PacBio

Application note

MAS-Seq for single-cell isoform sequencing

Introduction

Understanding cell heterogeneity at the isoform level is critical for both basic and disease research. Short reads can only capture gene-level information, while other long-read technologies lack the accuracy for accurate unique molecular identifiers (UMI) and cell barcode (CBC) identification. PacBio® HiFi reads sequence full-length RNA isoforms along with single-cell barcode and UMI information, revealing unprecedented insight into single-cell biology.

The *MAS-Seq for 10x Single Cell 3' kit* takes as input single-cell cDNA and outputs a sequencing-ready library that results in a 16-fold throughput increase compared to regular single-cell Iso-Seq® libraries. Combined with an isoform-aware single-cell analysis SMRT® Link software, PacBio offers cost-effective single-cell isoform sequencing that does not require orthogonal sequencing methods. The SMRT Link software supports MAS-Seq bioinformatics analysis to produce an isoform-level single-cell data matrix compatible with tertiary analysis software.

102-326-549 REV02 JAN2023

PacBio

Single-cell RNA sequencing

Single-cell RNA sequencing (scRNA-seq) emerged to characterize gene expression differences between individual cells derived from a complex tissue, allowing a higher-resolution view of the transcriptome.

Most single-cell experiments are done with short reads, which only capture the ends of molecules due to fragmentation. Sequencing fragments limits expression information to the gene level, missing important isoform diversity that could be important for disease or biological function.

PacBio HiFi reads sequence full-length RNA isoforms along with single-cell barcode and UMI information (Figure 1), revealing isoform diversity at the single-cell level.

HiFi sequencing advantages for single-cell RNA sequencing

- Full-length isoform information
- Accurate cell barcode and UMI detection
- Variant detection

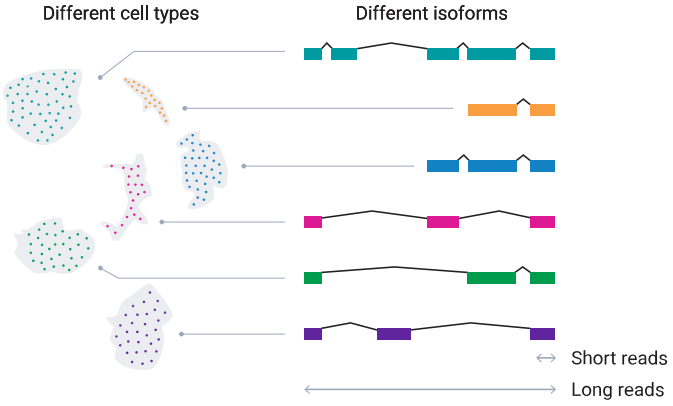

Figure 1. Single-cell isoform sequencing with PacBio long reads. Short reads only capture gene ends, missing isoform diversity. PacBio HiFi reads cover the entire isoform along with the single-cell barcode and UMI information with high accuracy.

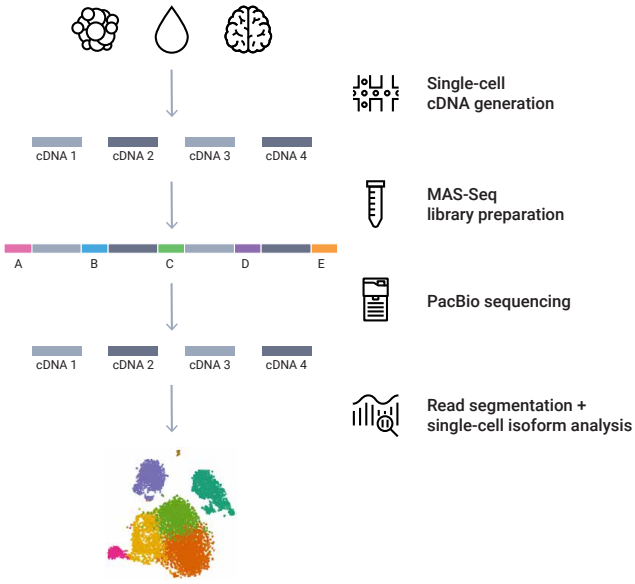

Figure 2. MAS-Seq for single-cell isoform sequencing. Single-cell cDNA molecules are concatenated into a larger insert library and sequenced, then processed using the PacBio software.

MAS-Seq for single-cell isoform sequencing

The Multiplexed Arrays Sequencing (MAS-Seq) method (Al'Khafaji et al., 2021) is a concatenation method for increasing throughput by joining cDNA molecules into longer concatenated fragments. HiFi reads generated from sequencing the concatenated molecules can then be bioinformatically broken up to retrieve the original cDNA sequences. The result is higher throughput and reduced sequencing needs for cost-effective single-cell isoform sequencing. Traditionally, orthogonal short-read scRNA-Seq is used to supplement the lower throughput of regular single-cell Iso-Seq method — with MAS-Seq, no orthogonal sequencing data are required.

The PacBio Single-cell Iso-Seq workflow processes the full-length cDNA sequences to classify them against a reference annotation (e.g., GENCODE) to identify novel genes and isoforms. The output consists of gene- and isoform-level count matrices that are compatible with tertiary analysis software.

The MAS-Seq kit is compatible with cDNA generated using the *10x Chromium Next GEM Single Cell 3' kit* (v3.1) and is intended for use on a 3,000 to 10,000 cell library with 15-75 ng of cDNA as input.

MAS-Seq library workflow

The MAS-Seq library workflow begins with single-cell cDNA and produces a MAS-Seq library that is ready for sequencing.

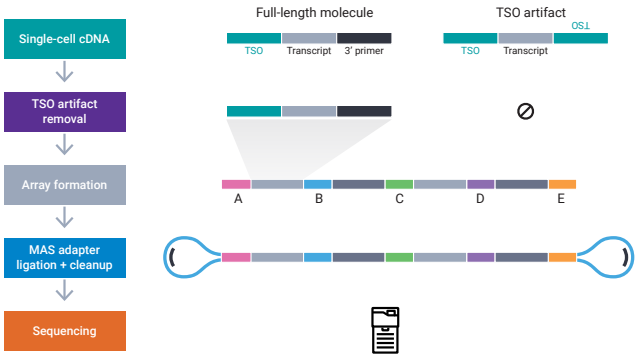

Figure 3. MAS-Seq library workflow.

Single-cell cDNA molecules are first removed of TSO (template-switching oligos) artifacts, then concatenated to form ordered arrays. MAS adapter ligation ensures full arrays are enriched, which are then sequenced on PacBio long-read sequencers (Figure 3).

With proper full array formation and adequate sequencing, one SMRT® Cell on the Sequel® II/IIe and Revio™ systems are expected to achieve 40 million and 100 million cDNA sequences, respectively (Table 1).

Table 1. Target MAS-Seq library performance

Metric	Performance
Sample preparation time	2 days
Expected library size	11,000–14,000 bp
Target P1 loading	60–80%
Expected HiFi yield per SMRT Cell	2.5 million HiFi reads (Sequel II/IIe) 6 million HiFi reads (Revio)
Expected full array %	85–92%
Expected read yield per SMRT Cell	40 million reads (Sequel II/IIe) 80 million reads (Revio)

MAS-Seq bioinformatics workflow

The SMRT Link *Read Segmentation and Single-cell Iso-Seq* workflow processes the HiFi reads generated from the MAS-Seq library to produce gene- and isoform-level count matrices that are compatible with tertiary single-cell analysis tools (Figure 4).

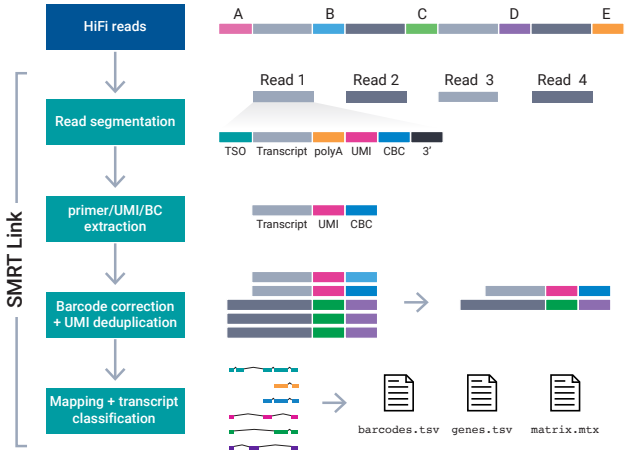

Figure 4. MAS-Seq analysis using *Read Segmentation and Single-cell Iso-Seq* workflow.

Read segmentation

HiFi reads are segmented into individual segmented reads (*S-reads*) that represent the original cDNA sequences.

Primer/UMI/BC extraction

Primers and polyA tails are removed, but also used to orient the read into 5' – 3' orientation. Single-cell barcode and UMI information are extracted.

Barcode correction & UMI deduplication

Cell barcodes are corrected given an expected barcode list. Real cells — cell barcodes that represent encapsulated single cells (as opposed to ambient RNA) are also identified at this step. Reads are then deduplicated based on cell barcodes and UMIs.

Mapping and transcript classification

Deduplicated reads are mapped to the reference genome and classified against a transcript annotation (e.g., GENCODE). Finally, a gene- and isoform-level single-cell matrix is output for tertiary analysis.

Metric	Sequel IIe PBMC 5k	Sequel IIe PBMC 10k	Revio PBMC 1	Revio PBMC 2
HiFi reads	2,622,891	2,789,502	7,261,030	7,439,608
Segmented reads (S-reads)	40,131,832	43,788,133	110,127,016	114,238,613
S-read length distribution	300–2,500 bp	300–2,500 bp	300–2,500 bp	300–2,500 bp
S-reads with valid barcodes	37,634,585 (94%)	41,369,598 (94%)	105,309,352 (96%)	109,213,167 (96%)
Deduplicated reads	23,883,685	30,903,830	60,421,293	61,831,404
Estimated number of cells	3,966	9,316	8,822	8,815
Reads in cells	91.6%	87.84%	94.99%	94.82%
Mean reads per cell	8,708	3,905	11,352	11,762
Median UMIs per cell	4,821	2,920	5,861	6,002
Median genes per cell	704	621	938	957
Median transcripts per cell	818	705	1,110	1,130

Table 2. Read, cell, and transcript statistics of PBMC MAS-Seq library runs on Sequel II/IIe and Revio systems. Each library was run on one SMRT Cell. Data can be downloaded here.

MAS-Seq example: PBMC dataset

PBMC cDNA generated using the *10x Chromium Next GEM Single Cell 3' kit* was made into MAS-Seq libraries and sequenced with one SMRT Cell each and analyzed using the *Read Segmentation and Single-Cell Iso-Seq* workflow in SMRT Link v11.1 (Table 2). The output results were then processed with tertiary tools to identify cell types (Figure 5).

MAS-Seq for single-cell isoform sequencing: summary

The *MAS-Seq for 10x Single Cell 3' kit* offers an end-to-end solution for single-cell RNA isoform sequencing from sample preparation to bioinformatics analysis.

- Supports cDNA generated from *10x Chromium Next GEM Single Cell 3' kit* (v3.1)
- 15-75 ng input cDNA
- Target 3,000 to 10,000 cell library
- 16-fold throughput increase compared to non-MAS-Seq methods
- No orthogonal sequencing data required

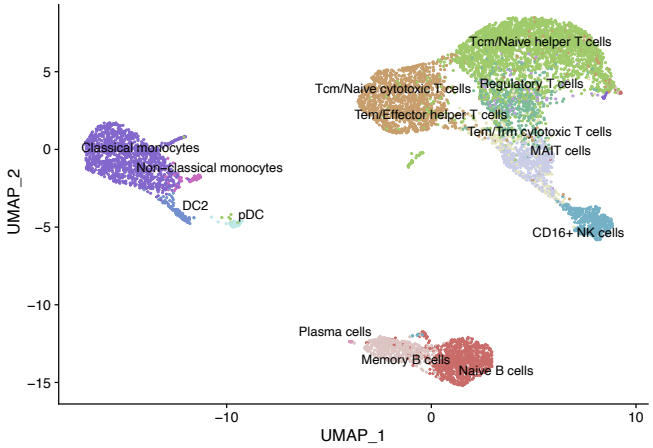

Figure 5. UMAP cell clustering of Revio-PBMC 2 using CellTypist after SMRT Link v11.1 analysis.

References

Al'Khafaji et al., (2021) **High-throughput RNA isoform sequencing using programmable cDNA concatenation.** *bioRxiv,* 10.01.462818, doi: https://doi.org/10.1101/2021.10.01.462818