SMART ligation-free tools for sequencing coding and noncoding RNA from liquid biopsy samples

Nathalie Bolduc*, Simon Lee, and Andrew Farmer

Takara Bio USA, Inc., Mountain View, CA 94043, USA *Corresponding Author: Nathalie_Bolduc@takarabio.com

Abstract

Small and long noncoding RNAs play a major role in the regulation of gene expression and disease. Obtaining an accurate portrait of their expression levels from small sample inputs carries potential for both the fulfillment of basic research objectives and the development of novel therapeutics and clinical diagnostic solutions. Towards this end, we have built upon the sensitivity of our SMART® technology to develop the SMARTer® Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian (Pico kit). The Pico kit relies on random priming to synthesize cDNA from both polyadenylated and non-polyadenylated RNA, thus ensuring a full RNA representation from even the most challenging samples, such as highly degraded RNA from FFPE or cfRNA. In addition, we developed a novel technology, ZapR, which allows for the removal of ribosomal RNA (rRNA)-derived cDNA after reverse transcription and library amplification. The Pico kit enables strand-specific transcriptome analysis from very low amounts of total RNA or cells, but it does not allow for the analysis of small RNAs, such as miRNAs, piRNAs, etc. To address this limitation, we developed the SMARTer smRNA-Seq Kit for Illumina®, a ligation-free approach for the preparation of small RNA sequencing libraries that leverages 3′ RNA polyadenylation followed by cDNA synthesis and template switching. This approach minimizes sample representation bias and is sensitive enough to accommodate inputs of as little as 1 ng of total RNA. The combination of the SMARTer Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian and the SMARTer smRNA-Seq Kit for Illumina provide a complete toolkit for accurate, sensitive, and reproducible detection of coding and noncoding RNAs of any size from the most challenging samples.

1 SMARTer Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian: Workflow

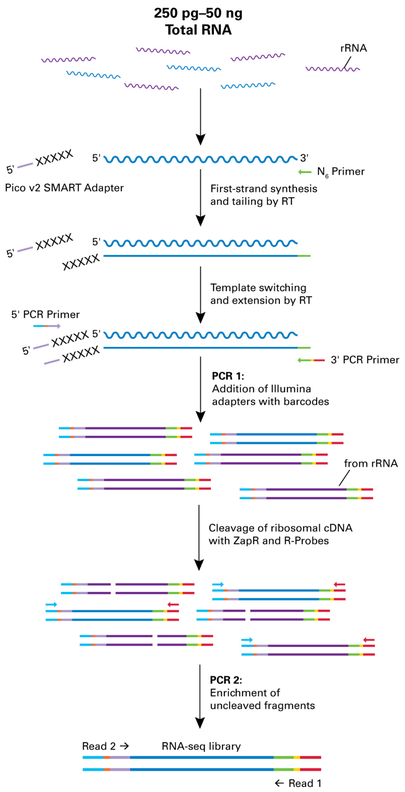

Figure 1. Schematic of technology used by the SMARTer Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian. SMART technology is used in this ligation-free protocol to preserve strand-of-origin information. Random priming allows for the generation of cDNA from all RNA fragments in the sample, including rRNA. When SMARTScribe™ Reverse Transcriptase (RT) reaches the 5′ end of an RNA template, the enzyme's terminal transferase activity adds a few non-templated nucleotides (shown as Xs) to the 3′ end of the cDNA. The Pico v2 SMART Adapter base-pairs with the non-templated nucleotides, creating an extended template that enables the RT to continue cDNA synthesis and add a priming site for subsequent PCR amplification. In the next step, a first round of PCR amplification (PCR 1) adds full-length Illumina adapters, including barcodes. Ribosomal cDNA (originating from rRNA) is then cleaved by ZapR in the presence of mammalian-specific R-Probes. This process leaves library fragments originating from non-rRNA molecules untouched, with priming sites available on both 5′ and 3′ ends for further PCR amplification. Final libraries are compatible with sequencing on any Illumina platform.

Takara Bio USA, Inc.
United States/Canada: +1.800.662.2566 • Asia Pacific: +1.650.919.7300 • Europe: +33.(0)1.3904.6880 • Japan: +81.(0)77.565.6999
For Research Use Only. Not for use in diagnostic procedures.

2 SMARTer Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian: Results

Sequencing alignment metrics for varying input amounts of mouse brain total RNA				
Input amount	**50 ng**	**10 ng**	**1 ng**	**250 pg**
Total number of reads	13,983,286	12,572,360	13,948,522	12,025,520
Number of transcripts (FPKM ≥1)	12,311	12,251	12,113	11,976
Proportion of total reads (%):				
Exonic	26.3	25.3	25.8	24.4
Intronic	40.9	40.1	41.0	39.6
Intergenic	9.6	9.7	9.9	9.6
rRNA	10.1	13.0	11.3	14.4
Mitochondrial	6.9	5.9	5.6	5.4
Duplicate rate (%)	17.9	19.6	38.9	56.0
Mapping to lncRNA:				
Proportion of total reads (%)	11.3	11.7	11.1	11.6
Number of transcripts (TPM ≥1)	10,647	10,363	10,153	9,372

Table 1. Evaluating the performance of the SMARTer Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian across input amounts. Libraries were generated using the indicated amounts of mouse brain total RNA and sequenced on an Illumina NextSeq™ 500 instrument (2 x 75 bp). Bioinformatic analyses were performed using CLC Genomics Workbench. Reads were trimmed and mapped to rRNA and the mitochondrial genome. Unmapped reads were then mapped to the mouse (mm10) genome with RefSeq masking. For analysis of lncRNA, reads were trimmed and mapped against the "Long non-coding RNA transcript sequences" Fasta file included in the GENCODE m12/m38.p5 release (containing 14,610 loci transcripts). FPKM and TPM refer to "fragments per kilobase of transcript per million mapped reads" and "transcripts per kilobase million", respectively.

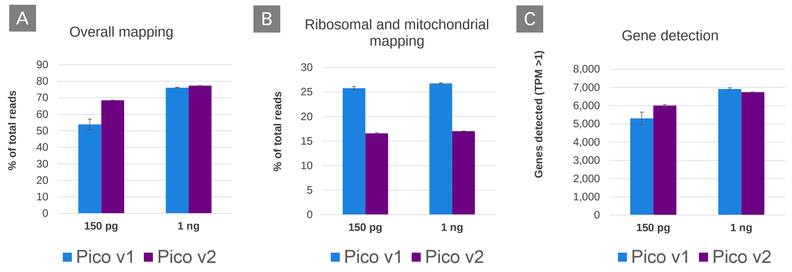

Figure 2. Evaluating the performance of the SMARTer Stranded Total RNA-Seq Kit v2 – Pico Input Mammalian with plasma RNA. A customer using the original Pico kit ("Pico v1") compared its performance to the Pico v2 kit using 1 ng and 150 pg of plasma RNA. All mapping statistics for Pico v2 were similar or better than for Pico v1, including higher mapping rate at the 150-pg input (**A**), lower quantities of ribosomal and mitochondrial mapping reads (**B**), and more genes detected at the 150-pg input (**C**). Data were generated and analyzed entirely by the customer.

Conclusions

SMART technology is a very sensitive and versatile tool for NGS library preparation from coding and noncoding RNAs of all sizes.

Whole transcriptome analysis:

- SMARTer stranded technology, included in the SMARTer Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian, is well-suited for RNA-seq applications that require exceptional sensitivity and strand-of-origin information, such as research involving lncRNAs
- The SMARTer Stranded Total RNA-Seq Kit v2 - Pico Input Mammalian can accommodate very degraded RNA samples, including cfRNA from plasma
- An interesting example of how SMARTer stranded technology can be used for RNA marker discovery in plasma samples was recently published: Ngo et al., Noninvasive blood tests for fetal development predict gestational age and preterm delivery. *Science* **360**, 1133-1136 (2018)

smRNA analysis:

- Using RNA 3′ polyadenylation and SMART template-switching technology, we developed a small RNA library preparation method that provides greater accuracy than approaches involving adapter ligation
- In addition to capturing miRNAs, the SMARTer smRNA-Seq Kit for Illumina enables analysis of piRNAs, snoRNAs, snRNAs, etc.
- Our approach successfully generated high-quality libraries from subnanogram amounts of RNA extracted from 200 µl of human plasma or serum
- Data generated with the SMARTer smRNA-Seq Kit for Illumina exhibits the robustness necessary for biomarker discovery from clinical samples

3 SMARTer smRNA-Seq Kit for Illumina: Overview

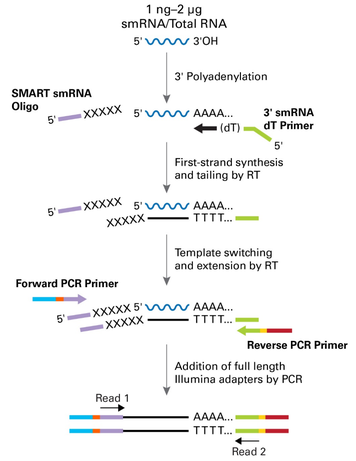

Figure 3. Schematic of technology used by the SMARTer smRNA-Seq Kit for Illumina. SMART technology is used in a ligation-free workflow to generate sequencing libraries for Illumina platforms. Input RNA is first polyadenylated in order to provide a priming sequence for an oligo(dT) primer (3′ smRNA dT Primer). 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 final step, full-length Illumina adapters (including indexes for sample multiplexing) are added during PCR amplification.

4 SMARTer smRNA-Seq Kit for Illumina: Results

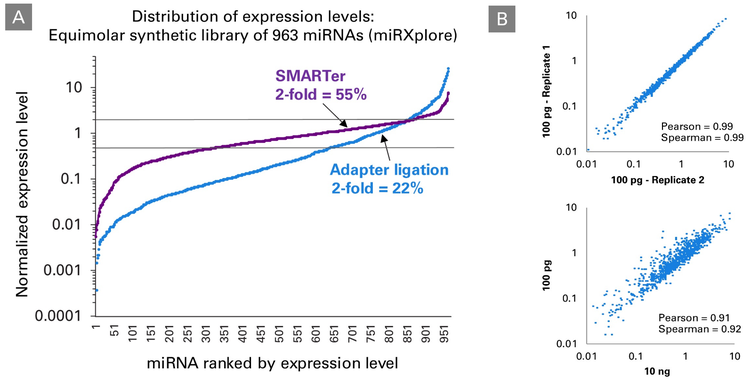

Figure 4. Accuracy and reproducibility of data generated with the SMARTer smRNA-Seq Kit for Illumina. Panel A. To gauge the accuracy of the SMARTer approach, sequencing libraries were generated from an equimolar pool of 963 synthetic miRNAs (miRXplore Universal Reference) using the SMARTer smRNA-Seq 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). 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. **Panel B.** To assess the reproducibility of data generated with the SMARTer approach, sequencing libraries were generated in parallel from the indicated input amounts of the miRXplore Universal Reference using the SMARTer smRNA-Seq Kit for Illumina. Expression levels of miRNAs identified for each library were quantified and plotted on correlation diagrams, and Pearson and Spearman correlation coefficients were calculated.

Sequencing alignment metrics for small RNA from placenta, brain, spleen, and blood										
RNA source	**Placenta**		**Brain**		**Spleen**		**Plasma-1**	**Plasma-2**	**Serum-1**	**Serum-2**
smRNA <200 nt (% of total RNA)	**13%**		**5%**		**2%**		-	-	-	-
Input amount	**2 µg**	**1 ng**	**2 µg**	**1 ng**	**2 µg**	**1 ng**	**400 pg**	**400 pg**	**1 ng**	**1 ng**
Total number of reads	4,342,213	4,744,519	4,764,574	4,275,787	3,796,263	4,254,142	7,493,889	5,629,347	10,571,256	7,305,083
Proportion of reads trimmed (%)	15.1	24.7	23.2	31.8	38.6	32.2	19.6	20.5	11.1	14.4
Number of reads mapped to GENCODE (%)	76.5	65.4	68.1	56.9	53.7	56.1	70.1	68.8	79.0	74.1
miRNA mapping (miRBase)										
Number of miRNA reads	485,331	509,129	657,282	340,814	326,870	353,553	256,128	197,537	204,627	144,229
Proportion of total reads	**11.2%**	**10.7%**	**13.8%**	**8.0%**	**8.6%**	**8.3%**	**3.4%**	**3.5%**	**1.9%**	**2.0%**
Unique miRNAs detected	260	263	286	253	198	221	186	167	151	125
Number of miRNAs in common	**247**		**243**		**187**		**158**		**120**	
Proportion of miRNAs in common	**89%**		**82%**		**81%**		**81%**		**77%**	
Other small RNA mapping (proportion of total reads, %)										
piRNA	3.5	3.7	8.9	5.1	4.1	3.1	2.9	2.8	3.2	1.7
snoRNA	1.0	0.7	0.8	0.5	1.1	1.3	0.2	0.2	0.4	0.4
snRNA	2.1	1.1	1.2	0.8	0.7	0.9	0.3	0.3	0.4	0.4
tRNA	2.1	3.4	4.0	2.9	0.9	0.9	1.0	1.0	4.5	2.3
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	9.6	10.0	34.5	31.7

Table 2. Evaluating the performance of the SMARTer method for small RNA-seq across RNA input types and amounts. Sequencing libraries were generated from 1 ng and 2 µg of human placenta, brain, and spleen total RNA, 400 pg of human plasma RNA, and 1 ng of human serum RNA. The miRNA fraction, corresponding to a final library size of about 175 bp, was enriched prior to sequencing using a BluePippin instrument. Following trimming, reads were mapped either to the GENCODE data set (for overall mapping), or to specific small RNA data sets, as indicated. Only miRNAs represented by at least five reads were included in count data for the number of miRNAs detected. Data from serum and plasma were generated in collaboration with University of Massachusetts Medical School.

800.662.2566
Visit us at takarabio.com/ngs

Clontech TaKaRa cellartis