample was given a unique index (Supplemental Table 1) and one to 12 samples were multiplexed within each lane (Fig. 1A). Between 58 and 85 million reads were obtained for each lane. As expected, the average number of total and aligned reads per sample decreased as the number of samples multiplexed per lane increased ($P < 0.001$) (Fig. 1B; Supplemental Table 2). No significant associations were observed between the percentage of aligned reads per sample and multiplex ($P = 0.83$) (Fig. 1C) or percentage of reads with a mismatch ($P = 0.92$) (Fig. 1D) showing that multiplexing does not affect the fraction of high-quality reads obtained per sample.

Assessing the effect of multiplexing on microRNA expression

Ideally, after normalization, expression of any given microRNA would be the same across replicate samples sequenced at different depths of coverage in different multiplexes. Using normalized microRNA expression values, we computed all pairwise correlations between replicate samples that were sequenced in different multiplexes at different depths of coverage and found that within-replicate correlation was very high ($R > 0.97$; Pearson correlation) (Fig. 2A;

FIGURE 1. Sequencing of multiplexed small RNA samples. (A) Using the Illumina TruSeq kit, lung tissue samples were given a unique index, placed in pools consisting of one, three, six, nine, or 12 samples, and sequenced using the Illumina HiSeq 2000. The same six samples were sequenced in the 6-plex, 9-plex, and 12-plex (gray) and used in differential expression analysis. (B) As expected, the total number of reads per sample decreased as more samples were included in the multiplex. (C) Percentage of aligned reads and (D) percentage of reads that aligned with a mismatch were not significantly associated with numbers of samples within a lane indicating that multiplexing does not affect read quality.

see Supplemental Fig. 1 for all correlations between the replicate samples sequenced in different multiplexes). We also sought to determine whether the proportion of reads for each microRNA in replicate samples sequenced in different multiplexes differed from what would be expected under random sampling using a methodology adapted from Marioni et al. (2008), which measured lane-to-lane variability for large RNA sequencing. In the absence of a multiplexing effect, the reads for a microRNA in one multiplex are expected to be a random sample of the reads for that microRNA from both multiplexes. We observed that on average 22.5 microRNAs ($\sim 1.1\%$) displayed significantly different proportions between replicates (FDR $q < 0.05$) (Fig. 2B; see Supplemental Fig. 2 for all comparisons between the replicate samples sequenced in different multiplexes). For example, miR-21-5p had 1.68% of total aligned reads in sample B in the 3-plex (corresponding to 14.039 $\log_2$ RPM) while the proportion of reads for this microRNA in the same sample significantly decreased to 1.65% in the 12-plex (corresponding to 14.012 $\log_2$ RPM; $P = 4.8 \times 10^{-6}$; Fisher's exact test).

As more microRNAs displayed significant differences in the proportion of reads between replicates than what is

FIGURE 2. Assessment of microRNA expression across different multiplexes. (A) All pairwise Pearson correlations were calculated between the samples in the 1-plex, 3-plex, 6-plex, 9-plex, and 12-plex. The correlation for sample B between the 3-plex and the 12-plex is shown as an example ($R = 0.98$). (B) For each sample sequenced in different multiplexes, we performed a test based on the hypergeometric distribution to determine if the proportions of reads for an individual microRNA were significantly different for a replicate sample sequenced in two multiplexes than what would be expected by chance. The x-axis shows the average RPM expression for each microRNA across the 3-plex and the 12-plex. The y-axis shows the difference in RPM expression for each microRNA between the 3-plex and the 12-plex. Positive and negative values on the y-axis indicate that the RPM values in the 12-plex were smaller or larger than RPM values in the 3-plex, respectively. Red indicates that the microRNA had a significant difference in the proportions of reads between multiplexes ($FDR\ q < 0.05$). (C) The distribution of Pearson correlation coefficients was significantly higher between replicate samples sequenced in different multiplexes (representing technical variability) compared with the correlations between different samples sequencing within the same multiplex (representing biological variability) and between different samples sequenced in different multiplexes (representing technical + biological variability). (D) Likewise, the numbers of microRNAs with a significant difference in the proportion of reads were lower when comparing replicate samples in different multiplexes than when comparing different biological samples. Asterisk indicates $P < 0.001$ from a Wilcoxon rank-sum test.

expected by chance, we next compared the within-replicate technical variability owing to multiplexing to the variability between different biological samples within and across multiplexes. We found that the within-replicate correlations were significantly higher than the correlations between different

biological samples sequenced within the same multiplex or sequenced in different multiplexes ($P < 0.001$) (Fig. 2C). Similarly, the numbers of microRNAs with a significant difference in the proportion of reads were lower between replicates than between different biological samples ($P < 0.001$) (Fig. 2D). Taken together, these results indicate that the technical variability introduced by multiplexing is less than the biological variability found between different lung tissue samples.

MicroRNA fold change estimates across different multiplexes

An important aspect of the genome-wide profiling of microRNA is the ability to detect reproducible differential expression estimates between two or more distinct biological conditions. We assessed the effect of multiplexing on differential expression by calculating the interaction effect for microRNA fold change estimates between the multiplex and class label (Fig. 3A). In the ideal case, the fold change estimate for any given microRNA between two groups of samples will be the same when calculated in either multiplex, resulting in an interaction effect of zero. We found that only nine microRNAs exhibited a significant interaction effect on any comparison ($P < 0.05$) (Fig. 3B) indicating that the fold changes for most microRNAs are unaffected by the multiplexing level. However, we did observe that the variability in the magnitude of the interaction effect was greater for microRNAs with lower average expression (i.e., more microRNAs with lower expression had an interaction effect farther away from zero compared with microRNAs with higher expression). The correlation between fold changes of the 6-plex and the 12-plex using all microRNAs was $R = 0.51$. To determine whether the correlation would improve by excluding microRNAs with lower expression levels,

we iteratively removed the microRNA with the lowest average RPM across both plexes, one at a time, and recalculated the correlation between the fold changes. We observed a rapid improvement in the correlation as more lowly expressed microRNAs were removed (Fig. 3C). Indeed, the correlation

FIGURE 3. Effect of multiplexing on fold change reproducibility. (A) Fold changes between all possible combinations of three versus three samples ($n = 10$) were calculated within the 6-plex and 12-plex using DESeq (Anders and Huber 2010). In addition, the interaction effect between sample class and multiplex was calculated which determines the degree to which fold changes in the 6-plex and 12-plex are discordant. (B) The difference between the fold change estimates between the two levels of multiplexing is plotted as a function of average expression level. While only nine microRNAs displayed a significant interaction effect on any comparison (black; $P < 0.05$), the microRNAs with lower average expression tended to show the greatest difference in fold change estimates between the multiplexes. (C) The correlation between fold changes of different multiplexes using all microRNAs was $R = 0.51$. However, when microRNAs with lower average expression were iteratively removed, the correlation between fold changes rapidly improved. (D) The fold change estimates in the 12-plex are plotted against the fold change estimates in the 6-plex for all microRNAs. MicroRNAs with average RPM > 1 across both plexes are colored black while those with average RPM < 1 across both plexes are gray. These results suggest that the fold change estimates for microRNAs with average RPM > 1 are largely reproducible when the samples are sequenced to greater depths.

increased to 0.88 when microRNAs with average expression RPM < 1 were removed from the correlation (Fig. 3D). Similar patterns were observed when comparing fold changes for the 6-plex versus the 9-plex and the 9-plex versus the 12-plex (Supplemental Fig. 3). Furthermore, we performed hierarchical clustering using microRNAs with average RPM > 1. We found that the same biological replicate samples sequenced in different multiplexes clustered together and the samples did not cluster according to multiplex or lane (Fig. 4). These results suggest that, given an approximately equal sequencing depth of ~ 5 million reads across samples, the differential expression estimates of microRNAs with an average expression of ~ 1 RPM will be largely reproducible at greater sequencing depths.

Detection of annotated microRNAs

In addition to performing differential expression between biological groups, small RNA sequencing can be used to characterize different types of RNAs present within a given sample. However, the ability to detect rare transcripts is dependent on the sequencing depth. We observed that the number of microRNAs detected in each sample (i.e., microRNAs with at least two reads) was strongly correlated with the total number of aligned reads aligning to a microRNA loci in each sample in $\log_{10}$ scale ($R = 0.96$) (Fig. 5A). The number of microRNAs detected with average expression RPM < 1 increased as the read coverage increased owing to multiplexing fewer samples (Fig. 5B). In contrast, the number of detected microRNAs with average RPM > 1 remained relatively constant across the multiplexes. In order to determine if sequencing depth had an effect on microRNA detection independently of multiplexing, we fit a linear model with the number of detected reads as the response and both the multiplex and number of aligned reads as predictors. Sequencing depth remained strongly associated with the number of detected microRNAs ($P = 4.2 \times 10^{-9}$) while controlling for multiplex suggesting that the primary factor in microRNA detection is sequencing depth. These results show that increasing the sequencing depth by reducing the number of samples multiplexed in each lane can result in the detection of additional lowly expressed microRNA, allowing for discovery and greater characterization of microRNA sequences.

The effect of sample size and sequencing depth on sensitivity to detect microRNA differential expression

In many studies where the total number of sequencing reads is fixed owing to a set budget, the question arises as to whether fewer samples should be sequenced to greater depths per sample or more samples should be sequenced with less depth per sample in order to achieve the most power to detect differential expression. To characterize the trade-off between sequencing depth and sample size, we performed a simulation analysis in which we “spiked in” a range of increasing fold changes into randomly selected microRNAs from five different sample sets. Each sample set had approximately

FIGURE 4. Clustering of replicate sequenced in different multiplexes. MicroRNAs with an average RPM > 1 were used to perform hierarchical clustering with average linkage. Replicate samples sequenced in different multiplexes clustered together rather than different biological samples from the same multiplex.

the same number of total microRNA counts, but the sets with more samples had proportionally fewer counts per sample. The sample sets included sample sizes of four versus four, six versus six, 12 versus 12, 24 versus 24, and 48 versus 48 with average sequencing depths of 7.6, 6.7, 3.4, 1.7, and 0.8 million reads, respectively. After applying DESeq and identifying microRNAs with FDR $q < 0.05$, we determined the smallest fold change which obtained a median sensitivity of $\geq 80\%$ within bins of microRNAs grouped together by overall expression levels.

In microRNA bins with higher average expression (RPM > 1), for a fixed quantity of sequencing, increasing the sample size led to the detection of smaller fold changes despite the decreased sequencing depth (Fig. 6; Supplemental Fig. 4). However, increasing the sequencing depth decreased power presumably owing to the decreased number of samples. For microRNAs with lower expression (RPM < 1), the two comparisons with the lowest sample sizes (four versus four and six versus six) and the comparison with the lowest sequencing depth (48 versus 48) failed to detect fold changes below 13 suggesting that both insufficient sequencing depth and lack of an adequate sample size can hinder differential expression detection at lower expression levels. These results demonstrate that the trade-off between sequencing depth and sample size is most pronounced in microRNAs with lower expression levels. In contrast, sample size is the most important factor in detecting differential expression among highly expressed microRNAs.

DISCUSSION

Many factors are important when determining a study design for sequencing small RNA samples including the total number of samples to be sequenced, the depth of coverage required per sample to meet the objectives of the study, and the cost of achieving that sequencing depth. The primary goal of this study was to assess the effect of varying levels of coverage owing to multiplexing on microRNA differential expression and detection. We did not assess technical variability introduced in library preparation instead assessed the effect of sequencing the same library to different depths of coverage in different multiplexes.

As expected, increasing the number of samples pooled within a lane caused the level of coverage to decrease for each sample within that pool. When comparing normalized microRNA expression for the same sample sequenced in different multiplexes, we found that the overall sample correlation was high ($R > 0.97$). More microRNAs displayed significant differences in proportion of reads between replicate samples sequenced in different multiplexes than what was expected by chance. This finding highlights the importance of careful study design to ensure that sequencing parameters such as number of reads per sample are not confounded with phenotypes of interest. Importantly, we observed that the sample correlations were higher when comparing replicate samples in different multiplexes versus different biological samples within a multiplex. A similar pattern was seen when examining the numbers of microRNAs with significantly different proportions of reads across replicate samples indicating that the technical variability owing to multiplexing is less than the variability between the different lung tissue samples we sequenced, a prerequisite necessary for being able to detect meaningful biological differences between groups of samples.

FIGURE 5. Detection of annotated microRNAs in samples across different levels of multiplexing. (A) The number of microRNAs detected with at least one read is correlated with the $\log_2$ number of aligned reads ($R = 0.96$) suggesting an exponential relationship. (B) The number of microRNAs detected in each multiplex is shown for microRNAs with average RPM < 1 (left) and average RPM > 1 (right). When sequencing depth is decreased by multiplexing more samples, the number of detected microRNAs with RPM < 1 decreased while the number of detected microRNAs with RPM > 1 remained relatively constant.

FIGURE 6. Examining the trade-off between having more samples at lower sequencing depth versus having fewer samples with higher sequencing depth. Fold changes ranging from 1.25 to 25 were spiked into randomly selected microRNAs in sample sets of different sizes with different levels of sequencing depth. Sets with more samples had fewer numbers of reads per sample on average. MicroRNAs were binned according to average expression across all samples in the 9-plex. The x-axis is the median $\log_2$ RPM for all microRNAs within a bin. The y-axis is the smallest fold change that had median sensitivity of at least 80% across all microRNAs within that bin. These results indicate that, for microRNAs with higher average expression ($\text{RPM} > 1$), larger sample sizes are better for detecting smaller fold changes compared with smaller sample sizes.

The ability to reliably detect microRNAs that are differentially expressed between biological conditions is a critical component of many studies. We performed a differential expression analysis by comparing three versus three samples where all six samples were sequenced to different depths of coverage in different multiplexes. Our results suggest that, given an average sequencing depth of ~ 5 million reads per sample, the differential expression estimates for microRNAs with an average expression level of $\text{RPM} \geq 1$ will be largely reproducible ($R > 0.88$) when the same samples are sequenced to greater depths. We note that in this study we examined all possible combinations of three versus three lung tissue samples. In a study where the biological differences between groups of samples are expected to be larger (e.g., when comparing cases versus controls), fold changes for microRNAs with an average $\text{RPM} < 1$ may also be reproducible between higher and lower sequencing depths. Overall, we found that small RNA sequencing can be used in combination with multiplexing to generate reproducible fold change estimates across different comparisons for reasonably expressed microRNAs.

One of the advantages of sequencing a sample to greater depth is the detection of additional microRNAs. Previous work by Alon et al. (2011) found that the number of microRNAs detected in mouse and human heart and brain samples, which were sequenced to a depth between 1 and 10 million, followed a Zipf's law with an exponential cutoff. In this study, we sequenced samples to an even greater depth (up to 49.4 million aligned reads in the 1-plex) and similarly found that the $\log_2$ number of reads aligning to microRNA loci was highly correlated to the number of detected

microRNAs in human lung ($R = 0.96$) suggesting a power-law relationship. Furthermore, the microRNA detection rate remained strongly associated with sequencing depth even while controlling for multiplex. These observations suggest that sequencing depth is the primary factor in low abundance microRNA detection. The additional microRNAs detected with increased sequencing depth reads tended to fall in the range of expression which had lower fold change reproducibility (e.g., average expression < 1 RPM). Therefore, independent validation may be required to be confident in their differential expression estimates with respect to biological phenotypes.

Finally, we examined the trade-off between using a fixed amount of sequencing to achieve either higher sequencing depth or profile larger numbers of samples by measuring our ability to detect differentially expressed microRNAs in different sample sets of varying sizes and average sequencing depths per sample. Higher average expression, larger fold changes, and larger sample sizes all contributed to increased sensitivity. For microRNAs with lower expression, only relatively large fold changes could be detected with high sensitivity potentially owing to the increased sampling noise inherent at this level of expression. In studies examining microRNA expression in complex and heterogeneous clinical cohorts, the fold changes between case and control groups may be relatively small and thus not easily detectable at lower expression levels even with greater sequencing depth. Additionally, when examining microRNAs with higher expression, larger sample sizes clearly out performed smaller sample sizes at detecting smaller fold changes. These observations collectively indicate that increasing the sample size is the most beneficial factor in increasing the power to detect the majority of differentially expressed microRNAs in clinical cohorts.

Overall, the findings in this study elucidate the trade-off between increasing the number of samples in a multiplex and decreasing sequencing depth. They will aid researchers in designing large-scale microRNA expression studies where the choice of sequencing depth and numbers of samples within a multiplexing are critical to feasibility in terms of cost and throughput.

MATERIALS AND METHODS

Sequencing the small RNA of lung tissue samples

Multiplexed small RNA sequencing was conducted on the Illumina HiSeq 2000 sequencer according to the manufacturer's protocol. Briefly, 1 μg of total RNA from each sample was used for library preparation with the TruSeq Small RNA Sample Prep Kit (Illumina). RNA adapters were ligated to 3' and 5' end of the RNA molecule and the adapter-ligated RNA was reverse transcribed into single-stranded cDNA. The cDNA was then PCR amplified using a common primer and a primer containing one of 12 index sequences. The introduction of the six-base index tag at the PCR step allowed multiplexed sequencing of different samples in a single lane of a

flowcell. One to 12 of these indexed cDNA libraries were pooled in equal amount and gel purified together. Each library was hybridized to one lane of the eight-lane single-read flowcell on a cBot Cluster Generation System (Illumina) using TruSeq Single-Read Cluster Kit (Illumina). The clustered flowcell is then loaded onto HiSeq 2000 sequencer for a multiplexed sequencing run that consists of a standard 36-cycle sequencing read with the addition of a seven-cycle index read. A PhiX library was sequenced in lane 4 and used for calibration.

Quantifying microRNA expression

The 3' adapter sequence was trimmed using the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). Reads longer than 15 nt were aligned to hg19 using Bowtie v0.12.7 (Langmead et al. 2009) allowing up to one mismatch and up to 10 genomic locations. MicroRNA expression was quantified by counting the number of reads aligning to mature microRNA loci (miRBase v18) using Bedtools v2.9.0 (Griffiths-Jones 2004; Quinlan and Hall 2010). MicroRNA counts within each sample were normalized to RPM values by adding a pseudocount of one to each microRNA, dividing by the total number of reads that aligned to all microRNA loci within that sample, multiplying by 1×10^6 , and then applying a $\log_2$ transformation.

Assessing multiplexing effects on microRNA expression

Based on the methodology used by Marioni et al. (2008) to examine lane-to-lane variability with large RNA sequencing data, we used the Fisher's exact test which is based on the hypergeometric distribution to determine if the proportion of reads for an individual microRNA significantly varied between the same biological sample sequenced in different multiplexes (replicate sample), between different biological samples sequenced within the same multiplex, or different biological samples sequenced in different multiplexes. If sample S was sequenced in lanes X and Y , each with different levels of multiplexing (e.g., 6-plex and 12-plex), then in the absence of a multiplexing effect, reads for microRNA M in sample S from lane X will be a random sample of reads for microRNA M in sample S from both lanes X and Y . Let T_X and T_Y be the total number of reads aligning to microRNA loci in lanes X and Y , respectively. For each microRNA M in sample S that was sequenced in both X and Y , the contingency table for the Fisher's exact test consisted of the counts for M in lane X ($C_{S,M,X}$), the counts for M in lane Y ($C_{S,M,Y}$), the total number of aligned reads in X not including M ($T_X - C_{S,M,X}$), and the total number of aligned reads in Y not including M ($T_Y - C_{S,M,Y}$). P -values were corrected for multiple hypothesis testing using the false discovery rate (FDR) (Benjamini and Hochberg 1995). MicroRNAs with FDR $q < 0.05$ were considered to have a significant difference in the proportion of reads.

Comparing fold change estimates across multiplexes

Six samples were sequenced in the 6-plex, 9-plex, and 12-plex including samples E, F, G, H, I, and J (Fig. 1A). We assessed the effect of multiplexing on differential expression by calculating the interaction effect for microRNA fold change estimates between the multiplex and class label using DESeq version 1.10.1 (Anders

and Huber 2010). This was performed for all 10 combinations of three versus three samples (i.e., Combination 1: samples E, F, G versus H, I, J; Combination 2: samples H, F, G versus samples E, I, J; etc.) in the comparison between each multiplex (6-plex versus 9-plex, 6-plex versus 12-plex, and 9-plex versus 12-plex). MicroRNAs whose fold changes could not be calculated owing to having only zero counts for all samples in one or both groups were excluded.

Assessing the trade-off between sample size and sequencing depth

Nine hundred sixty-two microRNAs that were detected with at least two reads in two samples in the 6-plex, 9-plex, or 12-plex were used in this analysis. We used two different approaches for generating sample sets of different sizes and average sequencing depths. First, the samples from the 9-plex and 12-plex were used for the four versus four and six versus six comparisons, respectively. One sample was randomly left out of the 9-plex. To achieve greater sample sizes of 12 versus 12, 24 versus 24, and 48 versus 48, the counts for the samples in the 12-plex were divided by 2, 4, or 8, respectively, and rounded to the nearest integer. This simulates the predicted sequencing depth that these samples would have had if they were included in a 24-plex, 48-plex, or 96-plex, respectively. Then for each microRNA, these depth-adjusted counts were sampled with replacement until the desired number of samples was obtained. This bootstrapping procedure was repeated in each of the 1000 iterations described below.

Fold changes ranging from 1.25 to 25 in increments of 0.25 were spiked into a subset of microRNAs using the following procedure:

1. Raw read counts were normalized to the percentage of microRNA reads within the sample.
2. Half of the samples were randomly chosen to be cases and while the other half to be controls.
3. The normalized expression values for 50 randomly selected microRNAs in the cases were scaled to have a mean X -fold greater than the mean of the controls (i.e., up-regulated in cases).
4. The normalized expression values for another 50 randomly selected microRNAs in the controls were scaled to have a mean X -fold greater than the mean of the cases (i.e., down-regulated in cases).
5. The normalized and scaled expression values within each sample were multiplied by the total number of reads originally obtained for that sample (same number used as the denominator in step 1) and rounded to the nearest integer to produce raw counts. These raw counts were used as input into DESeq.
6. This process was repeated 1000 times for each fold change examined.
7. Sensitivity for each microRNA was defined as the total number of times the microRNA was randomly selected and had an FDR $q < 0.05$ divided by the total number of times the microRNA was randomly selected across all iterations.

To more easily assess the effect of overall expression levels on sensitivity, microRNAs were placed into equally spaced bins based on their average expression in the 9-plex. The bins ranged from -3 to $13 \log_2$ RPM and increased in increments of 1. The median sensitivity for all microRNAs within a bin was used as the representative sensitivity for that expression level. The minimum fold change to

achieve a median of $\geq 80\%$ sensitivity was computed for each bin within each sample set.

DATA DEPOSITION

The data for these experiments are publicly available in the Gene Expression Omnibus (GEO) under the accession GSE56861.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

We thank the reviewers for their constructive feedback. This work was funded by the National Heart, Lung, and Blood Institute (RC2HL101715-01) and the National Cancer Institute's Early Detection Research Network (U01CA152751) of the National Institutes of Health.

Received April 28, 2014; accepted November 7, 2014.

REFERENCES

- Alon S, Vigneault F, Eminaga S, Christodoulou DC, Seidman JG, Church GM, Eisenberg E. 2011. Barcoding bias in high-throughput multiplex sequencing of miRNA. *Genome Res* **21**: 1506–1511.
- Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11**: R106.
- Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc* **57**: 289–300.
- Cloonan N, Wani S, Xu Q, Gu J, Lea K, Heater S, Barbacioru C, Steptoe AL, Martin HC, Nourbakhsh E, et al. 2011. MicroRNAs and their isomiRs function cooperatively to target common biological pathways. *Genome Biol* **12**: R126.
- Friedländer MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, Rajewsky N. 2008. Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol* **26**: 407–415.
- Griffiths-Jones S. 2004. The microRNA Registry. *Nucleic Acids Res* **32**: D109–D111.
- Hafner M, Renwick N, Farazi TA, Mihailović A, Pena JT, Tuschl T. 2012. Barcoded cDNA library preparation for small RNA profiling by next-generation sequencing. *Methods* **58**: 164–170.
- 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.
- Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* **18**: 1509–1517.
- Perdomo C, Campbell JD, Gerrein J, Tellez CS, Garrison CB, Walser TC, Drizik E, Si H, Gower AC, Vick J, et al. 2013. MicroRNA 4423 is a primate-specific regulator of airway epithelial cell differentiation and lung carcinogenesis. *Proc Natl Acad Sci* **110**: 18946–18951.
- Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.
- Thomas M, Lieberman J, Lal A. 2010. Desperately seeking microRNA targets. *Nat Struct Mol Biol* **17**: 1169–1174.
- Vigneault F, Ter-Ovanesyan D, Alon S, Eminaga S, Christodoulou DC, Seidman JG, Eisenberg E, Church GM. 2012. High-throughput multiplex sequencing of miRNA. *Curr Protoc Hum Genet* Chapter 11: Unit 11.12.1–11.12.10.
- Winter J, Jung S, Keller S, Gregory RI, Diederichs S. 2009. Many roads to maturity: microRNA biogenesis pathways and their regulation. *Nat Cell Biol* **11**: 228–234.

RNA

A PUBLICATION OF THE RNA SOCIETY

Assessment of microRNA differential expression and detection in multiplexed small RNA sequencing data

Joshua D. Campbell, Gang Liu, Lingqi Luo, et al.

RNA 2015 21: 164-171 originally published online December 17, 2014

Access the most recent version at doi:[10.1261/rna.046060.114](https://doi.org/10.1261/rna.046060.114)

Supplemental Material

<http://rnajournal.cshlp.org/content/suppl/2014/12/08/rna.046060.114.DC1>

References

This article cites 13 articles, 3 of which can be accessed free at:
<http://rnajournal.cshlp.org/content/21/2/164.full.html#ref-list-1>

Creative Commons License

This article is distributed exclusively by the RNA Society for the first 12 months after the full-issue publication date (see <http://rnajournal.cshlp.org/site/misc/terms.xhtml>). After 12 months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

To subscribe to RNA go to:

<http://rnajournal.cshlp.org/subscriptions>