Published online 2 September 2011

Nucleic Acids Research, 2011, Vol. 39, No. 21 ***e141***
doi:10.1093/nar/gkr693

Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing

Anitha D. Jayaprakash, Omar Jabado, Brian D. Brown and Ravi Sachidanandam*

Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, 1425 Madison Avenue, New York, NY 10029, USA

Received April 12, 2011; Revised August 3, 2011; Accepted August 8, 2011

ABSTRACT

Deep sequencing of small RNAs (sRNA-seq) is now the gold standard for small RNA profiling and discovery. Biases in sRNA-seq have been reported, but their etiology remains unidentified. Through a comprehensive series of sRNA-seq experiments, we establish that the predominant cause of the bias is the RNA ligases. We further demonstrate that RNA ligases have strong sequence-specific biases which distort the small RNA profiles considerably. We have devised a pooled adapter strategy to overcome this bias, and validated the method through data derived from microarray and qPCR. In light of our findings, published small RNA profiles, as well as barcoding strategies using adapter-end modifications, may need to be revisited. Importantly, by providing a wide spectrum of substrate for the ligase, the pooled-adapter strategy developed here provides a means to overcome issues of bias, and generate more accurate small RNA profiles.

INTRODUCTION

The advent of deep sequencing has now made it possible to sequence the full complement of small RNAs in a cell (1–4).

Small RNAs (15–30 nt), such as microRNAs, piRNAs and endogenous siRNAs, are crucial regulators of genetic activity (5–8). Though many methods like real-time (RT) PCR (9,10) and microarrays (11) can be used for profiling known small RNAs, identifying differences between closely related microRNAs and discovery of novel sequences can only be done through deep sequencing (12).

Deep sequencing is especially attractive for its sensitivity to low abundance transcripts (2,4). In light of this, a persistent mystery in the field of small RNA sequencing is the discrepancy between the results from deep sequencing, microarrays and qPCR (13,14), with certain miRNAs being under- or overrepresented in sRNA-seq (15). This calls into question quantitative data from deep sequencing, especially measurements of relative abundances of isoforms and variants.

Although other profiling platforms also exhibit biases, biases in sRNA-seq would undermine the incredible sensitivity and accuracy achievable by deep sequencing. For piRNAs, sequence features such as the T-bias at the 5′-end are inferred through profiling, and offer clues to their biogenesis (16). A data set that is biased by collection methods can, therefore, lead to erroneous conclusions.

The most widely used technique of sRNA-seq involves isolation of small RNAs (15–30 nt), ligation of 3′ and 5′ adapters onto the ends of the small RNAs using T4-RNA ligases (Rnl2 and Rnl1, respectively, Figure 1), followed by reverse transcription and amplification (7–21). Skews have been noticed in the profiles generated by this method of sample preparation, independent of the sequencing platform (13–15).

Thus, we set out to systematically investigate the presence and source of the biases in sRNA-seq. We deep-sequenced small RNAs from 293T human kidney-derived cells and mouse embryonic stem (mES) cells, using strategies aimed at identifying the source of bias. Upon careful investigation, we conclude that a reproducible discrepancy can arise only in the ligation or amplification steps.

The biases in the activity of RNA ligases have not been explored in the context of their use in deep sequencing (22,23). We show that the T4-RNA ligases used in sample preparation is the predominant cause of distortions and they mediate sequence-specific ligations. We show that this bias can be ameliorated using a pooled adapter strategy. Our results provide new insights into the activity of RNA ligases through deep sequencing,

*To whom correspondence should be addressed. Tel: +1 212 659 6735; Fax: +212 360 1809; Email: ravi.sachidanandam@mssm.edu

and an invaluable strategy to reduce biases and increase the accuracy of the profiles of the small RNA transcriptome generated through sRNA-seq.

MATERIALS AND METHODS

Library construction and sequencing

Total RNA was isolated from 293T cells and mouse embryonic fibroblasts using Trizol extraction (Invitrogen). Sequencing libraries enriched for micro-RNAs were constructed using a modified version of a small RNA protocol detailed by Pfeffer (21). Two RNA markers were synthesized *Spike 19* (CGUACGGUUUAAACUUC GA) and *Spike 24* (CGUACGGUUUAAACUUCGAA AUGU) (Sigma Aldrich); RNA was end-labeled using polynucleotide kinase and radioactive ATP (P32). Ten micrograms of total RNA was size fractionated by denaturing poly acrylamide gel electrophoresis (PAGE, 12% gel). miRNAs were excised from the gel using radiolabeled markers as guides. Purified miRNA was ligated to a 17 nt 3′ adapter with truncated T4 RNA ligase 2 (Rnl2) in an ATP-free buffer (Bioo Scientific). The ligated fragment of 36–41 nt was PAGE purified. A second RNA adapter was ligated to the 5′ side of the product using T4 RNA ligase 1 (Rnl1) and buffer containing ATP. The 72–78 nt ligated fragment was PAGE purified and then reverse transcribed using a specific primer (BanI-RT; ATTGATGGTGCCTACAG). cDNA was amplified by 22 cycles of PCR with primers that incorporate sequences compatible with the Illumina platform (Sol-5-SBS, AATGATACGGCGACCACCGA ACACTCTTTCCCTACACGACG and Sol-3-ModBan, CAAGCAGAAGACGGCATACGATTGATGGTGCC TACAG) (Figure 1). The library was sequenced using the Illumina Genome Analyzer IIx at 36 nt read length. Replicates were sequenced to verify the reproducibility of the results (Supplementary Figures S1–S3).

Microarray

miRNA abundance was assessed in 293T and mES RNA samples by oligonucleotide microarray using Affymetrix GeneChip (miRNA 1.0). One microgram of total RNA was labeled using the FlashTag Biotin 3DNA kit (Genisphere), as follows: polyadenylation of RNA by polymerase, ligation to a biotinylated 3DNA molecule mediated by an oligo with 5′ poly d(T) and 3′ 3DNA complementary adapter. Labeled RNA was hybridized to the microarray using standard Affymetrix methods. Fluorescence intensities were extracted using the R statistical package, using methods from the BioConductor module (http://www.bioconductor.org/).

Real-time PCR

Quantitative RT PCR was carried out using the Applied Biosystems (AB) microRNA specific reagents and a 7900HT thermocycler. Ten nanogram of total RNA was reverse transcribed with a miRNA specific hairpin primer using the AB micoRNA Reverse Transcription kit. Specific forward primers and universal reverse primers

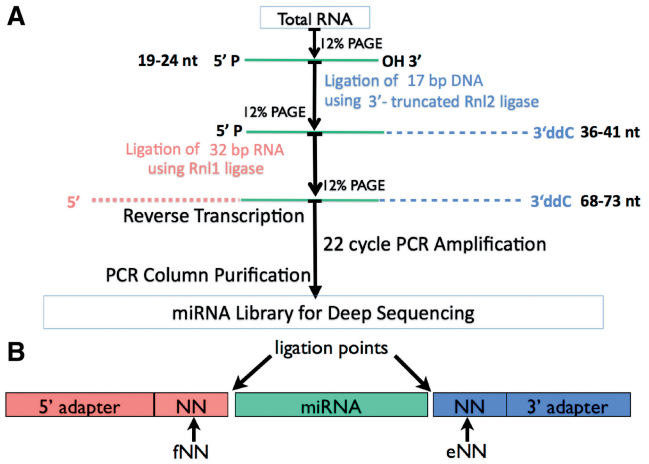

Figure 1. The protocol for preparing samples for small RNA sequencing. Total RNA is size fractionated by denaturing poly acrylamide gel electrophoresis (PAGE) and miRNAs are excised from the gel using radiolabeled markers as guides. Purified small RNAs are ligated, using a truncated T4 RNA ligase 2 (Rnl2) in an ATP-free buffer, to a 17-nt modified 3′ DNA adapter with dideoxy at the 3′-end and activated at the 5′-end by adenylation. The dideoxy prevents self-ligation of the adapter, while the truncated ligase prevents circularization of the small RNA inserts. The ligated fragment of 36–41 nt is then PAGE purified, to remove the unligated 3′ adapters. A 32-nt RNA adapter is ligated to the 5′ side of the product using T4 RNA ligase 1 (Rnl1). The 72–78 nt ligated fragment is PAGE purified again to remove the unligated 5′ adapters. The product is reverse transcribed using a specific primer and the resulting cDNA is amplified by PCR with primers that incorporate sequences compatible with a deep-sequencing platform.

were random with cDNA and AB Universal PCR Master Mix (no UNG) as recommended by the manufacturer. The following miRNAs were assayed: hsa-mir-18a, -20a, -106b, -92a, -103-2, -10, -16, -17 and hsa-let-7. Ct values were extracted from real-time data using the *auto threshold* setting.

Computational analysis

Analysis of such datasets is well established (25), but extracting the inserts from the libraries was complicated by three causes: (i) sequencing errors that mis-call a base, (ii) sequencing errors that miss a base and (iii) errors in the synthesis of the NN constructs. To mitigate problems from sequencing errors, we only accepted sequences where the 3′ adapter sequences were matched exactly. This eliminates most of the *problematic* reads, but does not solve the issue of point (iii) above. For that, we used the relative abundances of the various inserts in the small RNA library (from our analysis of data from several runs), to identify synthesis errors. Failure to synthesize a particular N, or a skew in a particular N, causes a mis-identification of the adapter on the sequence and its end modifications. Each sequence was binned into the appropriate NN category, as well as the appropriate version of the miRNA sequence (the canonical mature or an isomir, either derived from the original hairpin sequence or a non-template modification). Most of the processing was done using custom Perl scripts (which are available

from the authors on request). Custom R-scripts were used to generate the graphs and statistical analyses.

RESULTS

In order to establish the sequence dependence of the ligation of adapters to small RNAs, libraries were constructed for small RNA sequencing, using the standard protocol (Figure 1) with modified adapters. The set of samples that were sequenced are listed in the Supplementary Table S1.

Strategies using modified adapters

To understand the exact nature of the biases, we devised strategies using various 5′ and 3′ adapters with additions to the ligating ends (3′-end of the 5′ adapter and the 5′-end of the 3′ adapter). We devised six strategies using the adapter pools, as listed below and in Table 2.

(1) noNN: the standard modban 5′ and 3′ adapters
(2) 4-mer pool: the standard modban 3′ adapter, pool of twelve 5′ adapters with 4-mer additions.
(3) fNN: the standard 3′ adapter and a pool of 5′ adapters generated by adding random NN additions to the 3′-end of the 5′ modban adapter,
(4) eNN: the standard 5′ adapter with a pool of 3′ adapters that are modified at the 5′-end with NN additions (eNN),
(5) fNN_eNN: a pool of 5′ and 3′ adapters with the NN modifications described in (2) and (3), and
(6) fNNNN: a pool of 5′ modban adapters with the addition of random NNNN to the 3′-end and the standard 3′ adapter.

5′ adapter ligation efficiency is sequence dependent

In order to determine if there was sequence-dependent ligation of the 5′ adapters, we prepared small RNA samples from 293T cells, using a pool of twelve 5′ adapters, modified by the addition of 4-mers (TGAC, GAGT, GTAT, CGTC, GGAA, AAGG, GCTT, AACC, CCAA, AGCA, CTAG and TGTG).

The results showed big differences between data from different adapters (Figure 2). We also established in this experiment that the bias was not PCR-dependent, by reducing the number of PCR cycles down from 25 to 18, without any significant effect on the results (Figure 2D).

We prepared individual 293T cell samples using one adapter per sample, selecting five 4-mer ends (TGAC, CGTC, AACC, GTAT and GGAA). We found wide variations in the miRNA profiles, especially for highly expressed miRNAs such as hsa-mir-20a and hsa-mir-18. In Table 1, we list correlations between samples (293T-derived RNA) sequenced individually with adapters ending in TGAC, CGTC and GGAA and the same samples sequenced using pools of the five adapters listed above. The sequences with individual barcodes have poor correlation to each other, but pooling the adapters improves the concordance between the profiles for different replicates. This suggests that a pooled approach might reduce the effect of biases caused by adapter ligation on the 5′-end.

Nature of sequence dependence in the efficiency of 5′ adapter ligation

To identify the biases inherent in the 5′ adapter ligation, two samples from 293T and mES cells were prepared using the fNN strategy. The results showed that the profiles measured from the same sample can vary wildly for different adapters (Figure 3). Figure 3A (293T) and C (mES) depict the amount of miRNAs (*y*-axis) captured by each adapter (*x*-axis), suggesting some adapters are more efficient than others. If it were a simple matter of differing efficiencies for different adapters, then the miRNA profiles derived from each barcode should be scaled versions of each other. In fact, as shown in Figure 3B (293T) and D (mES), the profiles for different adapters are very dissimilar. In the Figure 3B and D, the *x*-axis shows different miRNAs, ranked by their overall occurence, which is the sum over all adapters. The *y*-axis shows, of all the miRNAs captured by a particular adapter, the fraction that each miRNA occupies. It is apparent there can be dramatic shifts in the rankings for the miRNAs (the profiles) between adapters.

We wanted to establish how much of the sequence proximal to the ligating end of the adapter determined the ligation efficiency. For this, we carried out an experiment using 5′ adapters with four terminal random nucleotides, the fNNNN strategy. Figure 4 shows that most of the ligation efficiency can be explained by the last two nucleotides. Only in one case, hsa-miR-106b, did we need the NNNN strategy to detect an abundant miRNA (see also Figures 2 and 10).

Biases in 3′ adapter ligation

We investigated the bias on the 3′-end independently, since the 3′ adapter ligation differs from the 5′ adapter ligation, in using a truncated RNA-ligase (rnl2) with an adenylated 3′ DNA (instead of RNA) adapter.

We designed a simple gel shift-based assay to test for the existence of a 3′ adapter ligation bias. We chose two radioactively labeled oligomers, a 19-mer (CGUACGGUUUA AACUUCGA) and a 24-mer that had a 5-mer (AAUGU) addition at the 3′-end of the 19-mer. Figure 5 shows that the 24-mer does not ligate efficiently to the standard 3′ adapter, in contrast to the 19-mer. On the other hand, the adapters from the **NN** strategy ligate well to both oligos. The 24-mer has good ligation only to certain sequences in the eNN pool, indicating the superiority of this strategy. While this may appear to be crude, the dramatic effect seen in the gel-shift suggests large differences in ligation efficiencies between different pairs of sequences, indicating that both the 5′ and 3′ adapter biases need to be taken into account in any sequencing experiment using T4-RNA ligases in the sample preparation.

We pursued a strategy similar to the case of 5′ adapter ligation, using the eNN strategy for 3′ adapters in order to systematically study the biases in the ligation of the 3′

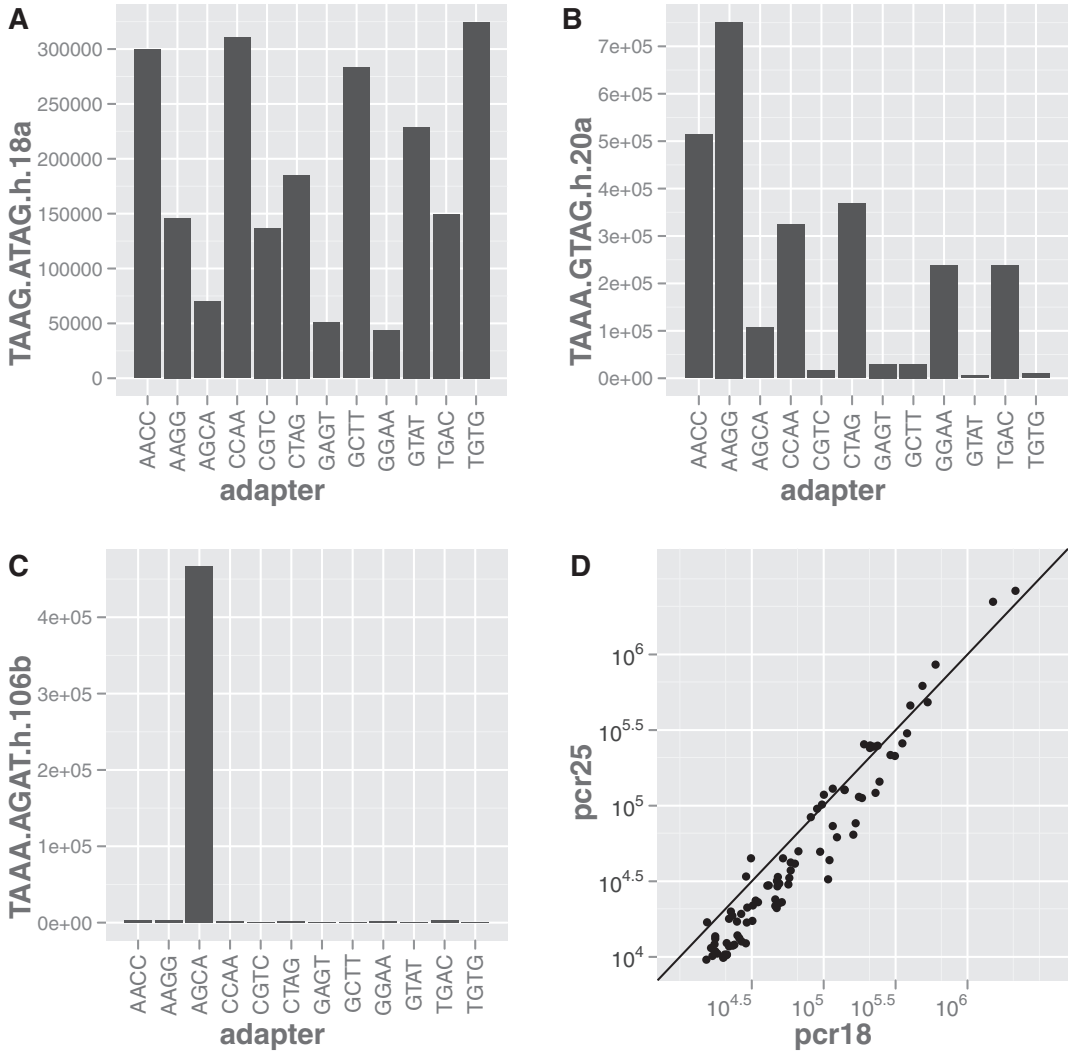

Figure 2. Choice of 5′ adapter ends determines miRNA abundance/ranking, not PCR cycles. Sequencing libraries were constructed from total RNA derived from 293-T cells, using a pooled set of twelve 5′ adapters that had different 4-mer 3′-ends, shown on the *x*-axis. There is great diversity in the capture of individual miRNAs by different 5′ adapters (**A**, **B** and **C** show data for miR-18a, miR-20a and miR-106b, respectively). (**C**) Shows an extreme case where miR-106b is captured well by only one adapter, ending in AGCA, out of the 12 combinations. These data are consistently reproduced in other experiments shown in Figure 4. To isolate the effect of PCR cycles, we prepared the samples twice, using 25 (*y*-axis) and 18 (*x*-axis) cycles of PCR (**D**). Each point represents a miRNA. The correlation between the two sets is high (coefficient of 0.95) and the best linear fit to the points is a line of slope 1, suggesting that the data are reproducible and PCR is not responsible for the biases.

Table 1. Correlations (spearman rank) between samples, based on the abundance of miRNA sequences

	293T_1 (TGAC) / pool	293T_2 (CGTC) / pool	293T_3 (GGAA) / pool
293T_1 (TGAC) / pool	1	(0.66) / 0.91	(0.68) / 0.92
293T_2 (CGTC) / pool	(0.66) / 0.91	1	(0.64) / 0.97
293T_3 (GGAA) / pool	(0.68) / 0.92	(0.64) / 0.97	1

The samples were prepared using either individual 5′ adapters that differ only at the 3′ terminus or a pooled set of 5′ adapters. The first number in each entry is correlation with the specific adapter, while the second number is for the pooled data. The relatively low correlations for individual adapters between biological replicates of 293T cells, in contrast to the results for the pooled data, suggest that the efficiency of ligation of the adapters to different miRNA sequences is variable. This suggests that mixed pools of adapters can help overcome the inherent biases in ligation efficiency.

adapter. In Figure 6, we show the efficiencies of the 5′ and 3′ adapter ligations in the form of a fluctuation graph. The area of the rectangles is proportional to the number of reads that come from a particular miRNA–adapter combination. The 3′ adapters show more variability, which is probably due to the greater diversity in the 3′-ends of the miRNAs compared with the 5′-ends, suggesting that the 3′ adapter ligation might be a bigger

Table 2. miRNA sequencing libraries were generated with the adapter combinations shown here

Strategy	5′ adapter RNA	3′ adapter DNA
noNN	ACACUCUUUCCCUACACGACGCUCUUCCGAUC	CTGTAGGCACCATCAAT
fNN	ACACUCUUUCCCUACACGACGCUCUUCCGAUC**NN**	CTGTAGGCACCATCAAT
fNNNN	ACACUCUUUCCCUACACGACGCUCUUCCGAUC**NNNN**	CTGTAGGCACCATCAAT
eNN	ACACUCUUUCCCUACACGACGCUCUUCCGAUC	**NN**CTGTAGGCACCATCAAT
fNN_eNN	ACACUCUUUCCCUACACGACGCUCUUCCGAUC**NN**	**NN**CTGTAGGCACCATCAAT
4-mer pool	ACACUCUUUCCCUACACGACGCUCUUCCGAUC**WXYZ**	CTGTAGGCACCATCAAT

4-mer pool is a mixture of 12 adapters, represented as WXYZ (CTAG, GAGT, CCAA, AGCA, AACC, AAGG, TGAC, CGTC, GCTT, GTAT, GGAA, TGTG). All 3′ DNA adapters have a 5′ rAPP and 3′ddC modifications to prevent self-ligation and circularization.
The bold sequences in this table indicate modifications to the standard adapter sequence.

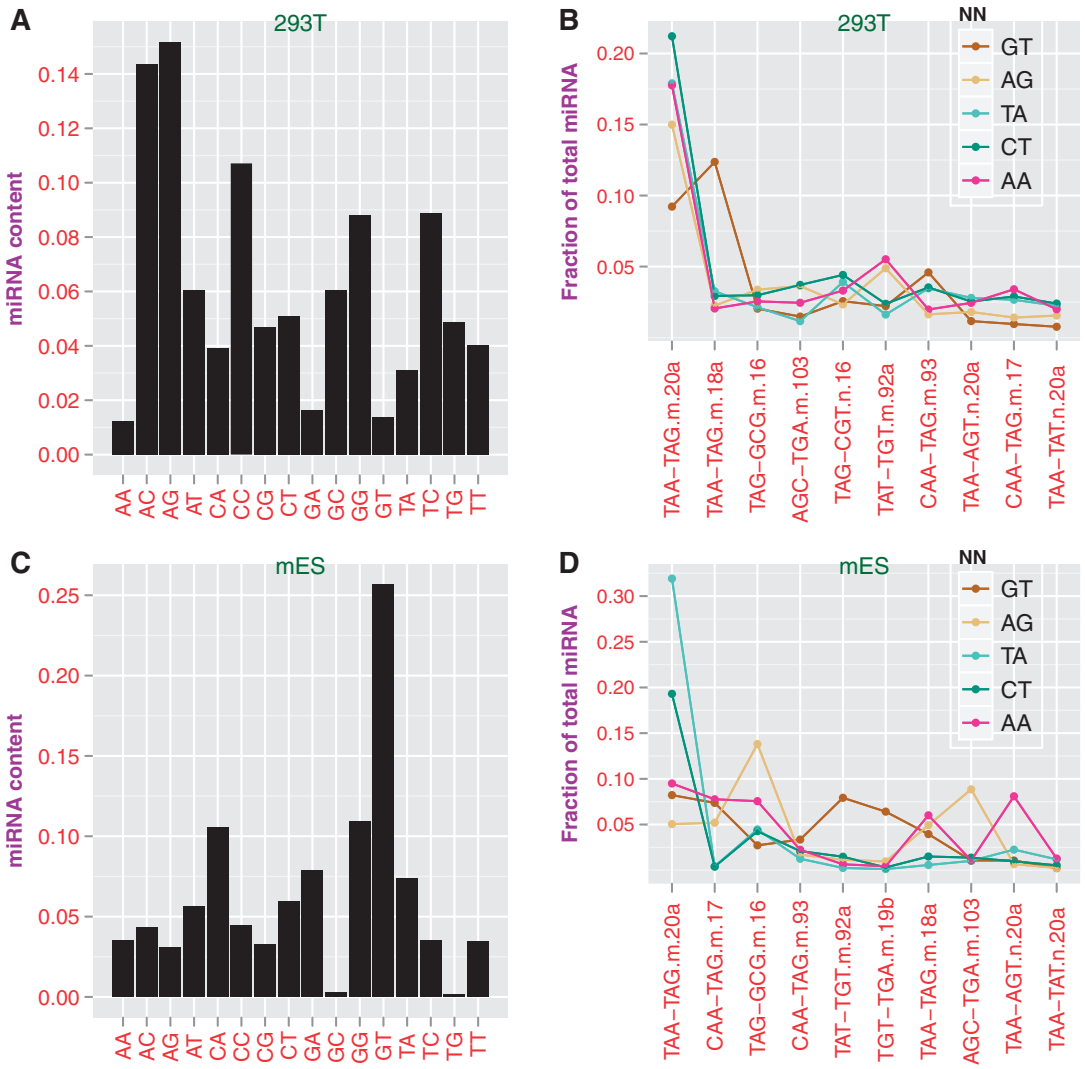

Figure 3. Measured miRNA abundance by the fNN strategy depends on both the adapter and the miRNA sequences. (**A**) and (**C**) show the fraction of miRNA in each adapter type, calculated by adding the total number of miRNA sequences (irrespective of identity) captured by each adapter type as a fraction of the total amount of miRNA captured by all the adapters combined. (**B**) and (**D**) show, for each adapter type (only 5 out of the 16 are shown here for clarity), the fraction occupied by the top miRNAs. The rankings of the miRNAs by relative abundance are dependent on the adapter. The A and C panels show differences in adapter efficiencies in capturing miRNAs, and the B and D panels show that these differences arise from variations in the efficiencies that depend on the miRNA–adapter combination.

source of biases in measurements. It is interesting to note that the efficiency of the standard modban adapters (the 5′ one ends in TC and the 3′ one starts with CT) is low, compared with some of the others, but there is no single adapter that is uniformly efficient across the miRNAs that we tested here. This again suggests that it is necessary to take a pooled approach on both adapters for an unbiased measurement.

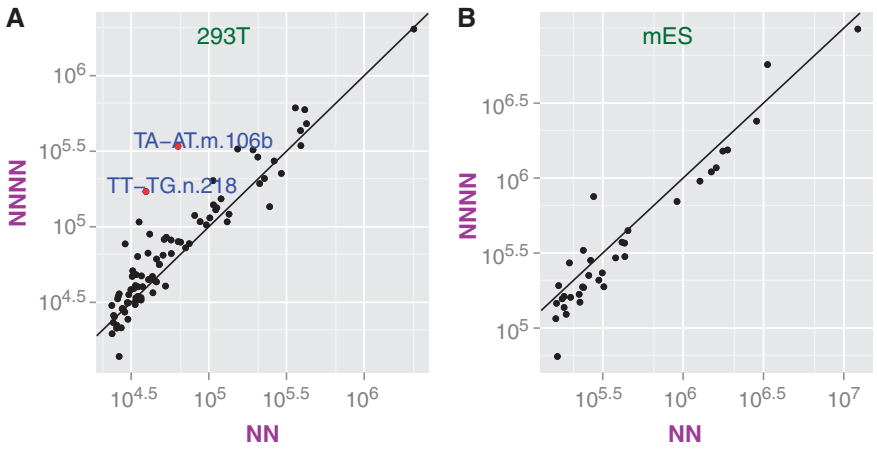

Figure 4. The two terminal 3′ bases of the 5′ adapter are the primary determinants of T4-RNA ligase 1 (Rnl1) ligation efficiency. Two distinct sets of 5′ adapters, one consisting of adapters with mixed bases in the last four bases (fNNNN) and another consisting of adapters with mixed bases in the last two bases (fNN), were used to generate a miRNA derived cDNA library for (**A**) human 293T and (**B**) mouse embryonic stem cell lines. miRNA abundance in read counts (dots) were plotted; the fNNNN data were compressed to NN, by combining values for AANN through TTNN for each NN. The high correlation between the compressed fNNNN and the fNN datasets indicates that the two terminal bases are dominant determinants of ligation efficiency. There are exceptions shown in red, which are systematic differences (106b, 181 in 293T cell), which we detected in an independent experiment described in Figure 2 suggesting that this is not a stochastic effect. The naming convention in all our figures is to show the beginning and end of the sequence followed by an *m* (for a canonical mature) or *n* for a non-canonical miRNA sequence followed by the name of the miRNA. Thus in the left we have a canonical mature hsa-miR-106-b and a non-canonical hsa-miR-218. The high abundance for hsa-miR-106b suggested by the fNNNN strategy (in contrast to the low values suggested by fNN and other strategies) seems real, as the microarray and RT–PCR results (Figure 9) are in concordance with the fNNNN values.

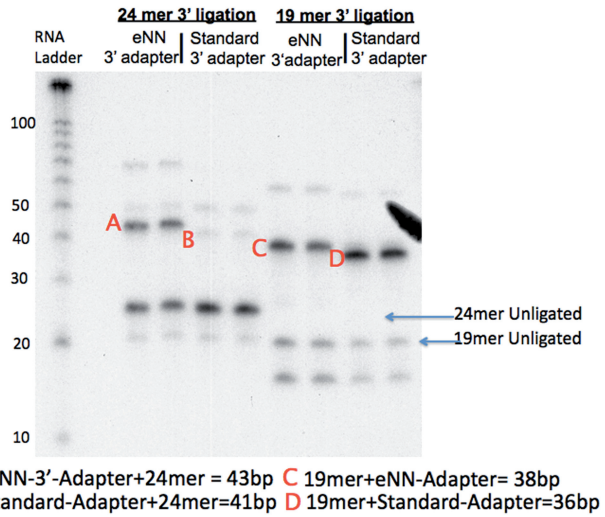

Figure 5. Synthetic RNA ligation to 3′ adapter is enhanced by using a pool of 3′ adapters with random NN at the 5′-end. Two RNA marker strands, 19 and 24-nt long, were synthesized. The 19-mer ends in UCGA, while the 24-mer has an extra 5 nt (AAUGU) on the 3′-end. The RNA markers were 5′-end-labeled with P32 and then ligated in duplicate to one of two sets adenylated 3′ DNA adapters; one set consisting of the standard adapter with a 5′ CTGT and the second set consisting of a mixture of adapters that differ from the standard adapter in having two extra mixed base positions on the 5′ side, with the start now becoming 5′ NNCTGT. After ligation, the RNA-DNA products were size fractionated on a 12% polyacrylamide gel. The 19 nt marker ligates efficiently, irrespective of the adapters used (lanes 5–8) while the ligated 24-mer product is low in abundance when the standard adapter is used (lanes 1–2), but is efficiently ligated (with abundant products) when the mixed-bases adapters are used (lanes 3–4).

A model for ligation efficiencies

To develop a unified picture of the ligation efficiency and show that the experiments are consistent with each other, we developed a model. The model assumes ligation efficiencies based on the 256 combinations at each ligation junction, determined by the two nucleotides (16 possible combinations, AA, AC…TG, TT) on the ligating end of the adapter and the two nucleotides (16 possible combinations) on the ligating ends of the miRNA. We define these as F_{ij} (i and j each varying from 1 through 16 where 1 stands for AA, 2 for AC going on to 15 for TG and 16 for TT) for the 5′ adapter ligation, and E_{mn} (m and n each varying from 1 through 16) for the 3′ adapter ligation. Let M^k be the actual abundance of a miRNA labeled k in the sample. Let m^k_{in} be the measured amount of miRNA labeled k using adapters with ends i and n on the 5′ and 3′ adapter, respectively. Then, the model suggests,

$$m^k_{in} = F_{ij} * M^k * E_{mn} \quad (1)$$

The various adapter combinations are in equimolar concentrations; so they do not enter the equation (other than a constant that can be absorbed in F and/or E). Figure 7 depicts the matrices F and E in a fluctuation graph, highlighting the variability. If this model is universal, we expect that the ratio between various F's (and various E's) from the fNN and eNN data sets should agree with the numbers derived from fNN-eNN. Since we do not know the M^k for an miRNA labeled k, we have to eliminate that from any quantity we measure. To do this, we pick the same value for eNN (CT) in the fNN_eNN set as

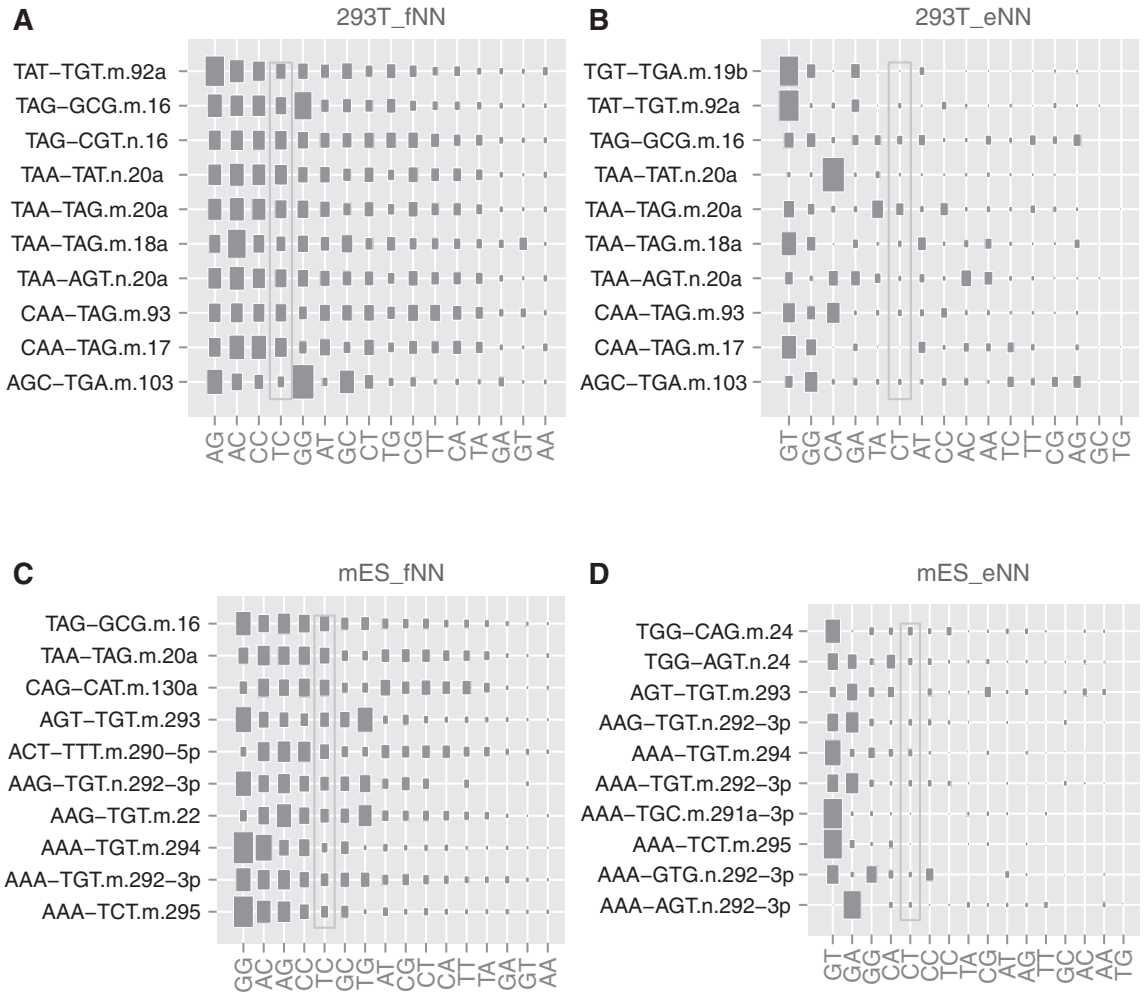

Figure 6. Fluctuation plots showing ligation efficiency for different fNN (A and C) and eNN (B and D) adapters against the most abundant miRNAs from 293T (**A** and **B**) and mES (**C** and **D**) cells. The naming convention in all our figures is to show the beginning and end of the sequence followed by an *m* (for a canonical mature) or *n* for a non-canonical miRNA sequence followed by the name of the miRNA. The area of the dark rectangles depicts the value for each combination of miRNA and adapter. The standard adapter ends (TC in fNN and CT in eNN, highlighted in gray boxes) are not very efficient in ligation to the most abundant miRNAs. Even the most efficient adapters show variability, suggesting that no single adapter can work well across all possible sequences. For the top miRNAs, most of the variability comes from the 3′ adapter ligation (the eNN adapters, B and D). In mES cells, there are two isomirs of mmu-miR-292-3p, the GT ending 3′ adapter captures the GAGT-ending isomir more efficiently, while the GA ending 3′ adapter captures the GAGTG-ending isomir more efficiently.

the 5′-end on the 3′ adapter in the fNN set. Within each experiment,we then define,

$$r^k_{ia} = m^k_{ij}/m^k_{aj} = F_{ij}/F_{aj}. \quad (2)$$

r^k_{ia}, which is the ratio between the number of miRNA k, captured by adapters with ends i and a, is now independent of M^k and it should be identical for the fNN_eNN (with eNN set at CT) and the fNN sets. We can do a similar comparison between the fNN_eNN and the eNN sets. These ratios, derived from independent experiments, are shown in the fluctuation plot in Figure 7. The agreement is visually striking, with high similarity between members of a pair. The numbers agree across miRNAs and across sample types, suggesting a level of universality for this model. The success of the model points to the reproducibility of the effects we have observed. Our enthusiasm for this model is tempered by the case of miR-106b, which shows a bias that depends on four nucleotides at the 3′-end of the 5′ adapter (Figure 2D and Figure 4).

Strategy to overcome the ligation biases

Based on all the evidence presented above, we devised the fNN_eNN strategy, described at the beginning of this section, to overcome the biases. Figure 8 shows the data for small RNA from the 293T and mES cells sequenced using the fNN_eNN strategy. Each microRNA seems to have a favored fNN-eNN pair that works best, once again suggesting the need for a pooled adapter approach. Supplementary Tables S2–S6 show the 50 most abundant sequences for each experiment in the 293T samples.

Validity of the fNN_eNN strategy

Our proposed strategy, fNN_eNN, is one that optimally picks up most sequences, and can help overcome the effect

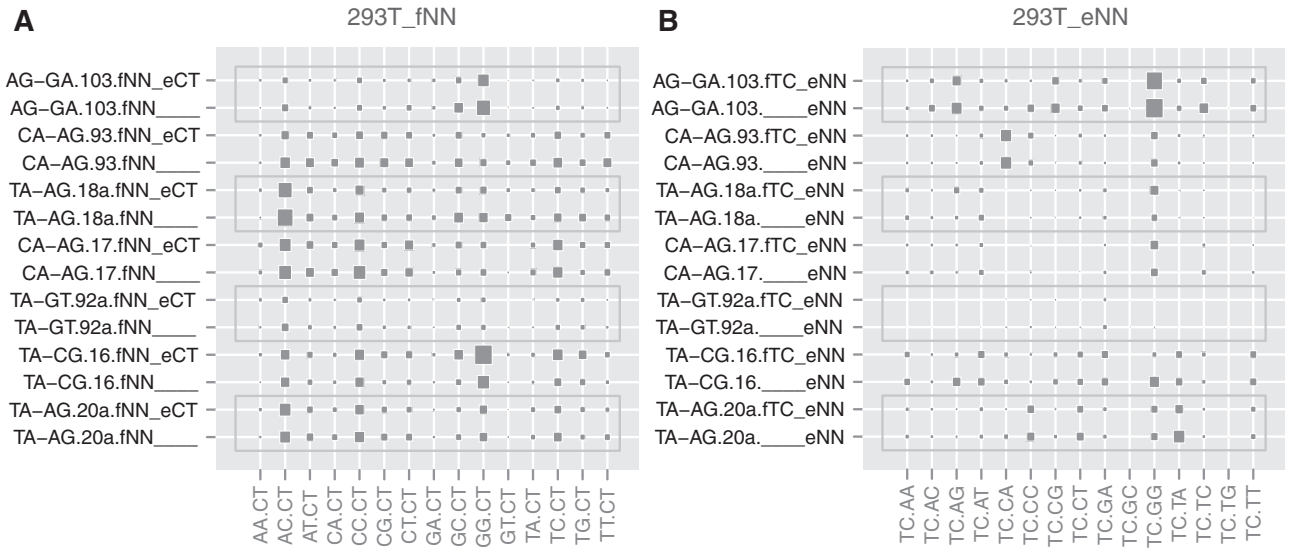

Figure 7. Comparison of parameters inferred from fNN (**A**) and eNN (**B**) against fNN_eNN data. The rows are miRNAs captured by different methods, alternate rows are data from the fNN_eNN. In the figure, fTC_eNN means the f end was the standard (TC) and the e end was varied while fNN_eCT means e end had a CT and the f end was varied. In the data for fNN_eCT versus fNN, the ratio to the AG–CT combination is depicted for each row. For the comparison of fNN_eNN against eNN, the ratio to the values for the TC–GT combination is considered. The pairs are highlighted (either light or dark shaded rectangles), and the numbers between members of a pair are expected to be similar, as explained in the text. There is a striking similarity between pairs of rows, suggesting that the fNN_eNN parameters are in concordance with separate measurements of parameters with fNN and eNN. The results section has an explanation for the model on which the calculations are based.

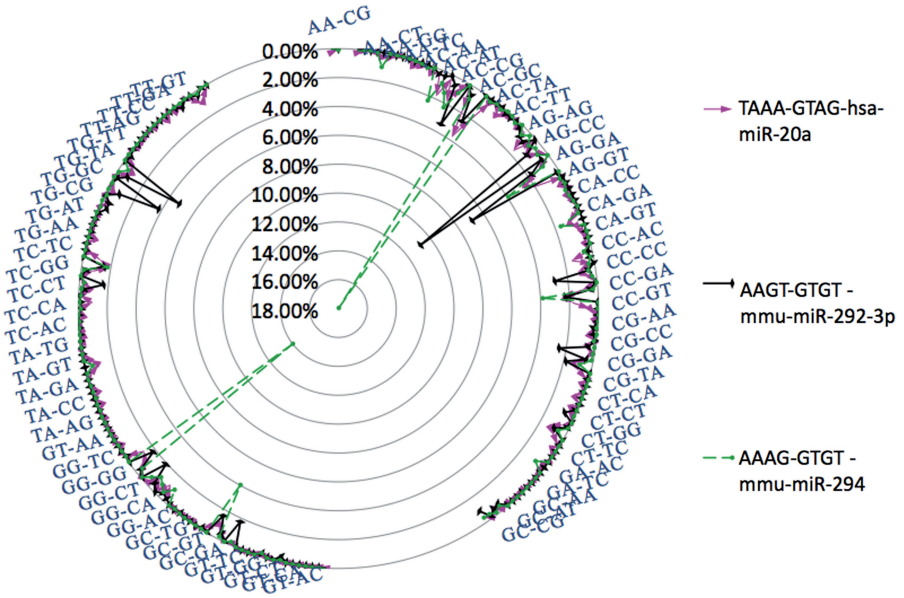

Figure 8. A radar plot showing the performance of different adapter termini combinations (fNN_eNN), shown outside the circle in blue. The inner circles represent percent contribution of each adapter combination to a particular miRNA that was sequenced. This plot shows data for the top miRNA (hsa-miR-20a) in 293T cells and two top miRNAs (mmu-miR-292-3p and mmu-miR-294) from mouse embryonic stem cells. There is large variation in the efficiency of capture between various combinations of 5′ and 3′ adapter end modifications. This emphasizes the need for a pooled strategy in sequencing.

of the biases and increase the efficiency of small RNA sequencing. We have compared data from various sequencing strategies to qPCR and microarray data (Figure 9). We can see that the best concordance is for fNN_eNN against arrays.

Using the fNN_eNN technique, we have identified several miRNAs in mouse embryonic stem cells and human kidney-derived 293T cells that are severely underrepresented in the current published profiles based on deep sequencing (Table 3 and Figure 10). Thus, we have established the existence of a pronounced, sequence-dependent bias in the ligation of 5′ and 3′ sequencing adapters to miRNAs. Our proposed strategy, fNN_eNN, will be able to overcome the limitations of the

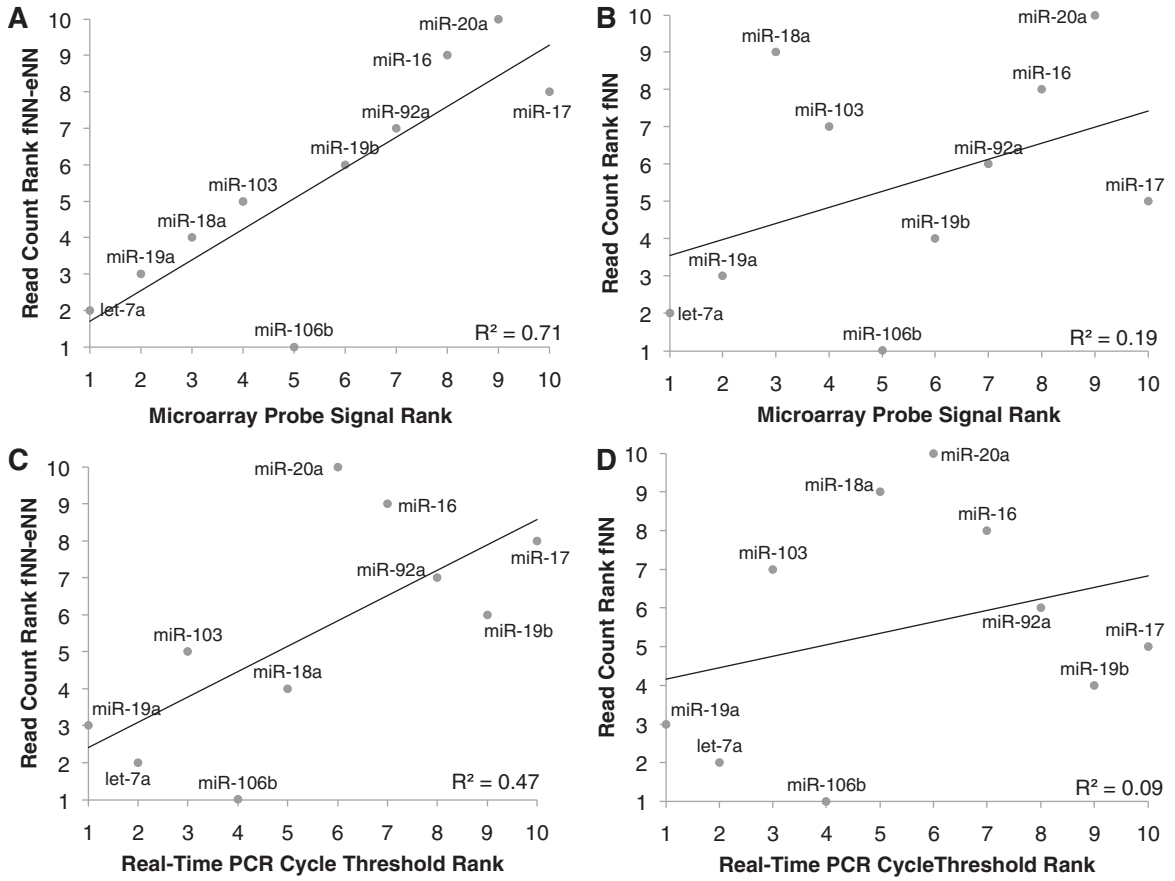

Figure 9. Comparison of sequencing against microarray (**A** and **B**) and RT–PCR (**C** and **D**) for mES (B and D) and 293T (A and C). There are outliers, such as miR-106b, which are only captured by the fNNNN strategy, but overall, there is significant correlation between the fNN_eNN strategy and the microarray data (A) and the fNN_eNN strategy and the RT–PCR data (C), while the fNN sequencing strategy does not give a good correlation to RT–PCR and array data (B and D).

Table 3. A few examples of miRNAs that show dramatic changes in their ranking, depending on the technique used to sequence them

Rank1	Adapter1	Rank2	Adapter2	Sequence	miRNA
41	noNN	10	fNNNN	TAAA-AGAT	hsa-miR-106b
24	noNN	110	fNN_eNN	TAAG-ATAG	mmu-miR-18a
4	noNN	10	fNN_eNN	TAAA-ATAG	hsa-miR-18a
6	noNN	2	fNN_eNN	AAAG-GTGT	mmu-miR-292-3p
5	noNN	21	fNN_eNN	ACTC-CTTT	mmu-miR-290-5p
7	noNN	20	fNN_eNN	TAAA-GTAG	mmu-miR-20a
9	noNN	57	fNN_eNN	TAGC-GGCG	mmu-miR-16
24	noNN	6	fNN_eNN	AAAG-GTGC	mmu-miR-291a-3p

In most cases, the qPCR and array data are in concordance with the fNN_eNN data, except in the case of miR-106b, which is more in accord with the fNNNN data.

bias in the RNA-ligase and help make sRNA-seq more representative of the profiles in the underlying samples.

DISCUSSION

Our experiments (and model) explain the source of biases observed with sRNA-seq. We have identified and quantified biases in the functioning of the T4-RNA ligases (Rnl1, Rnl2) (26,27) through deep sequencing, and the large numbers of ligated sequences in our experiments provide a measure of statistical reliability to our results.

Reasons for biases in the ligase activity

Bacteria, under viral attack, nick their tRNAs to block translation. T4-phage uses the ligases to repair the nick (26). Since the nicks are located at specific sequences in the tRNAs, efficient repair probably favors ligase structures exhibiting sequence specificity.

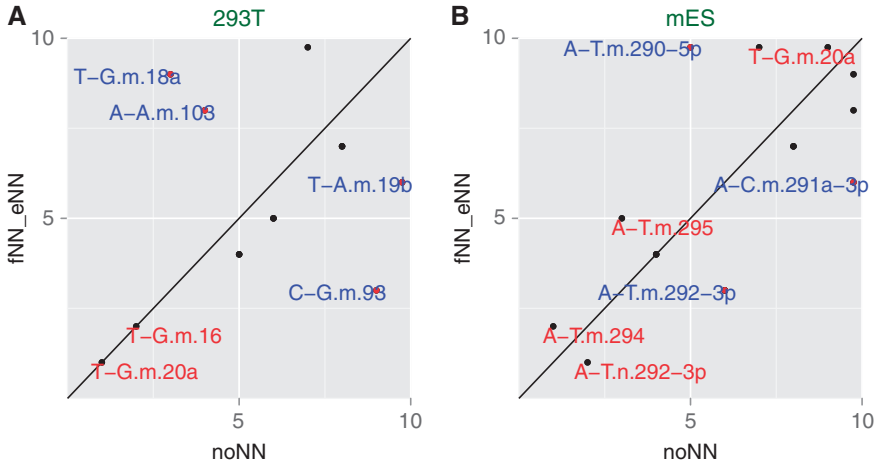

Figure 10. Comparison of rankings between the standard adapters (noNN,ranks along *x*-axis) versus fNN_eNN (ranks along *y*-axis) for 293T (**A**) and mES samples (**B**). A point above the diagonal represents a sequence that is overrepresented in noNN, while below the diagonal are points that are underrepresented in noNN. The hsa-miR-18a is overrepresented in the noNN case, where it is ranked 3, the array and qPCR data agree better with the fNN_eNN results which ranks it much lower [this skew is also seen in the mES samples, but the ranking in the noNN is 22 while the fNN_eNN is much lower (135)]. In the mES sample, mmu-miR-294 is first and a non-canonical form of mmu-mir-292-3p is second for noNN, while they switch ranks in the fNN_eNN case, the difference is very significant, because the abundances of the first and the second ranks are about 2-fold apart, suggesting a strong bias. mmu-miR-290-5p is very high at rank 5 in the case of noNN, it is outside the range of the graph in fNN_eNN, in accordance with the qPCR data. Thus, in every case that we can detect a difference between noNN and fNN_eNN, fNN_eNN seems to be more accurate in reflecting the profiles.

Profiling studies

Our results have important implications for miRNA profiling studies in cancer (24,28,29) and stem cells (30–33), that attempt to identify biomarkers for diagnosis and therapies.

Small changes in the ranking, from say 1 to 2, often reflect big changes in numbers, which can have important implications for the kinetics of the reactions mediated by the miRNA (34).

Of the several mES-specific miRNAs, mmu-miR-292-3p has two forms, a canonical form and a longer non-canonical form with an extra A at the 5′-end, which means the two forms have different targets. Thus, it is important to understand their relative abundances as it might have important implications for stem cell biology. In the normal protocol, with the standard adapters, the canonical form is ranked second, while the non-canonical form is about one-third as abundant (1 305 991 versus 552 573). In the fNN_eNN strategy, the two are the highest ranked, with the canonical form ranked first and the non-canonical form ranked second and about two-thirds as abundant (3 085 673 versus 2 356 385). The microarray ranks both as the most abundant miRNAs, but it cannot reliably distinguish between the two isoforms.

The Model

The ability of a unified model to predict the outcomes of sample preparations using different adapters suggests that the effects are not stochastic. The model for ligation bias seems to suggest that a single set of 5′ and 3′ adapters might suffice, with corrections helping generate the 'real' profile. This is illusory, since, for every adapter, we see at least one transcript that seems to be inefficiently ligated. Such pairs would need large corrections, resulting in excessive noise, reducing the reliability of the corrected results. Thus, in applications where it is critical to establish accurate profiles, using pooled adapters of fNN_eNN strategy is the best approach. We have made a pursuasive case for this through our series of experiments.

Another possible approach to testing and deriving parameters in our model is to use an equimolar distribution of synthetic miRNAs (35). Unfortunately, the miRNA collections do not have the diversity that a truly random set would have, due to various constraints on the composition of the miRNA (36). In addition, such libraries require amplification from small quantities, which leads to biases. Thus, such experiments need to be done with truly random collections of small RNA sequences.

fNNNN strategy

The miRNAs, mir-106b and mir-20a are identical at the first 9 nucleotides on the 5′-end. Despite this, mir-106b is efficiently captured only in the fNNNN strategy by a few adapters (as we already discussed in Figures 2 and 4), but the fNN strategy does not capture mir-106b very efficiently. In contrast, mir-20a is efficiently captured by both the fNNNN and the fNN strategies. This suggests there might be other factors such as secondary structures that could influence the ligation, but no obvious factor could explain the divergent ligation efficiencies in this case. It is certainly of biological interest to identify the distinct roles of the two miRNAs (especially as mir-20a seems to be abundant in many tissues) and if the inefficiencies in capturing mir-106b has led to its role being overlooked.

miRNA clusters

It is believed that all members of a miRNA cluster (miRNAs that are in close proximity, <1 kb apart from each other) are processed from a single transcriptional

unit, in which case, differential expression patterns within a cluster implies differential regulation. Thus, accurate measurement of the relative numbers for members of a cluster is biologically very relevant.

We can extract numbers for two clusters (miR-106b, miR-93, miR-25) that we label as 106b cluster. Depending on the strategy used, the relative amounts within each cluster are different. For the 106b cluster, the numbers relative to the miR-106b abundance are noNN (1.0, 4.8, 1.6) and fNN_eNN (1.0, 9.4, 1.95), and there is a big change in the relative abundance of miR-93.

Abundance of star sequences

The star sequences are captured only for a small set of miRNAs, they are usually not stable products of pre-miRNA processing. In the case of miR-17 in 293T cells, we find two star forms, the canonical one (*) and a form with an extra C at the 3′-end (*C). The relative ratios of the star forms (*,*C) versus the mature for different techniques are fNN_eNN(0.176, 0.2) and noNN(0.0672, 0.23). Thus, the star sequence abundance is strongly dependent on the sequencing method.

piRNA sequencing

piRNAs are small RNAs, 28–32 nt long, that are exclusively expressed in animal gonads. They are involved in transposon control and germline maintenance. A distinguishing feature of primary piRNAs is the bias for a T at position 1, and a change in this bias indicates piRNA processing defects. In one experiment, small RNA libraries were generated from wild-type and mutant samples using 5′ adapters with same ends (TC). The resultant sequence sets showed >80% T-bias. However, generation of biological replicates from additional mutant samples, but now using 5′ adapters with different 3′-ends, resulted in varying T-biases: 73% (for an adapter ending in GA), 69% (TA) and 57% (AA) (private communication, Pillai laboratory). This suggests that a careless choice of adapters can give rise to erroneous conclusions. Even comparisons between libraries generated with the same adapters could be misleading if different small RNAs in the two libraries have differing ligation efficiencies, masking changes in relative abundances.

Practical implications

The practical implications of our studies are as follows:

- Profiling by sequencing needs to be done using pools of adapter sequences.
- Isomir profiles generated using a single adapter sequences need to be revisited.
- Many studies have reported end-modifications of mature sequences, such as, uridylation (12). The modifications might have been under-(or over-)reported, because of the biases in the activity of the ligases.
- The isoforms identified as mature in mirBase (37) are usually the dominant ones, which may reflect the biases of the profiling methods rather than their natural biological enrichment.
- Barcoding of samples using adapters, for multiplexing sequencing, should be done carefully. The barcodes could be placed either in front of the NN ends on the 5′ adapter, or after the NN on the 3′ adapter, avoiding distortions in the results due to the ligation biases.

CONCLUSION

This study has proved that RNA ligases derived from T4-phage exhibit significant sequence specificity in their activity. The profiles of small RNAs are strongly dependent on the adapters used for sample preparation. In light of this, the current, popular, sRNA-seq protocols need revision. We find that a mix of adapters, with different sequence ends, permits a more accurate estimation of the amounts of individual miRNA sequences and their isoforms. The use of RNA ligases in other protocols, such as oligoribonucleotide circularization (38), should be reviewed for possible effects of the bias discussed in this study.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Ona Bloom, Benjamin Hubert and Lenny Teytelman gave many detailed comments and suggestions. Ramesh Pillai shared piRNA data, suggestions and gave enthusiastic support. The Lemischka lab at MSSM kindly donated the R1 mouse ES cells. Ben tenOever, Julius Brennecke, Colin Malone and Greg Hannon gave us encouragement, broad criticisms and comments. Numerous discussions with Andrew Chess were helpful in clarifying our message.

FUNDING

Genome Institute at Mount Sinai School of Medicine, NY; NIH Pathfinder Award (DP2DK083052-01, to B.D.B.); Juvenile Diabetes Research Foundation (17-2010-770). Funding for open access charge: The Genome Institute at Mount Sinai School of Medicine.

Conflict of interest statement. None declared.

REFERENCES

1. Berezikov,E., Thuemmler,F., van Laake,L.W., Kondova,I., Bontrop,R., Cuppen,E. and Plasterk,R.H.A. (2006) Diversity of microRNAs in human and chimpanzee brain. *Nat. Genet.*, **38**, 1375–1377.
2. Chiang,H.R., Schoenfeld,L.W., Ruby,J.G., Auyeung,V.C., Spies,N., Baek,D., Johnston,W.K., Russ,C., Luo,S., Babiarz,J.E. *et al.* (2010) Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes Dev.*, **24**, 992–1009.
3. Landgraf,P., Rusu,M., Sheridan,R., Sewer,A., Iovino,N., Aravin,A., Pfeffer,S., Rice,A., Kamphorst,A.O., Landthaler,M. *et al.* (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, **129**, 1401–1414.

4. Lee,L.W., Zhang,S., Etheridge,A., Ma,L., Martin,D., Galas,D. and Wang,K. (2010) Complexity of the microRNA repertoire revealed by next-generation sequencing. *RNA*, **16**, 2170–2180.
5. Aravin,A.A., Hannon,G.J. and Brennecke,J. (2007) The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science*, **318**, 761–764.
6. Cerutti,H. and Casas-Mollano,J.A. (2006) On the origin and functions of RNA-mediated silencing: from protists to man. *Curr. Genet.*, **50**, 81–99.
7. Czech,B., Malone,C.D., Zhou,R., Stark,A., Schlingeheyde,C., Dus,M., Perrimon,N., Kellis,M., Wohlschlegel,J.A., Sachidanandam,R. *et al.* (2008) An endogenous small interfering RNA pathway in drosophila. *Nature*, **453**, 798–802.
8. Tam,O.H., Aravin,A.A., Stein,P., Girard,A., Murchison,E.P., Cheloufi,S., Hodges,E., Anger,M., Sachidanandam,R., Schultz,R.M. *et al.* (2008) Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature*, **453**, 534–538.
9. Chen,C., Ridzon,D.A., Broomer,A.J., Zhou,Z., Lee,D.H., Nguyen,J.T., Barbisin,M., Xu,N.L., Mahuvakar,V.R., Andersen,M.R. *et al.* (2005) Real-time quantification of microRNAs by stem-loop RT–PCR. *Nucleic Acids Res.*, **33**, e179.
10. Shi,R. and Chiang,V.L. (2005) Facile means for quantifying microRNA expression by real-time PCR. *BioTechniques*, **39**, 519–525.
11. Goff,L.A., Yang,M., Bowers,J., Getts,R.C., Padgett,R.W. and Hart,R.P. (2005) Rational probe optimization and enhanced detection strategy for microRNAs using microarrays. *RNA Biol.*, **2**, 93–100.
12. Baccarini,A., Chauhan,H., Gardner,T.J., Jayaprakash,A.D., Sachidanandam,R. and Brown,B.D. (2011) Kinetic analysis reveals the fate of a MicroRNA following target regulation in mammalian cells. *Curr. Biol.*, **21**, 369–376.
13. Baker,M. (2010) MicroRNA profiling: separating signal from noise. *Nat. Methods*, **7**, 687–692.
14. Git,A., Dvinge,H., Salmon-Divon,M., Osborne,M., Kutter,C., Hadfield,J., Bertone,P. and Caldas,C. (2010) Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. *RNA*, **16**, 991–1006.
15. Linsen,S.E.V., de Wit,E., Janssens,G., Heater,S., Chapman,L., Parkin,R.K., Fritz,B., Wyman,S.K., de Bruijn,E., Voest,E.E. *et al.* (2009) Limitations and possibilities of small RNA digital gene expression profiling. *Nat. Methods*, **6**, 474–476.
16. Brennecke,J., Aravin,A.A., Stark,A., Dus,M., Kellis,M., Sachidanandam,R. and Hannon,G.J. (2007) Discrete small RNA-generating loci as master regulators of transposon activity in drosophila. *Cell*, **128**, 1089–1103.
17. Berezikov,E., Cuppen,E. and Plasterk,R.H.A. (2006) Approaches to microRNA discovery. *Nat. Genet.*, **38**, S2–S7.
18. Elbashir,S.M., Lendeckel,W. and Tuschl,T. (2001) RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev.*, **15**, 188–200.
19. Hafner,M., Landgraf,P., Ludwig,J., Rice,A., Ojo,T., Lin,C., Holoch,D., Lim,C. and Tuschl,T. (2008) Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods*, **44**, 3–12.
20. Lau,N.C., Lim,L.P., Weinstein,E.G. and Bartel,D.P. (2001) An abundant class of tiny RNAs with probable regulatory roles in caenorhabditis elegans. *Science*, **294**, 858–862.
21. Pfeffer,S., Lagos-Quintana,M. and Tuschl,T. (2005) Cloning of small RNA molecules. *Curr. Protocols Mol. Biol.*, Unit 26.4.
22. Nandakumar,J., Shuman,S. and Lima,C.D. (2006) RNA ligase structures reveal the basis for RNA specificity and conformational changes that drive ligation forward. *Cell*, **127**, 71–84.
23. Romaniuk,E., McLaughlin,L.W., Neilson,T. and Romaniuk,P.J. (1982) The effect of acceptor oligoribonucleotide sequence on the t4 RNA ligase reaction. *Eur. J. Biochem. / FEBS*, **125**, 639–643.
24. Kuchenbauer,F., Morin,R.D., Argiropoulos,B., Petriv,O.I., Griffith,M., Heuser,M., Yung,E., Piper,J., Delaney,A., Prabhu,A. *et al.* (2008) In-depth characterization of the microRNA transcriptome in a leukemia progression model. *Genome Res.*, **18**, 1787–1797.
25. Olson,A.J., Brennecke,J., Aravin,A.A., Hannon,G.J. and Sachidanandam,R. (2008) Analysis of large-scale sequencing of small RNAs. *Pac. Symp. Biocomput.*, 126–136.
26. Amitsur,M., Levitz,R. and Kaufmann,G. (1987) Bacteriophage t4 anticodon nuclease, polynucleotide kinase and RNA ligase reprocess the host lysine tRNA. *EMBO J.*, **6**, 2499–2503.
27. Wang,L.K., Ho,C.K., Pei,Y. and Shuman,S. (2003) Mutational analysis of bacteriophage t4 RNA ligase 1. *J. Biol. Chem.*, **278**, 29454–29462.
28. Farazi,T.A., Spitzer,J.I., Morozov,P. and Tuschl,T. (2011) miRNAs in human cancer. *J. Pathol.*, **223**, 102–115.
29. Jeffrey,S.S. (2008) Cancer biomarker profiling with microRNAs. *Nat. Biotechnol.*, **26**, 400–401.
30. Goff,L.A., Davila,J., Swerdel,M.R., Moore,J.C., Cohen,R.I., Wu,H., Sun,Y.E. and Hart,R.P. (2009) Ago2 immunoprecipitation identifies predicted microRNAs in human embryonic stem cells and neural precursors. *PLoS One*, **4**, e7192.
31. Marson,A., Levine,S.S., Cole,M.F., Frampton,G.M., Brambrink,T., Johnstone,S., Guenther,M.G., Johnston,W.K., Wernig,M., Newman,J. *et al.* (2008) Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell*, **134**, 521–533.
32. Morin,R.D., O'Connor,M.D., Griffith,M., Kuchenbauer,F., Delaney,A., Prabhu,A., 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.
33. Shi,Y., Zhao,X., Hsieh,J., Wichterle,H., Impey,S., Banerjee,S., Neveu,P. and Kosik,K.S. (2010) MicroRNA regulation of neural stem cells and neurogenesis. *J. Neurosci. Off. J. Soc. Neurosci.*, **30**, 14931–14936.
34. Brown,B.D., Gentner,B., Cantore,A., Colleoni,S., Amendola,M., Zingale,A., Baccarini,A., Lazzari,G., Galli,C. and Naldini,L. (2007) Endogenous microRNA can be broadly exploited to regulate transgene expression according to tissue, lineage and differentiation state. *Nat. Biotechnol.*, **25**, 1457–1467.
35. Bissels,U., Wild,S., Tomiuk,S., Holste,A., Hafner,M., Tuschl,T. and Bosio,A. (2009) Absolute quantification of microRNAs by using a universal reference. *RNA*, **15**, 2375–2384.
36. Silva,J.M., Sachidanandam,R. and Hannon,G.J. (2003) Free energy lights the path toward more effective RNAi. *Nat. Genet.*, **35**, 303–305.
37. Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.
38. Wang,L. and Ruffner,D.E. (1998) Oligoribonucleotide circularization by 'template-mediated' ligation with t4 RNA ligase: synthesis of circular hammerhead ribozymes. *Nucleic Acids Res.*, **26**, 2502–2504.