

Full-length transcriptomics in cancer research: Unraveling isoform complexity with SMART-Seq®



Yue Yun¹, Jackson Peterson¹, Peng Xu¹, Lisa Welter¹, Kazuo Tori¹, Alan Du¹, Shaveta Goyal¹, Eleanor Ziarnik¹, Mike Covington¹, Shuwen Chen¹, Elena Shagisultanova², Mohammad Fallahi¹, Bryan Bell¹, and Andrew Farmer^{1*}

¹Takara Bio USA, Inc., 2560 Orchard Pkwy, San Jose, CA

²Department of Medicine/Medical Oncology, University of Colorado, Anschutz Medical Campus, Aurora, Colorado, USA

*Corresponding Author

ABSTRACT

Cancer progression and therapeutic resistance are frequently associated with aberrant mRNA isoforms resulting from alternative splicing events. Accurately characterizing the full repertoire of these transcripts is critical for identifying novel biomarkers and therapeutic targets. However, the complexity and low abundance of certain isoforms require technologies with both high sensitivity and full-length coverage.

To address this need, we developed the SMART-Seq® mRNA Long Read (SS mRNA LR) kit for full-length mRNA library preparation. SS mRNA LR combines the high sensitivity and full-length coverage of SMART-Seq chemistry with the ultra-long sequencing capability of Oxford Nanopore Technologies (ONT). When we applied SS mRNA LR to cancer cell lines with evolved resistance to targeted therapies, we identified hundreds of novel isoforms and differential isoform usage associated with the evolution of therapeutic resistance. The newest version of this kit in development, SMART-Seq mRNA Long Read version 2 (SS mRNA LR v2), incorporates unique molecular identifiers (UMIs) and robust optimization to achieve 100% longer read lengths and 50% higher sensitivity than the current leading long-read ONT cDNA preparation method, allowing for improved isoform discovery and transcript abundance studies on longer and lower-abundance targets.

In addition to long-read only approaches, short-read sequencing can support long-read isoform discovery and quantification by supplying more cost-effective read depths, resulting in more accurate splice junction characterization and quantification. We evaluated the impact of short-read support using the SMART-Seq total RNA approach and the in-development SMART-Seq mRNA Stranded approach. Integration of short-read splice-junction evidence further improved the detection and confidence of novel isoforms identified by SS mRNA LR. All three technologies have been optimized for a broader input range (10 pg–2 µg) than traditional SMART-Seq low-input assays and expand SMART-Seq capabilities beyond ultra low-input profiling to establish a unified ecosystem for high-resolution splice isoform discovery.

1 SMART-Seq mRNA Long Read technology

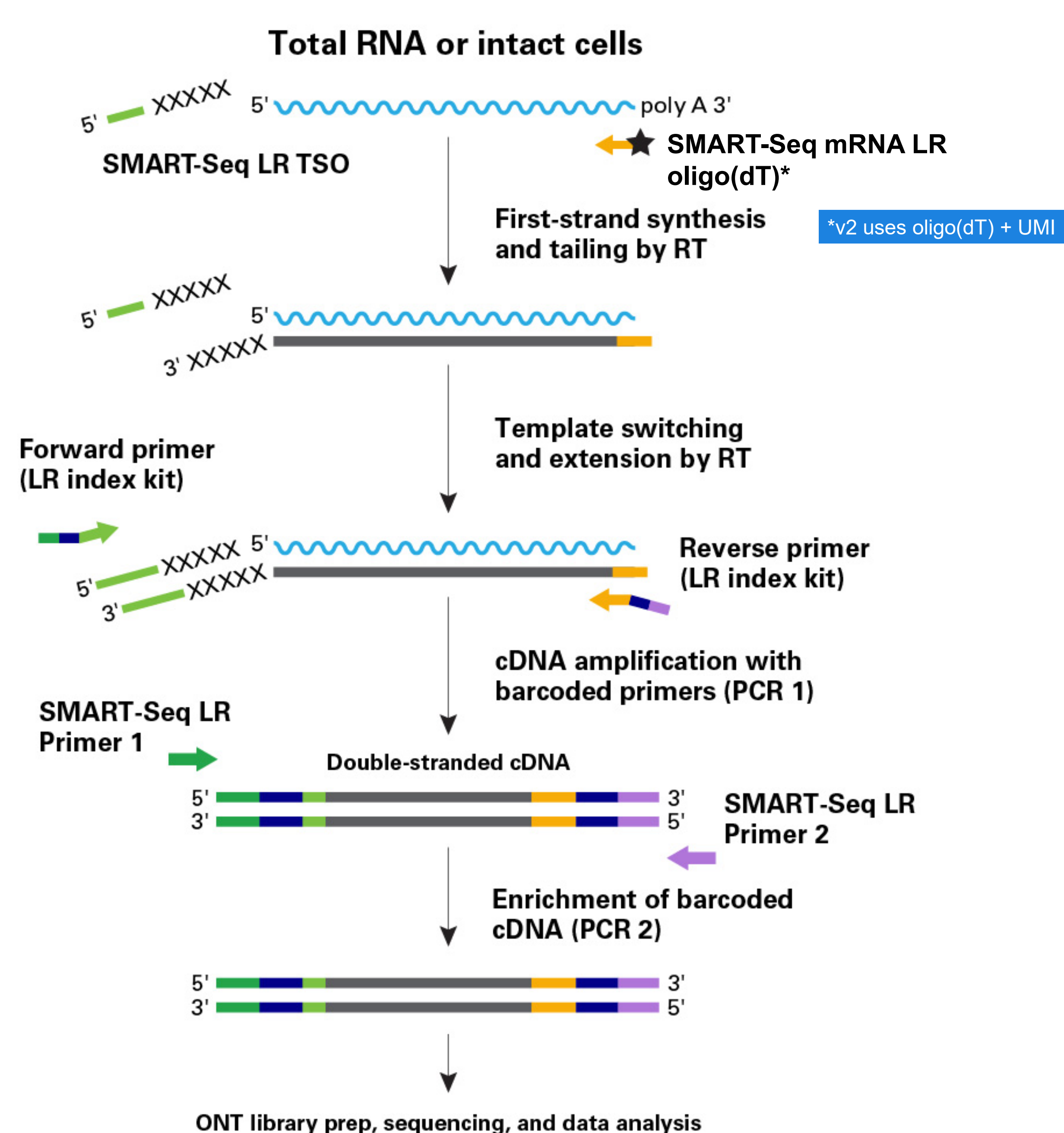


Figure 1. Library preparation workflow for the SMART-Seq mRNA Long Read (LR) kit. First-strand cDNA synthesis is primed by the SMART-Seq LR Primer (with UMIs [v2] or without UMIs [v1]), and performed by a reverse transcriptase (RT). Upon reaching the 5' end of each mRNA molecule, the RT adds non-templated 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 (ONT, LSK114) followed by sequencing by MinION or PromethION flow cells and basecalled with MinKnow sequencing software (ONT). Sequencing data were demultiplexed with Dorado (ONT) or with Cogent™ NGS Analysis Pipeline (CogentAP, Takara Bio). Downstream analysis was performed with CogentAP for alignment and gene counts, with isoform discovery and quantification analysis performed with FLAIR 3.0 (Tang et al. 2020).

2 SMART-Seq mRNA Long Read provides leading read-length performance

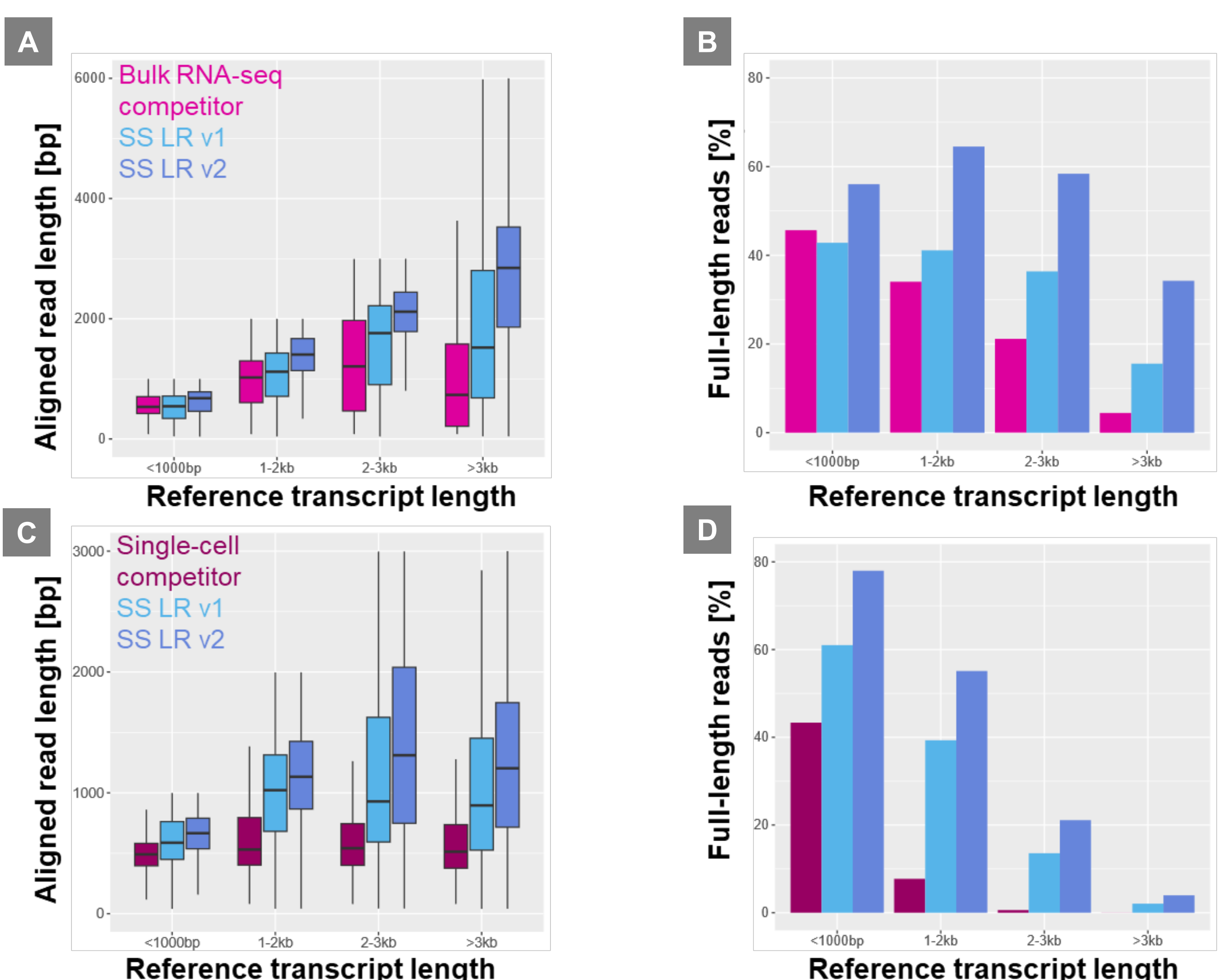


Figure 2. SMART-Seq mRNA Long Read technology demonstrates exceptional read-length performance. Sequencing data from bulk or single-cell libraries with SS mRNA LR ("SS LR v1"), SS mRNA LR v2 ("SS LR v2"), or competitor technologies were aligned to reference transcriptomes with minimap2. Aligned read length and the length of the reference transcript from each read was extracted, and resulting data matrices were analyzed in R. **Panel A–B.** Performance with bulk mouse brain RNA input was assessed for the SS mRNA LR kit (100 ng RNA input), SS mRNA LR v2 kit (100 ng RNA input), and the competitor bulk RNA-seq kit (500 ng RNA input). Displayed data are from one representative sample of three bulk RNA-seq datasets. **Panel C–D.** Single-cell libraries were generated from K562 cells and prepared for single-cell ONT long-read sequencing with SS LR v1, SS LR v2, or using the competitor's single-cell protocol. Datasets were downsampled to an equivalent read depth and analyzed in parallel. Box-and-whisker plot (**Panel A, C**) of aligned read lengths, with boxes and whiskers indicating the median, interquartile range (IQR), and 1.5x IQR of the read lengths within each bin. Bar graph (**Panel B, D**) showing frequency of full-length reads, defined as reads with aligned length of at least 90% of the length of the reference transcript.

3 Isoform discovery with SMART-Seq mRNA Long Read

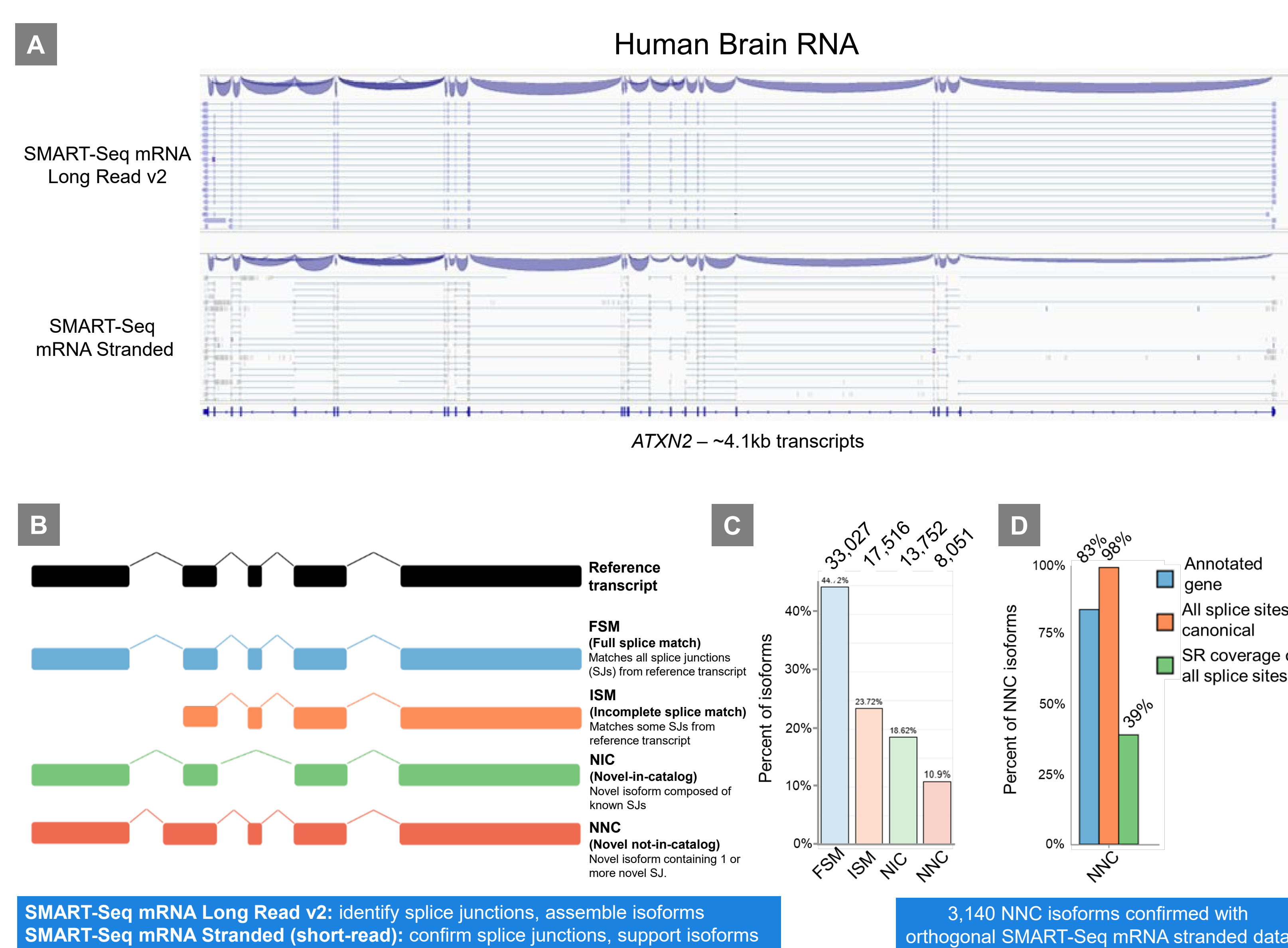


Figure 3. SMART-Seq mRNA technology enabled discovery of novel RNA isoforms. Libraries were produced from bulk human brain RNA using the in-development SMART-Seq mRNA Stranded and SS mRNA LR v2 kits. Short-read (SR) and long-read datasets were analyzed with CogentAP. Isoform discovery analysis was performed using long read data, with confirmatory support from short read data. Aligned BAM long-read data files were processed for transcript discovery using Flair. Isoforms identified with Flair were assessed with Sqaunt3 QC alongside SJ.out.tab files from short-read data generated with the SMART-Seq mRNA Stranded approach to produce QC reports describing transcript properties of novel isoforms. Sqaunt3 analysis with most up-to-date Gencodev49 reference annotation ensured detection of truly novel isoforms. **Panel A.** Splice junctions (SJs) and sequencing reads visualized in Integrative Genomics Viewer (IGV) long-read (top) and short-read (bottom) datasets. **Panel B.** Diagram of Sqaunt3 transcript model classifications. Transcript classification image was derived from the Sqaunt3 Github. **Panel C.** Bar chart displaying the frequency of full-splice match (FSM), incomplete splice-match (ISM), novel in catalog (NIC), and novel not-in-catalog (NNC) transcripts derived from Flair analysis. **Panel D.** Bar chart displaying the frequency of novel-not-in-catalog transcripts associated with known genes (blue), containing only canonical splice junctions (orange), and with all splice junctions fully supported by the splice junction data from SMART-Seq mRNA Stranded short-read sequencing (green). SMART-Seq mRNA Long Read version 2 sequencing and transcript discovery analysis uncovered 21,303 total novel isoforms (13,752 NIC and 8,051 NNC); additional analysis with SMART-Seq mRNA Stranded short-read data provided support for 3,140 of the NNC isoforms.

4 Cancer therapy resistance study

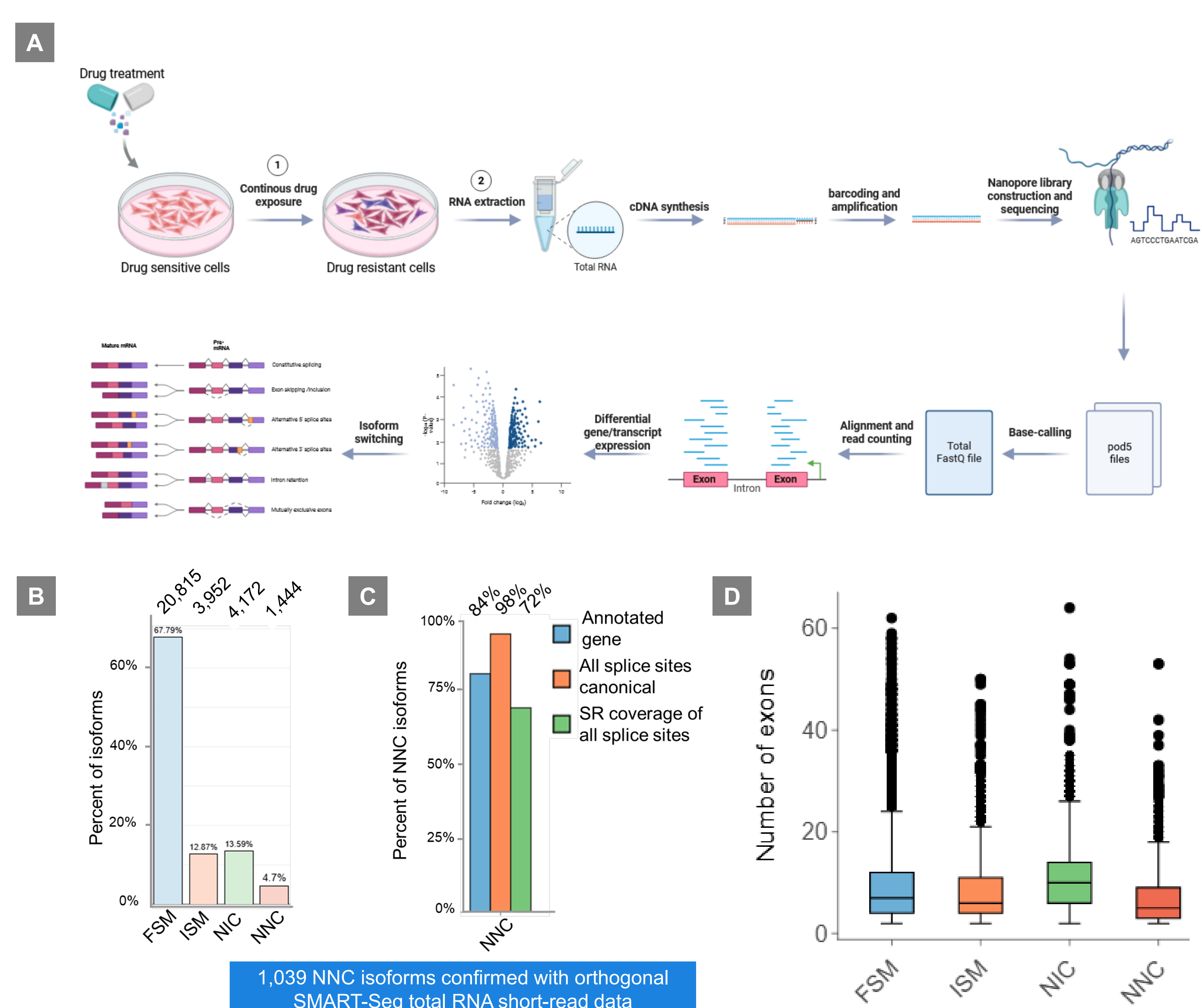


Figure 4. Novel splice junctions in cancer therapy-resistant cells were identified using SS mRNA LR and confirmed with supplementary short-read sequencing. **Panel A.** Experimental design for collaborative study using long-read technology. MDA-MB-361 breast cancer cell lines were treated long-term with 1) an anticancer therapy to generate therapy-resistant cells or 2) DMSO as a control condition. cDNA libraries were prepared from total RNA of both cultured conditions using SS mRNA LR or SMART-Seq Total RNA Library Prep with ZapR® Depletion 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. Analysis was done with CogentAP and FLAIR. Sqaunt3 QC reports were produced as in Figure 3 with supporting short-read data from the SMART-Seq total RNA assay. **Panel B–C.** Bar charts displaying the frequency of transcript classes (**Panel B**) and short-read splice junction support of novel-not-in-catalog transcripts (**Panel C**), as in Figure 3. **Panel D.** Box plot displaying the number of exons associated with assembled transcripts in each transcript category. Applying SS mRNA LR in this breast cancer therapy resistance study identified 5,616 total novel isoforms (4,172 NIC and 1,444 NNC), with 1,039 of the NNC isoforms supported with orthogonal SMART-Seq total RNA short-read data.

5 Differentially expressed genes and transcripts associated with therapeutic resistance in breast cancer cells

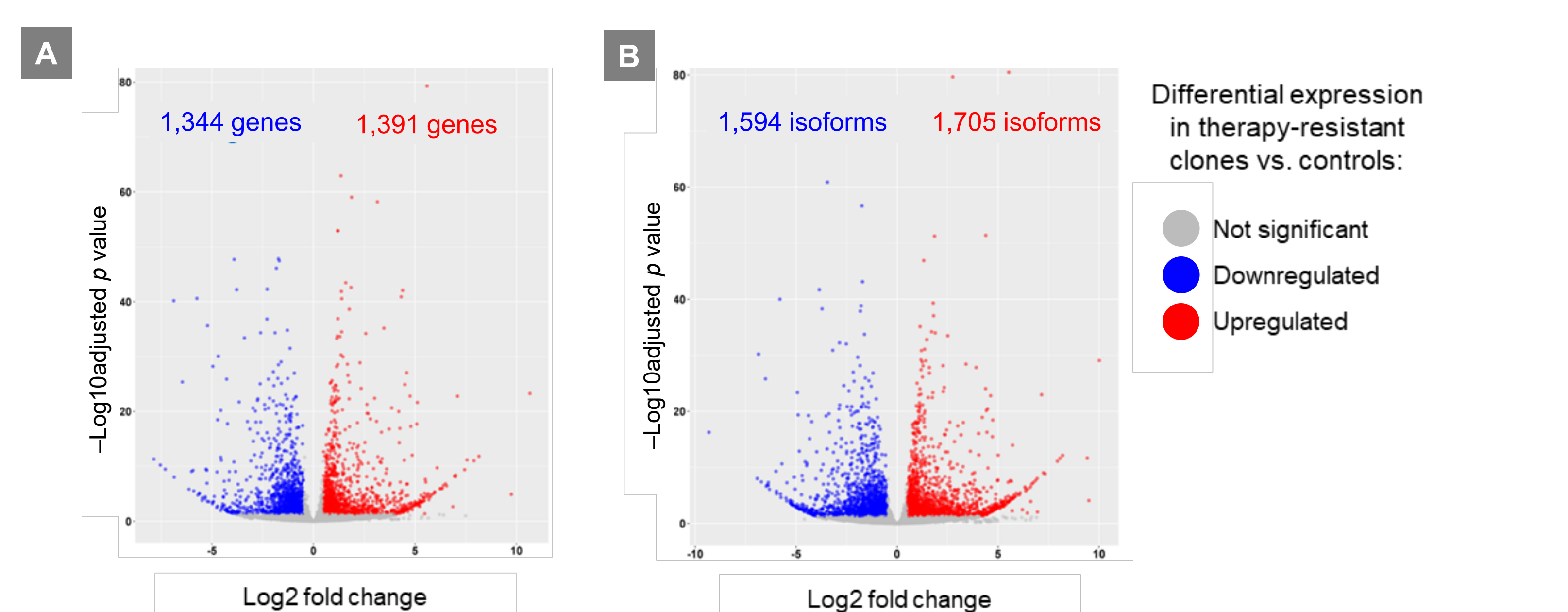


Figure 5. Long-read sequencing detected differentially expressed genes and transcripts associated with therapeutic resistance in breast cancer cells. Gene expression data (CogentAP, **Panel A**) and transcript expression data (FLAIR, **Panel B**) generated from SS mRNA LR were analyzed for differential expression; differential expression was visualized in volcano plots (CogentDS) using cutoffs of a log₂ fold change of 0.5 and an adjusted p value threshold of 0.05. SS mRNA LR enabled identification of 2,735 differentially expressed genes and 3,299 isoform transcripts associated with cancer therapy resistance.

6 Novel differentially expressed isoforms and isoform switching in breast cancer therapy resistance

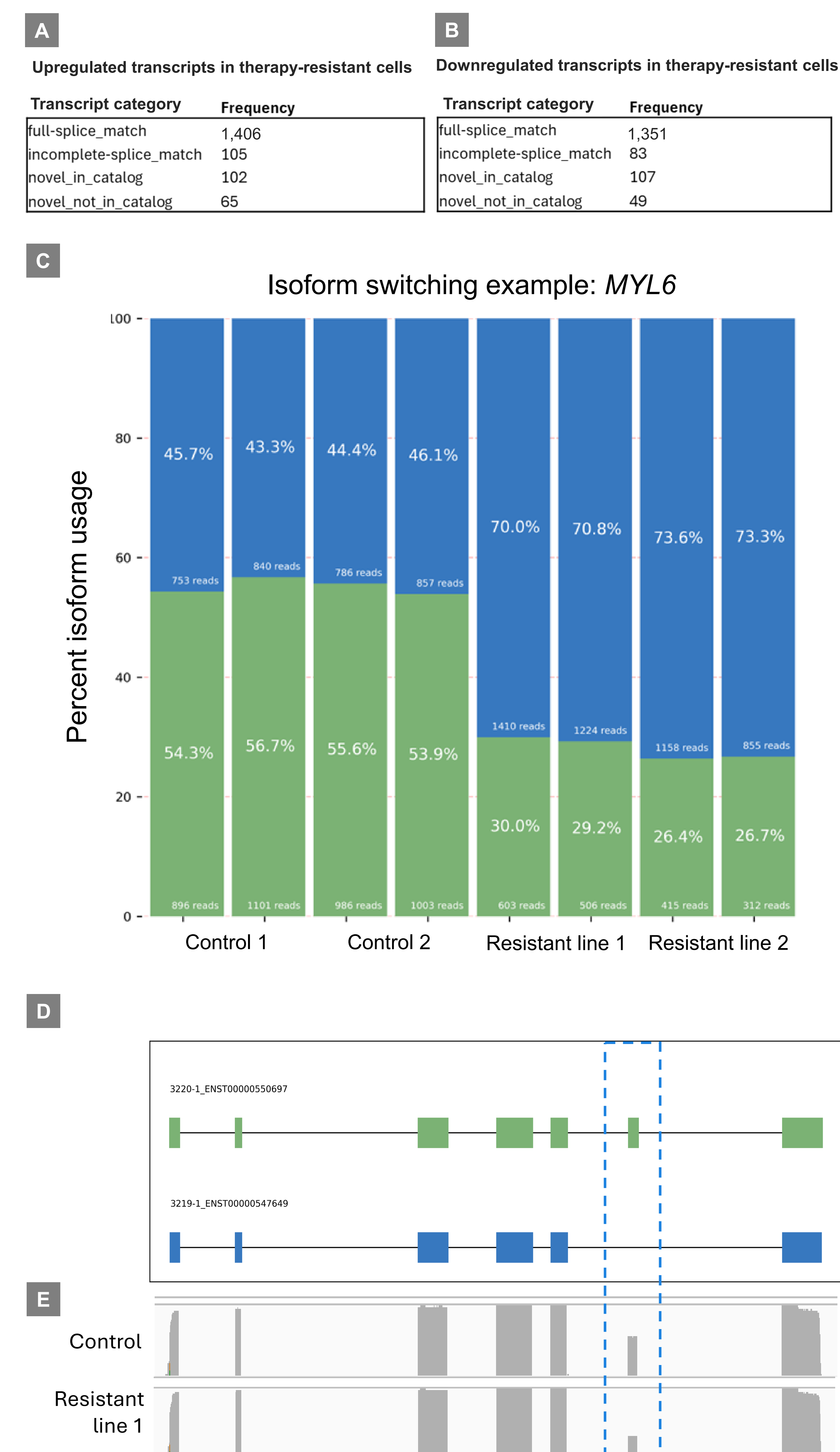


Figure 6. Long-read sequencing identified novel differentially expressed isoforms and isoform switching associated with breast cancer therapy resistance. **Panel A–B.** Frequency of differentially expressed transcripts in each major transcript category was derived from referencing the list of differentially expressed transcripts against the Sqaunt3 QC classifications. Green and blue colors represent two different FSM isoforms. **Panel C.** Isoform usage plot of MYL6 in control and therapy resistant samples (two technical replicates per sample). **Panel D.** Isoform structures of MYL6. Blue and green colors in Panel C correspond to isoforms in Panel D. **Panel E.** IGV coverage views of MYL6. Higher coverage in exon 6 (see blue box) of control samples reflects the isoform usage difference displayed in Panels C–D. SS mRNA LR identified novel isoforms associated with therapy resistance and differential transcript usage between control and resistant lines.

CONCLUSIONS

- SS mRNA LR technology generates high-quality barcoded cDNA from single cells and a wide range of bulk inputs for use with long-read sequencers.
- SS mRNA LR technology outperforms competing bulk and single-cell technologies in key read-length metrics.
- SS mRNA LR enables the discovery and quantification of novel isoforms. Read support with short-read data from SMART-Seq mRNA Stranded and SMART-Seq Total RNA Library Prep with ZapR Depletion technologies confirms novel splice junctions, supporting identification of novel-not-in-catalog isoforms.
- SS mRNA LR helped identify differential novel isoforms and differential isoform usage associated with breast cancer therapy resistance.

References

Tang, A. D. et al. Full-length transcript characterization of *SF3B1* mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat. Commun.* **11**, 1438 (2020).



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

800.662.2566
takarabio.com

Takara Bio USA, Inc. United States/Canada: +1 800 662 2566 • Asia Pacific: +1 650 910 7300 • Europe: +33 (0)1 3004 8880 • Japan: +81 (0)77 665 8999
FOR RESEARCH USE ONLY. NOT FOR USE IN DIAGNOSTIC PROCEDURES. © 2020 Takara Bio Inc. All Rights Reserved. All trademarks are the property of Takara Bio Inc. or its affiliates in the U.S. and/or other countries or their respective owners. Certain trademarks may not be registered in all jurisdictions. Additional product, intellectual property, and restricted use information is available at takarabio.com