High-throughput single-cell RNA-, DNA-, and ATAC-seq

Alain Mir, Ishminder Mann, Maggie Bostick, Marta Gonzalez, Kazuo Tori, Sangeetha Anandakrishnan, Christopher Hardy, Karthikeyan Swaminathan, and Andrew Farmer*

Takara Bio USA, Inc., Mountain View, CA 94043, USA

*Corresponding author: andrew_farmer@takarabio.com

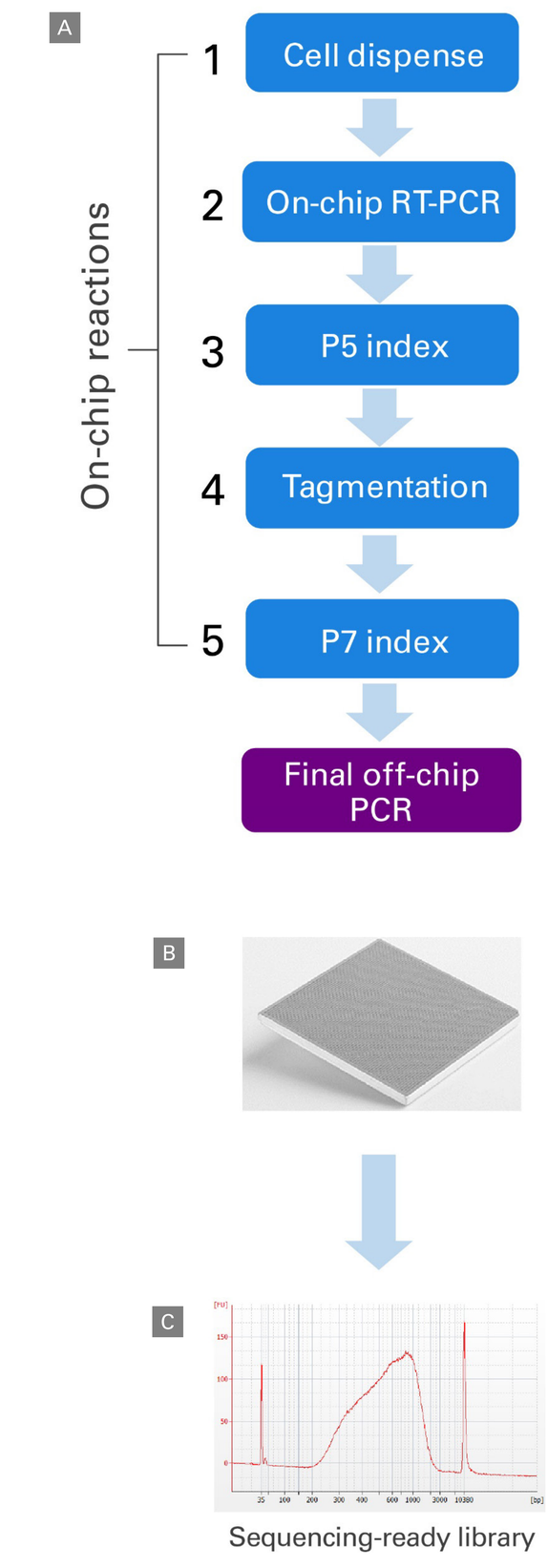
Abstract

The desire to examine single cells and resolve subpopulations such as those associated with drug resistance and cell renewal tracks alongside continual improvements for isolating single cells. Automation systems that resolve single cells regardless of cell type or size, coupled with high-quality reagents permitting robust NGS sample preparation, decreased batch effects, and improved reproducibility, continue to be desired. The SMARTer™ ICELL8® Single-Cell System meets those demands.

The SMARTer ICELL8 system is an automated platform specializing in single-cell isolation using high-quality NGS reagent kits. This platform utilizes simple Poisson dilution to segregate cells into singlets and resolve individual wells (5,184) bearing single cells using automated microscopy. The method is enabled via a large aperture dispenser (~125 microns) that allows users to rapidly dispense and characterize up to eight different samples while selectively delivering reagents to specific wells. A total of >1,700 single cells can be addressed per chip. All wells can be visually interrogated in up to three separate spectral channels (Hoechst, FITC, and propidium iodide), permitting live-dead and rare-cell measurements in individual wells.

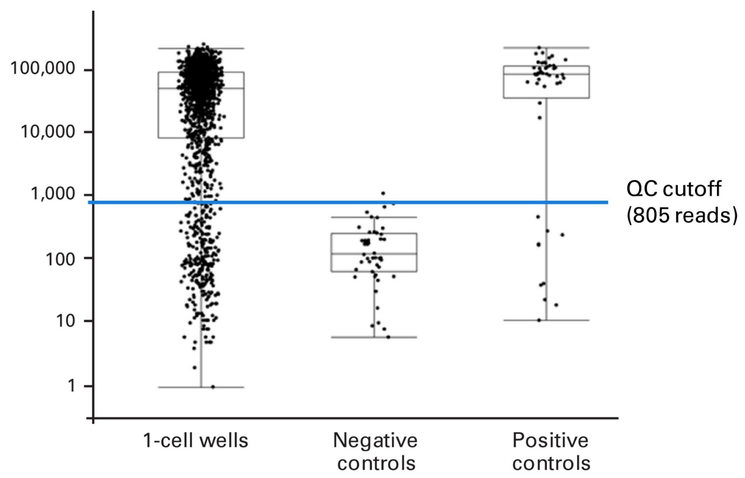
A key benefit of the SMARTer ICELL8 system is its open design, enabling rapid assay development for single-cell interrogation. For RNA analysis, we have developed 3′ mRNA end-capture kits and a kit that allows combined high-throughput T-cell receptor clonotype and 5′-end transcriptome profiling. Additionally, we developed a SMART-Seq® method for full-length RNA-seq that offers benefits in increased gene body coverage—results that will be described in detail in this poster. For DNA analysis, a protocol for ATAC-seq is now available, and a method for high-throughput CNV analysis is under development. The outputs from these methods are Illumina®-ready libraries. The SMARTer ICELL8 system's flexibility, combined with the availability of robust reagent sets, commends its use in a wide variety of single-cell analyses requiring both high sensitivity and reproducibility.

1 SMARTer ICELL8 SMART-Seq workflow



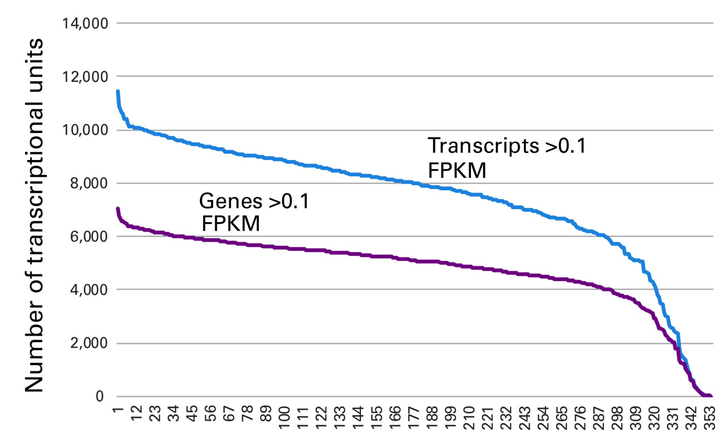
SMARTer® ICELL8 SMART-Seq workflow: full-length scRNA-seq on the SMARTer ICELL8 system. Panel A. This 8-hr protocol requires five dispensing steps. Cultured cells are dispensed into the wells of a SMARTer ICELL8 blank chip at an average of 1 cell/well (Step 1). CellSelect® Software is used to identify single cells. Cell lysis is followed by cDNA synthesis and amplification (Step 2). Full-length cDNA is tagmented with Illumina Nextera® Tn5 and amplified with Illumina-specific indexed adapters added in a grid-like fashion to generate uniquely indexed cDNA libraries in each well (Steps 3–5). The final libraries are pooled, further amplified, and purified prior to sequencing. **Panel B.** The cells, reagents, indexes, and Tn5 are dispensed into the 5,184-well chip. 72x72 indexing of the barcodes across the chip allows each well to have a unique combination of barcodes. **Panel C.** A representative Bioanalyzer trace of a sequencing-ready library.

2 Tight well-to-well read count distribution



Tight well-to-well read count distribution. K562 cells were dispensed across a chip together with 48 negative control wells (containing only buffer) and 48 positive control wells (containing 10 pg K562 RNA). CellSelect Software was used to identify candidate single-cell wells, and a total of 1,488 wells were identified. These single-cell wells, together with the positive and negative controls, were then processed according to the full-length RNA-seq (SMART-Seq) protocol outlined in Figure 1, and the resulting libraries were sequenced on an Illumina NextSeq® system to a median read depth of 58,000 reads/barcode. We chose a cutoff of 3 SD above the mean read count for the negative control wells to identify wells for deeper analysis. Of 1,488 cells analyzed, 1,224 (82%) passed this threshold and 89% of cell-containing wells had >10,000 reads per barcode.

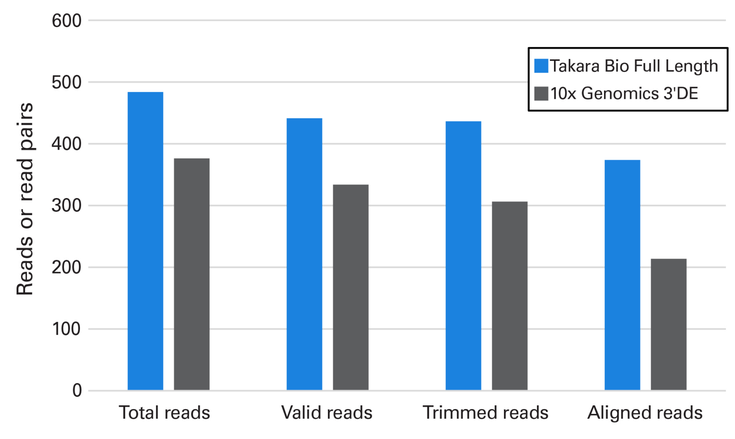
3 Highly complex sequencing libraries



Number of K562 single cells	366
Median read depth	234,304
rRNA	3.6%
Mitochondria	2.9%
Genome	92.8%
Exons	90.3%
Introns	6.3%
Intergenic	3.5%
Number of transcripts identified with >0.1 FPKM (median)	7,916
Number of genes identified with >0.1 FPKM (median)	5,066

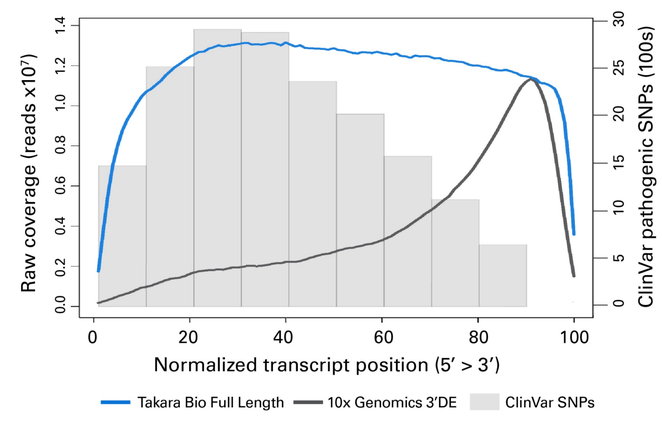
Sequencing libraries are highly complex. In a separate experiment, approximately 400 K562 cells were sequenced at higher sequencing depth (~230K reads/well). After applying a cutoff of 7,700 reads/cell, the reads from 366 cells were trimmed for poly(A), SMART® adapter, and Illumina adapters and subsequently mapped to hg19 with CLC, version 9.5.3. At this higher sequencing depth, a median of >5,000 genes (approx. 8,000 transcripts) was revealed. Other metrics are consistent with a highly specific amplification, with a low percentage of reads mapping to rRNA (3.6%) and mitochondrial (2.9%) regions and a high percentage of reads mapping to exons.

4 Transcriptome analysis on SMARTer ICELL8 and 10x Chromium systems



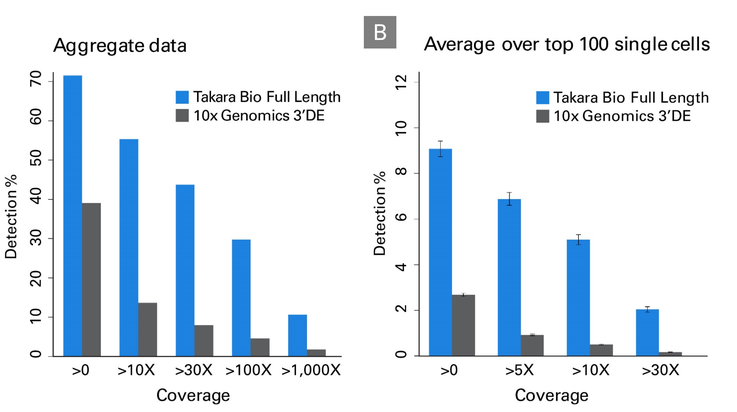
Comparison study with 10x Genomics 3′DE data. Processing of 1,000 HEK293 cells was performed using the 10x Chromium system with the Chromium Single Cell 3′ Library & Gel Bead Kit v2 ("10x Genomics 3′DE") or the SMARTer ICELL8 SMART-Seq workflow ("Takara Bio Full Length"). Both libraries were sequenced on a NextSeq high output cartridge, based on recommended loading and cycling conditions (10x library: 1.8 pM input, 1x26 and 1x76 cycles, and both indexes; Takara Bio library: 1.6 pM input, 2x75 cycles, and both indexes). The histogram illustrates the number of reads at each stage of the initial mapping process. Total reads are raw reads off of the sequencer; valid reads are reads that contain valid barcodes; trimmed reads are reads that remain after trimming for quality, poly(A) sequences, and adapter sequences; and aligned reads are reads that align to the human genome using STAR, version 2.5.2b. The Takara Bio Full Length library maintained more of the reads than the 10x Genomics 3′DE library throughout the analysis process; between this and the increased base pairs from the 2x75 reads, the Takara Bio library contained more coverage of the genome than the 10x library.

5 Transcriptome-based SNP coverage on SMARTer ICELL8 and 10x Chromium systems



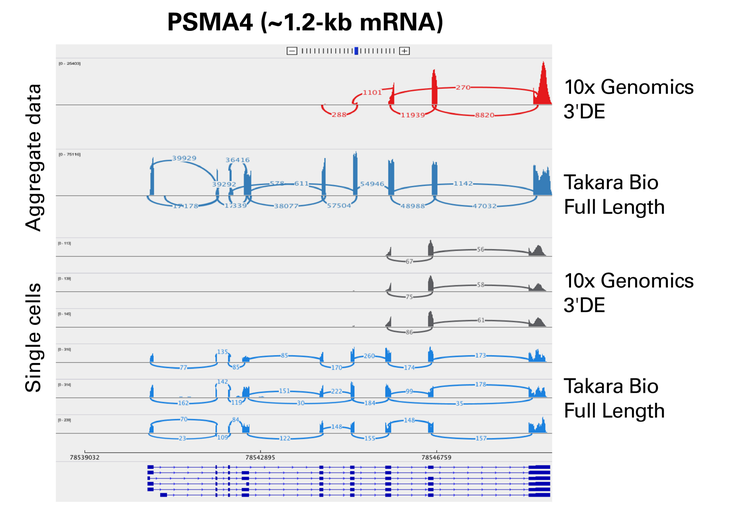
Distribution of sequencing reads and pathogenic SNPs annotated in ClinVar. Once the reads were mapped to the genome, the distribution of coverage across the annotated transcripts was determined using the `geneBody_coverage.py` script from RSeQC and the `hg38.RefSeq.union.nochr.bed` file. In this chart, the coverage is shown as raw reads, highlighting the overall difference in coverage between the 10x Genomics 3′DE data (black line) and the Takara Bio Full Length data (blue line). After normalizing all transcript positions, it is clear that the 10x data's coverage is predominantly in the region spanning the sequences closest to the 3′ ends. In contrast, the Takara Bio method provides high coverage over the entire gene body. This additional coverage can theoretically cover the regions of the transcriptome that contain pathogenic SNPs identified in the ClinVar database (~18,000), which preferentially map to the 5′ ends of the genes (gray columns, right axis).

6 Increased read coverage of positions annotated as pathogenic SNPs in ClinVar



Read coverage of annotated SNPs in HEK293 data. Panel A. Aggregate data. By mapping the combined reads from all 1,000 HEK293 cells together for each kit and investigating the read coverage at the loci that contain potential pathogenic SNPs according to the ClinVar database, we observed increased coverage of the SNPs with the Takara Bio Full Length method as compared to the 10x method, especially when looking for high coverage (>30X). **Panel B.** Average over the top 100 single cells. By taking the 100 cells with the highest number of reads from each library and using individual mapping, we observed fewer SNPs with coverage (<10%), but the coverage is still higher for the Takara Bio Full Length protocol.

7 Superior detection of differential splicing



Full-length data enables detection of differential splicing. The same HEK293 libraries were analyzed for splicing variation by generating Sashimi plots (IGV, version 2.4.10), using the Proteasome subunit alpha type-4 (PSMA4) locus (~1.2-kilobase mRNA) as an example. The aggregated data (all reads from each library) shows that multiple splicing variants are well supported in the Takara Bio Full Length data, but even with the short length of the mRNA, only the 3′ end of the gene was captured with the 10x data, and less splicing information was obtained. By zooming in to three single cells from each library, we can again see differences between the two libraries in their ability to capture splicing information from the 5′ end of the gene.

Conclusions

- The SMARTer ICELL8 SMART-Seq workflow provides full-length sequence information, high sensitivity, and high reproducibility.
- Data from Takara Bio's SMARTer ICELL8 SMART-Seq workflow outperforms 10x Genomics' 3′DE workflow in terms of gene body coverage, clustering, and read efficiency.
- The higher coverage obtained with Takara Bio's method results in greater read depth across positions annotated as pathogenic SNPs in ClinVar and ability to see splicing differences between cells—both in aggregate and at the single-cell level.

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

800.662.2566
Visit us at www.takarabio.com

Clontech TaKaRa cellartis