Generate UMI-labeled, stranded libraries for whole-transcriptome analyses, compatible with Illumina NGS platforms.

- Unique molecular identifiers for downstream data analysis
  Seamless integration of UMIs in the protocol help to mitigate PCR bias and accurately identify true variants and rare mutations, without any additional deduplication steps.
- Superior sequencing performance
  Reconfigured sequencing libraries perform well on all Illumina® platforms without the addition of PhiX adapter-ligated library as a control.

Introduction

Massively parallel cDNA sequencing, or RNA-seq, has become the gold standard for whole-transcriptome gene expression analysis and is widely used in numerous applications to study cell and tissue transcriptomes. However, despite its many advantages, RNA-seq can be challenging in some situations, including cases where input amounts are low or comprised of degraded RNA samples.

Takara Bio was a pioneer in the development of a low-input solution: RiboGone technology for rRNA removal from total RNA, which enables 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. With the release of the SMARTer Stranded Total RNA-Seq Kit - Pico Input Mammalian, we were able to be successful with even lower inputs by incorporating ZapR™, 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 latest update to the original kit, the SMARTer Stranded Total RNA-Seq Kit v3 - Pico Input Mammalian (referred to as “Pico v3," see workflow in Figure 1) now features the inclusion of unique molecular identifiers (UMIs). UMIs not only help mitigate PCR bias but also provide the user with the opportunity to identify true variance within the sample. Pico v3 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 7.5 hours.

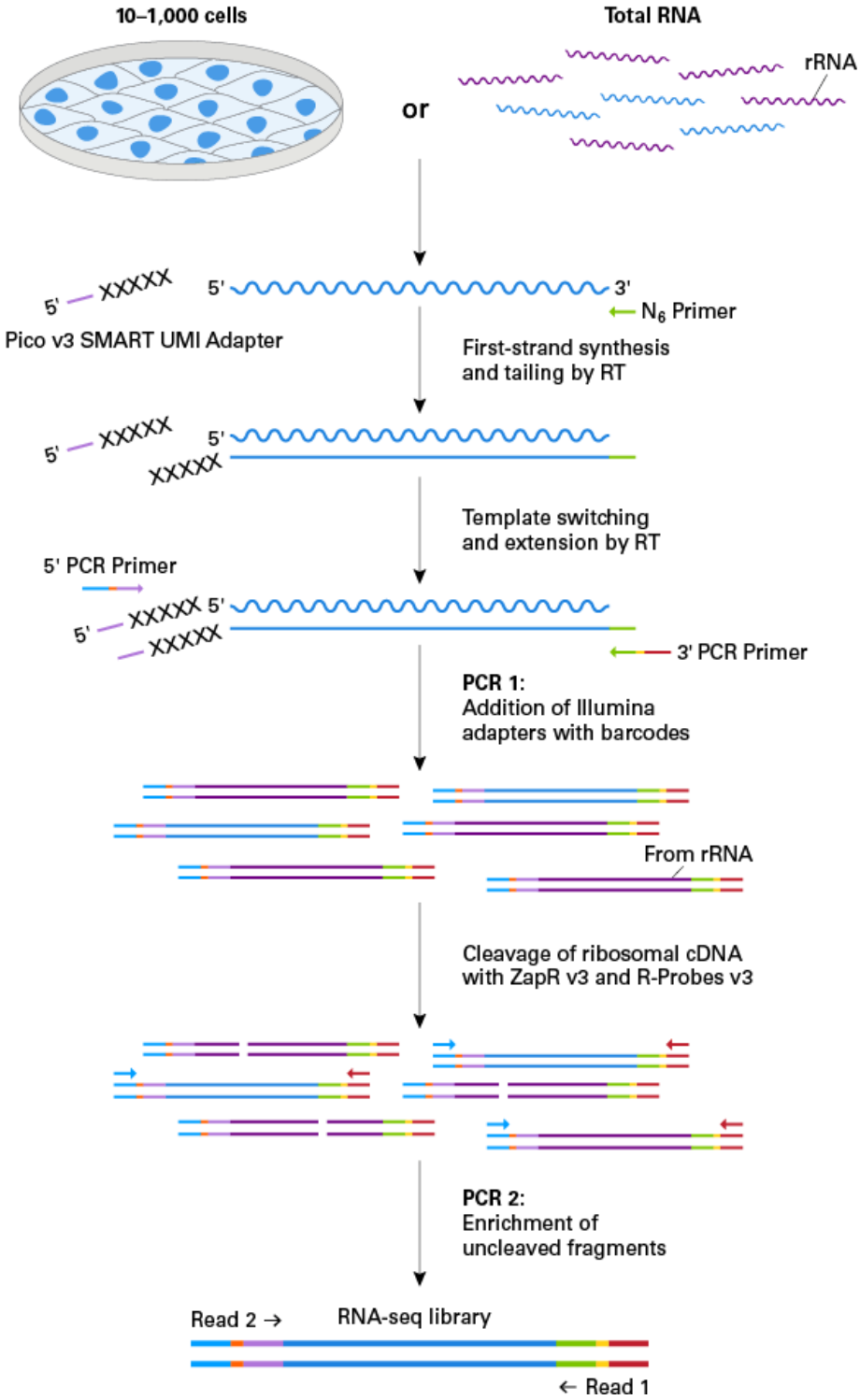

Figure 1. Schematic of technology in the SMARTer Stranded Total RNA-Seq Kit v3 - Pico Input Mammalian. SMART technology is used in this ligation-free protocol to preserve strand-of-origin information. Random priming (represented by the green N6 Primer) allows the generation of cDNA from all RNA fragments in the sample, including rRNA. When the SMARTScribe™ II Reverse Transcriptase (RT) reaches the 5' end of the RNA fragment, the enzyme's terminal transferase activity adds a few nontemplated nucleotides to the 3' end of the cDNA (shown as XXXXX). The Pico v3 SMART UMI Adapter (included in the SMART UMI-TSO Mix v3) base-pairs with the nontemplated nucleotide stretch, creating an extended template to enable the RT to continue replicating to the end of the oligonucleotide. The Pico v3 SMART UMI Adapter also contains 8 random nucleotides which function as the unique molecular identifiers for downstream data analysis. In the next step, the first round of PCR amplification (PCR 1) adds full-length Illumina adapters, including barcodes. The 5' PCR Primer binds to the Pico v3 SMART UMI Adapter sequence (light purple), while the 3' PCR Primer binds to the sequence associated with the random primer (green). The ribosomal cDNA is then cleaved by ZapR v3 in the presence of the mammalian-specific R-Probes. The resulting cDNA contains sequences derived from the random primer and the Pico v3 SMART UMI Adapter used in the reverse transcription reaction. 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 fragments are enriched via a second round of PCR amplification (PCR 2) using primers universal to all libraries. The final library contains sequences that allow clustering on any Illumina flow cell (see details in Figure 2).

Results

Unique molecular identifiers for downstream data analysis

Commonly used high-throughput sequencing platforms, including the Illumina NextSeq®, require PCR amplification during library construction to increase the number of cDNA molecules to an amount sufficient for sequencing and/or enrichment for fragments with successful adapter ligation (Cha and Thilly 1998; Dohm et al. 2008). However, PCR often stochastically introduces errors, such as sequencing artifacts and false mutations, that can propagate to late cycles. Biases in the PCR amplification step lead to particular sequences becoming overrepresented in the final library, resulting in an inaccurate fold-change measurement (Aird et al. 2011). What then, is the solution? This is where UMIs come in.

Unique molecular identifiers (UMIs) are molecular tags that are used to detect and quantify unique mRNA transcripts. The random sequence composition of UMIs assures that every fragment-UMI combination is unique in the library—much like how a barcode identifies an item in a grocery store or the Dewey decimal system identifies a book in a library. By using UMIs ligated to fragments of the input sample, PCR clones can be found by searching for non-unique fragment-UMI combinations, and therefore help to determine whether a sequence arises from truly distinct molecules, or from PCR amplification (Fu et al. 2018). UMIs provide the highest level of error correction and accuracy, allowing for superior representation of the transcripts.

UMIs are incorporated into all libraries produced with Pico v3 without additional steps required (Figure 2). An 8-nt UMI is introduced into the same location in each fragment during library preparation through the reverse transcription step (prior to PCR amplification). Thus, it is possible to accurately identify PCR duplicates and correct for specific preferentially amplified sequences. High-resolution reads and accurate detection of true variants are now achievable.

Figure 2. Structural features of final libraries generated with the SMARTer Stranded Total RNA-Seq Kit v3 - Pico Input Mammalian. The adapters added using the SMARTer RNA Unique Dual Index Kit - 24U (Cat. # 634451) contain sequences allowing clustering on any Illumina flow cell (P7 shown in light blue, P5 shown in red, Index 1 [i7] sequence shown in orange, and Index 2 [i5] sequence shown in yellow), as well as the regions recognized by sequencing primers Read Primer 2 (Read 2, purple) and Read Primer 1 (Read 1, green). Read 1 generates sequences antisense to the original RNA, while Read 2 yields sequences sense to the original RNA (orientation of original RNA denoted by 5' and 3' in dark blue). The first 8 nt of the second sequencing read (Read 2) are UMIs (dark purple) followed by 3 nucleotides of UMI-linker (shown as NNN) and 3 nucleotides derived from the Pico v3 SMART UMI Adapter (shown as XXX).

Superior sequencing performance

To test the consistency in the performance of Pico v3 over the recommended input range, libraries were generated from human lung cancer normal adjacent tissue (NAT FFPE) from a single donor (500 pg–10 ng), with two technical replicates per input amount (Figure 3).

Sequencing metrics were shown to be consistent between technical replicates, as well across the entire range of RNA input amounts. The average number of transcripts ≥0.1 TPM (transcripts per kilobase million) was reported as 43,778 ± 1,352 (mean ± s.d., standard deviation), CV=3.1%. Similarly, the average number of genes ≥1 TPM was reported as 18,690 ± 276, CV=1.5%. Comparison transcript expression levels also indicated a strong correlation across a range of input amounts. Proportions of reads mapped to various RNA species were comparable, regardless of RNA input amount. Of particular note were the relatively low proportions of reads mapping to nuclear and mitochondrial rRNA compared to those of the remaining RNA species reported.

Although we recommend using high-quality, gDNA-free RNA, we often notice residual levels of gDNA contamination in some challenging RNA samples (e.g., degraded RNA extracted from FFPE), resulting in high intergenic mapping rates and low strand specificities. Therefore, we validated an optional step (see the user manual, Appendix A) to remove potential gDNA contamination from RNA samples if such contamination is shown to be an issue. This optional step is seamlessly integrated into the Pico v3 workflow and significantly improves sequencing performance, as is indicated by a reduced fraction of reads mapped to intergenic RNA and a higher strand specificity versus no treatment (data not shown).

Sequencing alignment metrics from 500 pg–10 ng total RNA								
RNA source	**Human lung cancer total RNA (NAT FFPE)**							
Input amount (ng)	**0.5**		**1**		**5**		**10**	
Number of reads (millions)	4.0 (paired-end reads)							
Discarded reads (%)	5.9	3.6	2.7	2.3	1.2	1.1	4.7	4.3
Unique reads (%)	87	87	92	92	96	95	97	97
Overall mapping (%)	92	94	96	96	97	97	94	94
Number of transcripts ≥0.1 TPM	41,245	42,293	44,189	44,217	44,814	44,361	44,051	45,067
Number of genes ≥0.1 TPM	18,193	18,414	18,829	18,680	18,821	18,660	18,941	18,980
Strand specificity	92	92	92	92	92	92	92	92
Proportion of total reads (%)								
Exonic	21	22	22	22	21	21	14	21
Intronic	58	59	60	61	63	64	58	60
Intergenic	6.0	6.1	6.2	6.2	6.3	6.3	6.1	6.3
rRNA	5.8	6.4	6.1	6.2	5.5	5.1	8.5	6.6
Mitochondrial	0.9	1.0	1.0	1.0	1.0	0.9	0.9	0.8
Genomic	85	87	89	89	91	91	84	87
Duplicate	13	13	8	8	4	5	3	3

Table 1. Performance metrics for Pico v3. Sequencing libraries were generated from total RNA extracted from human lung FFPE tissue using the Pico v3 kit and sequenced on a MiniSeq™ Sequencing System. Sequencing metrics are shown for libraries generated from inputs of 0.5, 1, 5, and 10 ng, with two technical replicates per input amount. Sequences were analyzed as described in the Methods.

Conclusions

The SMARTer Stranded Total RNA-Seq Kit v3 - 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. The unique combination of SMART technology with ZapR for ribosomal cDNA depletion enables unparalleled sensitivity, consistency, and reproducibility 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. Additionally, design updates made to this version of the kit allow for seamless integration of unique molecular identifiers (UMIs) into the protocol, helping to mitigate PCR bias and accurately identify true variants and rare mutations within the sample. In around 7.5 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 noncoding RNA—a major development in library prep for next-gen RNA-seq.

Methods

Library preparation and sequencing for FFPE samples

To evaluate the performance of the Pico v3 kit with FFPE samples, total RNA was extracted from FFPE human lung cancer normal adjacent tissue (BioOption) using a NucleoSpin total RNA FFPE kit (Cat. # 740982.10). Prior to library preparation, RNA integrity was evaluated on an Agilent Bioanalyzer using an Agilent RNA 6000 Pico Kit (Agilent, Cat. # 5067-1513), yielding a DV200 value of 66%. Libraries were generated from the extracted RNA using the Pico v3 kit without additional RNA fragmentation (Protocol Option 2 in the user manual). Libraries were sequenced on a MiniSeq Sequencing System using a MiniSeq High Output Reagent Kit (150 cycles), (Illumina, Cat. # FC-420-1002). Sequencing reads were aligned to the human transcriptome and the human genome.

Sequencing data analysis

Reads from all libraries were trimmed and mapped to mammalian rRNA and the human mitochondrial genomes using CLC Genomics Workbench. The remaining reads were subsequently mapped to the human genome using CLC with ENSEMBL-GRCh38.81 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. The number of transcripts identified in each library was determined by the number of transcripts with an TPM greater than or equal to 0.1, as shown in Table 1.

References

Aird, D., Ross, M.G., Chen, W. et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12**:R18 (2011).

Cha R.S. & Thilly W.G. Specificity, efficiency, and fidelity of PCR. *PCR Methods Appl.* **3**:S18–S29 (1993).

Dohm J.C., Lottaz C., Borodina T., Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36**:e105 (2008).

Everaert C., et al. Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data. *Sci Rep.* **7**:1559 (2017).

Fu, Y., Wu, P., Beane, T. et al. Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. *BMC Genomics* **19**:531 (2018).

Mortazavi, A., Williams, B.A., McCue K., Schaeffer L., Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. *Methods* **5**:621–628 (2008).

Related Products

Cat. #	Product	Size	License	Quantity	Details
634485	SMARTer® Stranded Total RNA-Seq Kit v3 - Pico Input Mammalian	24 Rxns		*	

The SMARTer Stranded Total RNA-Seq Kit v3 - Pico Input Mammalian is used to generate strand-specific RNA-seq libraries for Illumina sequencing from 250 pg–10 ng inputs of purified total RNA or from 10–1,000 intact cells. This kit incorporates Takara Bio's proprietary SMART (Switching Mechanism at the 5' end of RNA Template) technology and includes refinements to the SMARTer method for stranded RNA-seq that simplify the library preparation workflow and improve sequencing performance. This method was developed to work with either high- or low-quality total RNA, does not require additional rRNA removal methods or kits, and produces sequencing libraries that retain strand-of-origin information. The integrated removal of cDNAs derived from rRNA—typically present in high abundance following cDNA synthesis from total RNA inputs—makes the workflow extremely sensitive, yielding data that is highly reproducible with low mapping to rRNA. The new library design featured in this updated kit adds an 8 nucleotide (nt) unique molecular identifier (UMI) through the reverse-transcription step to mitigate potential PCR bias as well as to provide customers with additional information for transcript quantification, specifically for true variants and rare mutations. This kit includes the SMARTer RNA Unique Dual Index Kit - 24U (Takara Bio, Cat. No. 634451). SMARTer RNA Unique Dual Index Kit – 96U Set A and B (Cat. #s 634452, and 634457) are available for purchase. Together they offer up to 192 unique dual indexes for multiplexing.

Documents	Components	Image Data

634486	SMARTer® Stranded Total RNA-Seq Kit v3 - Pico Input Mammalian	48 Rxns		*	
634487	SMARTer® Stranded Total RNA-Seq Kit v3 - Pico Input Mammalian	96 Rxns		*	
634488	SMARTer® Stranded Total RNA-Seq Kit v3 - Pico Input Mammalian	192 Rxns		*	

Add to Cart

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