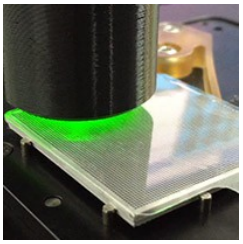
scientifix

PO Box 662 South Yarra
VIC 3141 Australia
www.scientifix.com.au

Free call: 1800 007 900
T: +61(0)3 8540 5900
F: +61(0)3 9543 7827
E: info@scientifix.com.au

Full-length scRNA-seq for better detection of gene fusions, SNPs, and alternative splicing



The use of next-generation sequencing for transcriptome analysis in clinical and applied spaces requires accurate, parallel processing of large numbers of single cells and the availability of chemistries that enable robust library preparation from the desired targets. Several automation platforms have emerged to meet this need, but variations in the underlying technology—such as microfluidics, droplet encapsulation, or multisample nanodispensing—influence how efficiently single cells are captured and processed. In addition to cell processing technology, the choice and sensitivity of the chemistry used for library prep—for example, 3'-end capture versus full-length, 5'→3' transcript capture—will yield different quality data to aid in answering specific questions about gene expression. Full-length capture, which provides more uniform coverage of the transcript, enables examination of gene fusions, SNP detection, and alternative splicing, whereas data on 3' differential expression (3'DE) mostly enables examination of gene expression regulation.

Our ICELL8 cx system, which utilizes a nanodispensing array, enables automated high-throughput processing of 1,000–2,000 single cells from a heterogeneous population per chip without the common microfluidic cell-size constraints or the imaging limitations of droplet-based systems. Advanced imaging technology driven by ICELL8 cx CellSelect Software allows the flexibility to choose exactly which wells to process or to fully automate this workflow—with both options being guided by cell staining to identify target cells of interest. This unique approach enables the user to trace sequencing data back to a particular well (containing a single cell, multinucleated cell, cluster of cells, or organoid). The application of our trusted, full-length SMART-seq chemistry on the ICELL8 cx system provides a high-throughput solution to obtaining richer data on single-cell transcriptomics. Along with high sensitivity in gene detection, it provides information related to splice variants, gene fusions, and mutations that are critical to a deeper understanding of cell biology.

In this technical note, we will compare results from the SMART-Seq ICELL8 application kit protocol, our automated method for full-length transcript capture (Figure 1), against results from a 3'DE method on a droplet-based system from 10x Genomics. We examine metrics such as gene body coverage, gene fusion and SNP detection, and alternative splicing identification. Although not a direct comparison of technologies, it is clear that the Takara Bio method provides full-length sequence information and a higher gene body coverage, resulting in improved detection of gene fusions, greater read depth across positions annotated as pathogenic SNPs in ClinVar and overall SNP detection, and ability to see splicing differences between cells.

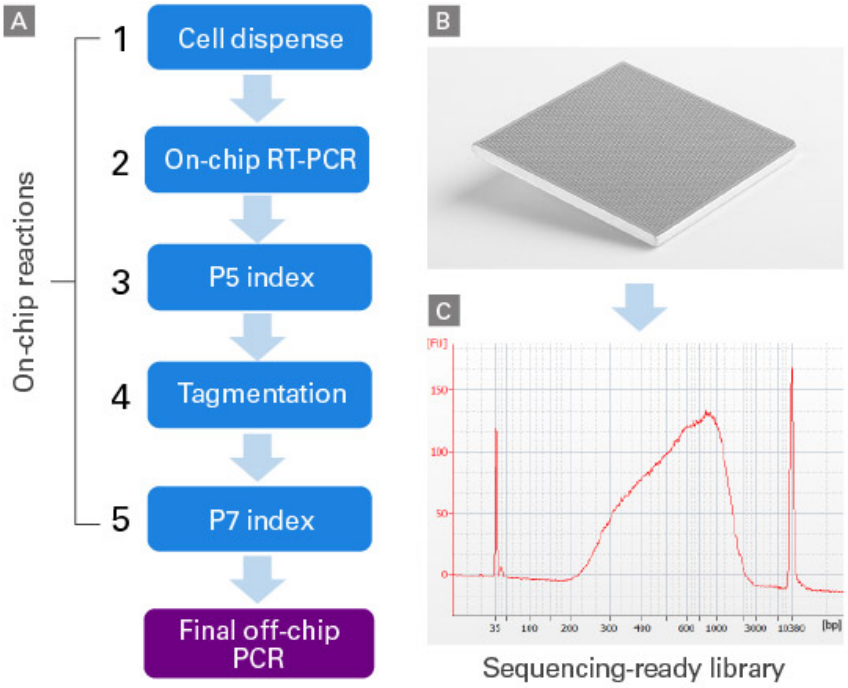


Figure 1. SMART-Seq ICELL8 application kit workflow. Panel A. This 10-hr protocol contains five dispensing steps. Cultured cells are dispensed into the wells of an ICELL8 blank chip at an average of 1 cell/well (Step 1). ICELL8 cx CellSelect Software is used to identify single cells. Cell lysis is followed by cDNA synthesis and amplification (Step 2). Full-length cDNA is tagmented with Illumina Nextera® Tagment DNA Enzyme and amplified with Illumina-specific indexed

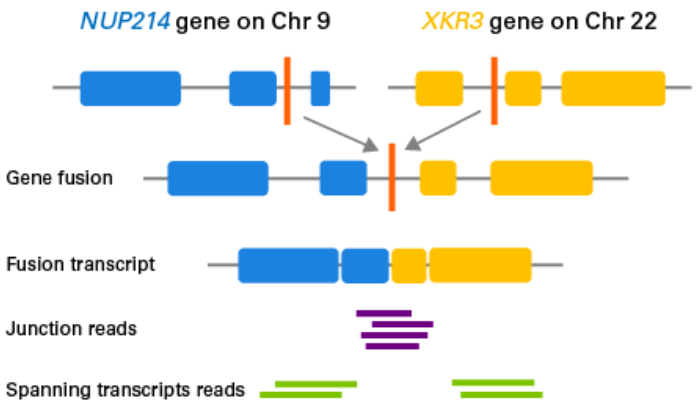
adapters added in a grid-like fashion (Steps 3–5). The final libraries are pooled, further amplified, and purified prior to sequencing. **Panel B.** The cells, reagents, indexes, and tagmentation enzyme are dispensed into the 5,184-well chip. 72 x 72 indexing of the barcodes across the chip allows each well to have a unique combination of barcodes. **Panel C.** A representative Bioanalyzer trace of a sequencing-ready library.

Gene fusions

To examine the detection of known gene fusion with the two methods, K562 cells were processed using the 10x Chromium system with the Chromium Single Cell 3′ Library & Gel Bead Kit ("10x Genomics 3′DE") or the SMART-Seq ICELL8 application kit workflow ("Takara Bio Full Length"). For all experiments, 10x libraries (single-end reads) were sequenced on a NovaSeq® platform according to 10x recommendations for loading and cycling conditions, and Takara Bio libraries (paired-end reads) were sequenced on a NextSeq® high-output cartridge, based on Illumina's recommended loading and cycling conditions.

Figure 2, Panel A depicts a fusion of *NUP214* and *XKR3*, with junction reads shown in purple and reads spanning transcripts in green. The table in Figure 2, Panel B shows that the 10x method was only able to detect one fusion gene out of five detected using the Takara Bio method. The Takara Bio method's paired-end read counts spanning fusion genes provide additional confirmation. A known *BCR--ABL1* fusion was also detected with the Takara Bio method.

A



B

Fusion name	Takara Bio Full Length		10x Genomics 3'DE		Left breakpoint	Right breakpoint
	Junction*	Spanning*	Junction*	Spanning*		
NUP214--XKR3	11	20	3	0	chr9:131199015:+	chr22:16808083:-
XACT--LRCH2	5	0	0	0	chrX:113938193:-	chrX:115163783:-
XACT--LRCH2	3	0	0	0	chrX:113938193:-	chrX:115156667:-
C16orf87--ORC6	1	9	0	0	chr16:46824386:-	chr16:46696017:+
C16orf87--ORC6	1	9	0	0	chr16:46824386:-	chr16:46693093:+
BCR--ABL1†	1	1	n/a	n/a	chr22:23290413:+	chr9:130854064:+

Figure 2. Detection of gene fusions. Data from K562 cells with >85% uniquely mapped reads from a total of ~23M reads. Here, the Takara Bio Full Length method was compared to the 10x Genomics 3'DE (v3) method. *Indicates read counts occurring across the junction ("junction") or spanning either side of the junction ("spanning"). †The known *BCR–ABL1* fusion was detected in K562 cells positive for the Philadelphia chromosome using the Takara Bio method. (This sample was not tested using the 10x method.) A K562 cell line negative for the *BCR–ABL1* fusion was used to generate the rest of the data.

SNP detection

To examine the potential for SNP detection in the two methods, 1,000 HEK293 cells were processed using the 10x Genomics 3′DE method or the Takara Bio Full Length method and libraries were sequenced on NovaSeq or NextSeq platforms, respectively.

In Figure 3, 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).

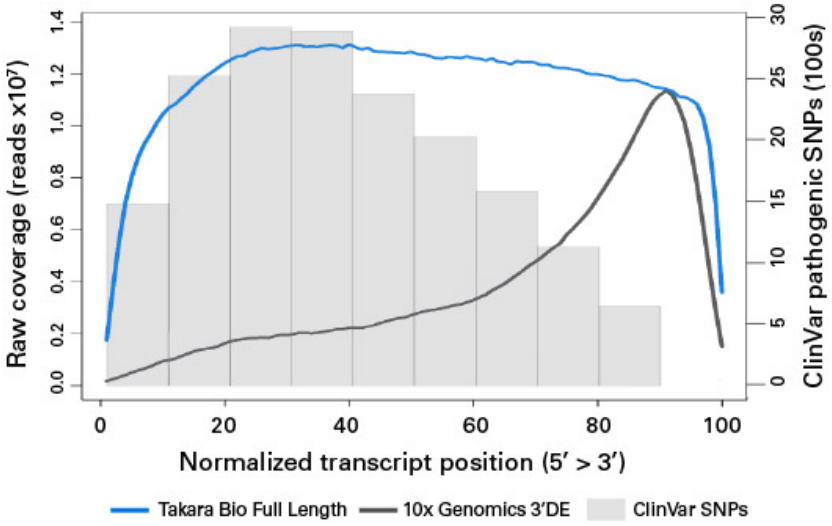


Figure 3. Distribution of sequencing reads and pathogenic SNPs annotated in ClinVar.

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 (Figure 4), especially when looking for high coverage (>30X).

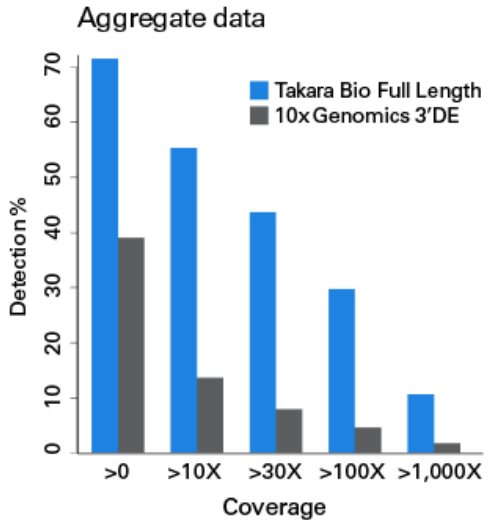


Figure 4. Read coverage of positions annotated as pathogenic SNPs in ClinVar.

We can visualize the difference in SNP detection by expressing it as a coverage ratio across the full length of the transcript. Figure 5 illustrates that across the normalized transcript, coverage of SNPs is approximately 32 times higher for the Takara Bio method than the 10x method. As expected, this ratio decreases at the last 10% of the normalized transcript due to the 3' bias of the 10x method, but the Takara Bio method still provides significantly more coverage of the 3' end.

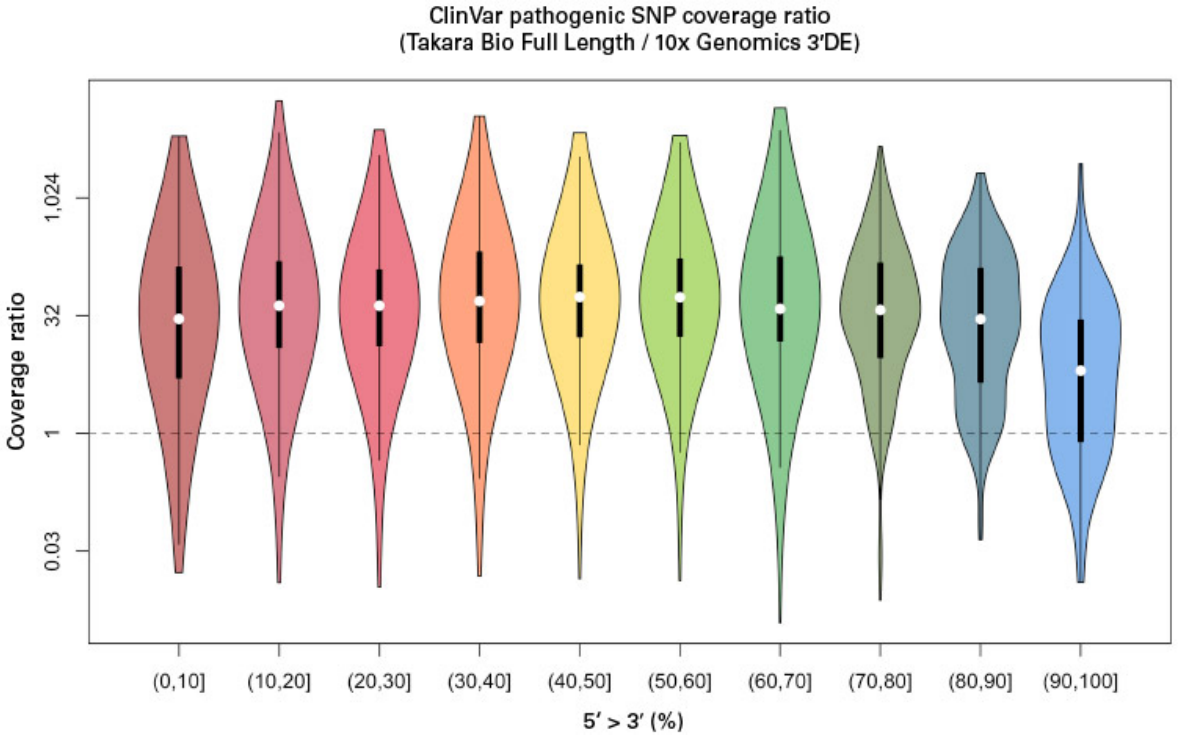
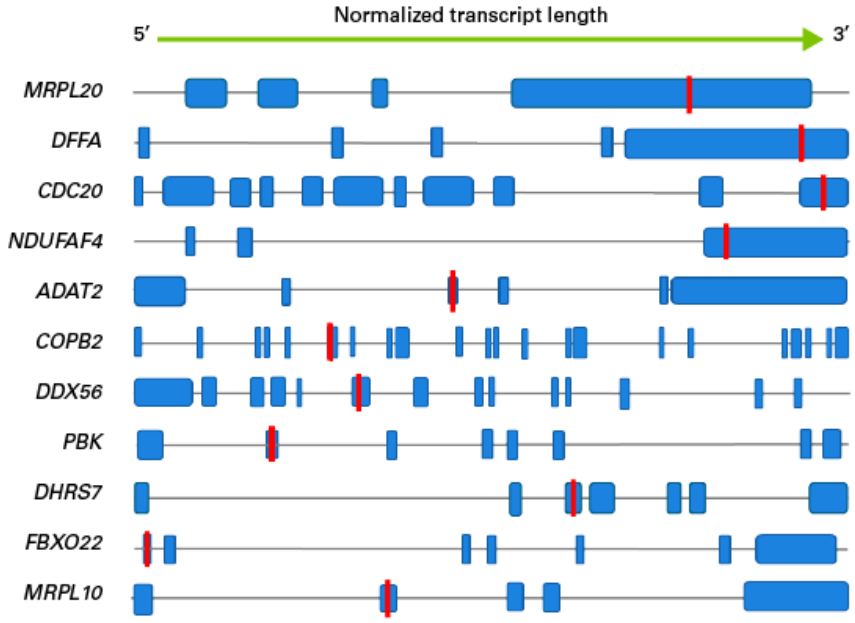


Figure 5. Read coverage of positions annotated as pathogenic SNPs ClinVar across the normalized transcript.

The advantage of the Takara Bio method's increased 5'-end coverage is apparent in the SNP detection data shown in Figure 6. Figure 6, Panel A provides a snapshot of the locations of SNPs (normalized length; exon map not drawn to scale) detected across the whole transcripts of 11 genes in K562 samples prepared with the Takara Bio Full Length method. The table in Figure 6, Panel B shows that the 10x method detected fewer SNPs occurring toward the 5' ends of genes than the Takara Bio method.

A



B

		Takara Bio Full Length		10x Genomics 3'DE		
Gene	**Wt / variant**	**Total reads**	**Variant reads**	**Total reads**	**Variant reads**	**Location**
MRPL20	G / C	662	163	127	27	Exon 4 (total 4)
DFFA	C / T	136	45	71	21	Exon 5 (total 5)
CDC20	G / A	161	84	144	86	Exon 11 (total 11)
NDUFAF4	A / C	225	74	98	42	Exon 3 (total 3)
ADAT2	A / C	107	25	17	5	Exon 3 (total 6)
COPB2	C / A	209	144	ND	ND	Exon 6 (total 22)
DDX56	G / A	108	32	ND	ND	Exon 6 (total 14)
PBK	C / G	160	55	ND	ND	Exon 2 (total 8)
DHRS7	T / C	175	173	ND	ND	Exon 3 (total 7)
FBXO22	C / T	169	55	ND	ND	Exon 1 (total 7)
MRLP10	G / T	231	137	ND	ND	Exon 2 (total 5)

Figure 6. Overall detection of SNPs. Here, the Takara Bio Full Length method was compared to the 10x Genomics 3'DE (v3) method. **Panel A.** Approximate depiction of the locations of SNPs across exons in 11 transcripts. Transcript lengths have been normalized. **Panel B.** While the Takara Bio method enabled SNP detection in the 11 genes listed, the 10x method could not detect (ND) most SNPs toward the 5' ends.

Alternative splicing

The aforementioned HEK293 libraries were analyzed for splicing variation by generating Sashimi plots (IGV, version 2.4.10), using the Proteasome subunit alpha type-4 (*PSMA4;* ~1.2-kb mRNA) and Branched chain amino acid transaminase 1 (*BCAT1;* ~9.6-kb mRNA) loci as examples. In Figure 7, Panel A, the aggregated data (all reads from each library) shows that multiple *PSMA4* 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. Figure 7, Panel B shows *BCAT1* splice variants discovered in five single cells using the Takara Bio method.

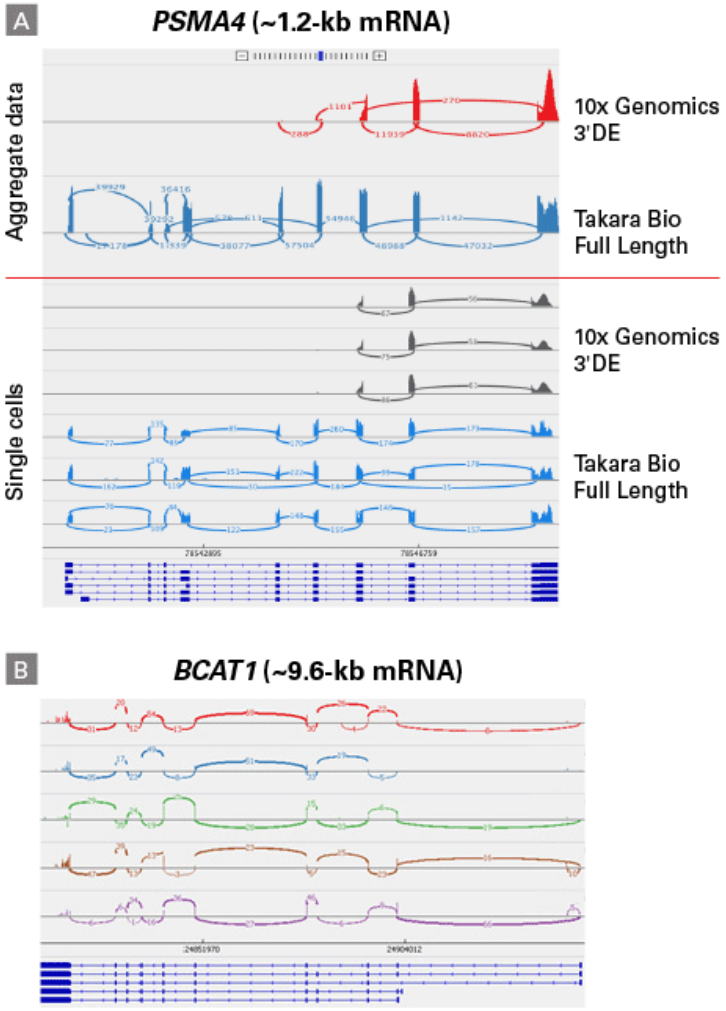


Figure 7. Detection of differential splicing. Panel A. Splice variants found in *PSMA4* at both aggregate and single-cell levels. **Panel B.** Splice variants found in *BCAT1* at the single-cell level.

Conclusions

The ICELL8 cx platform is a unique nanowell system that, unlike droplet-based systems, allows the multiple interactions between sample and reagents needed to support complex chemistry. Once the ICELL8 cx system dispenses cells from up to eight samples, it resolves individual wells bearing single cell candidates for downstream analysis using automated microscopy with up to three separate spectral channels, permitting live-dead and rare-cell measurements. Reagents, such as those in the SMART-Seq ICELL8 Reagent Kit, are distributed only to wells that are automatically or manually chosen for further processing. A key benefit of the ICELL8 cx system is its open design, enabling rapid assay development for single-cell interrogation in a high-throughput fashion.

Here, we described a sensitive SMART-Seq method for automated, full-length RNA-seq that offers benefits in increased gene body coverage, enabling improved detection of fusions, SNPs, and splice variants—applications that we hope will aid in the advancement of biomarker identification and the development of novel therapeutics (Zhao 2019). Additional applications for RNA analysis include 3′ mRNA end-capture kits and a kit that allows combined high-throughput T-cell receptor clonotype and 5′-end transcriptome profiling. For epigenetic analysis, a novel CUT&Tag technique has been developed, and for DNA analysis, a method for high-throughput CNV analysis is under development. The ICELL8 cx system's flexibility, combined with the availability of robust reagent sets and new bioinformatics tools for analysis, commends its use in a wide variety of single-cell analyses requiring both high sensitivity and reproducibility.

References

Zhao, S. Alternative splicing, RNA-seq and drug discovery. *Drug Discov. Today* (2019). https://doi.org/10.1016/j.drudis.2019.03.030

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