TECH NOTE

Pushing the Limit: A Complete Solution for Generating Stranded RNA-Seq Libraries from Picogram Inputs of Total Mammalian RNA

Stranded, Illumina®-ready library construction in <5 hr:
Obtain high representation of coding and noncoding regions due to a combination of random priming, SMART with LNA technology, and a new ribosomal cDNA depletion method >>

Excellent performance:
Sensitive, reproducible results with pico-scale inputs and across tissue types >>

Overview

Transcriptome analysis using RNA sequencing (RNA-seq) empowers a deeper understanding of genetics by enabling RNA expression analysis over the entire transcriptome with high sensitivity and a wide dynamic range. One powerful application within this field is stranded total RNA-seq, which makes it possible to distinguish overlapping genes and to conduct comprehensive annotation and quantification of long noncoding RNAs. Typical solutions for total RNA-seq library prep require relatively high input amounts, in the range of 100 ng to 1 µg, and it is standard practice to remove the ribosomal RNA (rRNA) from the total RNA sample prior to cDNA synthesis and library preparation. Clontech was a pioneer in the development of a low-input solution, RiboGone technology for rRNA removal from total RNA, enabling library construction from inputs spanning 10 ng to 100 ng. We integrated this technology into our SMARTer stranded RNA-seq kits, reducing the representation of rRNA in the final libraries and leading to exceptional performance with inputs as low as 10 ng.

Following our tradition of enabling good science, we recognized the need for a solution to process even smaller amounts of starting material. With challenging samples, like RNA isolated using laser capture microscopy, it is difficult to obtain 10 ng of starting material; additionally, rRNA depletion leads to sample loss, undermining robust library preparation. Our newest release surpasses the previously achieved lower input limit by incorporating a proprietary technology in which ribosomal cDNA is removed *after* creating the complete cDNA library, thereby enriching RNAs of interest—namely mRNA and non-polyadenylated RNA. The SMARTer Stranded Total RNA-Seq Kit - Pico Input Mammalian provides a unique, sensitive, and ligation-free method to generate stranded, Illumina-ready cDNA libraries from an input range of **250 pg–10 ng** of total mammalian RNA in about 5 hours. The kit performs equally well with high-quality, partially degraded, or low-quality total RNA.

Novel Combination of Technologies Enables Library Preparation from Picogram Inputs

Key technologies incorporated into the SMARTer Stranded Total RNA-Seq Kit - Pico Input Mammalian are:

- Random priming, which allows representation of both polyadenylated and non-polyadenylated RNA
- A blend of SMART and locked nucleic acid (LNA) technologies, which promotes high-sensitivity detection
- A new, proprietary technology to dramatically reduce abundant rRNA reads from picogram amounts of sample

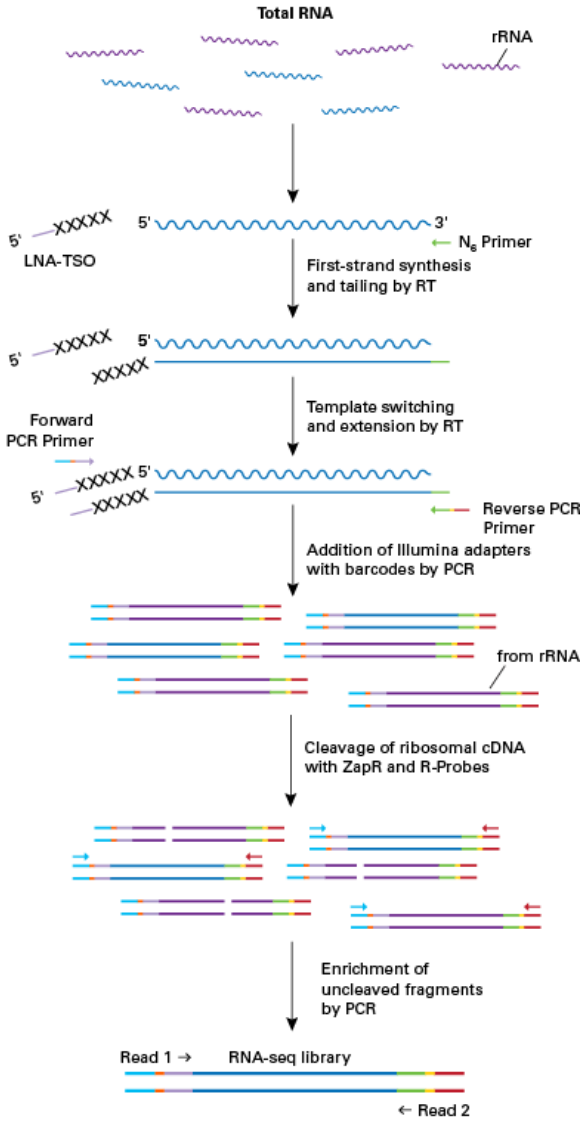
Random Priming Captures the Full Transcriptome

Compared to oligo dT priming, random priming enables a fuller representation of the transcriptome as well as better processing of degraded samples. Random priming also captures coding RNA *and*non-coding RNA, which are extremely important in gene expression and regulation, respectively (Mattick and Makunin, 2006; Kornienko, *et al.*, 2013), and in human disease development (Hindorff,*et al.*, 2009; Wapinski and Chang, 2011). This approach allows both polyadenylated and non-polyadenylated RNA to be captured—additional information to drive discoveries that might have been overlooked without random priming.

SMART, LNA, and Ribosomal cDNA Depletion Technologies Enable Highly Sensitive, Stranded Library Construction

This kit combines SMART (**S**witching **M**echanism **A**t 5' end of **R**NA **T**emplate) technology (Chenchik, *et al.*, 1998) with LNA technology, which is built into the template-switching oligo (LNA-TSO) to stabilize the interaction between the oligo and non-templated nucleotides added by the reverse transcriptase (Picelli, *et al.*, 2014). The directionality of the template-switching reaction preserves the strand orientation of the original RNA, making it possible to obtain strand-specific sequencing data from the synthesized cDNA.

This schematic shows the technologies integrated into and workflow of this product.



Cartoon of the technologies in the SMARTer Stranded Total RNA-Seq Kit - Pico Input Mammalian. This workflow allows users to generate Illumina-compatible libraries for RNA-seq experiments in approximately five hours.

- Random priming (represented as the green N_6 Primer) allows the generation of cDNA from all RNA fragments in the sample, including rRNA. When the SMARTScribe Reverse Transcriptase (RT) reaches the 5' end of the RNA fragment, the enzyme's terminal transferase activity adds a few non-templated nucleotides to the 3' end of the cDNA (shown as Xs). The carefully designed LNA-TSO (included in the Template Switching Oligo Mix) base-pairs with the non-templated nucleotide stretch, creating an extended template to enable the RT to continue replicating to the end of the oligonucleotide. The resulting cDNA contains sequences derived from the random primer and the LNA-TSO used in the reverse transcription reaction.
- In the next step, a first round of PCR amplification (PCR1) adds full-length Illumina adapters, including barcodes. The Forward PCR Primer binds to the LNA-TSO sequence (light purple), while the Reverse PCR Primer binds to sequence associated with the random primer (green). The ribosomal cDNA (originating from rRNA) is then cleaved by ZapR in the presence of the mammalian-specific R-Probes. This process leaves the library fragments originating from non-rRNA molecules untouched, with priming sites available on both 5' and 3' ends for further PCR amplification. These uncleaved fragments are enriched via a second round of PCR amplification (PCR2) using primers universal to all libraries.
- The final library contains sequences allowing clustering on an Illumina flow cell (P5 shown in light blue, P7 shown in red), Illumina (TruSeq® HT) indexes (Index 1 [i7] sequence shown in yellow, and Index 2 [i5] sequence shown in orange) as well as the regions recognized by sequencing primers Read Primer 1 (Read 1, purple) and Read Primer 2 (Read 2, green).

Optimal Performance over a Range of Input Amounts and Across Tissue Types

High Sensitivity and Reproducibility over a 100-Fold Input Range

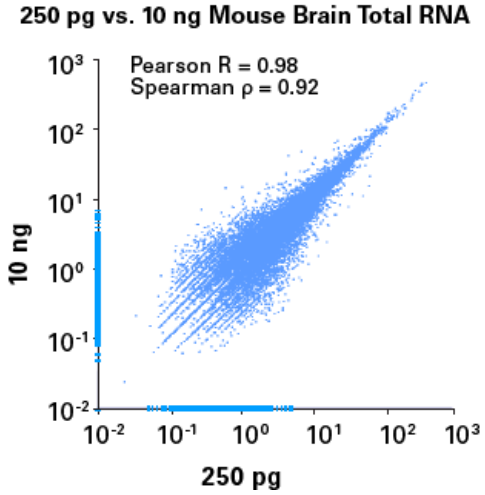
To test the consistency in performance of the SMARTer Stranded Total RNA-Seq Kit - Pico Input Mammalian over the recommended input range (and lower), libraries were generated from mouse brain total RNA (100 pg–10 ng), with two technical replicates per input amount.

Sequencing Alignment Metrics from 100 pg–10 ng Total RNA								
RNA source	Mouse brain total RNA							
Input amount (ng)	**10**		**1**		**0.25**		**0.1**	
Library yield (ng/µl)	10.5	14.8	9.93	8.3	6.91	7.48	5.76	7.26
Number of reads (millions)	2.6 (paired-end reads)							
Number of transcriptsFPKM >1	**12,714**	**12,709**	**12,744**	**12,725**	**12,540**	**12,615**	**12,286**	**12,528**
Pearson/Spearman correlations	**0.99/0.93**		**0.99/0.93**		**0.98/0.92**		**0.97/0.90**	
Correct strand per biological annotation (%)	97.7	97.7	97.7	97.7	97.7	97.7	97.7	97.6
Proportion of total reads (%):								
Exonic	22.6	22.8	23.4	23.5	23.3	23.1	23.1	22.8
Intronic	35.6	35.7	35.3	36.2	35.9	35.5	36.1	35.1
Intergenic	8.3	8.2	8.2	8.2	8.0	8.0	7.8	7.8
rRNA	**11.2**	**10.5**	**10.8**	**9.9**	**9.7**	**9.7**	**8.8**	**9.5**
Mitochondrial	8.8	8.7	8.3	8.5	8.3	8.4	7.5	7.9
Duplicate rate (%)	12.8	12.5	17.3	17.8	31.3	28.8	44.2	40.2

Sequencing metrics are consistent across RNA input amounts. Mouse brain total RNA (100 pg–10 ng) was used to generate RNA-seq libraries in duplicate with the SMARTer Stranded Total RNA-Seq Kit - Pico Input Mammalian. Libraries were amplified using 5 PCR cycles in "PCR1" and 10, 13, 15, or 16 PCR cycles in "PCR2" for the 10 ng, 1 ng, 250 pg, and 100 pg libraries, respectively. Sequences were analyzed as described in the methods.

Maintenance of Transcript Representation across a Range of Recommended Input Amounts

The sequencing statistics are very similar regardless of the input amount (table, above). The number of transcripts with FPKM >1 (fragments per kilobase of exon per million reads) is above 12,000 and strand-of-origin information is preserved for all inputs. Similarly, the proportion of total reads mapping to rRNA is between ~9% and 12%, compared to ~67% if libraries are prepared without removing ribosomal cDNA (data not shown). The correlations between replicates are consistently high, with Pearson (R) and Spearman (ρ) correlations averaging ~0.99 and ~0.92, respectively. As expected, the duplicate rate increases with decreasing input amounts. The data generated from 250 pg and 10 ng are highly consistent (scatter plot, below), with very few transcripts represented in only one of the libraries.

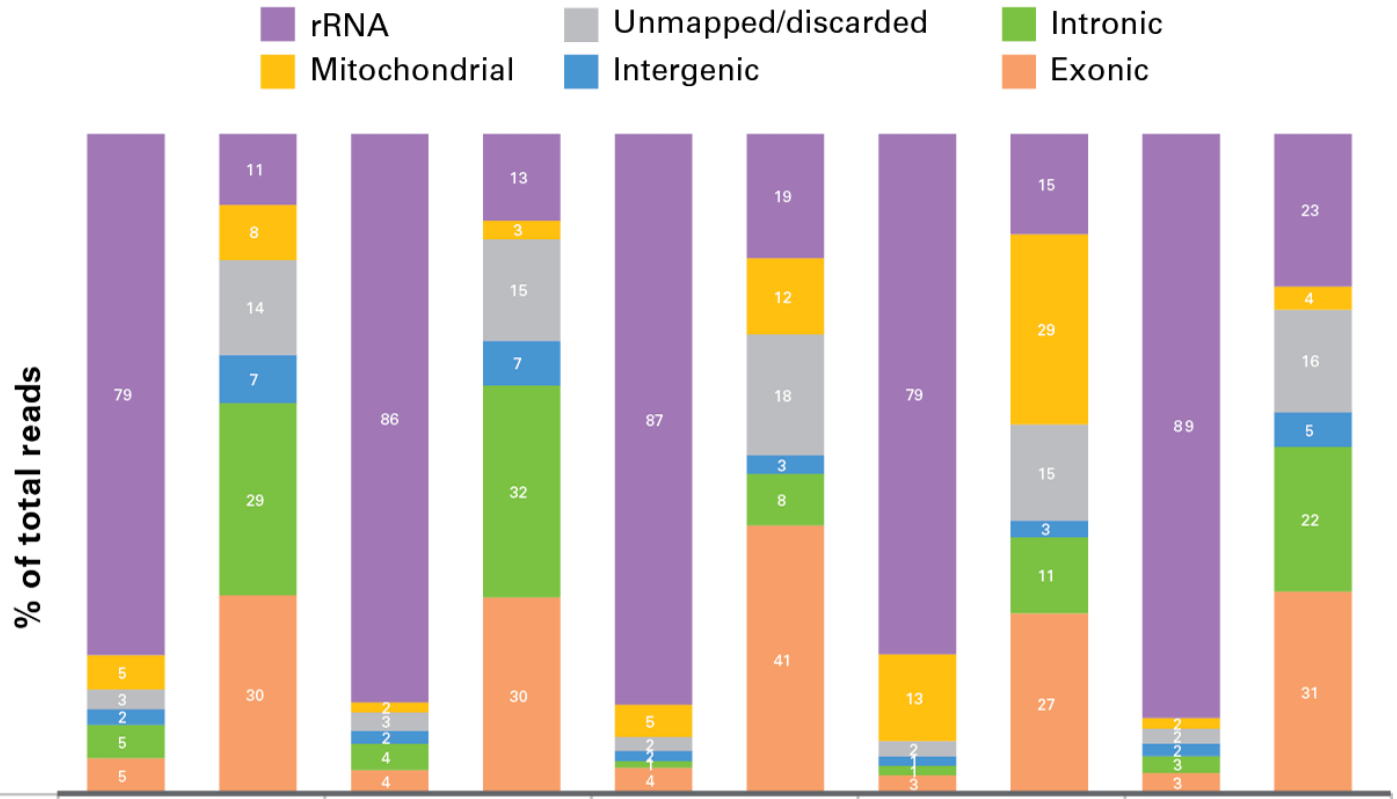


High reproducibility across the recommended input range. Comparison of FPKMs from libraries generated with 250 pg and 10 ng of mouse brain total RNA using the SMARTer Stranded Total RNA-Seq Kit - Pico Input Mammalian. FPKM values are shown on a Log_{10} scale. Transcripts represented in only one library can be seen along the X- and Y-axes of the scatter plot.

Efficient Reduction of Library Reads Mapping to rRNA Enhances Sensitivity across Tissue Types

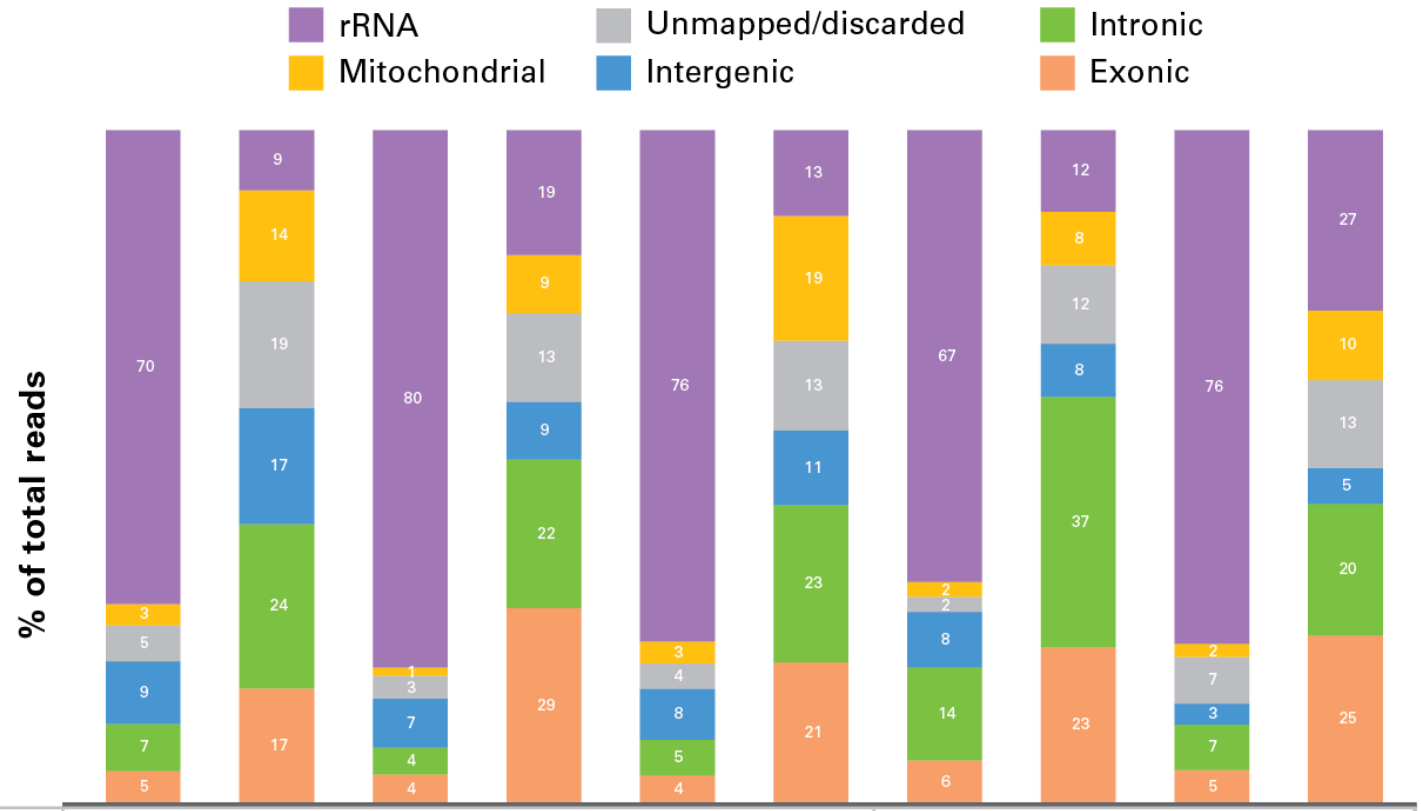
The effectiveness of the ribosomal cDNA depletion method employed in the SMARTer Stranded Total RNA-Seq Kit - Pico Input Mammalian was validated in a range of tissue types (human brain, placenta, skeletal muscle, heart, and spleen; rat brain, liver, and kidney; and mouse brain and liver). Samples were processed into libraries with or without the inclusion of R-Probes that allow the cleavage of ribosomal cDNA. In all cases, the number of reads mapping to rRNA was significantly reduced—down to approximately 10–30% of total reads, depending on the tissue type (see charts, below). This resulted in a 5- to 10-fold improvement in the number of reads useful for transcript identification (exons).

Distribution of Reads in Libraries from Human Tissues



Tissue type	Brain		Placenta		Skeletal Muscle		Heart		Spleen	
Ribosomal cDNA depletion	(–)	(+)	(–)	(+)	(–)	(+)	(–)	(+)	(–)	(+)
Library yield (ng/µl)	18.1	5.2	10.2	3.7	11.5	2.4	9.9	3.8	10.6	3.2
Number of transcripts FPKM >1	12,581	15,693	9,079	13,095	6,619	11,879	7,667	14,448	9,900	15,798
Correct strand per biological annotation (%)	97.6	97.6	97.8	97.8	98.5	98.5	98.4	98.4	98.3	98.1

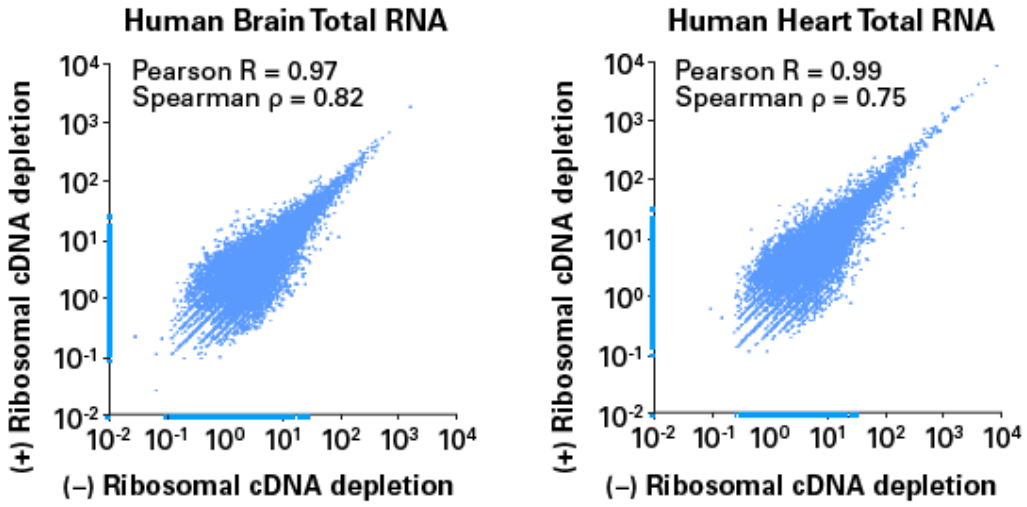
Distribution of Reads in Libraries from Rodent Tissues



Animal	Rat						Mouse			
Tissue type	Brain		Liver		Kidney		Brain		Liver	
Ribosomal cDNA depletion	(–)	(+)	(–)	(+)	(–)	(+)	(–)	(+)	(–)	(+)
Library yield (ng/µl)	18.7	4.3	20.8	2.7	18.8	3.7	15.8	14.2	18.0	5.2
Number of transcripts FPKM >1	8,618	12,477	5,645	10,647	7,939	13,197	10,465	12,627	8,001	10,922

Improved exon mapping and transcript identification from human and rodent tissues. Distribution of reads in libraries prepared with the SMARTer Stranded Total RNA-Seq Kit - Pico Input Mammalian in the absence (–) or presence (+) of R-Probes, using 250 pg of human total RNA (top panel) and 250 or 500 pg of rat or mouse total RNA (bottom panel). Sequences corresponding to 1 million paired-end reads per library were analyzed as described in the methods.

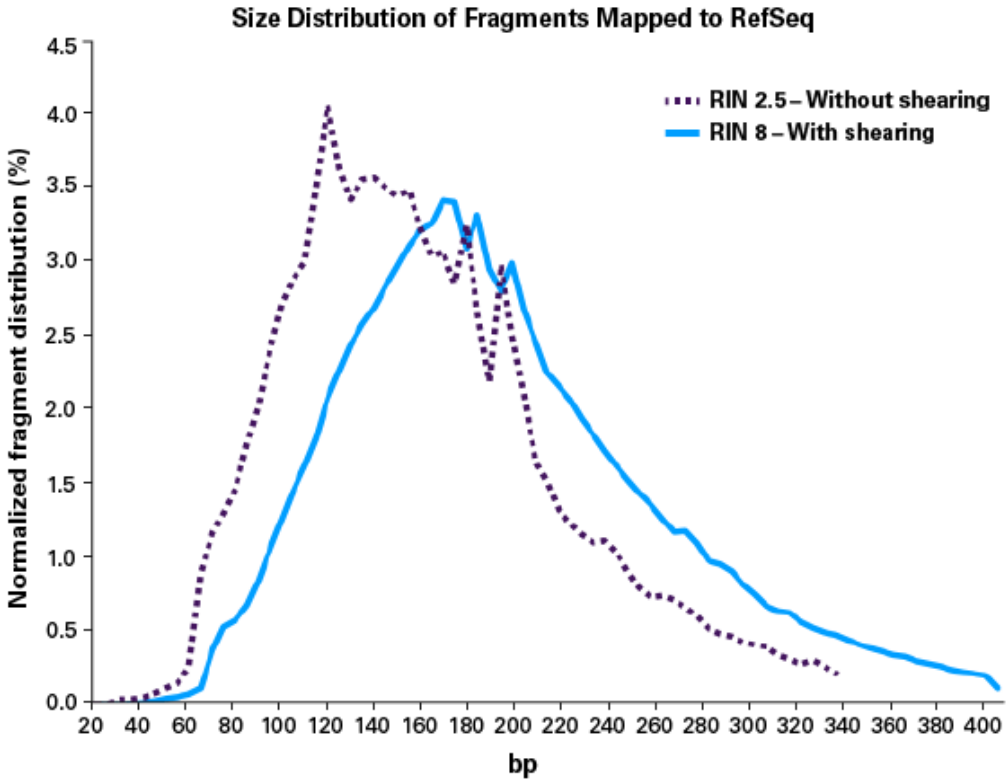
In addition to providing a significant improvement in gene detection from less than 10 ng of total RNA, another major advantage of our proprietary ribosomal cDNA depletion method is its ability to maintain the transcript representation found in the starting samples. When comparing libraries prepared with and without R-Probes (see tables above and scatter plots below), we observed excellent Pearson (R) and Spearman (ρ) correlations, averaging ~0.97 and ~0.82 for human brain and ~0.99 and ~0.75 for human heart, respectively. These data indicate that the ribosomal cDNA depletion method is specific to rRNA and does not have significant off-target effects on desirable transcripts.



High reproducibility between libraries with or without ribosomal cDNA depletion. Comparison of FPKMs from libraries generated with 250 pg human brain or heart total RNA with (+) or without (–) ribosomal cDNA depletion (i.e., with or without R-Probes, respectively) using the SMARTer Stranded Total RNA-Seq Kit - Pico Input Mammalian. FPKM values are shown on a Log_{10} scale. Transcripts represented in only one library can be seen along the X- and Y-axes of the scatter plots.

Complete Solution for Processing Samples of Varying Quality

Like our other SMARTer stranded kits, the new SMARTer Stranded Total RNA-Seq Kit - Pico Input Mammalian features an integrated RNA shearing step to reduce the RNA fragments to a size adequate for sequencing on Illumina platforms, plus an alternate no-shearing protocol to process samples that are already fragmented or degraded. This kit allows efficient capture of RNA fragments ~100 nt and above (see graph, below). With its robust accommodation of highly fragmented samples, the SMARTer Stranded Total RNA-Seq Kit - Pico Input Mammalian is a great choice for processing degraded samples that are only available in limited quantities.



Efficient capture of degraded RNA. Distribution of insert sizes in libraries generated from 250 pg of high-quality (RIN 8) or highly degraded (RIN 2.5) human total RNA using the SMARTer Stranded Total RNA-Seq Kit - Pico Input Mammalian with a 4-min shearing step (RIN 8) or with the alternate no-shearing protocol (RIN 2.5). RNA-seq libraries were sequenced on an Illumina MiSeq® instrument in paired-end mode. The number of mapped fragments with any given insert size was normalized to the total number of fragments in the library. Fragments mapping to rRNA or the mitochondrial genome were excluded. Sequences were analyzed as described in the methods.

Summary

The SMARTer Stranded Total RNA-Seq Kit - Pico Input Mammalian is a complete solution to the challenge of creating stranded, indexed cDNA libraries for RNA-seq from picogram amounts of total mammalian RNA. A unique combination of technologies, including SMART with LNA technology and a proprietary ribosomal cDNA depletion method, enables solid performance over the recommended range of 250 pg–10 ng, with demonstrated results from inputs as low as 100 pg. This kit excels with high-quality, partially degraded, and low-quality input RNA, enabling consistent, reproducible results from a broad range of sample types. In under five hours, using very low total RNA input amounts from samples of varying types and qualities, this kit can generate Illumina-ready libraries that accurately represent coding and non-coding RNA—a major development in library prep for next-gen RNA-seq.

Methods and References

Comparison of Different Input Amounts

For experiments comparing performance over different input amounts, 100 pg–10 ng of high-quality Mouse Brain Total RNA was used as the input for the SMARTer Stranded Total RNA-Seq Kit - Pico Input Mammalian, and was processed as described in the user manual (using a 4-min shearing time). For testing input RNA from different tissues, a selection of human, mouse, and rat high-quality total RNA was used at the input amount specified in the corresponding charts. This RNA was processed as described in the user manual, but the R-Probes were replaced with water for the "no ribosomal cDNA depletion" control. To evaluate the size distribution in libraries made from high-quality vs. degraded RNA samples, additional libraries were prepared with 250 pg of total RNA extracted from formalin-fixed, paraffin-embedded normal human liver with a RIN of 2.5 (BioChain). All libraries were sequenced using an Illumina MiSeq instrument, generating at least 2.6 million paired-end reads (2 x 75 bp) per library for the 100 pg–10 ng input comparison, and at least 1 million paired-end reads for all other libraries.

Sequence Analysis

Reads from all libraries were trimmed and mapped to mammalian rRNA and the human, mouse, or rat mitochondrial genomes, as appropriate, using CLC Genomics Workbench. The remaining reads were subsequently mapped using CLC to the human (hg19), mouse (mm10), or rat (Rnor_6.0) genomes with RefSeq annotation. All percentages shown, including the number of reads that map to introns, exons, or intergenic regions, are percentages of the total reads in the library. Reads unaccounted for in the mapping or discarded during trimming are shown as "unmapped/discarded." The number of genes identified in each library was determined by the number of transcripts with an FPKM greater than or equal to 1. The number of reads mapping to the correct strand (as defined in the current genome annotation) was determined using Picard analysis. Scatter plots were generated using FPKM values from CLC mapping to the transcriptome. In order to identify the transcripts found in only one replicate, 0.01 was added to each value prior to graphing.

References

1. Chenchik, A., *et al*. (1998) In RT-PCR Methods for Gene Cloning and Analysis. (BioTechniques Books, MA) pp. 305–319.
2. Hindorff, L. A., *et al.* (2009) *Proc. Natl. Acad. Sci. U. S. A.* **106**(23):9362–9367.
3. Kornienko, A. E., *et al.* (2013) *BMC Biol.* **11**:59.
4. Mattick, J. S. and Makunin, I. V. (2006) *Hum. Mol. Genet.* **15**:R17–29.
5. Picelli, S., *et al.* (2013) *Nat. Methods* **10**(11):1096–1098.
6. Wapinski, O. and Chang H. Y. (2011) *Trends Cell Biol.* **21**(6):354–61.

View web page >>
http://www.clontech.com/US/Products/cDNA_Synthesis_and_Library_Construction/NGS_Learning_Resources/Technical_Notes/Pico_Input_Total_RNA_Seq_Illumina