RNA isoform discovery and quantification with SMART-Seq® mRNA Long Read (v1 and v2) kits

Yue Yun[1], Jackson Peterson[1], Lisa Welter[1], Kazuo Tori[1], Alan Du[1], Yana Ryan[1], Shiyi Yin[1], Mike Covington[1], Shuwen Chen[1], Elena Shagisultanova[2], Mohammad Fallahi[1], Bryan Bell[1], and Andrew Farmer[1]*

[1]Takara Bio USA, Inc., 2560 Orchard Pkwy, San Jose, CA 95131 USA
[2]Department of Medicine/Medical Oncology, University of Colorado, Anschutz Medical Campus, Aurora, Colorado, USA
*Corresponding author

ABSTRACT

RNA serves a central role in biology by converting genomic information into effector molecules, either as functional non-coding RNAs or as protein-coding mRNAs. While it has long been appreciated that complex RNA transcript profiles can be produced through alternative splicing, numerous discoveries have highlighted the essential role of alternative splicing in cell and developmental biology. Furthermore, aberrant splicing has recently been linked to diseases like cancer, neurodegeneration, and autoimmunity. Thus, understanding the expression and function of RNA isoforms has become increasingly important in basic and applied molecular biology research.

A current limitation of long-read RNA sequencing (LR-RNA-seq) is the requirement of large amounts of input RNA, which can be unachievable for samples such as resected tumors or sorted single cells. Here, we describe SMART-Seq mRNA Long Read kit, a new LR-RNA-seq library preparation technology that enables full-length RNA sequencing from single cells (~10 pg RNA/cell) up to 2 µg total RNA. A comparison study at the single-cell level shows that SMART-Seq mRNA Long Read technology substantially outperforms existing single-cell LR-RNA-seq methods. With high-quality bulk RNA inputs, we demonstrate the ability to reliably sequence at an average read length (N50) of 2 kb and to detect full-length transcripts as long as 8 kb. With the ability to process up to 96 samples at a time, this technology will accelerate discovery as scientists catalog and study splice isoforms in rare or valuable cell samples.

1 SMART-Seq mRNA Long Read (LR) workflow

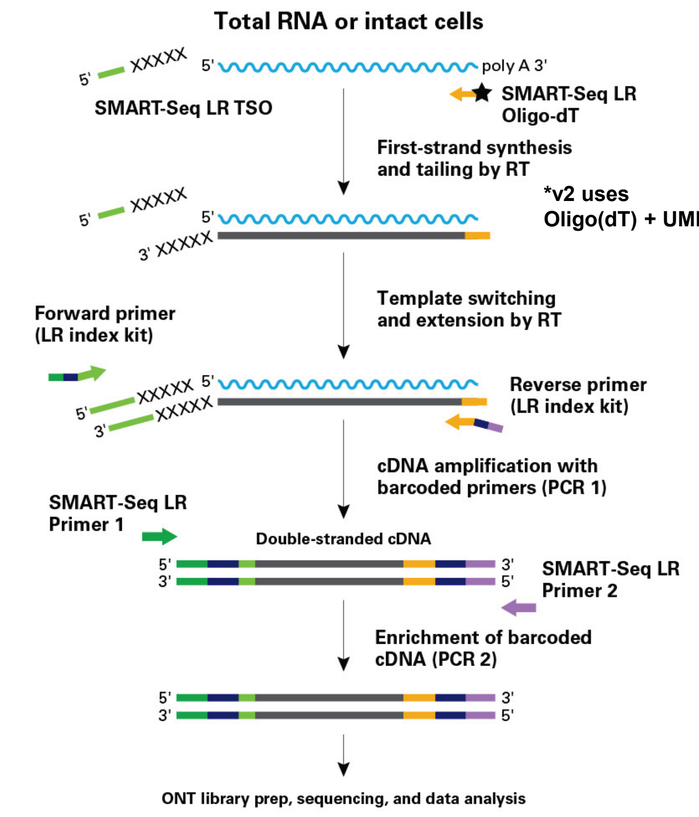


Figure 1. Library preparation workflow for the SMART-Seq mRNA Long Read (LR) kit. First-strand cDNA synthesis is primed by the SMART-Seq LR Primer (with [v2] or without [v1] UMIs) and performed by a Moloney murine leukemia virus (MMLV)-derived reverse transcriptase (RT). Upon reaching the 5' end of each 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, LSK114) followed by sequencing by MinIon or PromethIon flow cells. After sequencing, samples were basecalled and demultiplexed using Guppy or Dorado (Oxford Nanopore Technologies). Downstream analysis is performed using FLAIR for analyses involving novel isoforms or with minimap2, SAMtools, and Salmon to analyze known isoforms.

2 SMART-Seq mRNA LR kit shows high sensitivity across single cells and a broad range of total RNA inputs

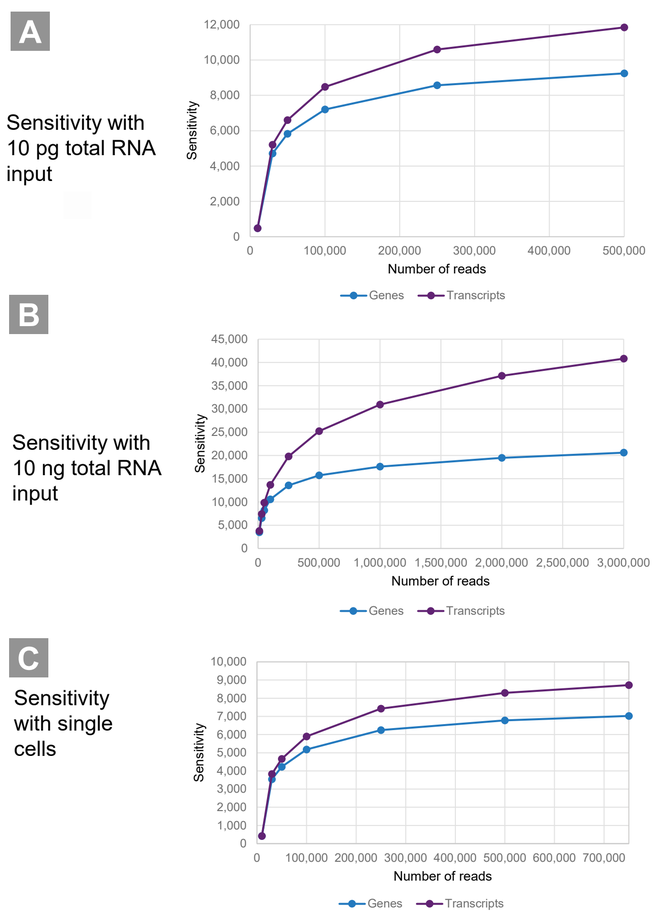


Figure 2. SMART-Seq mRNA Long Read demonstrates high sensitivity across a broad range of RNA inputs. To evaluate the performance of the SMART-Seq mRNA LR workflow, cDNA was generated from 10 pg or 10 ng total mouse brain RNA (MBR) or single K562 cells using the workflow described in Figure 1. After sequencing, data were 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.

3 SMART-Seq mRNA LR kit performance with long SIRV reference standards

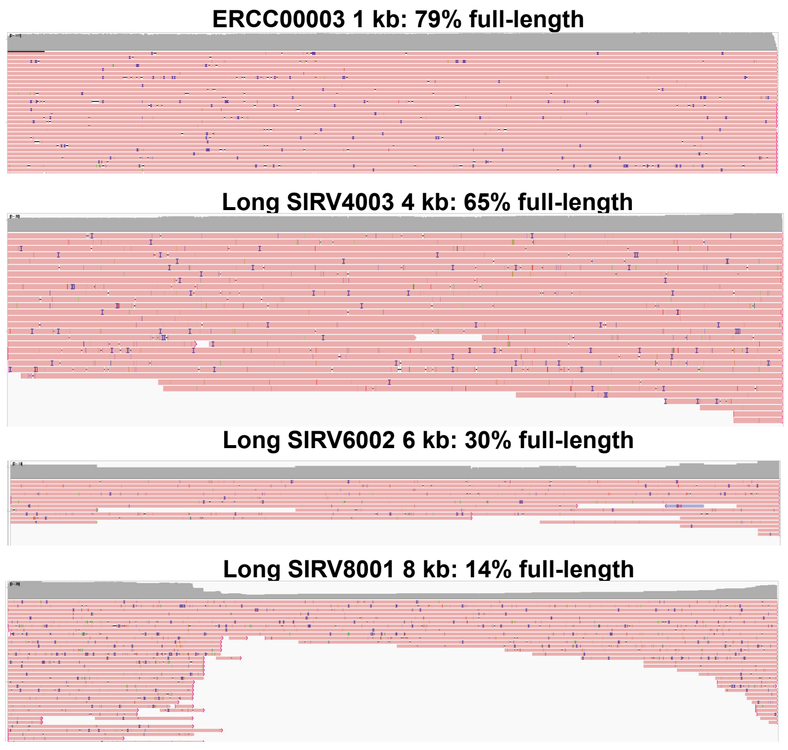


Figure 3. SMART-Seq mRNA LR kit performance with reference standards. To evaluate the performance of the SMART-Seq mRNA LR kit, cDNA was generated from SIRV-Set 4 (Lexogen) mRNA reference standards, which contains both ERCC quantification controls and long SIRV mRNA standards. 10 ng of mouse brain total RNA was prepared with SIRV spike-ins added to account for approximately 5% of reads. IGV plots show data from 1, 4, 6, and 8 kb long transcripts from the ERCC and long SIRV isoform set. Red-colored reads indicate positive-strand reads; blue indicates very rare negative-strand reads; 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.

4 SMART-Seq mRNA LR technology identifies novel isoforms for biomarker discovery

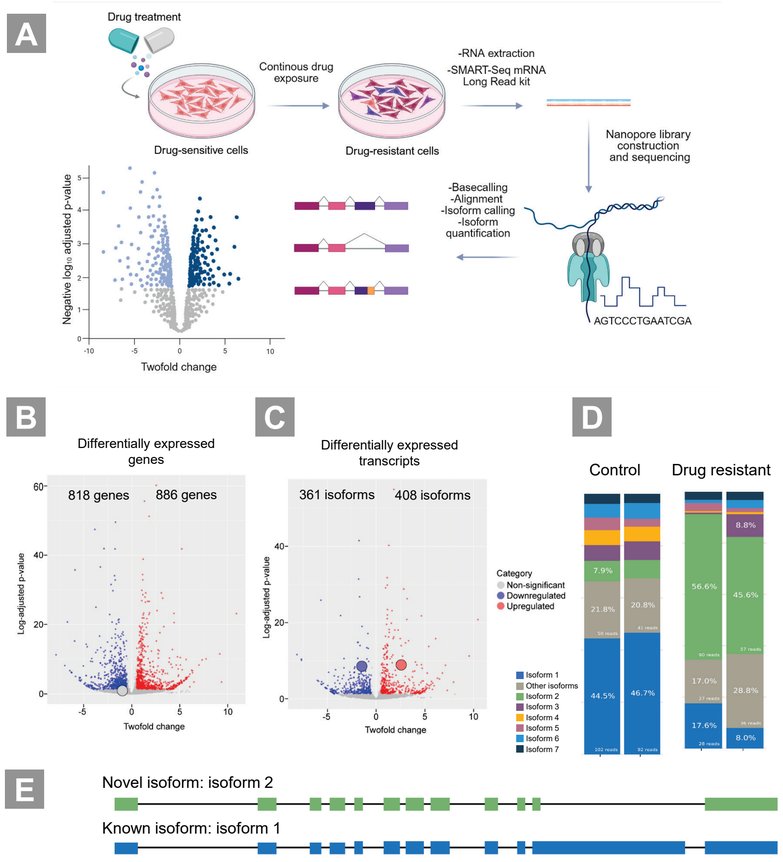
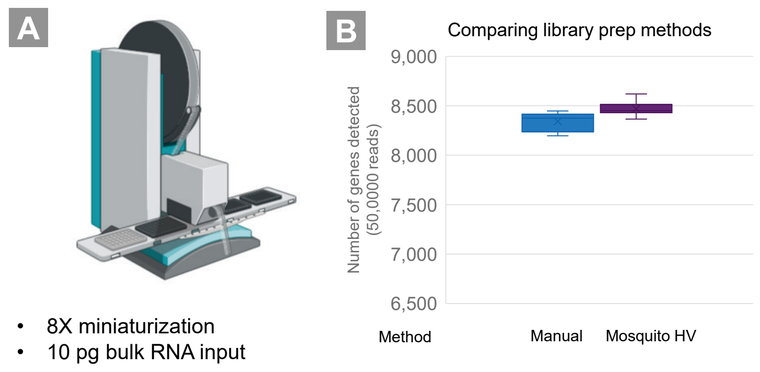


Figure 4. SMART-Seq mRNA LR kit identifies novel isoforms. Panel A. Experimental design for project performed as a collaboration between Takara Bio USA, Inc. and academia. Cell lines were cultured long-term in 1) the presence of an anticancer therapy, or 2) in a control condition with DMSO, to generate lines resistant to the anticancer therapy. cDNA libraries were prepared using SMART-Seq mRNA LR kit from both cultured conditions' total RNA and sequenced (PromethION). For each condition, two biological replicates (separate cell cultures) were sequenced by two technical replicates (distinct libraries). All four replicates were used in isoform and gene expression analyses. **Panels B & C.** Volcano plots showing differentially expressed genes (**Panel B**) and transcripts (**Panel C**). (Counts were tabulated with featureCounts or FLAIR, respectively, and differential expression was performed with Cogent™ Discovery Software. A gene of interest is highlighted with larger circles in **Panels B & C** and further elaborated on in **Panels D & E**. **Panel B.** Differential gene expression found 818 downregulated and 886 upregulated genes in the resistant clones. **Panel C.** Differential isoform expression analysis identified 361 isoforms and 408 isoforms differentially expressed. **Panel D.** Isoform usage of gene of interest (FLAIR). One of the two technical replicates is shown for each sample. The gray "Other isoforms" box indicates isoforms that were detected at an abundance below six individual reads. **Panel E**. Diagrams of green and blue isoforms from **Panel D**. Based on the data presented in Figure 4, it is clear that the SMART-Seq mRNA LR technology identifies changes in transcript expression, even when gene expression remains the same. Since literature supports the role of mRNA transcripts in therapeutic resistance, identifying novel isoforms will aid in better understanding resistance mechanisms.

5 SMART-Seq mRNA LR technology is compatible with automation for high-throughput library preparation

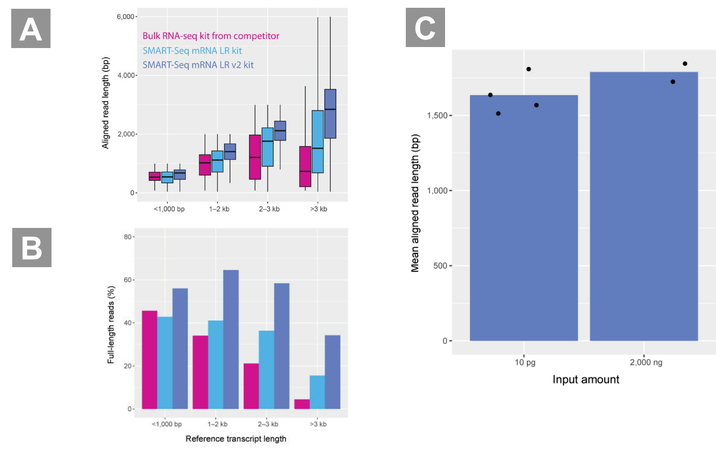


C

Method	Mean size of read	Median size of read (bp)	Mapped reads (%)	Uniquely mapped reads (%)
Manual	1,509.5	1,209.6	97.4	92.8
mosquito HV	1,480.9	1,262.9	98.8	94.9

Figure 5. SMART-Seq mRNA LR technology is compatible with automation and miniaturization. The SMART-Seq mRNA LR workflow was used to create cDNA from 10 ng of mouse brain RNA. 96 replicates were prepared via two conditions: 1) manually with the full reaction volumes and 2) automated at 1/8X on a mosquito HV (SPT Labtech, **Panel A**). Barcoded cDNA of all 96 samples per experiment was pooled, and libraries were generated and sequenced (MinION). **Panel B.** Gene counts were assessed (Salmon) and indicate a similar performance between the two methods. **Panel C.** Mapping stats were calculated (samtools) and indicate a similar performance between the two methods.

6 Introducing the SMART-Seq mRNA LR v2 kit (in development) with improved performance



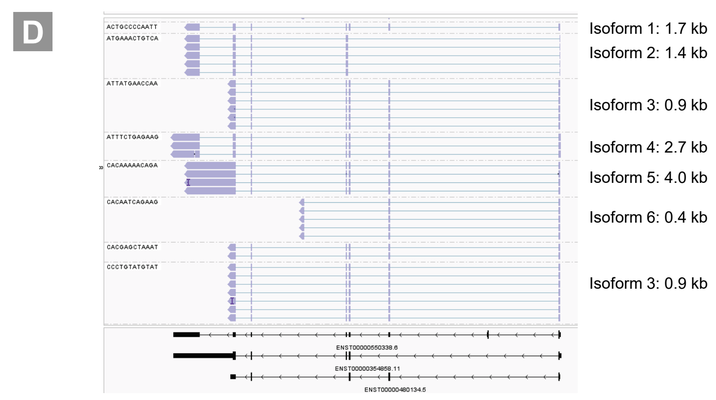


Figure 6. SMART-Seq mRNA LR v2 kit (in development) improves performance, expands input range, and includes UMIs. Here we disclose the development of SMART-Seq mRNA LR kit v2, a new improved version of the SMART-Seq mRNA LR kit. Performance of the v2 kit (darker blue) was assessed with 100 ng bulk mouse brain RNA (MBR) input and compared to the original (v1) kit (100 ng input, light blue) and a bulk RNA-seq kit from a competitor (500 ng input, pink). After sequencing, data were aligned (minimap2) to the mouse transcriptome. Using custom Python tools, the length of the reference transcript and the aligned read length of each read were extracted; the resulting data matrices were analyzed (R). **Panel A**. Box and whiskers plots indicate the median, IQR, and 1.5X IQR of the read lengths within each bin. Data shown from one representative sample of three bulk RNA-seq datasets. Read lengths were longer than the overall read lengths from the competitor by 42% (v1) or 129% (v2). **Panel B**. Frequency of full-length reads, defined as reads with aligned length of at least 90% of the length of the reference transcript. **Panel C.** Mean aligned read lengths from 10 pg and 2,000 ng bulk MBR samples. The points represent technical replicates. **Panel D.** Reads visualized (IGV) after UMI analysis and grouping of reads by UMI tag. Small colored marks indicate indels above a five bp threshold.

7 SMART-Seq mRNA LR technology empowers full-length isoform analysis in single cells

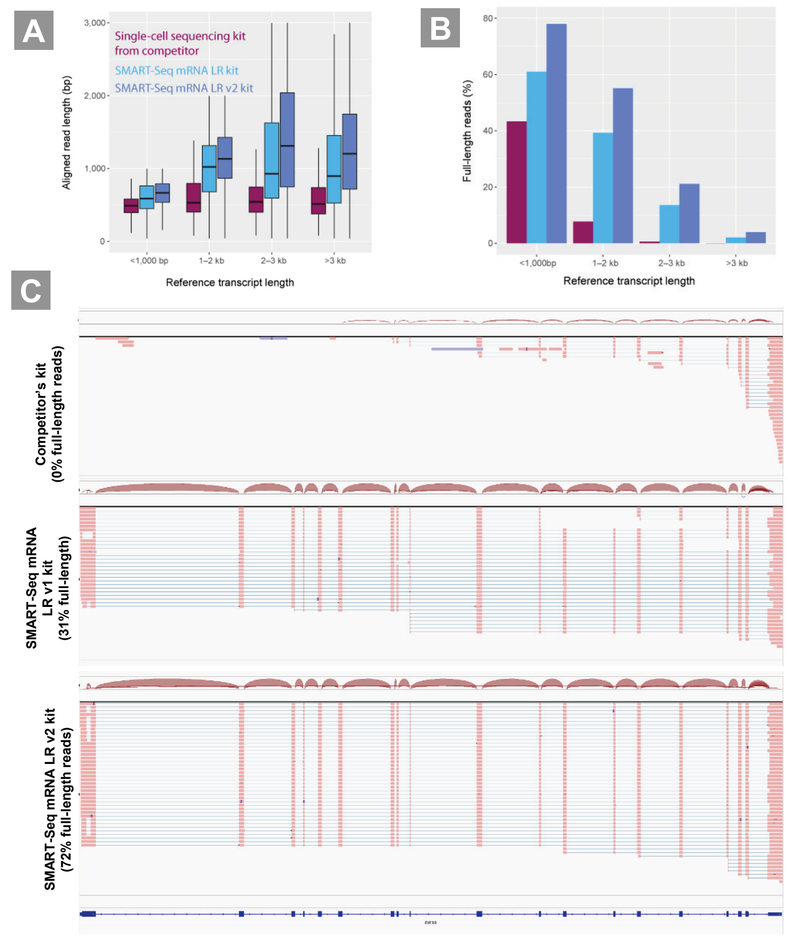


Figure 7. Performance of SMART-Seq mRNA LR kit versus the major competitor's protocol using single-cell inputs. Panel A. Single-cell libraries were generated from K562 cells with SMART-Seq mRNA LR v1 kit (light blue), v2 kit (darker blue), or a major competitor's single-cell protocol (red), and prepared for single-cell Oxford Nanopore long-read sequencing. Datasets were downsampled to an equivalent read depth and analyzed in parallel. **Panel A.** Average aligned read length from each technology. Read lengths were longer than those of the competitor's by 60% (v1) and 90% (v2). Box and whiskers plots indicate the median, IQR, and 1.5X IQR of the read lengths. **Panel B.** Frequency of full-length reads, defined as reads with aligned length of at least 90% of the length of the reference transcript. **Panel C.** IGV view of 75 reads from each dataset aligning to the *EIF3B* gene. Small colored marks indicate indels above a five bp threshold.

CONCLUSIONS

- Our new SMART-Seq mRNA Long Read workflow generates high-quality barcoded cDNA from single cells and a wide range of bulk inputs for sequencing on ONT sequencers
- SMART-Seq mRNA Long Read kits outperform competing bulk and single-cell technologies in key read-length metrics
- SMART-Seq mRNA Long Read v2 kit (in development) will improve read-length performance, enable long-read mRNA seq library prep from an expanded input range, & use UMIs to allow grouping/de-duplication of reads

Download the poster and learn more about the SMART-Seq mRNA LR kits:
takarabio.com/mRNA_LR

800.662.2566
takarabio.com