Efficient full-length RNA single-cell isoform sequencing using Sequencing by Expansion

Christophe H. Georgescu[1,2], Houlin Yu[1,2], Brian J. Haas[1,2], Guoping Wang[1,2], Megan Rogers-Peckham[3], Stephanie Yaung[4], Michael Kartje[4], Charlotte Yang[3], Ghadman Al-Eryani[1,2], Akanksha Khorgade[1,2], Allison Brookhart[1,2], Dan A. Bartlett[1,2], Lacey McGee[3], Kendall Berg[3], Cynthia Cech[3], Salka Barrett[3], Anasha Arryman[3], Taylor Lehmann[3], Chen Zhao[4], Marc Prindle[3], Melud Nabavi[3], Niall J. Lennon[1,2], Mark Kokoris[3], Aziz M. Al'Khafaji[1,2]

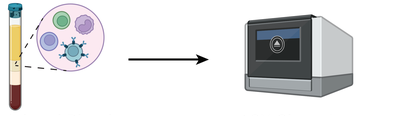
[1] Broad Clinical Labs, Cambridge, MA ; [2] Broad Institute, Cambridge, MA ; [3] Roche Sequencing Solutions, Inc, Seattle, WA, USA ; [4] Roche Sequencing Solutions, Inc, Santa Clara, CA, USA

Introduction

Recent advances in long-read sequencing have enabled capture of full-length RNA isoforms at scale, allowing researchers greater resolution into the diversity and transcriptional dynamics of alternative splicing. In tandem, innovations in single-cell technology have dramatically increased the scale of cell capture, leading to a consequent gap in long-read sequencing approaches with sufficient throughput and depth for RNA isoform capture. To address this outstanding limitation we evaluated the use of Sequencing by Expansion (SBX) which has the relevant advantages of immense throughput.

Materials & Methods

To evaluate SBX technology for single-cell RNA isoform sequencing, we captured ~3000 Peripheral Blood Mononuclear Cells using the 10x Chromium 5' Gene Expression Kit and sequenced the intermediate full-length cDNA library. We compared results to both Illumina short-read sequencing for gene expression level analyses and PacBio long-read for isoform level analyses.



Peripheral Blood Mononuclear Cells (PBMCs)

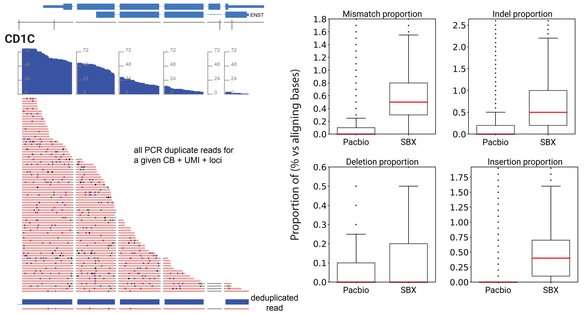
10X Chromium Single Cell 5'

Illumina | Roche SBX | Pacbio Kinnex

Original run	**NovaSeq 6000** SP flowcell 1.5B PE reads 72 x 246 - 38h	***Prototype SBX platform used*** 2B reads - 1h	**Revio** 2 flowcells (multiplexed) 70M full length reads 24h
Current highest throughput option	**NovaSeq X Plus** 25B flowcell 25B PE reads 45 x 72 - 25h	***Prototype SBX platform used*** 2B reads/h	**Revio** 1 flowcell 100M full length reads 24h

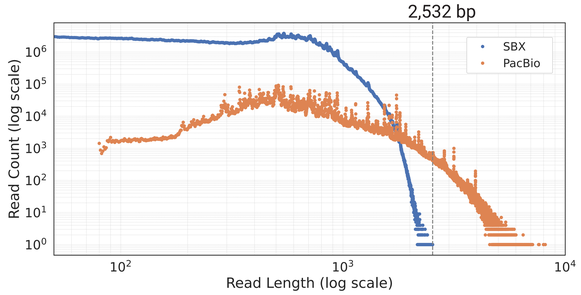
Results

SBX reads have unique length properties. Unlike fixed-length Illumina reads or full-length Pacbio reads, SBX reads are of variable length, always originating from the adapter-ligated end. This process generates reads that are less than or equal to the length of the parent cDNA molecule. Consequently, PCR duplicates from a single molecule yield a population of reads with a range of lengths, as shown in the figure. To accommodate this, our deduplication pipeline retains only the longest read for each unique combination of Cell Barcode (CB), UMI, and locus.



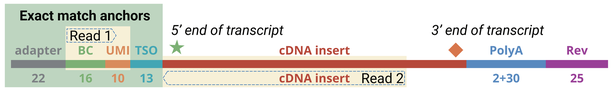
While Illumina produces fixed length reads (246 bp read2 in our run), SBX and Pacbio reads vary in lengths due to a range of cDNA sizes and a different base principale. The plot shows the actual read length distribution, while the table summarizes statistics.

Single cell RNA-seq	SBX	Pacbio
Median read length (bp)	512	632
Top quartile read length (bp)	690	934
Mismatch rate (mean)	0.006	0.001
Indel rate (mean)	0.006	0.001

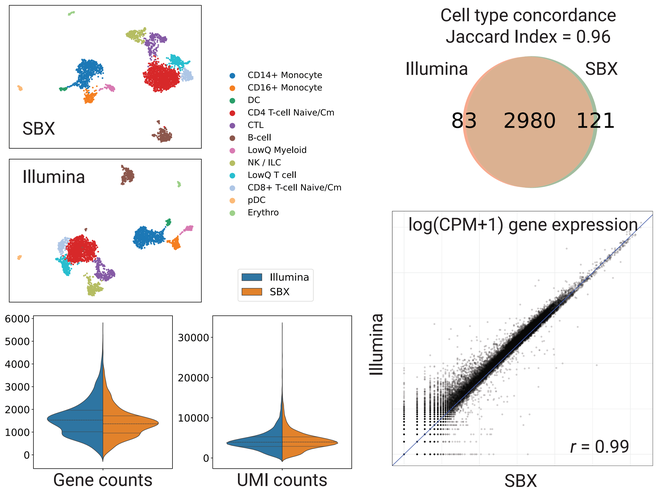


To enable a direct comparison with Illumina data, single-end SBX reads were adapted for use with the paired-end CellRanger pipeline. First, reads were stringently filtered to require exact matches to the adapter and TSO sequences, as well as a complete 26 bp CB and UMI. A pseudo-paired-end structure was then created: the 26 bp CB+UMI sequence was assigned as Read 1, while Read 2 was defined from the reverse complement of the first 649 bp of the cDNA sequence (the input limit for the STAR aligner).

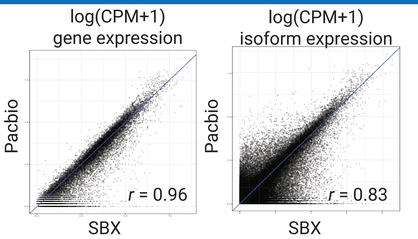
	Illumina	SBX
Reads sequenced	1.70B	3.1B
Reads into CellRanger	1.70B	2.09B
Num cells called	3,080	3,114
Valid Barcodes	87.7%	99,1%
Reads Mapped Conf. Transcriptome	60.7%	76.6%
Useful reads (sum read counts in molecule_info.h5)	0.99B	1.57B
% useful reads	58.2%	75.2%

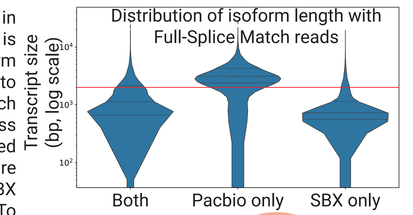


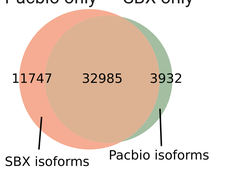
Across all three sequencing platforms, after downsampling to equal UMI counts, we identify 3103-2822 valid cell barcodes based on both the 10X barcodes whitelist and UMI counts, with 2795 cells overlapping between all platforms. For the SBX and Illumina comparison, we clustered cells using Scanpy and obtained the same consistent 11 cell type clusters, with a 0.96 Jaccard index for cells present in the same cell type clusters, and highly concordant gene expression on pseudobulked data (0.99 Pearson correlation coefficient).



Comparing SBX and Pacbio, we focused on isoforms as this is where read length brings the most benefits. After deduplicating reads, we ran LRAA to quantify reference gene and isoform expression. Pseudobulked gene expression correlation is lower than with Illumina due to some dropout in low expression genes, and with the increased diversity in isoforms, the isoform correlation is significantly lower. Looking at isoform differences, we limited ourselves to isoforms with Full-Splice Match (FSM) reads which are less ambiguous. Although most detected isoforms are shared (33k), there are isoforms uniquely identified by SBX (11.7k) and Pacbio (3.9k). To investigate what could be driving the differences, we compared the length of the reference transcripts and a clear pattern appears. SBX is able to identify more isoforms below 1kbp due to increased sequencing depth, while Pacbio is able to identify isoforms larger than 2kbp.







Conclusion

SBX has the capacity to sequence at very high throughput in short runtimes, with this analysis demonstrating that results obtained at the gene level are very similar compared to Illumina sequencing. Although SBX has a higher error rate, single-cell applications use amplified cDNA and sequence the same PCR duplicates multiple times, which allows error correction for applications such as mutation calling. Meanwhile, the increased read length (half the SBX reads are at least twice as long as the Illumina reads) enables looking at isoform level expression which is not possible with Illumina.

Looking at the Pacbio comparison, the balance is different. The much higher sequencing depth of SBX enables the capture of more isoforms and scaling experiments to higher number of cells. It however is not set to replace Pacbio entirely as it is not currently able to achieve sufficient read depth above 2kbp or go above 2.5kbp which is required for unambiguous identification of numerous isoforms.

Acknowledgments

We received support for this research from Roche* and 10x Genomics.

*Data shown in this presentation was generated on a prototype SBX platform (in development).