Comprehensive assessment of multiple biases in small RNA sequencing reveals significant differences in the performance of widely used methods

Authors: Carrie Wright*[1,2], Anandita Rajpurohit*[1], Emily E. Burke[1], Courtney Williams[1], Leonardo Collado-Torres[1], Martha Kimos[1], Nicholas J. Brandon[3], Alan J. Cross[3], Andrew E. Jaffe[1, 4-11], Daniel R. Weinberger[1, 6-9], and Joo Heon Shin[1,7]

*These authors contributed equally to this work

Affiliations

1. Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore, MD, 21205 USA
2. AstraZeneca Postdoc Program, Innovative Medicines and Early Development Biotech Unit, Cambridge, MA, USA
3. AstraZeneca Neuroscience, Innovative Medicines and Early Development Biotech Unit, Cambridge, MA, USA
4. Center for Computational Biology, Johns Hopkins University, Baltimore, MD, USA
5. Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
6. Department of Psychiatry and Behavioral Sciences, Johns Hopkins School of Medicine, Baltimore, MD, USA
7. Department of Neurology, Johns Hopkins School of Medicine, MD, USA
8. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA
9. The Solomon H. Snyder Department of Neuroscience, Johns Hopkins School of Medicine, MD, USA
10. Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
11. Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA; [7]Baltimore, MD, 21205 USA

Corresponding author contact information: Daniel R. Weinberger, address: 855 North Wolfe Street, Suite 300, 3rd Floor, Baltimore, MD, 21205, email: drweinberger@libd.org, phone: 410-955-1000, Fax: 410-955-1044

Abstract

High-throughput sequencing offers advantages over other quantification methods for microRNA (miRNA), yet numerous biases make reliable quantification challenging. Previous evaluations of reverse transcription or amplification bias in small RNA sequencing have been limited. Furthermore, little work has evaluated quantifications of isomiRs (miRNA isoforms) or the influence of starting amount on performance. We therefore evaluated quantifications of canonical miRNA and isomiRs using four library preparation kits, with various starting amounts (100ng to 2000ng), as well as quantifications following removal of duplicate reads using unique molecular identifiers (UMIs) to mitigate reverse transcription and amplification biases. Randomized adapter and adapter-free methods mitigated bias; however, the adapter-free method was especially prone to false isomiR detection. We demonstrate that using UMIs improves accuracy and we provide a guide for input amounts to improve consistency. Our data show differences and limitations of current methods, thus raising concerns about the validity of quantification of miRNA and isomiRs.

Research of miRNA expression has been instrumental in identifying miRNAs involved in development and diseases [1], and identifying expression-signatures for use as biomarkers [2–4]. Small RNA sequencing (sRNA-seq) allows for detection of novel miRNAs and altered canonical miRNA sequences, termed isomiRs [5–7]. Despite this enhanced capability, sRNA-seq miRNA quantifications are often inconsistent across studies [8]. This is likely in part due to differences between methods and/or variation in the detection by individual methods [9] (from library preparation to preprocessing to normalization, etc.). Furthermore, several aspects of sRNA-seq can lead to the preferential quantification of some miRNAs and reduced or completely lacking quantification of others, thus introducing biases that lead to misrepresentations of true miRNA expression levels. Evaluations and comparisons of the accuracy (how close measurements are to the truth) and consistency (how close measurements are across replicates) associated with current methods are critical for proper cross-study interpretation and for guiding methodological improvement.

Evidence suggests that biases and inconsistencies in sRNA-seq based quantifications and group comparisons are largely based on study design and library preparation methods [9]. The details of these issues have been reviewed elsewhere [8,10–16]. Some of these issues are avoidable with proper study design. However, bias and inconsistency related to adapter ligation, cDNA synthesis, and amplification may principally be dependent on library preparation and preprocessing methods, which are less readily controlled.

A considerable number of studies have evaluated adapter ligation bias in quantifications from several commercially available kits [15,17–19]; however, to our knowledge, only one study has directly compared the performance of randomized adapter methods and adapter ligation-free methods [15]. Furthermore, limited studies have investigated the influence of reverse transcription or amplification bias in sRNA-seq [20,21] and no study to date has evaluated the use of unique molecular identifiers (UMIs) in order to identify and remove duplicate reads to mitigate such biases in sRNA-seq biological samples. While a couple of studies have used UMIs in sRNA-seq [20,21], only one report has evaluated the reproducibility of sRNA-seq quantifications obtained from utilizing UMIs to those without [20], in which the authors concluded that biological technical replicates had less variation when UMIs were used to remove duplicate reads compared to when either all or no duplicates were removed. However, no statistical tests were performed in this assessment, and no evaluation of the influence of the UMI deduplication on the accuracy of biological samples was performed. In the sRNA-seq literature there has also been little assessment of the influence of starting amount on the consistency of quantifications. While RNA editing detection has been evaluated [17], other aspects of isomiR analysis have not yet been performed.

To complete the gap left by previous studies, we comprehensively evaluated and compared miRNA and isomiR quantifications from four commercially available library preparation methods, as well as those obtained following the removal of duplicate reads using UMIs. We also evaluated the consistency of the results using a variety of starting amounts. We assessed the similarity of the quantifications from each method, the diversity of the detection of different types of small RNAs by each method, as well as the accuracy and the consistency of the results obtained from each method within and across batch. Such evaluations are critical for optimizing sRNA-seq methods to obtain both reliably consistent and accurate results across batches and studies, and to therefore allow for more accurate and reproducible miRNA quantifications in disease states and conditions. Based on these results, we offer suggestions for future study designs.

Results

Study Design

In this study we evaluated the influence of several potential sources of bias and inconsistency on miRNA quantifications (**Fig. 1.A**) by comparing the performance of four commercially available kits (**Fig. 1.B**) and two preprocessing methods (**Fig. 1.C**) using various starting amounts (100ng to 2,000ng) for each method (**Fig. 1.D**). The following library preparation kits were compared: 1) the Clontech SMARTer smRNA-Seq Kit for Illumina, now owned by Takara Bio (Clontech), which incorporates adapter and index sequences during reverse transcription and amplification and is therefore ligase-free; 2) the Bioo Scientific NEXTflex Illumina Small RNA Sequencing Kit v3 (NEXTflex), now owned by Perkin Elmer and called NEXTFLEX, which utilizes adapter sequences with random nucleotide sequences adjacent to the miRNA binding location giving each miRNA a variety of adapter sequences to bind; 3) the Illumina TruSeq Small RNA Library Prep Kit (Illumina); and 4) the New England BioLabs Next Multiplex small RNA kit (NEB). Based on our literature search Illumina and NEB sRNA-seq kits appear to be the first and second most widely used kits to date, respectively. The NEB and the NEXTflex kits include polyethylene glycol in an effort to reduce adapter ligation bias by improving overall ligation efficiency.

We also evaluated the influence of reverse transcription and amplification bias by utilizing the random sequences within the adapters of the NEXTflex kit (that are added prior to the cDNA synthesis and PCR amplification steps), as UMIs. These UMIs allow for the removal of duplicate reads introduced during amplification and possible mitigation for sequences that may have been preferentially reverse transcribed (**Fig. 1.C**). We will hereinafter refer to these data as "Deduped". To determine if differences identified between the Deduped data and the NEXTflex data were simply due to a reduction in the number of reads (as the UMI-based deduplication process reduces the data down to 5% of the original), we also included a random 5 % subset of the NEXTflex reads, hereinafter referred to as "Fivepercent," for comparison.

We evaluated two types of samples (**Fig. 1.D-E**) and processed the data following the methods outlined in **Fig. 1.E**. See the **online methods** for more details of our experimental approach. We then evaluated several questions shown in **Fig. 1.F** about the similarity of the quantifications obtained from the 6 tested methods (**Fig. 1. B**), the accuracy of those quantifications using synthetic miRNAs in equimolar concentration, the ability of each method to detect a variety of miRNAs and isomiRs, and the consistency of the quantifications by each method of technical replicates within the same batch and across two batches.

Similarity - Overall quantifications are similar, yet results for individual miRNAs are quite divergent across methods

We first performed a general evaluation of the similarity of the resulting miRNA quantifications from each method (**Fig. 1.B**) and of major contributors to overall variability using the data derived from the same human brain sample across technical replicates (**Fig. 1.F.Similarity**). Hierarchical cluster analysis indicated that the samples generally clustered by method and starting amount (**Fig. 2.A)**. Differential analysis and correlation analysis (**Supplementary Fig. 1**) of the miRNA expression estimates revealed that the methods (**Fig. 1.B**) produce overall relatively similar results, however some individual miRNAs showed very different quantifications with intensity ratios ranging as extreme as -9 to 6 (**Fig. 2.B**).

Evaluating the top 20 abundant miRNAs from each method (**Supplementary Table 1**), only 6 miRNAs (30%) overlapped across all methods (however the top 20 for Fivepercent were identical to the top 20 from the raw full NEXTflex data). Thus, emphasizing only the most abundant miRNAs for further study may be problematic. The overlap between the most abundant miRNAs detected by Clontech and the other methods was lower (45% to 55%) than the overlap between Illumina, NEB, and NEXTflex (60-65%). The Deduped method resulted in an 85% overlap with the raw NEXTflex data.

Sum of squares analysis revealed that method choice was the largest contributor to miRNA count variability (on average 82% variance explained for individual miRNAs (**Fig. 2.C**) when evaluating the data from all methods (excluding the Fivepercent control). This further exemplifies the lack of consistency in quantifications that may occur when different methods are utilized.

Accuracy - Reduction of numerous biases improves accuracy

To assess the accuracy of each method (**Fig. 1.F.Accuracy**), we investigated how consistently each kit detected 962 equimolar synthetic miRNA sequences. We calculated the difference of each miRNA count from the mean count for all miRNAs for each method, which we called "accuracy error". The six methods showed significant differences in accuracy (F = 40.00, p < 2.2e-16). The Deduped data had significantly less accuracy error compared to all other methods (up to ≈ 8% less error, with an effect size Hedge's g of 0.59), followed by comparable accuracy for Clontech, NEXTflex, and Fivepercent methods (which did not significantly differ from one another in post hoc analysis), but worse accuracy for the NEB and Illumina methods. This suggests that the Illumina and NEB methods detect different sequences with less validity than the other methods. This was expected, given the known adapter ligation bias associated with these methods. Our results suggest that the methods utilized by the Clontech and NEXTflex kits both diminish bias – likely due to a reduction in adapter ligation bias. Using UMI sequences for deduping resulted in additional error reduction (the raw NEXTflex data had 2.81% more error) **(Fig. 3.A** and **Supplementary Table 2),** which may be due to a reduction in reverse transcription and/or amplification bias. This is consistent with our analysis of the overall variance of the counts for these synthetic sequences (**Fig. 3.B)**. The concordance of the rank of the sequences with higher accuracy error across the methods was poor (**Supplementary Fig. 2** and **Supplementary Table 3)**, suggesting that different sequences were prone to bias for each of the methods**.**

We thus analyzed the overall contribution of different sequence characteristics to the variance of the count estimates of the synthetic miRNAs and found that indeed different characteristics were associated with variability for the different methods **(Fig. 3.C, Fig. 3.D, Supplementary Table 4).** The secondary structure Gibbs free energy (**Supplementary Fig. 3**) was highly influential for Clontech (explaining ≈ 7% of the variance), and the NEXTflex-based methods (explaining ≈ 10 % of the variance for each). The identity of the last 2 bases was influential for all methods (**Supplementary Fig. 4**), but in particular for the NEB and Illumina methods (explaining ≈ 6 % of the variance for each), suggesting that adapter ligation of the 3' end particularly introduces bias of miRNA quantifications, in agreement with previous work [22]. The identity of the first 2 bases (5') was most influential for Clontech and explained ≈ 8% of the variance **(Supplementary Fig. 5**), suggesting that the SMART template-switching of the 5' end may introduce more bias. The number of Cs within a sequence also accounted for a relatively large percentage of the variance (2.5-5% for all methods except for Clontech) (**Supplementary Fig. 6**). Interestingly, GC content only accounted for ≈ 1% of the variance for each method.

Detection of RNA classes - Libraries generated using the Clontech Kit had very low miRNA mapping rates

We next assessed the percentage of reads that mapped to miRNA or other small RNA species for each of our brain-derived samples using *bowtie* [23] (**Fig 1.F. Detection Diversity**). We excluded the Deduped data and its control, as alignment was required to produce these data. There was a significant difference in the miRNA mapping rate of the 1000ng starting input data across the kits (F = 108.9, p-value = 5.73e-09). The NEXTflex and NEB methods had the highest rates, while the Clontech method had the lowest mapping rate, with only 1-2% of all reads mapping to miRNAs (**Fig. 4.A-B)**, as previously described [15] (**Supplementary Table 5**). There was a significant difference for all the tested types of RNA across the methods except for small Cajal body-specific RNAs (scaRNA) after multiple testing correction. The Clontech reads largely aligned to ribosomal RNA (rRNA) and had significantly higher rates of small nucleolar RNAs (snoRNA) and small nuclear (snRNA) mapping than the other methods, while the NEXTflex method resulted in the largest number of P-element induced wimpy testis (PIWI)-interacting RNA (piRNA) reads (**Supplementary Table 6)**. All of the kits had quite consistent mapping rates across the various starting amounts (**Fig. 4.B**). Mapping rates of the synthetic RNA were much more comparable among the methods, suggesting that the differences seen with the biological samples are largely due to differences in detection of other biological RNAs (**Supplementary Table 7**).

Detection of unique miRNAs - The Deduped data and Clontech data had the best detection rates

To discern if any of the methods have an advantage in detecting a diversity of unique miRNAs, we compared the detection rate of miRNA sequences (**Fig 1.F. Detection Diversity**). Here we define a miRNA as detected if the miRNA was present with at least 10 normalized reads in the quantifications for each of the triplicates of the 1000ng batch 1 brain data. The number of detected unique miRNAs was highest in the Deduped data, and lowest in the Fivepercent and Illumina data (**Fig. 4.C**)**,** which was consistent when including the second batch **(Supplementary Table 8)**. Despite the low mapping rate of the Clontech samples, the miRNA diversity detected by this kit was relatively comparable to that of the other methods tested. Since both the Deduped and the Fivepercent data also include only 5% of the total raw NEXTflex reads, both of these methods also resulted in a much lower number of reads that could map to miRNA. The similarity of the detection rates of all the methods despite the large difference in miRNA mapping rates is due to the DESeq2[24] normalization strategy utilized, which accounts for differences in library composition, and the high sequencing depth. An analysis of subsamples containing only 10 million, 5 million or 1 million reads of the Clontech data resulted in lower detection diversity **(Supplementary Table 9).**

Using the data from all starting amounts, there was a significant difference in the number of detected miRNAs across methods (F = 7.69, p-value = 0.00017), however pairwise comparisons were largely nonsignificant (**Supplementary Table 10**). There was a weak but significant positive relationship (r = 0.4, p-value = 0.027) between detection diversity and input amount (**Supplementary Fig. 7**). Thus, as anticipated, larger inputs resulted in a more diverse pool of detected unique miRNAs; however, the pool size did not differ greatly (**Supplementary Table 11**). When evaluating each kit individually, only the Deduped method had a significant (p-value = 0.009) and strongly positive relationship between starting amount and the number of unique detected miRNAs (r = 0.92).

Detection consistency - The Clontech method was significantly worse than the others

To determine how well each method consistently detected the same miRNAs (**Fig. 1.F.**

Detection Diversity), we calculated the proportion of miRNAs detected for each sample that were not detected by the other two samples within the triplicates as a measure of detection inconsistency using the 1000ng input data (**Fig. 4.D**). There was a significant difference in the inconsistency of detection overall between the methods (F = 12.27, p-value = 0.0002), and although no individual contrasts between pairs of methods were significant in post-hoc analysis, the Clontech data resulted in the highest level of inconsistency, and NEB performed the best, with the lowest level of inconsistency.

Analysis of the full set of data including all starting amounts (**Supplementary Fig. 8, Supplementary Table 12**) demonstrated a significant difference in the inconsistency across methods (F = 14.83, p-value = 1.65e-10) and starting amounts (t = -3, p-value = 0.00257, Pearson r = -0.31). There was significantly more inconsistency for the Clontech method compared to all other methods (up to 240%) except the Fivepercent control method. This suggests that although the Clontech level of detection may have been rescued by the high depth of sequencing, the low mapping rate may still result in much poorer consistency of detection.

Detection overlap - Despite different miRNA mapping rates, all the methods capture overlapping miRNAs

We next characterized the overlap of unique miRNA sequences captured by each method (**Fig. 1.F. Detection Diversity**). Evaluating the miRNAs consistently detected by all 1000ng triplicates of the first batch, we determined that in general a large proportion of the miRNAs were commonly detected by all of the methods (on average 74%), and only a small fraction of miRNAs was uniquely detected by a single method (4.7% on average) (**Fig. 4.E**). In contrast, on average only 5% of the detected isomiRs by each method overlapped those detected by all the other methods (**Supplementary Fig. 9**).

Detection of isomiRs - The methods detected significantly different numbers of isomiRs - Clontech method the most

We next evaluated the isomiR detection rate of each method (**Fig. 1.F. Detection Diversity**). We define an isomiR as detected if it had greater than 100 normalized reads in all triplicates for each method of the 1000ng input data. We observed the largest number of unique isomiR sequences in the Clontech data and the lowest in the NEXTflex data (**Fig. 4.F**). When evaluating detection across both batches, the Clontech data remained the most diverse, while the Fivepercent control detected the lowest number of unique isomiRs. Using all the data derived from all the starting amounts, we determined that there was a significant difference in the number of isomiRs detected across the methods (F = 83.5, p-value = 8.89e-15), but not across starting amounts (**Supplementary Fig. 10**). Clontech detected the largest number (up to 250% more), followed by NEB (up to 169% more) and Illumina (up to 147% more), while the NEXTflex based methods similarly detected the least (**Supplementary Table 13**).

When evaluating the consistency of isomiR detection (**Fig. 4.G**), there was a significant difference in the consistency of detection (F = 5.9, p-value = 0.006), but again no individual contrasts between pairs of methods were significant. The Illumina data had the highest inconsistency, while the NEB data had the least.

Detection of false positive isomiRs - All methods detected false isomiRs, especially Clontech and NEB

To assess the possibility that the methods may result in false isomiR detections, we evaluated the presence of isomiRs in the synthetic miRNA data (which has no designed isomiRs) (**Fig. 1.F. Detection Diversity**). False isomiRs were detected by all of the tested methods. We compared the methods based on: 1) the number of overall unique detected isomiR sequences, 2) the number of unique isomiR sequences detected for each individual synthetic sequence, and 3) the quantifications of the individual false isomiRs. There was a significant difference in the number of isomiRs detected for each synthetic sequence by the methods (F = 176.37, p-value = <2.2e-16). The Clontech method detected more unique isomiRs overall than all the other tested methods (on average 401% more false isomiRs) (**Fig. 4.H**). The number of unique isomiRs observed for each synthetic sequence was also significantly higher. On average the Clontech method resulted in 14 isomiRs per synthetic sequence, while the NEXTflex-based methods (the raw NEXTflex data, the Deduped, and the Fivepercent) resulted in roughly 2 isomiRs per synthetic sequence (**Fig. 4.I, Supplementary Table 14**). The counts observed for the individual isomiRs detected were significantly higher for NEB than all the other methods (with 60% higher expression than the isomiRs detected by the NEXTflex based methods) (**Fig. 4.J**). The NEXTflex methods (raw, Deduped, and Fivepercent) resulted in the fewest isomiRs detected, with the fewest isomiR counts per synthetic sequences, and with the lowest expression. There was no difference between the Deduped and the raw NEXTflex data for the expression of the isomiRs or in the number detected per synthetic sequence (**Supplementary Table 15**).

Sequence feature analysis revealed that the identity of the first two bases of the 5' end accounted for most of the variance in the number of isomiRs detected for each synthetic sequence for the Clontech kit (accounting for nearly 9% of the variance) (**Fig. 4.K, Supplementary Table 16**). Therefore, false positive isomiRs may be generated during the reverse transcription step of the library preparation for this method. This is consistent with other studies that suggest that the template-switching reverse transcription method utilized by Clontech of the 5' end can lead to shortened miRNA transcripts in a process called strand invasion [25] and potentially longer miRNA transcripts due to concatamers of the template-switching oligo [26]. In contrast, the last 2 bases on the 3' end accounted for the largest amount of variance of the other methods (on average 5.3%).

Consistency across Batch - Illumina had the lowest consistency, while the other methods performed similarly

To determine the consistency of results obtained across batches for each method, we compared the mean of the triplicates in one batch to a second batch of a single sample of the same human brain (**Fig. 1.F.Consistency**). Using the normalized and log transformed reads for miRNAs that were found to be detected by each kit when filtering for greater than 10 reads across all 4 samples for each kit, we calculated the distance from the mean of the two batches for each detected miRNA individually. Overall, method choice had a weak significant association with error across batch (F = 2.39, p-value =0.036). This association was driven by the batch error of the Illumina method which was significantly higher than the other methods, with up to 74% more error than other methods (**Fig. 5.A, Supplementary Table 17**). NEB, NEXTflex, and the Deduped data were the most consistent across batch, with no significant difference in the performance of these methods (p>0.05). The top miRNAs showed some level of concordance across the methods (**Supplementary Fig. 11**).

Consistency across triplicates - Clontech and Illumina had the lowest consistency

We then evaluated the consistency of the triplicates (**Fig. 1.F.Consistency**) within the 1000ng data, by calculating the distance of each triplicate from the mean of all three triplicates. We determined that there was a significant difference across the methods (F = 36.7, p-value = <2.2e-16). Consistency was significantly higher for the NEB and Deduped methods, while Clontech, Illumina, and the random Fivepercent had the lowest consistency (with ≈ 20-40% more error, **Fig. 5.B, Supplementary Table 18)**. Deduping of the NEXTflex data improved consistency. The raw data had 14% more error between triplicates.

We then calculated the triplicate consistency for each starting amount. We determined that using all starting amount data, there was still a significant difference in triplicate consistency between the methods (F = 79.7, p-value =<2.2e-16), but there was no relationship with starting amount (Pearson r = -0.11) (**Fig. 5.C**). All pairs of methods were significantly different, except for the contrasts between NEB and Deduped and between Clontech and Illumina (**Fig. 5.C, Supplementary Table 19**). NEB and Deduped again had the greatest consistency (up to 23% less inconsistency) and Clontech and Illumina had the least (≈17% more inconsistency).

Factors affecting consistency - The most abundant miRNAs were the most inconsistently detected for each method

To determine if any aspect of the miRNA sequences was associated with more or less consistency across batch (**Fig. 1.F.Consistency**), we evaluated the association of various sequence factors with the batch error estimate. For each method, the expression of each individual miRNA was the largest contributor to variance of batch error (**Fig. 5.D-E)**. All methods showed a significant and positive relationship between expression and inconsistency across batch (Pearson r >= 0.83 for all methods), **Fig. 5.F**.

Evaluating sequence characteristic associations with triplicate consistency, again, expression was the largest contributor to variance of error estimates (**Fig. 5.G-H, Supplementary Table 21**) and again all methods showed a significant positive association with expression and inconsistency across triplicates (Pearson r >0.82 for all methods), **Fig. 5.I**.

To determine if the same miRNAs showed high error across the starting amounts or methods, we ranked the triplicate consistency error estimates and compared the ranks between the starting amounts of a given method and between methods (**Supplementary Fig. 12**, **Supplementary Fig. 13**). The concordance of the ranks between starting amounts and methods was highest among the sequences with the highest error with roughly 40% concordance.

Consistency and its relationship to starting amount - There was no improvement in consistency beyond 500ng of total RNA for most methods

Using data normalized and filtered for greater than 10 normalized reads for each method individually, we further evaluated the influence of starting amount on consistency across triplicates (**Fig. 1.F.Consistency**). Overall, starting amount was significantly associated ($p<0.05$) with triplicate consistency error for each method except for Illumina, which is likely due to the fact that fewer starting amounts were assessed for this method, **Fig. 5.C, Supplementary Table 22.** The results suggest that a larger starting amount will generally improve consistency, see **Supplementary Table 23** for specific guidance for each kit. For most methods the highest consistency with the lowest starting amount was achieved with 500ng, however, 1000ng improved the consistency of the Deduped data. The consistency was relatively similar for all the Clontech kit samples regardless of starting amount.

Discussion

We report an extensive comparison of commonly used sRNA-seq kits for their performance in identifying and quantifying miRNAs, as well as the results obtained with the use of a UMI and a UMI control. Our detailed analyses identify critical factors that influence their performance. Prior performance evaluations of current sRNA-seq methods have been very limited and adapter ligation bias has largely been the focus of earlier reports [17,27–29]. Several studies have compared the NEXTflex kit with the Illumina and NEB kits [15,17–19,30,31], and most suggest that the NEXTflex kit offers advantages due to reduced adapter ligation bias by including randomized adapters. We compared the NEXTflex kit with the Clontech kit which was also designed to mitigate adapter ligation bias, but by using an adapter ligation-free method. Only one prior study has compared the performance of these two kits using a previous and now discontinued version of the NEXTflex kit [15], which demonstrated that the Clontech kit resulted in less bias, however, only 6 synthetic miRNAs were utilized in their accuracy assessment. A recent study performed at the same time as ours agreed with our findings that these two kits perform similarly for accuracy [32]. A similar UMI method is utilized by a recent library preparation kit by Qiagen. However, this kit was released after the data collection of our analysis. In addition, this kit, similarly to the NEB and Illumina kits, does not include methods to reduce adapter ligation bias, and the UMI is added after reverse transcription, which therefore does not allow for any reduction in bias associated with this step. The results of a recent study, which performed a similar analysis as ours, further suggest that the Qiagen kit has more bias and is less accurate than the Clontech and NEXTflex Kits [32].

We have compared the quantifications from each method using a variety of metrics including: 1) Similarity – how similar are the quantifications across methods (**Fig. 1.F.Similarity**) ; 2)

Accuracy – how well does each method equally quantify different equimolar miRNAs (**Fig. 1.F.Accuracy**); 3) Detection diversity – what capacity does each method have to capture a diverse range of unique small RNAs (**Fig. 1.F.Detection Diversity**); and 4) Consistency – how similar are results across batches, triplicates, and different starting inputs (**Fig. 1.F.Consistency**). Our analysis of individual sequences using the metric tests provide information about potential bias due to adapter ligation, reverse transcription, and amplification. **Table 1** summarizes our results. Overall, there are clear and important differences between the methods tested and all show performance limitations in real world small RNA sequencing. Based on our results, we propose a number of suggestions for future studies.

First, cross-study comparisons using different methods should be viewed with skepticism, because although the kits resulted in fairly similar results overall, quantifications of individual miRNAs, including the most abundant miRNAs, varied widely across methods (**Supplementary Table 1, Fig. 2.B, Supplementary Fig. 1**). In particular, the Clontech methods resulted in the most dissimilar results (**Supplementary Fig. 1**). More research is required to determine how to best utilize data derived from different sRNA-seq methods for mega- and meta-analyses. We also advise against further study of only the top expressed miRNAs from a single sRNA-seq study, particularly when a more biased method is utilized, as the top observed miRNAs may not be truly among the most abundant or influential, but instead those that are preferentially observed by the method. This issue has previously been discussed at length[16]. It is important to note that it remains unclear how relatively abundant a miRNA needs to be to exert biological importance in different contexts.

We suggest the use of a degenerate base method, such as NEXTflex or a ligation-free method, to improve accuracy. These methods appeared to equally improve accuracy, likely due to a reduction of adapter ligation bias (**Fig. 3.A**-**B**). We suggest that future small RNA studies utilize a UMI strategy, as our NEXTflex data preprocessed to account for UMI duplicates was even more accurate, reducing the overall variance of the $\log_2$ transformed and normalized quantifications by 68%, or on average the difference from the mean for each miRNA by nearly 3% (**Fig. 3.A-B, Supplementary Table 2**). We speculate that our deduplication method led to such improvements due to reduced reverse transcription and/or amplification bias. Our sequence-specific analysis further indicated that secondary structure of miRNAs was one of the largest contributors to error of the Clontech and NEXTflex kits for the accuracy assessment, which appeared to be mitigated in the UMI deduped data for the NEXTflex kit (**Fig. 3.C-D**). This suggests that the secondary structure of miRNA sequences may be particularly influential for reverse transcription and/or amplification bias, in agreement with previous work that indicates that secondary structure can indeed influence reverse transcription [33]. More work is required to determine the extent that amplification or reverse transcription are particularly contributing to bias, and to what extent each are mitigated by the use of UMIs. Furthermore, it is unclear if the use of deduplication would improve other methods beyond the performance level of the current NEXTflex kit. However, the UMIs are inherently already included in the NEXTflex adapters, making this one of the best current options to mitigate bias in sRNA-seq.

All of the methods tested were capable of detecting a diverse range of miRNA sequences and there was a high degree of overlap in the identity of the miRNAs detected by each method (**Fig. 4.C-E**). Therefore, any of the tested methods may be appropriate for assessments about general miRNA diversity. However, the identity of miRNAs detected by Illumina varied greatly across batch, **Supplementary Table 8**). We observed greater resolution for detection of a larger variety of miRNAs with greater sequencing depth. We did not evaluate depths above 20 million reads, so it remains unknown if even greater resolution can be achieved beyond this depth, however subsets of our 20 million depth data resulted in a reduction of diversity. We also observed that a more diverse pool was detected with larger input amounts; therefore, for the best diversity of detection, we recommend using the largest input possible.

The Clontech kit resulted in the largest percentage of reads mapping to snoRNAs and snRNAs, while the NEXTflex kit resulted in the largest percentage of piRNA mapping reads (nearly 4.2 times higher than Clontech) (**Fig. 4.A-B, Supplementary Table 6**). Therefore, if these particular small RNAs are of interest, we would suggest the use of these two kits respectively. We did not evaluate the diversity of these other classes of small RNAs; however, given the results of our miRNA analysis, we predict that deeper sequencing will result in greater resolution of diversity.

We especially suggest using randomized adapter methods, such as NEXTflex, for studies involving isomiR analysis. We suggest that all isomiR studies utilize an additional method for validation, as all methods resulted in the observation of false isomiRs. In particular, the Clontech method resulted in the highest level, thus we do not suggest that others utilize this method for studies that aim to evaluate isomiR expression (**Fig. 4.H-J**). Furthermore, because this method utilizes polyadenylation of the 3' end, it is impossible to truly distinguish isomiRs that terminate with 3' adenine bases. In all, the Deduped method resulted in the highest number of detected miRNAs with the lowest false isomiR detection (**Fig. 4.H-J**). Therefore, of the tested methods, we suggest that the Deduped method may be the best for detecting the most diverse and reliable set of miRNAs.

The Deduped method was also the most consistent for individual miRNA quantifications across triplicates within the same batch (**Fig. 5.B**) Therefore, we suggest the use of this method for optimal consistency. In general, we particularly caution against the use of Illumina when multiple batches of sequencing will be involved in a study, as this method resulted in significantly more inconsistent results across batches relative to all the other tested methods (**Fig. 5.A, Supplementary Table 8**)

An earlier analysis determined that detection consistency was poorer with much smaller starting amounts (10ng)[18]. Agreeing with this, our results indicate that larger starting amounts for some methods may mitigate inconsistent quantifications of miRNAs and isomiRs.
Overall, we observed the most consistent results across triplicates when utilizing 500ng or greater of starting input. In most cases, 500ng was sufficient, and no improvement was achieved with higher input amounts. However, the Deduped method performed best with at least 1000ng and the Clontech method resulted in similar levels of consistency despite the use of smaller inputs (**Supplementary Table 23**). Thus, if differing starting amounts or smaller starting amounts are required, and interest in isomiRs is limited, the Clontech method may be

the best choice.

Additional studies of UMI use for other library preparation methods and across biological samples are necessary to further understand the ability of UMIs to improve the consistency and reproducibility of sRNA-seq studies. Further work is also necessary to optimize the length of the UMI. In some cases, all UMIs will become saturated if a given small RNA is very highly expressed. Our calculations suggest that this UMI length is sufficient for the brain (using our current methods), in which miRNA make up a very small fraction of the total RNA and in which our data suggested that the most abundant miRNA represented only 11% of the miRNA reads. However longer UMIs may be required for tissues with greater enrichment of miRNAs or greater enrichment of other small RNAs of interest, where single RNAs may have more than 65,536 individual copies before amplification (see **Supplementary Note 1,** which refers to **Supplementary Table 24**).

In conclusion, we observed significant differences in the accuracy, detection, and consistency of the various sRNA-seq methods tested. Our results underscore the importance of the choice of sRNA-seq method and suggest that with moderately large starting amounts, the NEXTflex kit with deduplication may produce the least biased and most consistent results within and across studies. Our results suggest that bias is introduced in sRNA-seq due to reverse transcription and/or amplification and that the use of UMIs should be considered for further optimization to mitigate these biases in future sRNA-seq studies. Additional work is needed to decipher the role of these biases in sRNA-seq in order to guide more accurate sequencing methods. Ultimately, additional standardization of sRNA-seq data generation and analysis will improve our ability to understand the expression and regulatory role of these small but important RNAs in conditions and disease.

Methods

Library preparation and sequencing:

Two sample types were used to evaluate the performance of the sRNA-seq methods, (**Fig. 1.D** and **Supplementary Table 25**). To evaluate detection and consistency we used various starting amounts in triplicate of total RNA from a homogenate human brain sample, purchased from Ambion and derived from a 74-year-old Caucasian female. The cause of death of this individual was respiratory failure. To evaluate accuracy, we used 300ng of the Miltenyi Biotec miRXplore Universal Reference equimolar pool of 962 synthetic sequences corresponding to human, rat, mouse, and virus miRNA.

Each library preparation was performed by the same two lab scientists using the same equipment. Each protocol was followed exactly as provided by the vendor for each kit. The number of PCR cycles for each sample was determined based on the recommendations of each kit for the various starting input amounts (**Supplementary Table 26**). Size selection using PAGE gels was recommended by three of the manufacturers (Illumina, NEXTflex, and NEB kits) and was performed for these kits for better comparisons. We used AMPure XP beads for size selection for the Clontech samples, as the vendor does not recommend PAGE gel size selection.

A Qubit Fluorometer was used to determine the concentration of the final libraries. The library preparations were sequenced using single-end sequencing on the Illumina HiSeq 3000 with the Illumina Real Time Analysis (RTA) module and the bcl2fastq2 v2.17 to generate 51 base pair reads.

Unique Molecular Identifier deduplication:

In order to test the use of UMIs to mitigate reverse transcription and amplification bias, reads derived from the NEXTflex kit were collapsed based on random sequences of 4 bases in length contained within both the 3' and 5' adapter sequences as a UMI using UMI-tools [34] (**Fig. 1.C**). In this method, the adapters are ligated before reverse transcription and PCR amplification, therefore allowing for the estimation of the abundance of the sequences present in the sample before these steps. In the collapsed NEXTflex data referred to as "Deduped", only reads that contained the same pair of a unique transcript with a unique UMI were maintained, while duplicate pairs were discarded. Therefore, each unique sequence had the opportunity to bind up to 65,536 different UMIs. As a control, we compared the performance of the Deduped data to a random 5% subset of the reads, referred to as "Fivepercent". This was necessary as only 5% of the total reads remained following the collapsing method which required a preliminary alignment step. Thus this data was also produced with the preliminary alignment step, all preprocessing was the same except for the use of UMI-tools[34].

We utilized an in-house script to extract the degenerate bases from the adapters to determine the UMI sequence for each read and to add it to the identifier line of the FASTQ files for each sample. In this script we also removed reads which contained any unknown bases within the UMI. We then used *bowtie* [23] (v1.2.2) with a seed length of 15 allowing for 2 mismatches to produce a liberally aligned bam file to be used with UMI-tools [34] for deduping. We utilized the directional method in UMI-tools to remove duplicate reads from the bam file. We then converted the bam file to a FASTQ file for alignment with miRge [35] with the other method samples. Our script to prepare NEXTflex samples for UMI-tools [34] is available on GitHub at https://github.com/LieberInstitute/miRNA_Kit_Comparison.

Adapter and degenerate base trimming and alignment:

For the NEXTflex (and therefore the Deduped and Fivepercent), NEB, and Illumina FASTQ files the 3' adapter sequences and all bases 3' of the adapter were trimmed from the ends of the reads using cutadapt [36] version 1.8.3. For the NEXTflex samples the first and last four bases, which correspond to the random bases included in the adapter sequences, were also trimmed. In the case of the Deduped samples these sequences were added to the identifier line prior to trimming. These bases correspond to the random adapter sequences because sequencing begins at the location of the 4 random bases in the 5' adapter for this kit.

Unlike the other kits, the Clontech kit is stranded. Read 1 corresponds to the sense strand of the input RNA and the first three bases correspond to the nucleotides added during the SMART template-switching method. Then 10 Adenine bases were removed from the 3' end, as well as all potential bases 3' of this stretch of bases.

When trimming the synthetic sample FASTQ files, a lower length limit of 16 bases was used (as this was the shortest synthetic RNA), while a lower length limit of 18 bases was used for the brain samples (as human miRNAs are generally longer than 16 bases), to reduce the inclusion of reads that were too short in the data.

After trimming (and deduping in the case of the Deduped method) samples were aligned to the miRBase human miRNA sequences [37] and the Miltenyi synthetic sequences using miRge [35].

Similarity Analysis

To perform the hierarchical cluster analysis, we used the miRNA quantifications from all brain libraries with all starting amounts using both batches (total of 99 libraries, 19 for the Clontech, NEXTflex, Deduped and Fivepercent methods and 10 for Illumina and 13 for NEB). We normalized the data using the DESeq2[24] method with the method as the design, using the DESeqDataSetFromMatrix(), estimateSizeFactors(), and counts() functions of the Bioconductor package DESeq2 [24](v 1.18.1). The DESeq2[24] method was chosen for normalization as we assume little difference between the individual synthetic miRNAs, the replicates across batches, and the triplicates within a given batch given that the samples are biologically the same. Normalization for small RNA sequencing is a debated topic and further studies are needed to confirm the best method for different small RNA sources. We then determined which normalized expression estimates were greater than one for all 99 samples. This resulted in 151 common miRNAs above the threshold. We then $\log_2$ transformed these estimates. We determined the distance between the samples using the hclust() function of the stats package (v 3.4.0). We also used these quantification estimates in a sum of squares analysis to determine the percent of variance explained by method, starting amount, batch, and the number of reads that mapped to miRNA. To do this we used the Anova() type II function of the CRAN car package(v 3.0-0). To create the MA plots we used only the 1000ng brain samples from both batches (a total of 24 samples, 4 for each method). We normalized this subset of samples using again using DESeq2[24] and method as the design. We again restricted our analysis to miRNAs with greater than one normalized count in all 24 samples. This resulted in 174 common miRNAs above the threshold. We then manually created the MA plots. We then ranked the $\log_2$ normalized quantifications and determined the overlap of the most abundant miRNAs.

Accuracy Assessment

To perform the accuracy analyses, we used equimolar pools of 962 synthetic miRNAs purchased from Miltenyi Biotec. The Gibbs minimum free energy of the secondary structures for each synthetic miRNA was determined using RNAfold as part of the ViennaRNA package 2 (version 2.3.5) [38,39]. GC content was determined using the letterFrequency() function of the Bioconductor package Biostrings [40] (v 2.46.0). Alignments were performed using the miRge program. The miRge raw count estimates were normalized using DESeq2 [24] (v 1.18.1) with method as the design. The difference of each miRNA count estimate was then calculated from the mean of all synthetic sequences. The absolute of this difference was then $\log_2$ transformed for statistical comparisons and is referred to as "accuracy error".

A linear model was used to evaluate the influence of method on accuracy error, using the lm() function of the stats package, and paired t-tests using the t.test() function of the stats package (v 3.4.0) were used for pairwise comparisons of each method. The Bonferroni method was used to for multiple testing correction. The omega squared value was calculated using the anova_stats() function of the CRAN package sjstats [41] (v 0.16.0). Hedge's g was calculated using the tes() function of the CRAN package compute.es (v 0.2-4). Catplots to evaluate concordance of error rank were created using the CATplot() function of the Bioconductor package ffpe (v 1.22.0).

Detection Diversity Assessment

To assess mapping rates to various classes of RNAs, we collected fasta files for a variety of RNAs: miRNA, piRNA, rRNA, scaRNA, snoRNA, snRNA, and transfer RNA (tRNA) and then created a merged fasta file from each of the smaller fasta files. We used the miRNA fasta file included in miRge. The piRNA data was acquired from piRNAQuest [42] (http://bicresources.jcbose.ac.in/zhumur/pirnaquest/). The rRNA, tRNA, and snRNA data came from the hg19 assembly from the UCSC genome table browser [43] (http://genome.ucsc.edu). The snoRNA and scaRNA data came from snoRNABase [44] (https://www-snorna.biotoul.fr/browse.php). Only the C/D box snoRNAs were included as all the H/ACA box snoRNAs overlapped with the snRNA data from UCSC. Six of the C/D snoRNA sequences and the snRNA overlapped in our merged fasta file. Additionally, all of the scaRNAs overlapped the snoRNA C/D box sequences, but we maintained them in order to analyze scaRNA. Exact matches of miRNA sequences and piRNA were removed from the piRNA portion of the fasta file. *bowtie* [23] was used for alignment to all the sequences simultaneously allowing for zero mismatches within the default seed length of 28 bases to better distinguish similar sequences of different RNA classes. We then determined the count of reads that mapped to each RNA class.

miRNAs were considered detected if they were observed with > 10 normalized reads in all triplicates of a given starting amount. Raw counts from miRge for the brain batch 1 samples (93 in total) were normalized with DESeq2[24] but were not log transformed. Another analysis was performed using both batches and normalizing with DESeq2[24] using all brain samples and a threshold of >10 normalized reads for all samples of a given starting amount. The percent of detected miRNAs that were inconsistently detected was calculated as follows:

$$\left(\frac{U_X - U_{1,2,3}}{U_X}\right) \times 100$$

Where U_X = number of unique miRNAs detected with >10 normalized reads in a given triplicate and where $U_{1,2,3}$ = number of unique miRNAs detected with >10 normalized reads in all triplicates.

The same methods were used for the isomiR analysis, however a threshold of 100 normalized reads was used instead of 10.

Statistical analyses of the detection and detection inconsistency were performed as in the Accuracy assessment with lm() and t.test() of stats package (v 3.4.0) and the tes() function of the compute.es package (v 0.2-4) and the anova_stats() function of the sjstats package (v 0.16.0) to calculate effect sizes. Percent variance explained analyses of sequence characteristics were performed using the Anova() function of the CRAN package car (v 3.0-0). Concordance was evaluated using the CATplot() function of the Bioconductor package ffpe (v 1.22.0). The evaluation of false isomiRs used the synthetic miRNA data. An isomiR was considered as detected if over 100 normalized reads were observed.

Consistency Assessment

Consistency across batch was determined using all 1000 ng brain samples (24 samples). DESeq2[24] (v 1.18.1) was used to normalize these samples with method as the design. Quantifications were filtered for those with >10 normalized reads in all samples of a given method. The mean of the quantifications from the first batch triplicates was compared with that of the second batch quantifications. The $\log_2$ transformed value of the absolute difference between these two quantifications was used to compare the batch consistency of the methods. Again the lm() of the stats package (v 3.4.0) was used for global analyses, while the t.test() function with Bonferroni correction was used to compare pairs of methods. Evaluating the intersection of all miRNAs detected across both batches for each method (total of 162 miRNAs), we determined the percent of variance in triplicate error for sequence characteristics.

To evaluate the consistency of triplicates, we used all 93 brain samples of the first batch. This data was normalized using DESeq2[24] using method as the design. The quantification estimates were filtered for those with >10 normalized counts in all samples for a given starting amount. "Triplicate error" was determined as the difference of the value of each triplicate relative to the mean of all triplicates. The absolute value of these differences was then $\log_2$ transformed and the mean error value of triplicates was determined for each miRNA detected by each method for statistical comparisons. Evaluating just the intersection of all miRNAs detected for each starting amount and method (total of 228 miRNAs), we determined the percent of variance in triplicate error for sequence characteristics. The consistency of triplicates was then used to compare the various starting amounts. The Bonferroni method was used for multiple testing correction.

Code and data availability:

The code for all of the analyses performed in this manuscript is will be publically available at https://github.com/LieberInstitute/miRNA_Kit_Comparison. The data will be made available at the National Institutes of Health (NIH) Sequence Read Archive (SRA), and the accession number will be listed on the GitHub readme for the repository.

Acknowledgements:

We are grateful for the generosity of the Lieber and Maltz families in establishing an institute dedicated to understanding the basis of developmental brain disorders. This research was funded by the Lieber Institute for Brain Development and the AstraZeneca postdoctoral fellowship program.

Author Contributions:

C.W. (Carrie Wright) designed the study, performed the data analysis, and wrote the manuscript. C.W. (Carrie Wright) and A.R. performed the library preparations and data preprocessing. A.R. also edited the manuscript and assisted with the study design and the design for Fig. 1. A.R. wrote the script to prepare the FASTQ files for UMI-tools with the assistance of C.W. (Carrie Wright). E.E.B. assisted with creating the figures and edited the manuscript. C.W. (Courtney Williams) and M.K. performed the sequencing. L.C.-T. and A.E.J assisted with the statistical analysis design. L.C.-T, A.E.J, D.R.W., and J.H.S. also edited the manuscript. J.H.S. supervised this project and assisted with the study design and the design of Fig. 1. A.J.C., N.J.B., and D.R.W. secured funding for this project and contributed to the overall direction.

References:

1. Bandiera, S., Hatem, E., Lyonnet, S. & Henrion-Caude, A. microRNAs in diseases: from candidate to modifier genes. *Clinical Genetics* **77,** 306–313 (2010).
2. Miller, B. H. & Wahlestedt, C. MicroRNA dysregulation in psychiatric disease. *Brain Research* **1338,** 89–99 (2010).
3. Basak, I., Patil, K. S., Alves, G., Larsen, J. P. & Møller, S. G. microRNAs as neuroregulators, biomarkers and therapeutic agents in neurodegenerative diseases. *Cellular and Molecular Life Sciences* **73,** 811–827 (2016).
4. Reid, G., Kirschner, M. B. & van Zandwijk, N. Circulating microRNAs: Association with disease and potential use as biomarkers. *Critical Reviews in Oncology/Hematology* **80,** 193–208 (2011).
5. Eminaga, S., Christodoulou, D. C., Vigneault, F., Church, G. M. & Seidman, J. G. Quantification of microRNA Expression with Next-Generation Sequencing. in *Current Protocols in Molecular Biology* (eds. Ausubel, F. M. et al.) (John Wiley & Sons, Inc., 2013). doi:10.1002/0471142727.mb0417s103
6. Neilsen, C. T., Goodall, G. J. & Bracken, C. P. IsomiRs – the overlooked repertoire in the dynamic microRNAome. *Trends in Genetics* **28,** 544–549 (2012).
7. Morin, R. D. *et al.* Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Research* **18,** 610–621 (2008).
8. Witwer, K. W. & Halushka, M. K. Toward the promise of microRNAs – Enhancing reproducibility and rigor in microRNA research. *RNA Biology* **13,** 1103–1116 (2016).
9. Linsen, S. E. V. *et al.* Limitations and possibilities of small RNA digital gene expression

profiling. *Nature Methods* **6,** 4734–476 (2009).

10. Buschmann, D. *et al.* Toward reliable biomarker signatures in the age of liquid biopsies - how to standardize the small RNA-Seq workflow. *Nucleic Acids Research* **44,** 5995–6018 (2016).

11. Lopez, J. P. *et al.* Biomarker discovery: quantification of microRNAs and other small non-coding RNAs using next generation sequencing. *BMC Medical Genomics* **8,** (2015).

12. Head, S. R. *et al.* Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques* **56,** (2014).

13. Pritchard, C. C., Cheng, H. H. & Tewari, M. MicroRNA profiling: approaches and considerations. *Nature Reviews Genetics* **13,** 358–369 (2012).

14. Kim, Y.-K., Yeo, J., Kim, B., Ha, M. & Kim, V. N. Short structured RNAs with low GC content are selectively lost during extraction from a small number of cells. *Molecular cell* **46,** 893–895 (2012).

15. Dard-Dascot, C. *et al.* Systematic comparison of small RNA library preparation protocols for next-generation sequencing. *BMC Genomics* **19,** (2018).

16. Sorefan, K. *et al.* Reducing ligation bias of small RNAs in libraries for next generation sequencing. *Silence* **3,** 4 (2012).

17. Giraldez, M. D. *et al.* Comprehensive multi-center assessment of small RNA-seq methods for quantitative miRNA profiling. *Nature Biotechnology* (2018). doi:10.1038/nbt.4183

18. Yeri, A. *et al.* Evaluation of commercially available small RNASeq library preparation kits using low input RNA. *BMC Genomics* **19,** (2018).

19. Baran-Gale, J. *et al.* Addressing Bias in Small RNA Library Preparation for Sequencing: A New Protocol Recovers MicroRNAs that Evade Capture by Current Methods. *Frontiers in Genetics* **6,** (2015).

20. Fu, Y., Wu, P.-H., Beane, T., Zamore, P. D. & Weng, Z. Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. *BMC Genomics* **19,** (2018).

21. Faridani, O. R. *et al.* Single-cell sequencing of the small-RNA transcriptome. *Nature Biotechnology* **34,** 1264–1266 (2016).

22. Jayaprakash, A. D., Jabado, O., Brown, B. D. & Sachidanandam, R. Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. *Nucleic Acids Research* **39,** e141–e141 (2011).

23. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10,** R25 (2009).

24. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15,** (2014).

25. Tang, D. T. P. *et al.* Suppression of artifacts and barcode bias in high-throughput transcriptome analyses utilizing template switching. *Nucleic Acids Research* **41,** e44–e44 (2013).

26. Kapteyn, J., He, R., McDowell, E. T. & Gang, D. R. Incorporation of non-natural nucleotides into template-switching oligonucleotides reduces background and improves cDNA synthesis from very small RNA samples. *BMC Genomics* **11,** 413 (2010).

27. Fuchs, R. T., Sun, Z., Zhuang, F. & Robb, G. B. Bias in Ligation-Based Small RNA Sequencing Library Construction Is Determined by Adaptor and RNA Structure. *PLOS ONE* **10,**

e0126049 (2015).

28. Hafner, M. *et al.* RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA* **17,** 1697–1712 (2011).

29. Zhuang, F., Fuchs, R. T., Sun, Z., Zheng, Y. & Robb, G. B. Structural bias in T4 RNA ligase-mediated 3′-adapter ligation. *Nucleic Acids Research* **40,** e54–e54 (2012).

30. Huang, C.-J. *et al.* Frequent Co-Expression of miRNA-5p and -3p Species and Cross-Targeting in Induced Pluripotent Stem Cells. *International Journal of Medical Sciences* **11,** 824–833 (2014).

31. Huang, X. *et al.* Characterization of human plasma-derived exosomal RNAs by deep sequencing. *BMC genomics* **14,** 319 (2013).

32. Barberán-Soler, S. *et al.* Decreasing miRNA sequencing bias using a single adapter and circularization approach. *Genome Biology* **19,** (2018).

33. Stahlberg, A. Properties of the Reverse Transcription Reaction in mRNA Quantification. *Clinical Chemistry* **50,** 509–515 (2004).

34. Smith, T. S., Heger, A. & Sudbery, I. UMI-tools: Modelling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. (2016). doi:10.1101/051755

35. Baras, A. S. *et al.* miRge - A Multiplexed Method of Processing Small RNA-Seq Data to Determine MicroRNA Entropy. *PLOS ONE* **10,** e0143066 (2015).

36. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10–12 (2011).

37. Griffiths-Jones, S. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research* **34,** D140–D144 (2006).

38. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms for Molecular Biology* **6,** 26 (2011).

39. Hofacker, I. L. *et al.* Fast folding and comparison of RNA secondary structures. *Monatshefte fur Chemie Chemical Monthly* **125,** 167–188 (1994).

40. Pagès, H., Aboyoun, P., Gentleman, R. & DebRoy, S. *Biostrings: Efficient manipulation of biological strings.* (2018).

41. Lüdecke, D. *sjstats: Statistical Functions for Regression Models*. (2018).

42. Sarkar, A., Maj, R., Saha, S. & Ghosh, Z. piRNAQuest: searching the piRNAome for silencers. *BMC Genomics* **15,** 555 (2014).

43. Karolchik, D. The UCSC Table Browser data retrieval tool. *Nucleic Acids Research* **32,** 493D – 496 (2004).

44. Lestrade, L. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Research* **34,** D158–D162 (2006).

Table Legends

Table 1. Summary of results
The table depicts the results of the various assessments performed on the 6 small RNA sequencing methods. *The mapping rate of the raw NEXTflex data should reflect the quality for the data used for the Deduped and Fivepercent methods as these are derived from the raw NEXTflex data.

Figure Legends

Fig. 1. Study Design

(**A**) We focused on several sources of inconsistency and bias in small RNA sequencing. We evaluated the influence of various starting amounts on the consistency of the results, as well as the accuracy of the results obtained when using a variety of methods including those intended to reduce adapter ligation bias, reverse transcription bias, and amplification bias. (**B**) We compared four commercially available kits and two preprocessing methods to address reverse transcription (RT) and PCR amplification bias. (**C**) In the Deduped method we collapsed duplicate reads based on a unique molecular identifier (UMI) that came from the degenerate bases in the adapter sequences (bases within the black boxes). We also compared the collapsed data with a random 5% subset (called Fivepercent) of the data to determine if performance differences were due to collapsing the reads based on the UMI or simply due to having fewer reads. (**D**) We evaluated two types of data: miRNA quantifications from homogenate whole brain total RNA from a single female and miRNA quantifications from a pool of 962 equimolar synthetic RNAs with sequences that correspond to human, rat, mouse, and virus miRNA. We had two batches of the human brain data. The first batch included triplicates of many different starting amounts based on the range of inputs suggested by the manufacturers of the library preparation kits. The second batch included a single sample of the same human brain with 1000ng of input. We used 300ng of the synthetic miRNAs for each tested method. (**E**) This flowchart depicts our processing pipelines for the two types of RNA studied. (**F**) We evaluated the 6 small RNA sequencing methods using 4 major assessments. The brain icon indicates when we utilized brain samples to assess a question, while the red tube indicates when we utilized the synthetic miRNA samples.

Fig. 2. Similarity Assessment

(**A**) Dendrogram depicting cluster analysis shows that samples largely cluster by method and starting amount. (**B**) Individual points represent the miRNAs quantified by all of the methods; the y-axis of each plot shows the log ratio of the normalized quantification estimates between the two methods, while the x-axis shows the average expression. These plots are referred to as MA plots. (**C**) The percent of variance explained by method, starting amount, batch, the number of reads mapped to miRNA, and the variance unaccounted for by these factors. Each point represents the variance explained by each factor for an individual miRNA sequence that was quantified by all of the tested methods.

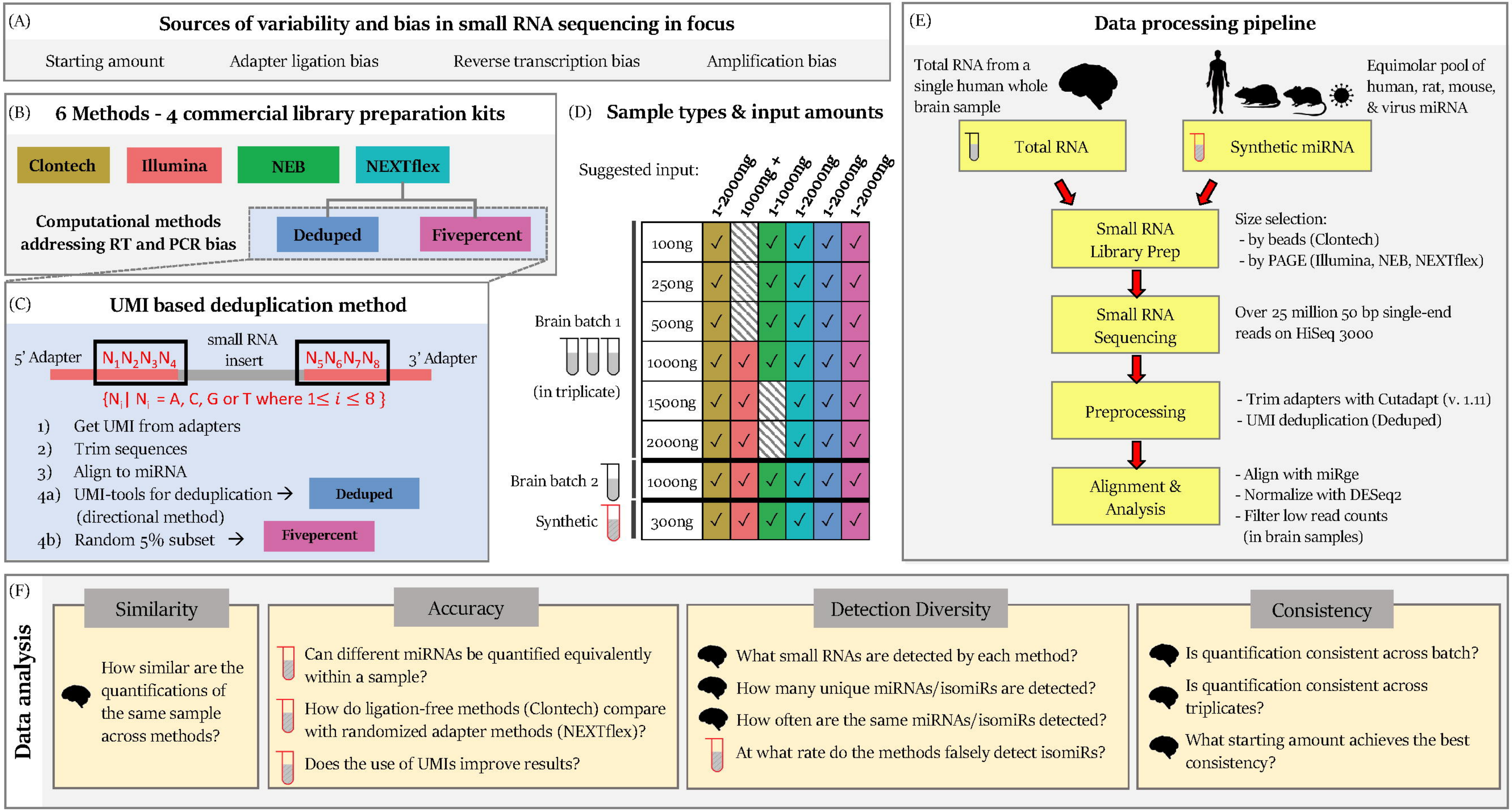
Fig. 3. Accuracy Assessment

(**A**) Individual points represent the absolute difference of each synthetic miRNA quantification from the mean of all quantifications of the equimolar synthetic sequences for each small RNA sequencing method. (**B**) The variance of all the quantification estimates for the synthetic sequences. (**C**) The percent variance of synthetic sequence quantifications explained by each of these sequence characteristics: GC content, length, Gibbs free energy of the predicted secondary structure (FoldG), identity of the first and last two bases, the count of individual bases, and the presence of repeat sequences, such as duplets of the same base or quadruplets of the same base. The heatmap legend shows the percentage of variance from 0 to 10 percent. (**D**) The percent variance explained by each of the sequence characteristics but weighted by the overall variance of each method, as shown in B. The heatmap legend shows the percentage of variance from 0 to 10 percent.

Fig. 4. Detection Diversity Assessment

(**A**) Mapping rate of various small RNAs utilizing the 1000ng input human brain data for each method. Undetermined indicates that the read did not map to the annotations of the evaluated small RNA classes. (**B**) The mapping rate of small RNAs for all starting input amounts for each method. The Y-axis shows the percentage of reads of each category and the X-axis shows each tested brain sample. (**C**) The number of unique miRNAs with greater than 10 normalized reads in all triplicates for the 1000ng data of the first batch. (**D**) The percentage of miRNAs that had quantifications above 10 in only 1 or 2 of the triplicates. (**E**) The overlap of the unique miRNAs with greater than 10 normalized reads in all triplicates for the 1000ng data of the first batch. (**F**) The number of unique isomiRs with greater than 100 normalized reads in all triplicates for the 1000ng data of the first batch. (**G**) The percentage of isomiRs that had quantifications above 100 in only 1 or 2 of the triplicates. (**H**) The number of false isomiRs observed in the synthetic data with over 100 normalized reads. (**I**) The number of false isomiRs detected for each of the 962 synthetic sequences. (**J**) The number of normalized reads (expression) of the false isomiRs. (**K**) The percent variance of the number of isomiRs observed for each synthetic sequence explained by various sequence characteristics. The heatmap legend shows the percentage of variance from 0 to 9 percent.

Fig. 5. Consistency Assessment
(**A**) Absolute difference of the normalized and $\log_2$ transformed quantifications of the second batch from the mean of the triplicates of first batch for each quantified miRNA of the 1000ng input data. (**B**) Absolute difference of each normalized and $\log_2$ transformed quantification for each quantified miRNA from a given triplicate to that of the mean of all three triplicates of the 1000ng input data. (**C**) Absolute difference of each normalized and $\log_2$ transformed quantification for each quantified miRNA from a given triplicate to that of the mean of all three triplicates of the data for all the starting inputs. (**D**) Percent variance of batch inconsistency (A) explained by various sequence factors. The heatmap legend shows the percentage of variance from 0 to 75 percent. (**E**) Percent variance of batch inconsistency (A) explained by various sequence factors weighted by the overall batch variance of each method. The heatmap legend shows the percentage of variance from 0 to 75 percent. (**F**) Plots of the association of expression and batch error. (**G**) Percent variance explained by various sequence factors of the triplicate inconsistency plotted in (C). The heatmap legend shows the percentage of variance from 0 to 100 percent. (**H**) Percent variance explained by various sequence factors of the triplicate inconsistency plotted in C and weighted by the overall variance of triplicate error for each method. (**I**) Plots of the association of expression and triplicate inconsistency using all starting input data in (C). The heatmap legend shows the percentage of variance from 0 to 100 percent.



(A) **Sources of variability and bias in small RNA sequencing in focus**

Starting amount — Adapter ligation bias — Reverse transcription bias — Amplification bias

(B) **6 Methods - 4 commercial library preparation kits**

Clontech — Illumina — NEB — NEXTflex

Computational methods addressing RT and PCR bias: Deduped — Fivepercent

(C) **UMI based deduplication method**

5' Adapter $N_1N_2N_3N_4$ small RNA insert $N_5N_6N_7N_8$ 3' Adapter

$\{N_i \mid N_i = A, C, G \text{ or } T \text{ where } 1 \le i \le 8\}$

1) Get UMI from adapters
2) Trim sequences
3) Align to miRNA
4a) UMI-tools for deduplication (directional method) → Deduped
4b) Random 5% subset → Fivepercent

(D) **Sample types & input amounts**

Suggested input:		1-2000ng	1000ng +	1-1000ng	1-2000ng	1-2000ng	1-2000ng
Brain batch 1 (in triplicate)	100ng	✓	▨	✓	✓	✓	✓
	250ng	✓	▨	✓	✓	✓	✓
	500ng	✓	▨	✓	✓	✓	✓
	1000ng	✓	✓	✓	✓	✓	✓
	1500ng	✓	✓	▨	✓	✓	✓
	2000ng	✓	✓	▨	✓	✓	✓
Brain batch 2	1000ng	✓	✓	✓	✓	✓	✓
Synthetic	300ng	✓	✓	✓	✓	✓	✓

(E) **Data processing pipeline**

Total RNA from a single human whole brain sample → Total RNA

Equimolar pool of human, rat, mouse, & virus miRNA → Synthetic miRNA

Small RNA Library Prep — Size selection:
- by beads (Clontech)
- by PAGE (Illumina, NEB, NEXTflex)

Small RNA Sequencing — Over 25 million 50 bp single-end reads on HiSeq 3000

Preprocessing
- Trim adapters with Cutadapt (v. 1.11)
- UMI deduplication (Deduped)

Alignment & Analysis
- Align with miRge
- Normalize with DESeq2
- Filter low read counts (in brain samples)

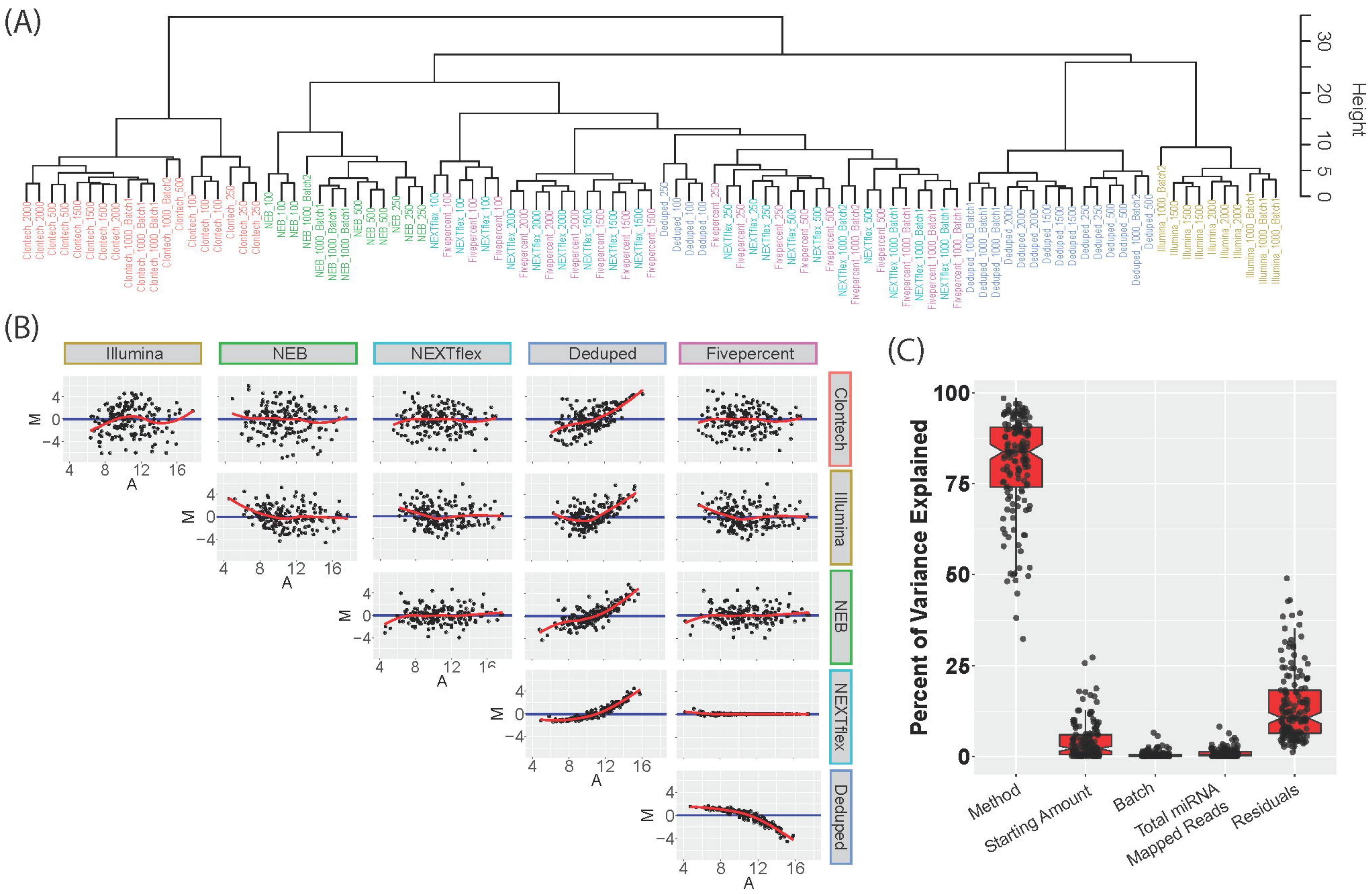
(F) **Data analysis**

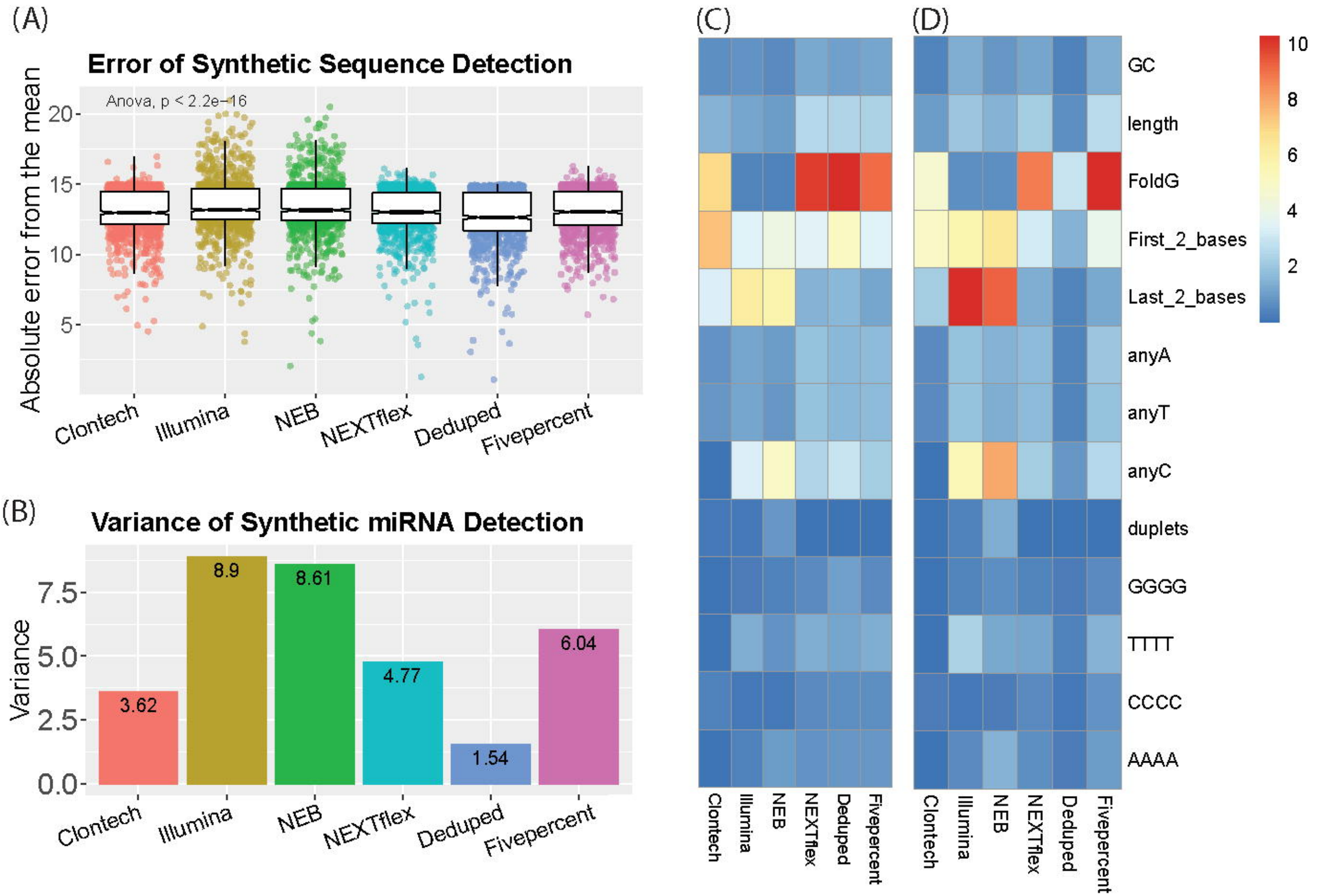
Similarity
- How similar are the quantifications of the same sample across methods?

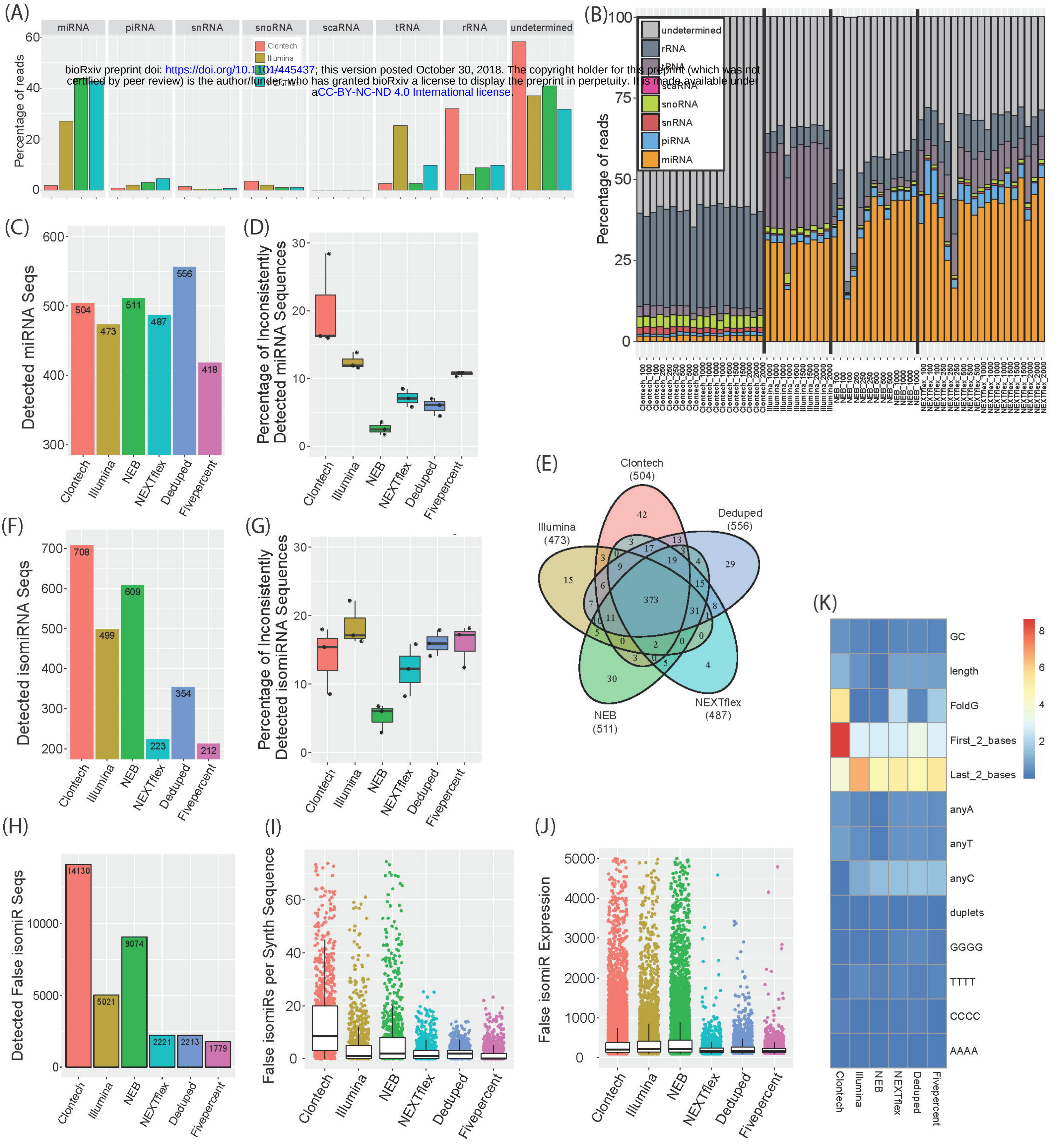
Accuracy
- Can different miRNAs be quantified equivalently within a sample?
- How do ligation-free methods (Clontech) compare with randomized adapter methods (NEXTflex)?
- Does the use of UMIs improve results?

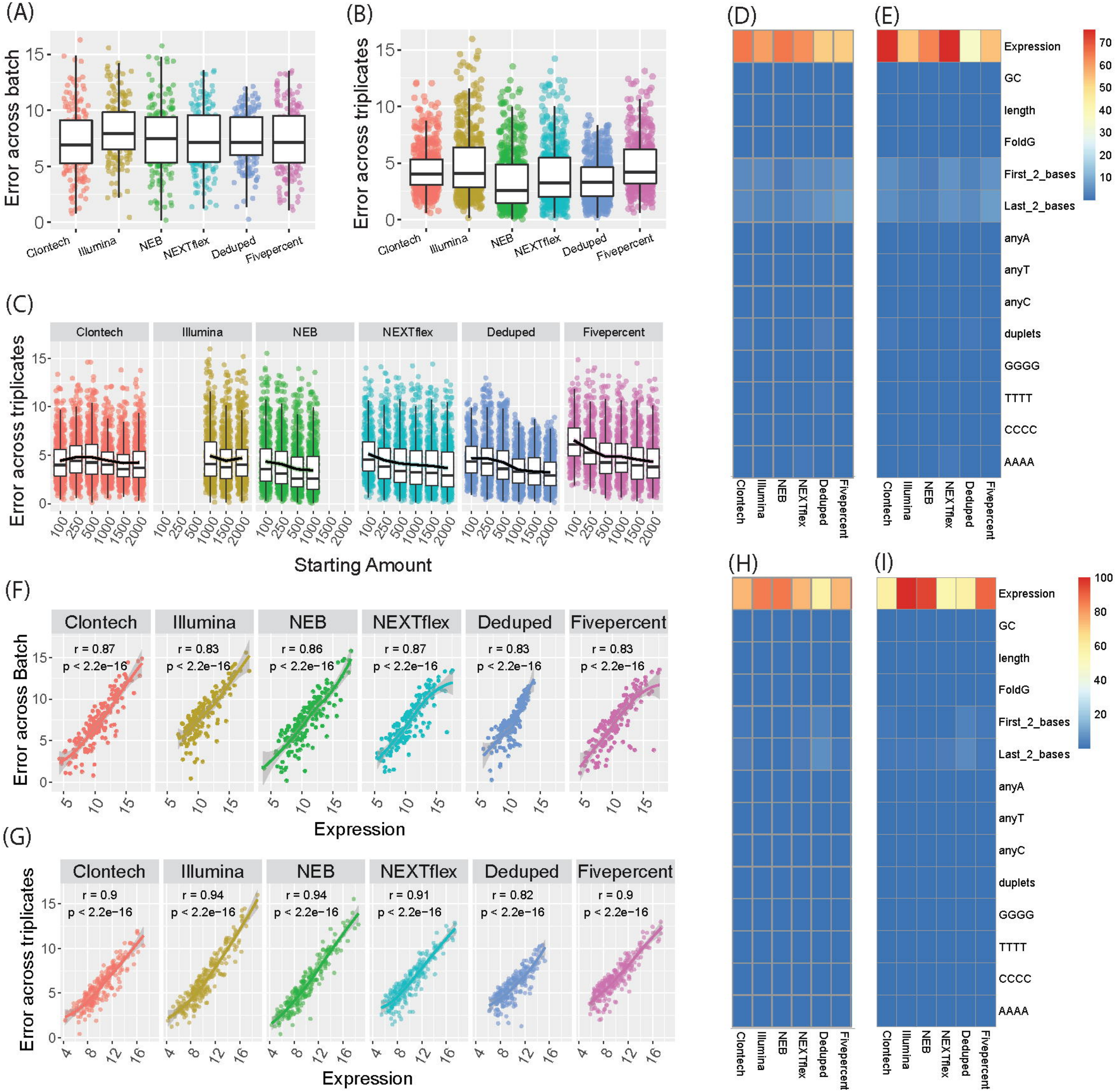
Detection Diversity
- What small RNAs are detected by each method?
- How many unique miRNAs/isomiRs are detected?
- How often are the same miRNAs/isomiRs detected?
- At what rate do the methods falsely detect isomiRs?

Consistency
- Is quantification consistent across batch?
- Is quantification consistent across triplicates?
- What starting amount achieves the best consistency?


(A)
Height
30
20
10
5
0
Clontech_2000
Clontech_2000
Clontech_500
Clontech_500
Clontech_1500
Clontech_1500
Clontech_1500
Clontech_2000
Clontech_1000_Batch1
Clontech_1000_Batch1
Clontech_1000_Batch1
Clontech_1000_Batch2
Clontech_500
Clontech_100
Clontech_100
Clontech_100
Clontech_250
Clontech_250
Clontech_250
NEB_100
NEB_100
NEB_100
NEB_1000_Batch2
NEB_1000_Batch1
NEB_1000_Batch1
NEB_1000_Batch1
NEB_500
NEB_500
NEB_500
NEB_250
NEB_250
NEB_250
NEXTflex_100
Fivepercent_100
NEXTflex_100
Fivepercent_100
NEXTflex_100
Fivepercent_100
NEXTflex_2000
Fivepercent_2000
NEXTflex_2000
Fivepercent_2000
NEXTflex_2000
Fivepercent_2000
NEXTflex_1500
Fivepercent_1500
NEXTflex_1500
Fivepercent_1500
NEXTflex_1500
Fivepercent_1500
Deduped_250
Deduped_100
Deduped_100
Deduped_100
Fivepercent_250
NEXTflex_250
Fivepercent_250
NEXTflex_250
NEXTflex_250
Fivepercent_250
NEXTflex_500
Fivepercent_500
NEXTflex_500
Fivepercent_500
NEXTflex_1000_Batch2
Fivepercent_1000_Batch2
NEXTflex_500
Fivepercent_500
NEXTflex_1000_Batch1
Fivepercent_1000_Batch1
NEXTflex_1000_Batch1
Fivepercent_1000_Batch1
NEXTflex_1000_Batch1
Fivepercent_1000_Batch1
Deduped_1000_Batch1
Deduped_1000_Batch1
Deduped_1000_Batch1
Deduped_2000
Deduped_2000
Deduped_2000
Deduped_1500
Deduped_1500
Deduped_1500
Deduped_250
Deduped_250
Deduped_500
Deduped_500
Deduped_1000_Batch2
Deduped_500
Illumina_1000_Batch2
Illumina_1500
Illumina_1500
Illumina_1500
Illumina_2000
Illumina_2000
Illumina_2000
Illumina_1000_Batch1
Illumina_1000_Batch1
Illumina_1000_Batch1
(B)
Illumina
NEB
NEXTflex
Deduped
Fivepercent
Clontech
Illumina
NEB
NEXTflex
Deduped
M
A
4
0
−4
4
8
12
16
(C)
Percent of Variance Explained
100
75
50
25
0
Method
Starting Amount
Batch
Total miRNA Mapped Reads
Residuals


(A)
Error of Synthetic Sequence Detection
Anova, p < 2.2e−16
Absolute error from the mean
20
15
10
5
Clontech
Illumina
NEB
NEXTflex
Deduped
Fivepercent
(B)
Variance of Synthetic miRNA Detection
Variance
7.5
5.0
2.5
0.0
3.62
8.9
8.61
4.77
1.54
6.04
Clontech
Illumina
NEB
NEXTflex
Deduped
Fivepercent
(C)
(D)
GC
length
FoldG
First_2_bases
Last_2_bases
anyA
anyT
anyC
duplets
GGGG
TTTT
CCCC
AAAA
10
8
6
4
2
Clontech
Illumina
NEB
NEXTflex
Deduped
Fivepercent
Clontech
Illumina
NEB
NEXTflex
Deduped
Fivepercent


(A)
miRNA
piRNA
snRNA
snoRNA
scaRNA
tRNA
rRNA
undetermined
Percentage of reads
0
20
40
60
Clontech
Illumina
NEB
NEXTflex
(B)
undetermined
rRNA
tRNA
scaRNA
snoRNA
snRNA
piRNA
miRNA
Percentage of reads
0
25
50
75
100
Clontech_100
Clontech_250
Clontech_500
Clontech_1000
Clontech_1500
Clontech_2000
Illumina_1000
Illumina_1500
Illumina_2000
NEB_100
NEB_250
NEB_500
NEB_1000
NEXTflex_100
NEXTflex_250
NEXTflex_500
NEXTflex_1000
NEXTflex_1500
NEXTflex_2000
(C)
Detected miRNA Seqs
300
400
500
600
504
473
511
487
556
418
Clontech
Illumina
NEB
NEXTflex
Deduped
Fivepercent
(D)
Percentage of Inconsistently Detected miRNA Sequences
0
10
20
30
Clontech
Illumina
NEB
NEXTflex
Deduped
Fivepercent
(E)
Clontech (504)
Deduped (556)
Illumina (473)
NEB (511)
NEXTflex (487)
42
29
15
30
4
373
3
17
13
19
3
4
3
0
9
6
7
10
11
5
15
31
8
1
0
0
2
0
3
0
5
(F)
Detected isomiRNA Seqs
200
300
400
500
600
700
708
499
609
223
354
212
Clontech
Illumina
NEB
NEXTflex
Deduped
Fivepercent
(G)
Percentage of Inconsistently Detected isomiRNA Sequences
0
10
20
30
Clontech
Illumina
NEB
NEXTflex
Deduped
Fivepercent
(H)
Detected False isomiR Seqs
0
5000
10000
14130
5021
9074
2221
2213
1779
Clontech
Illumina
NEB
NEXTflex
Deduped
Fivepercent
(I)
False isomiRs per Synth Sequence
0
20
40
60
Clontech
Illumina
NEB
NEXTflex
Deduped
Fivepercent
(J)
False isomiR Expression
0
1000
2000
3000
4000
5000
Clontech
Illumina
NEB
NEXTflex
Deduped
Fivepercent
(K)
GC
length
FoldG
First_2_bases
Last_2_bases
anyA
anyT
anyC
duplets
GGGG
TTTT
CCCC
AAAA
2
4
6
8
Clontech
Illumina
NEB
NEXTflex
Deduped
Fivepercent


(A)
Error across batch
0
5
10
15
Clontech
Illumina
NEB
NEXTflex
Deduped
Fivepercent
(B)
Error across triplicates
0
5
10
15
Clontech
Illumina
NEB
NEXTflex
Deduped
Fivepercent
(C)
Clontech
Illumina
NEB
NEXTflex
Deduped
Fivepercent
Error across triplicates
0
5
10
15
100
250
500
1000
1500
2000
Starting Amount
(D)
(E)
Expression
GC
length
FoldG
First_2_bases
Last_2_bases
anyA
anyT
anyC
duplets
GGGG
TTTT
CCCC
AAAA
70
60
50
40
30
20
10
Clontech
Illumina
NEB
NEXTflex
Deduped
Fivepercent
(F)
Clontech
Illumina
NEB
NEXTflex
Deduped
Fivepercent
r = 0.87
p < 2.2e-16
r = 0.83
p < 2.2e-16
r = 0.86
p < 2.2e-16
r = 0.87
p < 2.2e-16
r = 0.83
p < 2.2e-16
r = 0.83
p < 2.2e-16
Error across Batch
0
5
10
15
5
10
15
Expression
(G)
Clontech
Illumina
NEB
NEXTflex
Deduped
Fivepercent
r = 0.9
p < 2.2e-16
r = 0.94
p < 2.2e-16
r = 0.94
p < 2.2e-16
r = 0.91
p < 2.2e-16
r = 0.82
p < 2.2e-16
r = 0.9
p < 2.2e-16
Error across triplicates
0
5
10
15
4
8
12
16
Expression
(H)
(I)
Expression
GC
length
FoldG
First_2_bases
Last_2_bases
anyA
anyT
anyC
duplets
GGGG
TTTT
CCCC
AAAA
100
80
60
40
20
Clontech
Illumina
NEB
NEXTflex
Deduped
Fivepercent