that's GOOD science!

TECH NOTE

A SMARTer Approach to Small RNA Sequencing

SMARTer smRNA-Seq kit for Illumina

Small RNA library preparation with reduced bias:
RNA 3′ polyadenylation and SMART template-switching technology capture small RNAs with greater accuracy than approaches involving adapter ligation >>

Analysis of diverse small RNA species
In addition to miRNAs, this approach allows for capture of piRNAs, snoRNAs, snRNAs, etc >>

Consistent performance across a range of input amounts and sample types
Comparable sequencing results are obtained for 1 ng–2 µg inputs of total RNA or enriched small RNA >>

Overview

Small non-coding RNAs (smRNAs) regulate gene expression via diverse mechanisms and facilitate fundamental cellular processes such as transcript splicing and protein translation. Accordingly, obtaining an accurate portrait of small RNA expression levels from small sample inputs carries potential both for the fulfillment of basic research objectives and the development of novel therapeutics and clinical diagnostic solutions. Towards this end, we have developed a novel approach for the preparation of small RNA sequencing libraries that leverages RNA 3′ polyadenylation and template switching during cDNA synthesis that minimizes sample representation bias and is sensitive enough to accommodate as little as 1 ng of total RNA. Here we present data demonstrating the accuracy, sensitivity, and reproducibility afforded by our SMARTer approach to small RNA transcriptomics.

Introduction

The term "small non-coding RNA" broadly refers to diverse RNA species ~15–150 nucleotides (nt) in size that fulfill biological functions without being translated into proteins. While the involvement of small RNAs in cellular housekeeping processes such as transcript splicing and protein translation has been known since the 1960s, research over the past twenty years has revealed that small RNAs play vital roles in the regulation of gene expression, via both transcriptional and post-transcriptional mechanisms (Choudhuri, 2010).

Of the small RNAs involved in gene regulation, the most well-studied are microRNAs (miRNAs; ~22 nt in size), which facilitate post-transcriptional gene silencing by binding specific target mRNAs via base-pair complementarity, and either blocking translation or triggering transcript degradation (Ha and Kim, 2014). Another group of small RNAs that have been well characterized are Piwi-interacting RNAs (piRNAs), which silence transposons using a miRNA-like mechanism, in addition to inducing epigenetic modifications that influence the transcription of both transposons and protein-coding genes (Weick and Miska, 2014).

Tremendous progress has been made in the identification and characterization of small RNAs, and the current rate of discovery in this field suggests that much more remains to be elucidated. The development of next-

generation sequencing (NGS) technology has proven instrumental to this progress, in part because it allows for identification of small RNAs without prior knowledge of their existence (in contrast with array-based or qPCR methods), and can discriminate between small RNA variants that differ by a single nucleotide. However, small RNA-seq library preparation is not without its challenges, which may include time-consuming enrichment steps prior to cDNA synthesis, and sample misrepresentation due to biases in small RNA end modification, reverse transcription, and PCR amplification.

A major source of bias in small RNA-seq data involves the manner in which small RNAs are captured during library construction (reviewed in Raabe *et al.*, 2014). The most common method involves using a T4 RNA ligase (T4Rnl) to attach adapters to RNA 5′ and 3′ ends. However, T4Rnl exhibits sequence-specific substrate preferences, such that certain combinations of adapters and small RNAs are more readily incorporated than others, leading to sample misrepresentation in small RNA-seq libraries (Jayaprakash *et al.*, 2011; Hafner *et al.*, 2011). An alternative to adapter ligation is RNA 3′ polyadenylation, in which a poly(A) polymerase is used to add a stretch of repeated nucleotides to RNA 3′ ends. In contrast with adapter ligation, RNA polyadenylation occurs in a sequence-independent manner. While RNA 3′ polyadenylation was previously reported to generate small RNA-seq libraries (Berezikov, *et al.*, 2006), this approach involved ligation of RNA 5′ ends, and was still susceptible to sequence-specific biases.

Here we present data from the SMARTer smRNA-Seq Kit for Illumina, which employs RNA 3′ polyadenylation and SMART (**S**witching **M**echanism **a**t the 5′ end of **R**NA **T**emplate) technology (Chenchik *et al.*, 1998) to generate sequencing libraries in a ligation-independent manner. Rather than ligating adapters to small RNAs, this method incorporates adapters at both ends of nascent cDNAs during first-strand synthesis (**Figure 1**). Following polyadenylation of input RNA, first-strand cDNA synthesis is dT-primed (3′ smRNA dT Primer) and performed by the MMLV-derived PrimeScript Reverse Transcriptase (RT), which adds non-templated nucleotides upon reaching the 5′ end of each RNA template. The SMART smRNA Oligo then anneals to the non-templated nucleotides, and serves as a template for the incorporation of an additional sequence of nucleotides to the first-strand cDNA by the RT. Sequences incorporated at the 5′ and 3′ ends of each cDNA molecule serve as primer-annealing sites for PCR, which is performed using oligos that incorporate Illumina-compatible adapters and indexes during library amplification.
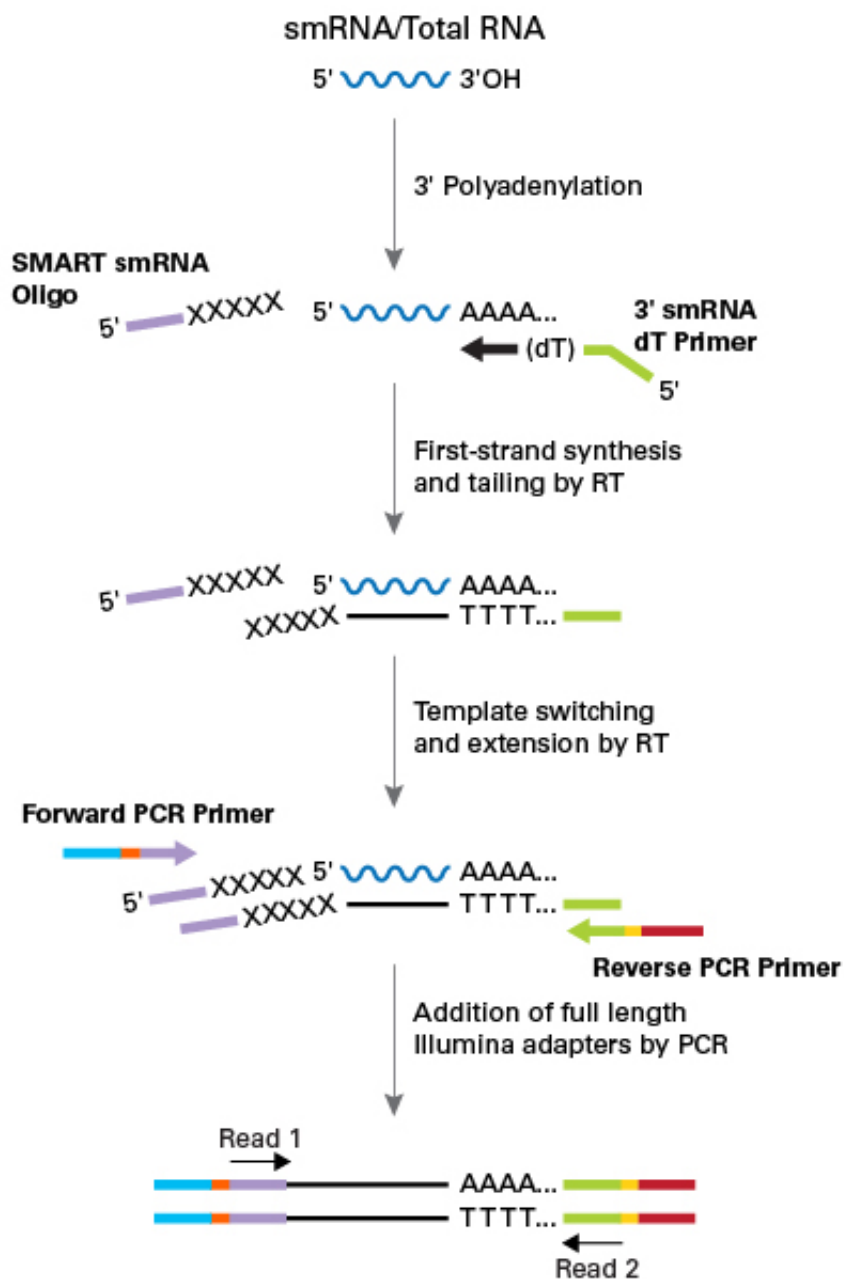
Figure 1. Schematic of technology used by the SMARTer smRNA-Seq Kit for Illumina. SMART technology is used in a ligation-free workflow to generate sequencing libraries for Illumina platforms. Input RNA is first polyadenylated in order to provide a priming sequence for an oligo(dT) primer. cDNA synthesis is primed by the 3′ smRNA dT Primer, which incorporates an adapter sequence (green) at the 5′ end of each first-strand cDNA molecule. When the MMLV-derived PrimeScript Reverse Transcriptase (RT) reaches the 5′ end of each RNA template, it adds non-templated nucleotides which are bound by the SMART smRNA Oligo—enhanced with locked nucleic acid (LNA) technology for greater sensitivity. In the template-switching step, PrimeScript RT uses the SMART smRNA Oligo as a template for the addition of a second adapter sequence (purple) to the 3′end of each first-strand cDNA molecule. In the next step, full-length Illumina adapters (including indexes for sample multiplexing) are added during PCR amplification. The Forward PCR Primer binds to the sequence added by the SMART smRNA Oligo, while the Reverse PCR Primer binds to the sequence added by the 3′ smRNA dT Primer. Resulting library cDNA molecules include adapters required for clustering on an Illumina flow cell (P5 shown in light blue, P7 shown in red), Illumina TruSeq® HT indexes (Index 2 [i5] shown in orange, Index 1 [i7] shown in yellow), and regions bound by the Read Primer 1 or Read Primer 2 sequencing primers

(shown in purple and green, respectively). Note that adapters included in the final library add 153 bp to the size of RNA-derived insert sequences.

Following PCR and column-based purification of PCR products, library profiles are analyzed using an Agilent Bioanalyzer (or similar device) to confirm that small RNA sequences were successfully incorporated and amplified. The combined length of 5′ and 3′ library adapters is 153 bp. Therefore, library molecules containing miRNA-derived sequences typically yield a discrete peak in the ~172–178 bp size range in resulting electropherograms (**Figure 2, Panels A and B**). For most applications, a size selection step is required; for example, libraries generated from total RNA typically include a substantial amount of large molecular weight products and yield a peak at ~1,000 bp (**Figure 2, Panel A**) due to dT-primed capture of mRNAs (which are naturally polyadenylated) during cDNA synthesis. For libraries that require size selection, there are two options: a gel-free, bead-based approach that retains library molecules including inserts ≤150 bp in size (**Figure 2, Panel C**), or size selection with the BluePippin system, which allows for enrichment of specific small RNA species (**Figure 2, Panel D**). Following size selection and validation, libraries are ready for sequencing on an Illumina platform.
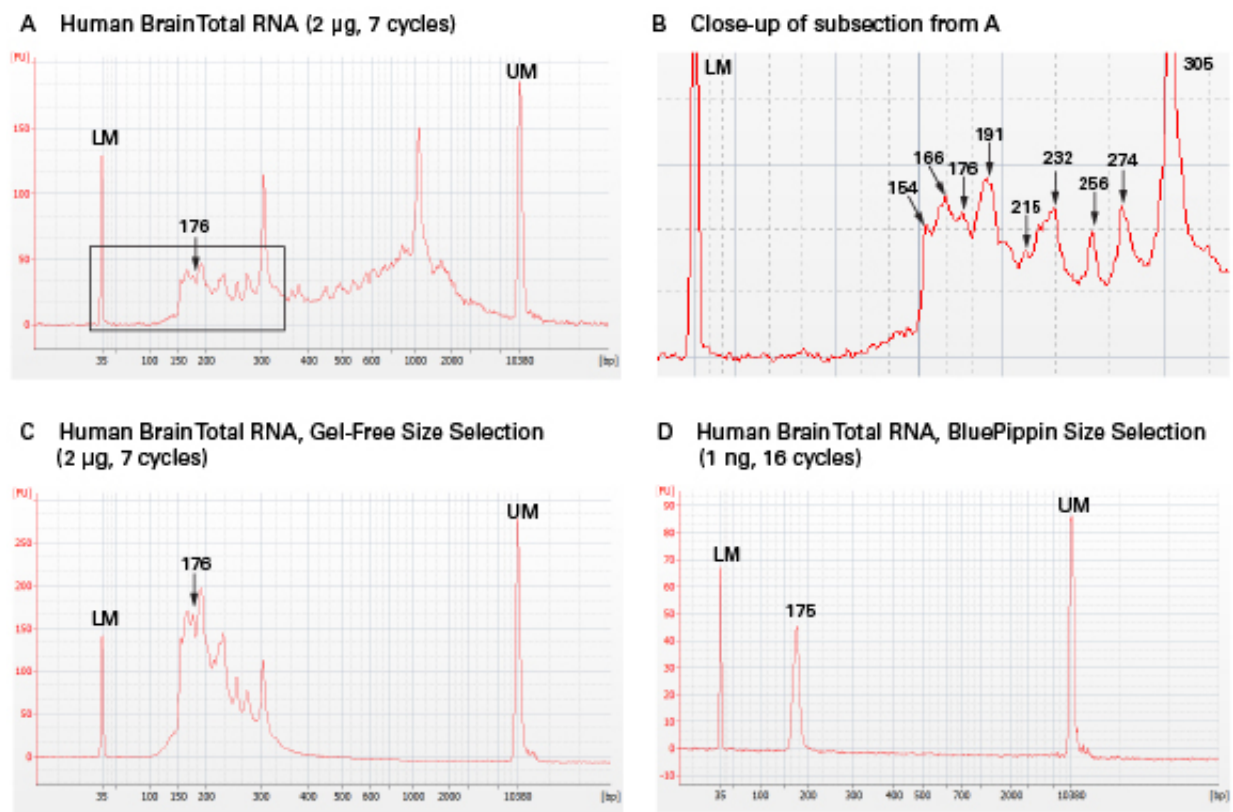

Figure 2. Small RNA-seq library profiles before and after size selection. Libraries were generated using the SMARTer smRNA-Seq Kit for Illumina with the indicated inputs and cycling parameters, and analyzed on an Agilent 2100 Bioanalyzer. Peaks labeled "LM" and "UM" correspond to DNA reference markers included in each analysis. **Panel A.** Typical result for a library generated from total RNA, prior to size selection. The peak at 176 bp corresponds with the predicted combined size of miRNAs plus adapters. **Panel B**. Blowup of the boxed region in Panel A, with individual peaks labeled by size (bp). **Panel C.** Typical result following gel-free, bead-based size selection of the library profiled in Panels A and B. Visible peaks fall within the size range of ~153–300 bp, which corresponds with inserts of 0–150 bp. **Panel D.** Typical result following BluePippin size selection, which affords greater stringency than the bead-based approach. The peak at 175 bp corresponds with the predicted combined size of miRNAs plus adapters.

Results

Assessing the accuracy of the SMARTer smRNA-Seq Kit for Illumina

To measure the accuracy of the SMARTer approach, the SMARTer smRNA-Seq Kit for Illumina was used to generate a sequencing library from the miRXplore Universal Reference (Miltenyi Biotec Inc.), an equimolar pool of 963 synthetic miRNAs. For comparison purposes, a sequencing library was prepared in parallel from the same starting material using an adapter-ligation-based method from Competitor N.

Following sequencing and mapping, the number of reads corresponding to each miRNA was counted and normalized such that each miRNA in the library was expected to have a normalized value of "1". Given that the various miRNAs are present in equal quantities, each should be represented equally in the sequencing output, whereas observed differences in expression are conceivably due to biases imposed during library construction.

For the sequencing library generated using the adapter ligation kit from Competitor N, only ~22% of the miRNAs presented an expression level within a 2-fold cutoff of the expected expression level (**Figure 3**), similar to previously reported data (Fuchs *et al.*, 2015). In contrast, with the SMARTer approach, expression levels for ~55% of the miRNAs fell within the cutoff range. Furthermore, whereas a 7.3×10^{4}-fold range of expression levels was observed using adapter ligation, the SMARTer approach yielded a 1.5×10^{3}-fold range, an improvement of greater than one order of magnitude. These results suggest that libraries generated with the SMARTer smRNA-Seq Kit for Illumina much more accurately portray the expression of miRNA species relative to libraries produced using adapter ligation.
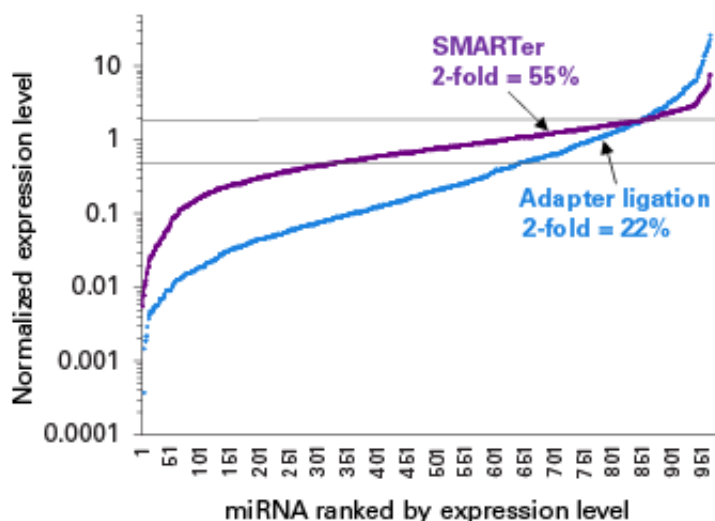

Figure 3. Demonstrating the accuracy of the SMARTer approach for small RNA-seq. Sequencing libraries were generated from an equimolar pool of 963 synthetic miRNAs (miRXplore Universal Reference) using the SMARTer smRNA-Seq Kit for Illumina (1 ng input; purple), or a small RNA-seq kit from a different vendor (Competitor N) employing an adapter ligation method (100 ng input; blue). Following sequencing, mapping, and counting of reads, miRNA expression levels (Y axis, log scale) were normalized, resulting in an expected expression level equal to 1 for each miRNA, and a 2-fold cutoff was assigned both above and below the expected expression level (indicated by two horizontal lines). For visualization purposes, miRNAs are ranked along the X axis in order of expression level.

Evaluating performance across total RNA sources and input amounts

To gauge the performance of the SMARTer approach, the SMARTer smRNA-Seq Kit for Illumina was used to generate sequencing libraries from 1 ng and 2 µg inputs of human total RNA obtained from brain, placenta, and spleen tissue. These three different tissues exhibit a range of small RNA (<200 nt) expression levels, varying

from 2–13% as shown in the table below (Table 1). Consistent with its relatively high proportion of small RNAs (~13%), placental total RNA yielded the most robust libraries, with the lowest number of reads lost in trimming and the best overall mapping to the transcriptome of all three RNA sample types tested (mapping to GENCODE). In addition, libraries generated from placental total RNA demonstrated the most overlap between 1 ng and 2 µg input amounts in terms of which miRNAs were identified. Nevertheless, consistent sequencing metrics were also obtained for libraries generated from brain and spleen total RNA (Table 1 and Figure 4), indicating that the SMARTer smRNA-Seq Kit for Illumina generates reliable small RNA-seq data for RNA samples containing various proportions of small RNAs. Lastly, the data also demonstrate that the SMARTer approach is capable of capturing diverse small RNA species other than miRNAs, including but not limited to piRNAs, small nucleolar RNAs (snoRNAs), and small nuclear RNAs (snRNAs) (Table 1), while keeping representation of less desirable ribosomal RNAs (rRNAs) relatively low (typically between 10–20%).

Sequencing Alignment Metrics for Small RNA from Placenta, Brain, and Spleen						
RNA source	**Placenta**		**Brain**		**Spleen**	
smRNA <200 nt (% of total RNA)	13		5		2	
Input Amount	**2 µg**	**1 ng**	**2 µg**	**1 ng**	**2 µg**	**1 ng**
Total number of reads	4,342,213	4,744,519	4,764,574	4,275,787	3,796,263	4,254,142
Proportion of reads trimmed (%)	15.1	24.7	23.2	31.8	38.6	32.2
Number of reads mapped to GENCODE	3,323,785	3,104,277	3,244,272	2,434,156	2,038,302	2,389,960
Proportion of reads mapped to GENCODE (%)	76.5	65.4	68.1	56.9	53.7	56.2
miRNA						
Proportion of total reads (%)	11.2	10.7	13.8	8.0	8.6	8.3
Number of miRNAs detected	260	263	286	253	198	221
Number of miRNAs in common	247		243		187	
Proportion of miRNAs in common (%)	89		82		81	
Other smRNA (proportion of total reads, %)						
piRNA	3.5	3.7	8.9	5.1	4.1	3.1
snoRNA	1.0	0.7	0.8	0.5	1.1	1.3
snRNA	2.1	1.1	1.2	0.8	0.7	0.9
Other RNAs (proportion of total reads, %)						
rRNA						

(5, 5.8, 18 and 28S)	19.2	12.8	12.1	10.1	14.9	13.8

Table 1. Evaluating the performance of the SMARTer approach across total RNA input types and amounts. Sequencing libraries were generated from 1 ng and 2 µg of human placenta, brain, and spleen total RNA using the SMARTer smRNA-Seq Kit for Illumina, and size selected using the BluePippin system for further enrichment of the miRNA fraction. Following trimming of 3′ A-tail sequences attributed to polyadenylation and 5′ sequences derived from template switching, reads shorter than 15 nt were eliminated, resulting in trimming of 15–38% of reads. Remaining reads were mapped either to GENCODE for overall mapping, or specific small RNA datasets, as indicated. Only miRNAs represented by at least five reads were included in count data for the number of miRNAs detected.
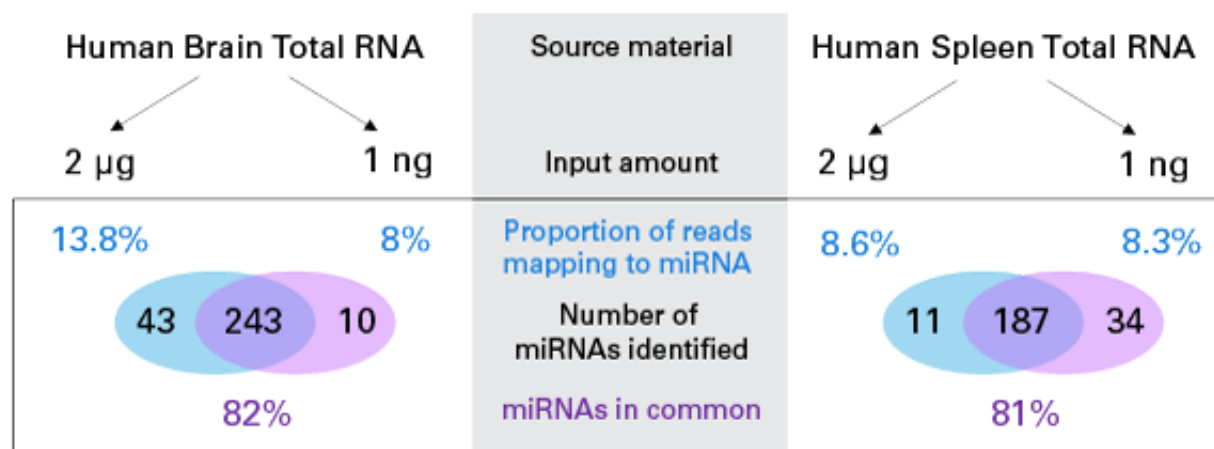

Figure 4. Overlap of miRNAs identified across input amounts. miRNA mapping data included in Table 1 (brain and spleen total RNA samples) were displayed as Venn diagrams indicating the number of miRNAs identified. The percentage of miRNAs that were identified from both input amounts is indicated below each Venn diagram.

Measuring data reproducibility with the SMARTer approach

To specifically evaluate the reproducibility of the SMARTer smRNA-Seq Kit for Illumina, miRNA expression levels measured from different technical replicates were plotted and statistical analyses were performed (**Figure 5**). The results indicate a strong correlation in miRNA expression levels between replicates performed using 1 ng of human brain total RNA (Pearson coefficient = 0.99, Spearman coefficient = 0.86), indicating that data generated using the SMARTer smRNA-Seq Kit for Illumina is highly reproducible. Consistent with this result and data presented above, miRNA expression levels for 1 ng and 2 µg input amounts of human brain total RNA were also found to be highly correlated (Pearson = 0.94, Spearman = 0.87; (**Figure 5**).
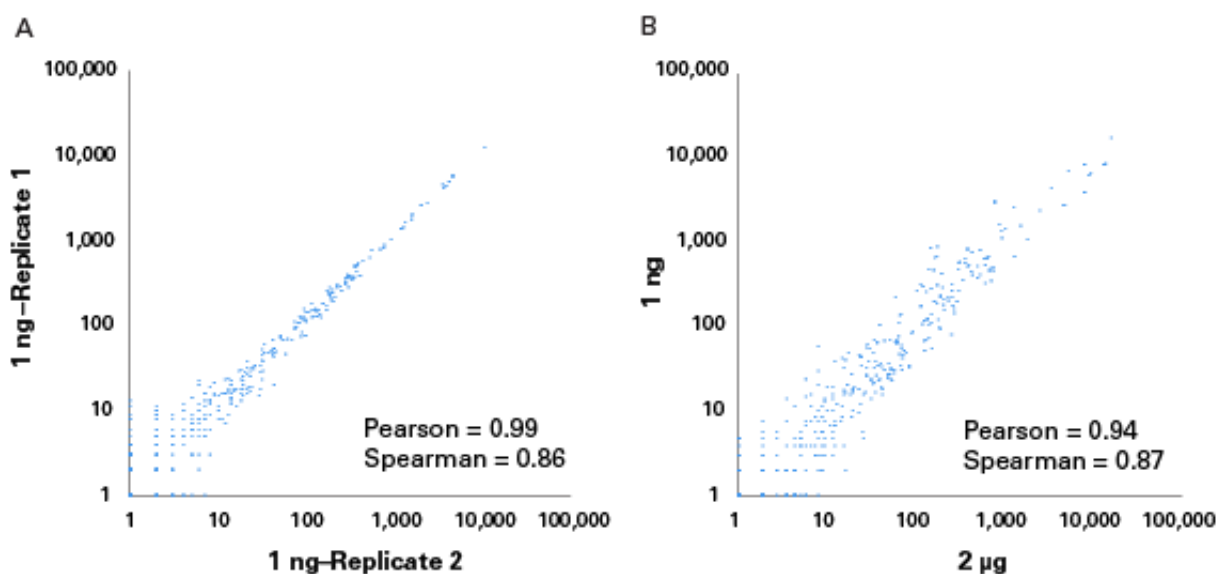
Figure 5. Reproducibility of SMARTer small RNA-seq data. Sequencing libraries were generated in parallel from the indicated input amounts of human brain total RNA using the SMARTer smRNA-Seq Kit for Illumina, and size selected using the BluePippin system. Following sequencing, data processing, and mapping, expression levels of miRNAs identified for each library were quantified and plotted on correlation diagrams.**Panel A.** Correlation of miRNA expression levels for experimental replicates involving 1 ng inputs. **Panel B.**Correlation of miRNA expression levels for 2 µg vs. 1 ng inputs.

Summary

The development of NGS technology has greatly accelerated the discovery and characterization of diverse small RNA species. However, the ability to accurately measure small RNA expression levels on a genome-wide scale remains elusive, largely due to biases imposed during sequencing library construction. Furthermore, study of small RNAs typically requires large amounts of starting material. To help researchers overcome these hurdles, we have developed the SMARTer smRNA-Seq Kit for Illumina, a new library-preparation kit that leverages RNA 3′ polyadenylation and SMART template-switching technology during cDNA synthesis. Through the application of these two techniques, our approach avoids biases associated with adapter ligation methods, and provides the sensitivity necessary to generate high quality sequencing libraries from as little as 1 ng of RNA.

Methods

Input material:

To assess the accuracy of the SMARTer approach (**Figure 3**), sequencing libraries were generated from the miRXplore Universal Reference (Miltenyi Biotech Inc., Cat. No. 130-093-521) using a 1 ng input for the SMARTer approach, and a 100 ng input for the adapter-ligation approach. Starting material for other experiments (**Table 1, Figures 4 and 5**) consisted of the following commercially available preparations of total RNA: Human Brain Total RNA, Human Placenta Total RNA, and Human Spleen Total RNA (Thermo Fisher; Cat. Nos. AM7962, AM7950, and AM7970).

Library preparation

Sequencing libraries were generated as specified in the user manual for the SMARTer smRNA-Seq Kit for Illumina. Preparation of a sequencing library using an adapter ligation approach was performed using a small

RNA-seq kit from another vendor, according to its accompanying protocol.

Library purification

As specified in the SMARTer smRNA-Seq Kit for Illumina user manual, following PCR amplification, all libraries were purified using the Macherey-Nagel NucleoSpin Gel and PCR Clean-Up kit (Clontech, sold as part of the SMARTer smRNA-Seq Kit for Illumina; also sold separately as Cat. Nos. 740609.50 and 740609.250).

Library size selection

Libraries included in the accuracy analysis (**Figure 3**) were not size-selected following PCR amplification. For all other libraries, post-PCR size selection was performed using the BluePippin Size Selection System (Sage Science, BLU0001) and 3% Agarose Gel Cassettes (Sage Science, BDF3010), as specified in the SMARTer smRNA-Seq Kit for Illumina user manual, but with a size selection range of 148–184 bp.

Library quantification and validation

Prior to validation and sequencing, all libraries were quantified using a Qubit Fluorometer (Thermo Fisher Scientific) and a Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific, Cat. No. 32851). Validation was performed using an Agilent Bioanalyzer and the Agilent High Sensitivity DNA Kit (Agilent, Cat. No. 5067-4626).

Illumina sequencing

All libraries were sequenced on the Illumina MiSeq® platform using single end reads (50 bp), generating at least 1 million reads per library.

Data analysis

Sequencing reads for all libraries were trimmed and annotated using CLC Genomics Workbench 8.5.1 (Qiagen) and Small RNA Analysis tools, allowing no more than one mismatch during mapping. Overall mapping was performed against GENCODE (GRCH38). miRNA sequences were mapped to miRBase (release 21), while all other small RNA species were analyzed by mapping to the NONCODE v3.0 dataset. For analysis of the miRXplore libraries, reads were annotated using a reference file containing sequences present in the synthetic pool. Normalization was performed by determining, for each miRNA, the ratio between the observed number of reads and the predicted number of reads.

References

Berezikov, E., Cuppen, E., & Plasterk, R. H. (2006) Approaches to microRNA discovery. *Nat. Genet.* **38**:S2–S7.

Chenchik, A., Zhu, Y.Y., Diatchenko, L., Li, R., Hill, J., & Siebert, P.D. (1998) Generation and use of high quality cDNA from small amounts of total RNA by SMART PCR. In *Gene Cloning and Analysis of RT-PCR.* Eds. Siebert, P. & Larrick, J. (BioTechniques Books, MA):305–319.

Choudhuri, S. (2010) Small noncoding RNAs: biogenesis, function, and emerging significance in toxicology. *J. Biochem. Mol. Toxic.* **24**(3):195–216.

Fuchs, R. T., Sun, Z., Zhuang, F., & Robb, G. B. (2015) Bias in ligation-based small RNA sequencing library construction is determined by adaptor and RNA structure. *PloS One* **10**(5):e0126049.

Ha, M., & Kim, V.N. (2014) Regulation of microRNA biogenesis. *Nat. Rev. Mol. Cell Bio.* **15**(8):509–524.

Hafner, M., Renwick, N., Brown, M., Mihailović, A., Holoch, D., Lin, C., Pena, J.T., Nusbaum, J.D., Morozov,

P., Ludwig, J. and Ojo, T. (2011) RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA* **17**(9):1697–1712.

Jayaprakash, A. D., Jabado, O., Brown, B. D., & Sachidanandam, R. (2011) Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. *Nucleic Acids Res.***39**(21):e141.

Raabe, C. A., Tang, T. H., Brosius, J., & Rozhdestvensky, T. S. (2014) Biases in small RNA deep sequencing data. *Nucleic Acids Res.* **42**(3):1414–1426.

Weick, E.M., Sarkies, P., Silva, N., Chen, R.A., Moss, S.M., Cording, A.C., Ahringer, J., Martinez-Perez, E. & Miska, E.A. (2014) PRDE-1 is a nuclear factor essential for the biogenesis of Ruby motif-dependent piRNAs in *C. elegans*. *Genes Dev.* **28**(7):783–796.

View web page >>

http://www.clontech.com/US/Products/cDNA_Synthesis_and_Library_Construction/
NGS_Learning_Resources/Technical_Notes/smRNA-seq