Enabling long-read mRNA-seq for biomarker discovery using limited clinical sample inputs

Yue Yun, Lisa Welter, Kazuo Tori, Jackson Peterson, Alan Du, Gilma Sevilla, Yana Ryan, Ploy Setthasap, Sherry Wei, Tomoya Uchiyama, Shiyi Yin, Mike Covington, Mohammad Fallahi, Saloni Pasta, Bryan Bell, and Andrew Farmer*

Takara Bio USA, Inc., 2560 Orchard Pkwy, San Jose, CA 95131 USA *Corresponding Author

Abstract

RNA serves a central role in cellular biology by converting genomic information into effector molecules, either as protein-coding mRNAs or as functional non-coding RNAs. As a result, RNA sequencing has become an important method for understanding biological processes. While it has long been appreciated that a vast diversity of RNA transcripts can be produced through alternative splicing, more recent work has identified contributions of alternate and aberrant splicing to diseases like cancer, neurodegeneration, and autoimmunity. Therefore, understanding the diversity of RNA transcripts has become increasingly important in pathophysiology and biomarker discovery.

Third-generation sequencing technologies provide the opportunity to sequence full-length cDNA without the need for fragmentation and thus provide a more complete picture of isoform structure and transcript abundance. However, a current limitation of long-read RNA sequencing (LR-RNA-seq) is the requirement for large amounts of input RNA, which can be unachievable for some primary sample types, such as resected tumor samples or sorted blood cancer cells.

Here we describe a new LR-RNA-seq product that enables full-length RNA sequencing from single cells (~10 pg) up to 100 ng total RNA. Using a 10 ng total RNA input we demonstrate the ability of this technology to reliably sequence at an average length (N50) of 2 kb and detect full-length transcripts as long as 10 kb. Performance at the single-cell level is substantially better than existing single-cell LR-RNA-seq methods. Furthermore, our data provides a more complete picture of isoform-specific changes compared to other commercially available technologies. With the ability to process up to 96 samples at a time, this technology will enable the processing of rare or valuable samples to uncover novel biomarkers beyond gene expression.

1 SMART-Seq® mRNA Long Read (LR) workflow

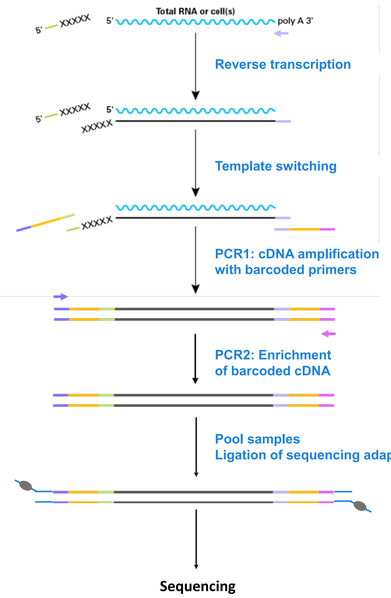

Figure 1. Library preparation workflow for the SMART-Seq mRNA Long Read kit. First-strand cDNA synthesis is primed by the SMART-Seq LR Primer and performed by an Moloney murine leukemia virus (MMLV)-derived reverse transcriptase (RT). Upon reaching the 5' end of each mRNA molecule, the RT adds non-templated nucleotides to the first-strand cDNA facilitating hybridization with a template-switching oligonucleotide (TSO). In the template-switching step, the RT uses the remainder of the SMART-Seq LR TSO as a template for the incorporation of an additional sequence on the end of the first-strand cDNA. The first-strand cDNA is then barcoded and amplified by the first round of PCR (PCR1); after clean-up of the PCR1 product a second round of PCR enriches for barcoded fragments. Samples are pooled and end-prepped, and sequencing adapters are ligated using the Ligation Sequencing Kit V14 (Oxford Nanopore Technologies or ONT). The SMART-Seq Library Prep Kit is used to generate sequencing-ready libraries. After sequencing, samples are basecalled and demultiplexed using Guppy (ONT). Downstream analysis is performed using minimap2 (Github), SAMtools, and Salmon.

2 SMART-Seq mRNA LR library size distribution

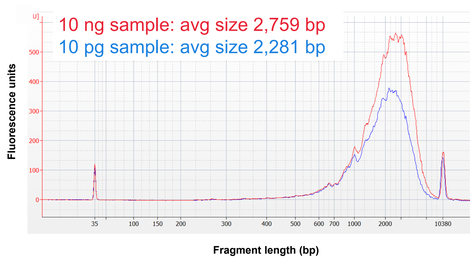

Figure 2. SMART-Seq mRNA LR generates long barcoded cDNA across a wide range of inputs. The SMART-Seq mRNA LR workflow was used to create cDNA from 10 pg or 10 ng total mouse brain RNA. The cDNA size distribution was measured on a 2,100 Bioanalyzer (Agilent Technologies) using an Agilent High Sensitivity DNA Kit. Multiple replicates produced similar results, so data from representative 10 ng and 10 pg samples are shown.

3 SMART-Seq mRNA LR shows high sensitivity across broad total RNA input ranges and for single cells

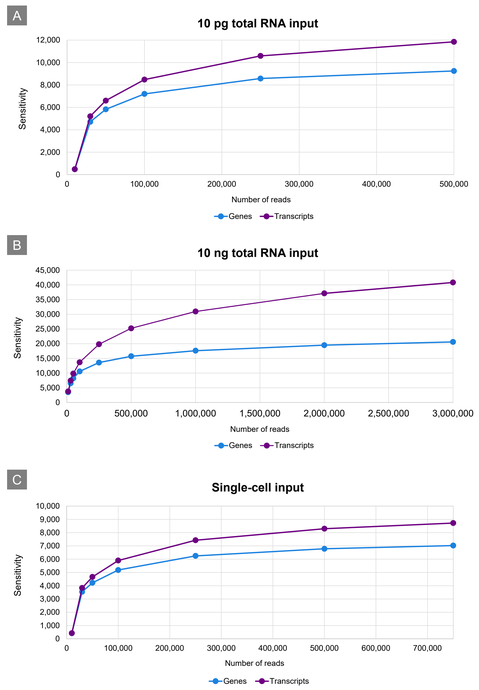

Figure 3. SMART-Seq mRNA Long Read demonstrates high sensitivity across a broad range of RNA inputs. To evaluate the performance of the SMART-Seq mRNA LR workflow, cDNA was generated from 10 pg or 10 ng total mouse brain RNA (MBR) or single K562 cells using the workflow described in Figure 1. After sequencing, data was basecalled and demultiplexed using Guppy, and reads were downsampled to the indicated read counts. Downsampling analysis of the 10 pg MBR dataset (**Panel A**), 10 ng MBR dataset (**Panel B**), and single K562 cells (**Panel C**) demonstrates the gene and transcript sensitivity of the workflow.

4 SMART-Seq mRNA LR detects diverse full-length isoforms

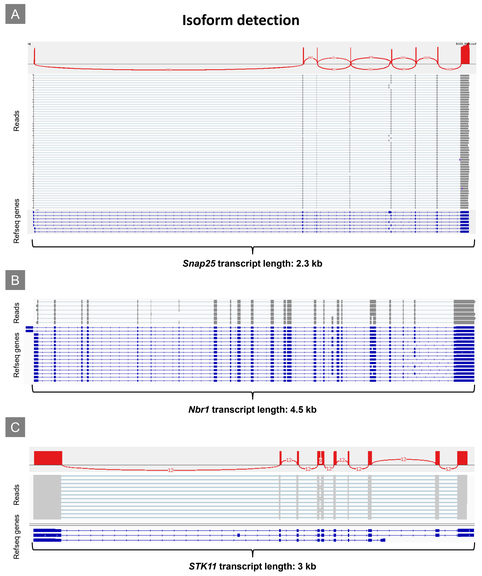

Figure 4. SMART-Seq mRNA LR detects full-length isoforms. cDNA was generated using the SMART-Seq mRNA LR workflow described in Figure 1. Basecalling and demultiplexing was performed using Guppy, and reads were aligned using minimap2. Isoforms of *Snap25* (**Panel A**) and *Nbr1* (**Panel B**) detected from 10 pg mouse brain RNA input are visualized in Integrative Genomics Viewer (IGV). **Panel C.** Transcripts of *STK11* detected from 10 pg of RNA isolated from primary human lung cancer samples.

5 SMART-Seq mRNA LR generates even gene-body coverage

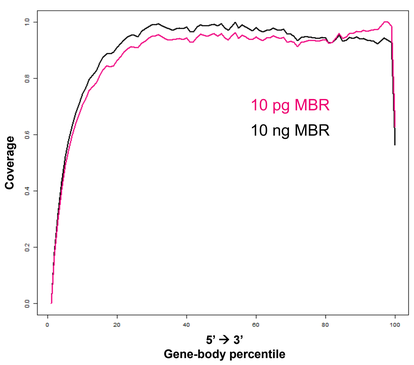

Figure 5. SMART-Seq mRNA LR yields consistent gene-body coverage, even for low input total RNA samples. 10 pg and 10 ng MBR samples underwent the SMART-Seq mRNA LR workflow. Gene-body coverage was assessed for the average of 8 replicates of 10 pg and 10 ng MBR samples.

6 SMART-Seq mRNA LR shows highly reproducible gene expression patterns across replicates

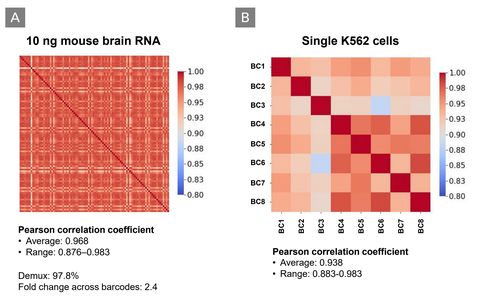

Figure 6. SMART-Seq mRNA LR is highly reproducible across replicates. The SMART-Seq mRNA LR workflow was used to create cDNA from 96 replicates of 10 ng MBR (**Panel A**) or 8 single K562 cells (**Panel B**). Barcoded cDNA was pooled, and libraries were generated and sequenced according to the workflow. Samples were basecalled, demultiplexed using Guppy, aligned with minimap2, and counts matrices were generated using feature counts. Pairwise correlation matrices were calculated from the counts matrices in R. High Pearson correlation values between samples indicate high technical reproducibility.

7 Performance of SMART-Seq mRNA LR on single cells vs. major competitor's protocol

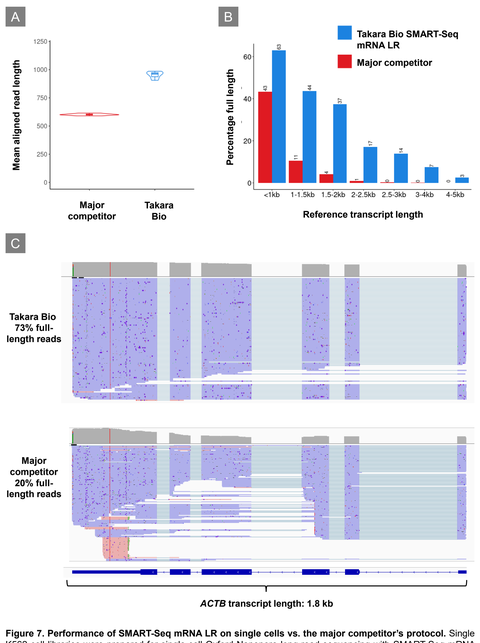

Figure 7. Performance of SMART-Seq mRNA LR on single cells vs. the major competitor's protocol. Single K562 cell libraries were prepared for single-cell Oxford Nanopore long-read sequencing with SMART-Seq mRNA LR (Figure 1) or using the major competitor's single-cell protocol. Datasets were downsampled to an equivalent read depth and analyzed in parallel. **Panel A.** Average aligned read length from each technology, calculated from the CIGAR string after alignment with minimap2. **Panel B.** Aligned read lengths binned by reference transcript length, with the frequency of full-length reads in each bin indicated. **Panel C.** IGV view of 600 reads from each dataset aligning to the Beta-actin (ACTB) gene. Blue read color indicates genomic negative strand reads (corresponding to the Beta-actin sense strand), red color indicates unexpected genomic positive strand reads resulting from library preparation artifacts. Small colored marks indicate small indels or variants common in ONT-sequencing. Full-length is defined as reads that cover at least 90% of the expected mRNA length.

8 SMART-Seq mRNA LR performance with SIRV and ERCC mRNA reference standards

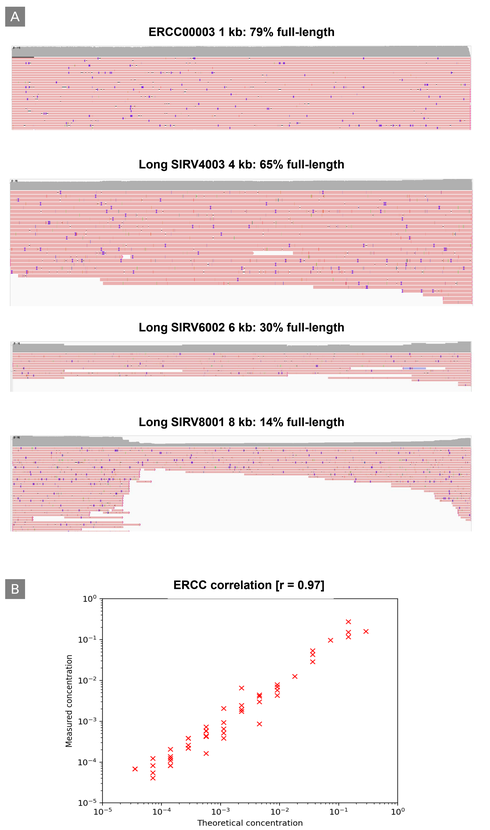

Figure 8. SMART-Seq mRNA LR performance with mRNA reference standards. To evaluate the performance of the SMART-Seq mRNA LR workflow, cDNA was generated from SIRV-Set 4 (Lexogen) mRNA reference standards which contains both ERCC quantification controls and Long SIRV mRNA standards. 10 ng of mouse brain total RNA was prepared—with SIRV spike-ins added to account for approximately 5% of reads—using the workflow described in Figure 1. Libraries were sequenced by ONT MinION, FASTQ data was read-strand corrected using the Restrander tool, and data were aligned with minimap2. **Panel A.** IGV plots of data from 1 kb, 4 kb, 6 kb, and 8 kb long transcripts from the ERCC and long SIRV isoform set. Red colored reads indicate positive strand reads, and small purple or red marks indicate small indels/variations common in ONT-sequencing. Full-length is defined as reads that cover at least 90% of expected mRNA length. **Panel B.** Plot of ERCC standard abundance for measured vs. theoretical concentration.

Conclusions

- Our new SMART-Seq mRNA Long Read workflow generates high-quality barcoded cDNA from ultra-low inputs for sequencing on Oxford Nanopore Technologies devices.
- SMART-Seq mRNA Long Read empowers multiplexed single-tube Oxford Nanopore Technologies library preparation.
- SMART-Seq mRNA Long Read enables reproducible full-length sequencing of long cDNAs.
- SMART-Seq mRNA Long Read is compatible with single-cell inputs and outperforms competing single-cell technologies.

Download poster & learn about SMART-Seq mRNA Long Read
takarabio.com/mRNA_LR

800.662.2566
www.takarabio.com