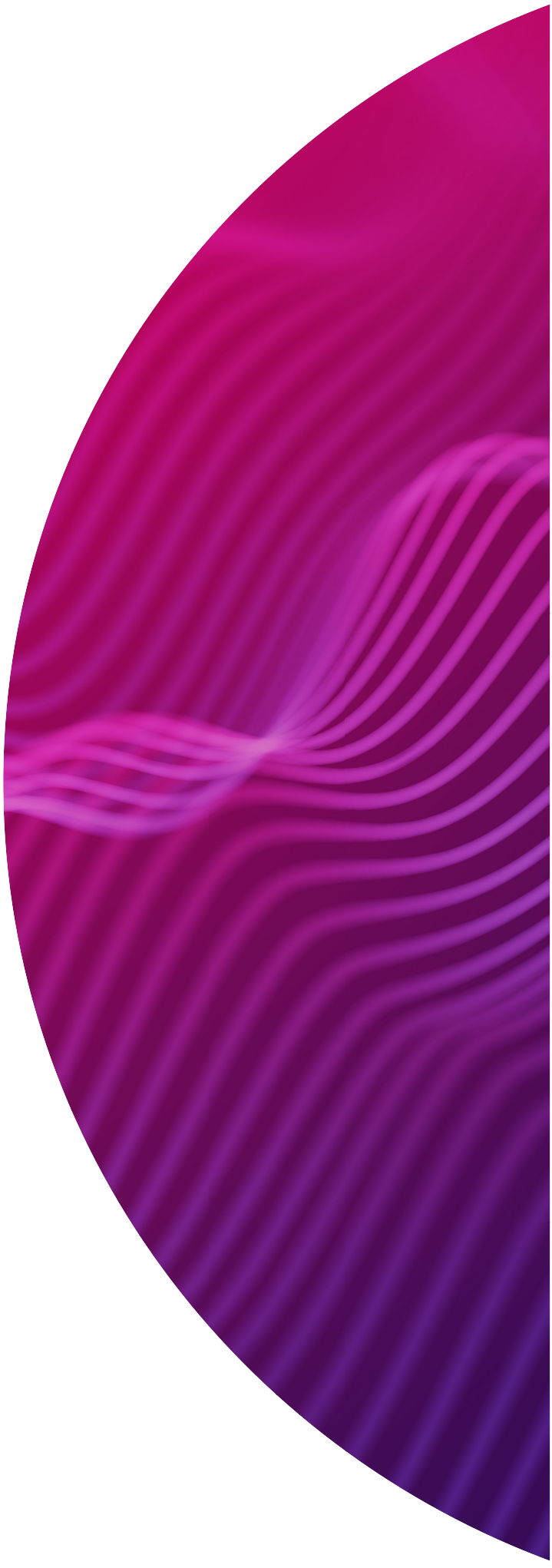
PacBio

SMRT® Link MAS-Seq Single-Cell troubleshooting guide (v12.0)



Research use only. Not for use in diagnostic procedures.

P/N 102-994-400 Version 02 (June 2023)

PacBio
1305 O'Brien Drive
Menlo Park, CA 94025
www.pacb.com

SMRT® Link MAS-Seq Single-Cell troubleshooting guide (v12.0)

Introduction

This document describes the metrics generated by the **Read Segmentation** and **Single-Cell Iso-Seq®** workflow in SMRT Link v12.0. The document also describes possible issues that can occur when using the **MAS-Seq for 10x Single Cell 3'** kit, for both supported and unsupported use cases.

Note: Everything in this document **also** applies to SMRT Link v11.1.

- Example data sets (PBMC 5k and PBMC 10k cells) are available here.
- Additional command-line information, example commands, and suggestions for tertiary analyses are described here.

SMRT Link Read Segmentation

The SMRT Link Read Segmentation workflow can be invoked either as a standalone Data Utility workflow, or in combination with Single-Cell Iso-Seq as an Analysis workflow. For MAS-Seq single cell users using the **MAS-Seq for 10x Single Cell 3'** kit, **Read Segmentation and Single-Cell Iso-Seq** is the recommended workflow.

Read Segmentation deconcatenates HiFi reads into segmented reads (S-reads) based on segmentation adapters, using the command-line `skera` tool. (See here for details.)

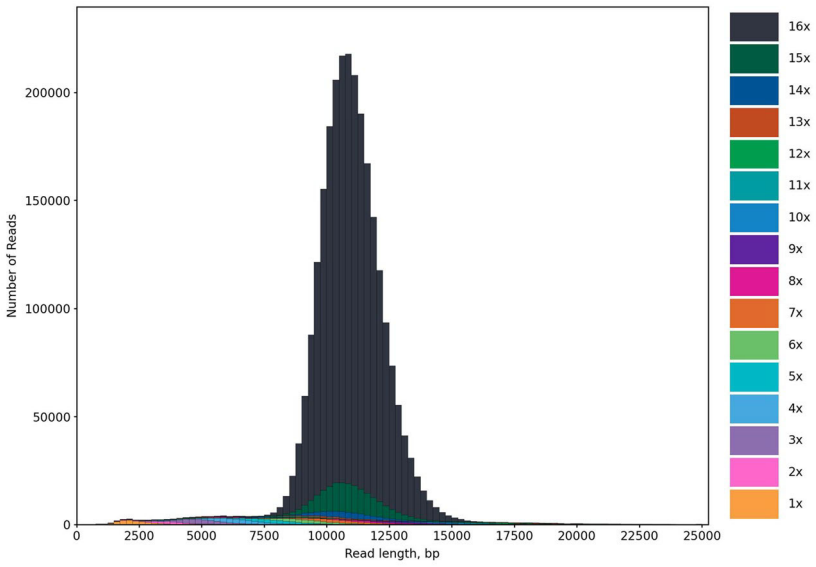
The MAS-Seq kit enriches for full (16-fold) arrays, while most 10x cDNA libraries using the 3' kit are 600-1000 bp. Therefore, the percentage of full array and concatenation factors should have typical values as shown below.

Metric	Explanation	Typical value
Reads	Number of HiFi reads	Depends on sequencing yield
S-reads	Number of segmented reads	Depends on HiFi read yield and concatenation success
Mean Length of S-reads	Mean read length of S-reads	600-800bp for 10x cDNA
Percent of Reads with Full Arrays	Percent of HiFi reads with full MAS arrays	85-90%
Mean Array Size	Concatenation factor	~15.xx

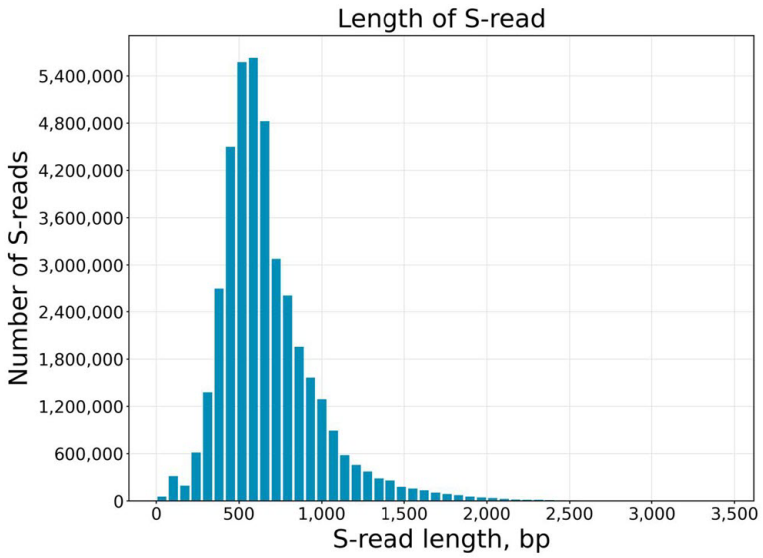
Read Segmentation

Value	Analysis Metric	
2,622,891	Reads	← Input # HiFi reads. Depends on loading (P1), Pol RL, HiFi conversion rate
40,131,832	Segmented reads (S-reads)	
672	Mean length of S-reads	← Depends on input 10x cDNA size, but generally 600-800 bp
86.32 %	Percent of reads with full arrays	← Regardless of input cDNA size and input reads, should be at least ~8x%
15.30	Mean array size (concatenation factor)	← Regardless of input cDNA size and input reads, should always be ~15.xx

A clean peak between 10,000 – 14,000 bp indicates good MAS array formation and successful enrichment of full arrays:



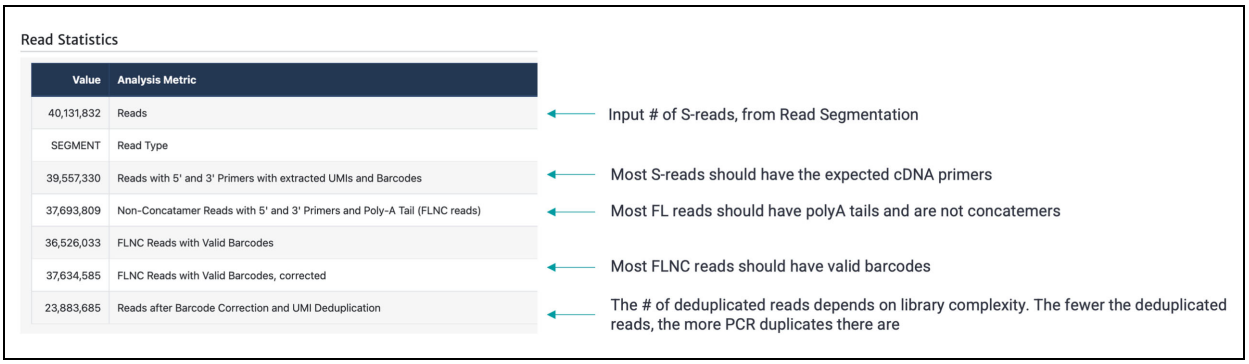
S-read read length should largely reflect the original 10x cDNA library size:



SMRT Link Single-Cell Iso-Seq workflow: Read statistics

cDNA primers and polyA tails are removed from S-reads, then UMI/BC are extracted and reads are deduplicated. This is performed using the command `isoseq3 tag/refine/correct/groupdedup`.
(See here for the high-level workflow.)

Metric	Explanation	Typical value
Reads	Number of S-reads	Depends on sequencing yield
Read Type	CCS or SEGMENT	CCS or SEGMENT
Reads with 5' and 3' Primers with Extracted UMIs and Barcodes	Full-Length (FL) tagged reads	>95% of reads should be FL tagged
Non-Concatemer Reads with 5' and 3' Primers and PolyA Tail	Full-Length Non-Concatemer (FLNC) tagged reads	>90% of reads should be FLNC tagged
FLNC Reads with Valid Barcodes	FLNC reads matching a barcode white list	>90% of reads should match barcodes in the white list
FLNC Reads with Valid Barcodes, Corrected	FLNC reads matching the barcode white list after correction	>90% of reads should match barcodes in the white list after correction
Reads After Barcode Correction and UMI Deduplication	Deduplicated reads	Deduplicated read yield depends on the 10x library complexity and PCR duplication rate



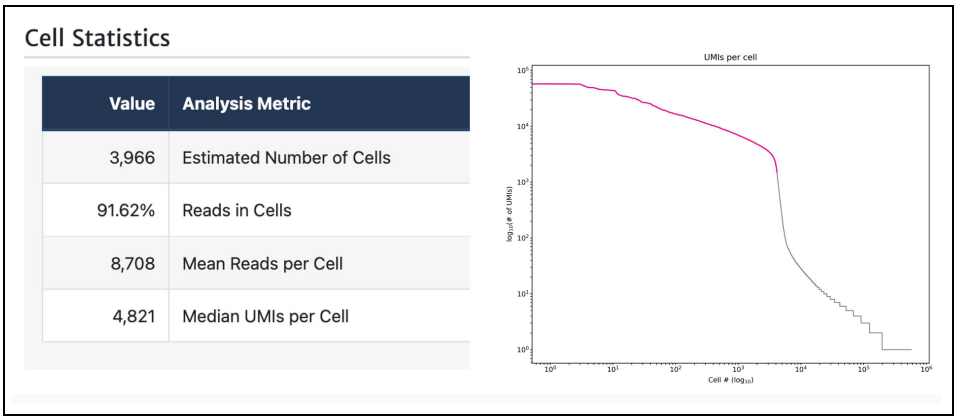
SMRT Link Single-Cell Iso-Seq workflow: Cell statistics

The number of estimated cells ("real cells") varies by experiment. The estimation is performed using the `isoseq3 bcstats` command. (See here for information.)

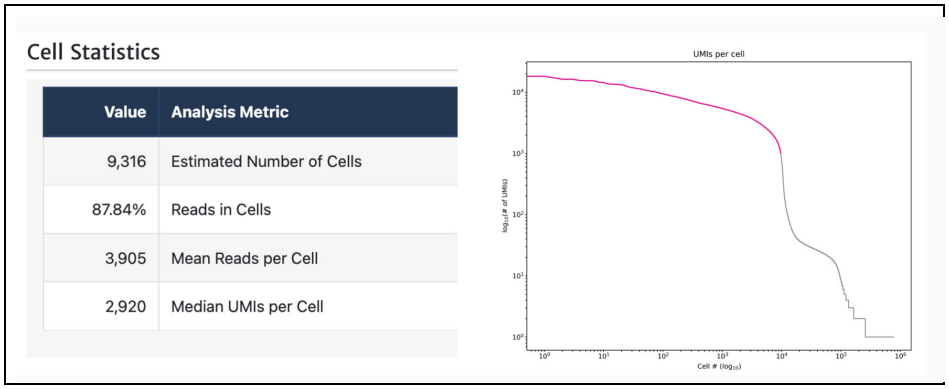
Metric	Explanation	Typical value
Estimated Number of Cells	The number of real cells	Depends on the 10x library
Reads in Cells	The percent of reads in real cells	>85%
Mean Reads per Cell	The mean reads per real cell	Depends on the 10x library and read yield
Median UMIs per Cell	The median UMI per real cell	Depends on the 10x library, read yield, and PCR duplication rate

The estimated number of cells, mean reads per cell and median UMIs per cell are highly dependent on the single-cell library and sample complexity. If you suspect that the cell estimation is incorrect using the default `knee` method for `isoseq3 correct`, the cells can be re-estimated using the alternative `percentile` method. (See here for details.)

Example 1: PBMC 5k cells - Cell statistics



Example 2: PBMC 10k cells - Cell statistics



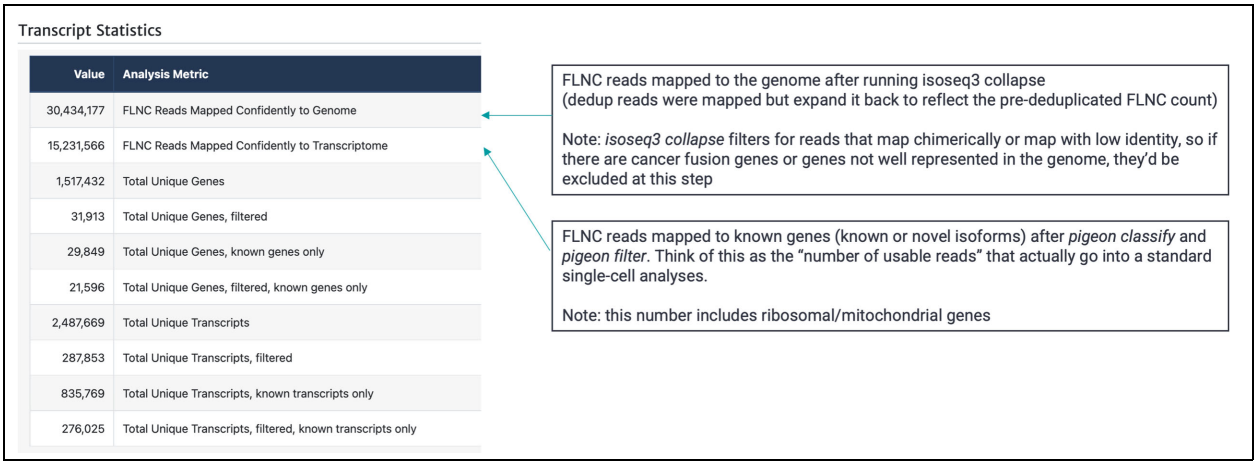
SMRT Link Single-Cell Iso-Seq workflow: Transcript statistics

Deduplicated reads are mapped to a genome, classified and filtered using `pigeon` software (SQANTI3). This is performed using the command `pbmm2/isoseq3 collapse/pigeon`. (See here for information.)

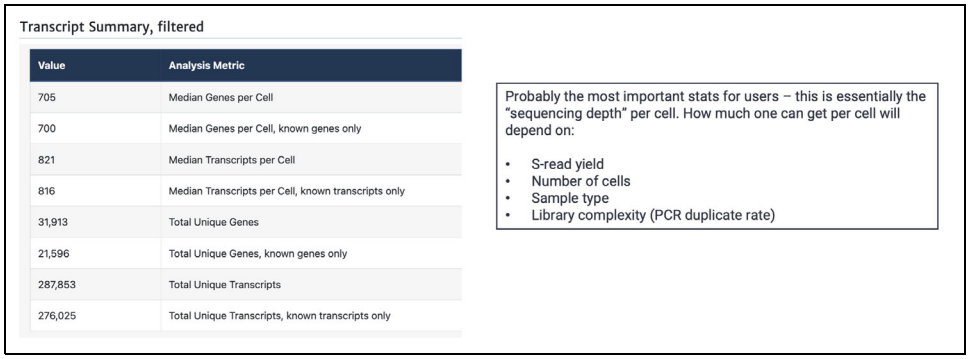
Metric	Explanation	Typical value
FLNC Reads Mapped Confidently to Genome	FLNC reads (before deduplication) mapped to the genome.[a]	~80%
FLNC Reads Mapped Confidently to Transcriptome	FLNC reads (before deduplication) mapped to transcriptome[b]	30-50%
Total Unique Genes	Total unique genes **before** `pigeon` filtering[c]	Sample-dependent
Total Unique Genes, filtered	Total unique genes **after** `pigeon` filtering[c]	Sample-dependent
Total Unique Genes, known genes only	Total unique known genes **before** `pigeon` filtering[c]	Sample-dependent
Total Unique Genes, filtered, known genes only	Total unique known genes **after** `pigeon` filtering[c]	Sample-dependent
Total Unique Transcripts	Total unique transcripts **before** `pigeon` filtering	Sample-dependent
Total Unique Transcripts, filtered	Total unique transcripts **after** `pigeon` filtering	Sample-dependent
Total Unique Transcripts, known transcripts only	Total unique known transcripts **before** `pigeon` filtering	Sample-dependent
Total Unique Transcripts, filtered, known transcripts only	Total unique known transcripts **after** `pigeon` filtering	Sample-dependent

a. FLNC reads mapped to the genome after running `isoseq3 collapse`. Though actual mapping is done with deduplicated reads, UMI count is summarized post-mapping to reflect the pre-deduplicated FLNC count. Note that `isoseq3 collapse` filters for reads that map chimerically or map with low identity, so if there are cancer fusion genes or genes not well represented in the genome, they would be **excluded** at this step. In general, one should expect most (~80%) FLNC reads to map to the genome, even if they end up mapping to, say, intergenic regions.

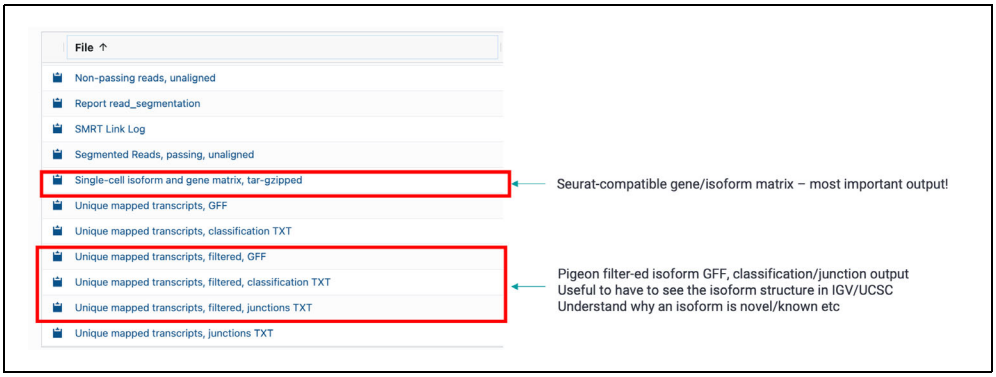
b. FLNC reads mapped to known genes (known or novel isoforms) after `pigeon classify` and `pigeon filter`. This number more likely represents the "number of usable reads" that actually go into a standard single-cell analysis. This number includes ribosomal/mitochondrial genes. It is typical to see 30-50% FLNC reads map to the transcriptome, which is consistent with equivalent 10x short read sequencing data. Most of the non-transcriptomic but genomically-mapped reads are attributed to intergenic regions and are filtered out by `pigeon filter`.
c. It is typical to see a very high number of "total number of genes/transcripts" before `pigeon filter`. This is due to the high number of loci that are intergenic and still being assigned a "novel gene" status before `pigeon filter`.



After `pigeon` filtering, the number of genes/isoforms per cell:



SMRT Link Read Segmentation and Single-Cell Iso-Seq Workflow: File downloads



Possible issues when using the MAS-Seq for 10x Single Cell 3' kit for supported use cases

The currently-supported use case for the MAS-Seq kit is a single-cell library produced using the 10x Single Cell 3' kit, with a 3000-10,000 cell targeted recovery.

Observed issue	Likely cause	Solution
• Good concatenation factor • Low S-read yield	Low P1 loading or HiFi conversion	Perform additional sequencing
• Good S-read yield • Poor FLNC yield and beyond	Not using the 10x 3' kit (v3.1).	Reanalyze with proper cDNA primer, UMI/BC design and barcode white list. Additional 10x cDNA primers and barcode white list can be found here.
• Good S-read yield • Good cell statistics • Poor read mapping and low gene counts	The wrong reference was selected.	Choose correct reference genome and annotation. SMRT Link supports **only** human and mouse reference genome + Gencode annotation (available here). If using different genomes or annotations, refer to the `pigeon` documentation for command line analysis. (See here for details.)
• Good S-read yield • Poor cell recovery	The algorithm underestimated the number of cells.	Reanalyze using the `percentile` method in SMRT Link or using the command line. (See here for details.)
• Analysis experienced an error, but was able to recover and complete successfully; High Barcode Errors	Incorrect barcode white list	Reanalyze using the correct barcode white list. The error message **Analysis experienced an error, but was able to recover and complete successfully; High Barcode Errors** indicates that the barcode white list provided is **incorrect**. Note that SMRT Link expects a barcode white list that is reverse-complemented, which is not how the 10x white list is typically provided. A list of common barcode white list in reverse-complement can be found here.

Troubleshooting Example 1: Wrong reference selected, poor gene/ transcript recovery

Inputs

Data Type	Name	Import Complete
BarcodeSet	Barcode Sets: 10x Chromium single cell 3' cDNA primers	Yes
BarcodeSet	Barcode Sets: MAS-Seq Adapter v1 (MAS16)	Yes
ConsensusReadSet	HiFi Reads: JacksonLab_1-Cell4 (CCS)	Yes
ReferenceSet	References: Human Genome hg38, with Gencode v39 annotations	Yes

Transcript Summary, filtered

Value	Analysis Metric
32	Median Genes per Cell
32	Median Genes per Cell, known genes only
34	Median Transcripts per Cell
34	Median Transcripts per Cell, known transcripts only
2,394	Total Unique Genes
2,359	Total Unique Genes, known genes only
10,781	Total Unique Transcripts
10,731	Total Unique Transcripts, known transcripts only

Correct reference selected, good gene/transcript recovery

Inputs

Data Type	Name	Import Complete
BarcodeSet	Barcode Sets: 10x Chromium single cell 3' cDNA primers	Yes
ConsensusReadSet	HiFi Reads: JacksonLab_1-Cell4 (CCS) Segmented Reads	Yes
ReferenceSet	References: Mouse Genome mm39, with Gencode vM28 annotati...	Yes

Transcript Summary, filtered

Value	Analysis Metric
757	Median Genes per Cell
751	Median Genes per Cell, known genes only
842	Median Transcripts per Cell
837	Median Transcripts per Cell, known transcripts only
28,754	Total Unique Genes
20,446	Total Unique Genes, known genes only
340,640	Total Unique Transcripts
330,490	Total Unique Transcripts, known transcripts only

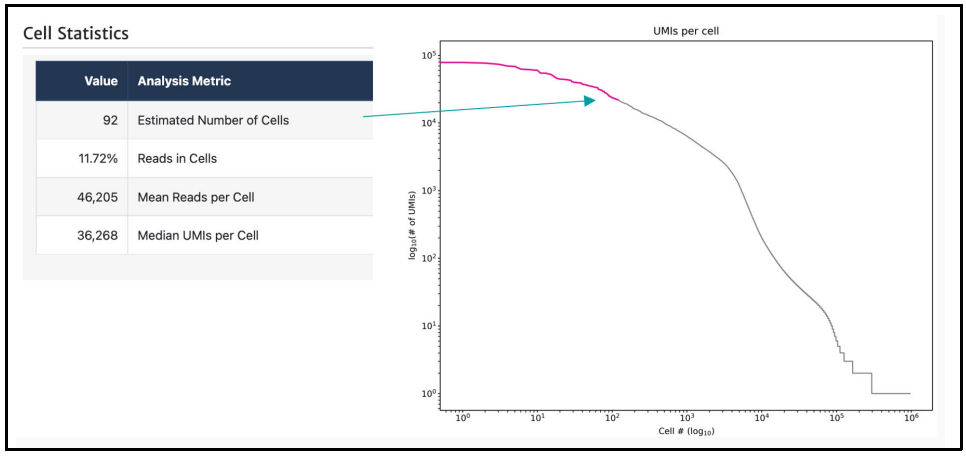
Troubleshooting Example 2: Underestimating the number of cells

If you generated matching short read data or have an expected target cell recovery, you might identify cases in which the cell barcode calling algorithm **underestimated** the number of cells. This affects:

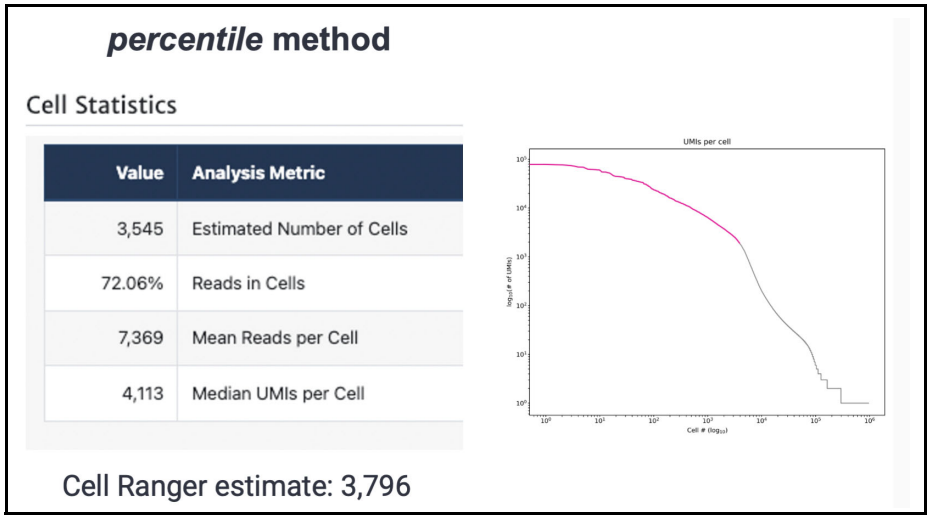
- Cell statistics
- Transcript statistics
- Output count matrix

It does **not** affect:

- Segmentation statistics
- Read statistics



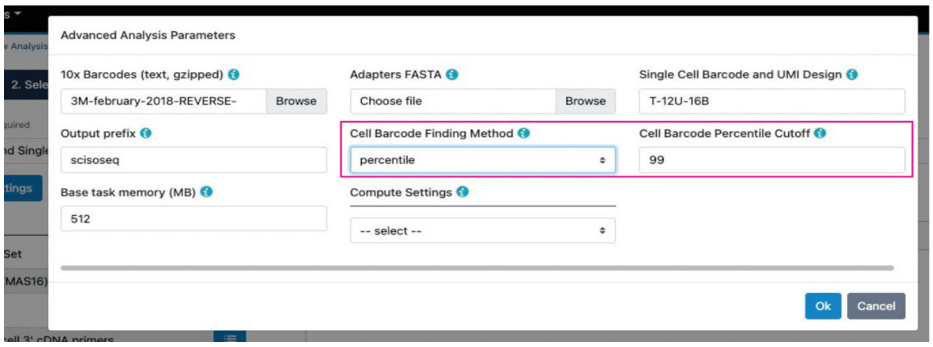
In most cases, the `knee` method is successful in estimating the number of real cells. Following are examples where the `knee` method was **not** successful, and the `percentile` method (with 97% or 99% cutoff) was used to achieve cell recovery.



Correct estimation of cells increases the number of usable FLNC.

Metric	Percentile	Knee
FLNC Reads Mapped Confidently to Genome	26,220,947	4,005,710
FLNC Reads Mapped Confidently to Transcriptome	9,105,973	2,089,836
Median Genes per Cell	239	3,888
Median Genes per Cell, known genes only	235	3,872
Median Transcripts per Cell	271	5,750
Median Transcripts per Cell, known transcripts only	269	5,726
Total Unique Genes	33,038	19,152
Total Unique Genes, known genes only	24,087	17,265
Total Unique Transcripts	336,099	118,066
Total Unique Transcripts, known transcripts only	326,154	116,022

SMRT Link v12.0 now supports the optional **percentile** method:



Possible issues when using the MAS-Seq for 10x Single Cell 3' kit for unsupported use cases

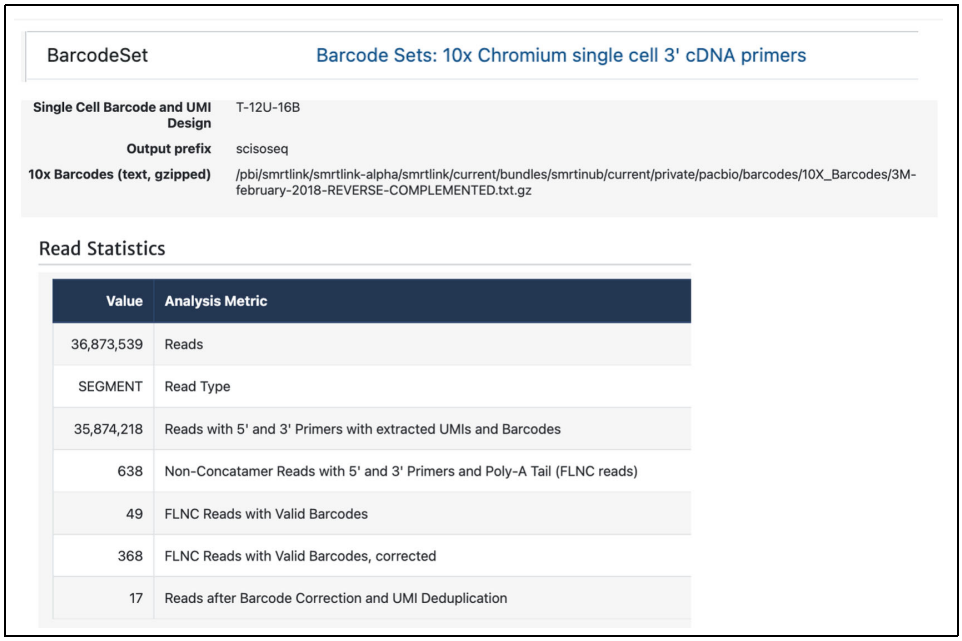
The following are **unsupported use cases** for the MAS-Seq kit that are commonly observed. Note that PacBio **cannot** offer official support for library preparations, sequencing, or analyses for use of MAS-Seq kit in unsupported scenarios including those described below. The unsupported use cases described herein have not been validated by PacBio® and are provided as-is and without any warranty. Use of these unsupported use cases is offered to those customers who understand and accept the associated terms and conditions and wish to take advantage of their potential for use of their samples for analysis using the PacBio system. If any of part of these unsupported use cases is to be used in a production environment, it is the responsibility of the end user to perform the required validation.

Observed issue	Likely cause	Solution
• Good S-read yield • Poor FLNC yield and beyond	Using the MAS-Seq kit with a 10x 5' library, often with further changes to library preparation based on this preprint.	Rerun the analysis using modified (1) cDNA primer; (2) cell barcode list; (3) barcode and UMI design.
• Good S-read and FLNC yield • Poor FLNC with barcodes and beyond.	Using the MAS-Seq kit with a Visium (spatial) library	Rerun the analysis using (1) cell barcode list; (2) barcode and UMI design.
• Poor S-read yield	Using SMRT Link with a homebrew method based on this preprint.	Rerun the analysis using modified segmentation adapter FASTA file.

Example unsupported use: MAS-Seq kit with 10x 5' library

In some cases, users can use the MAS-Seq kit to work with 10x 5' libraries by modifying the TSO depletion step with a custom oligo (not sold in kit). Changes to SMRT Link workflow parameters are **required**.

Proposed parameters for MAS 5' unsupported use case



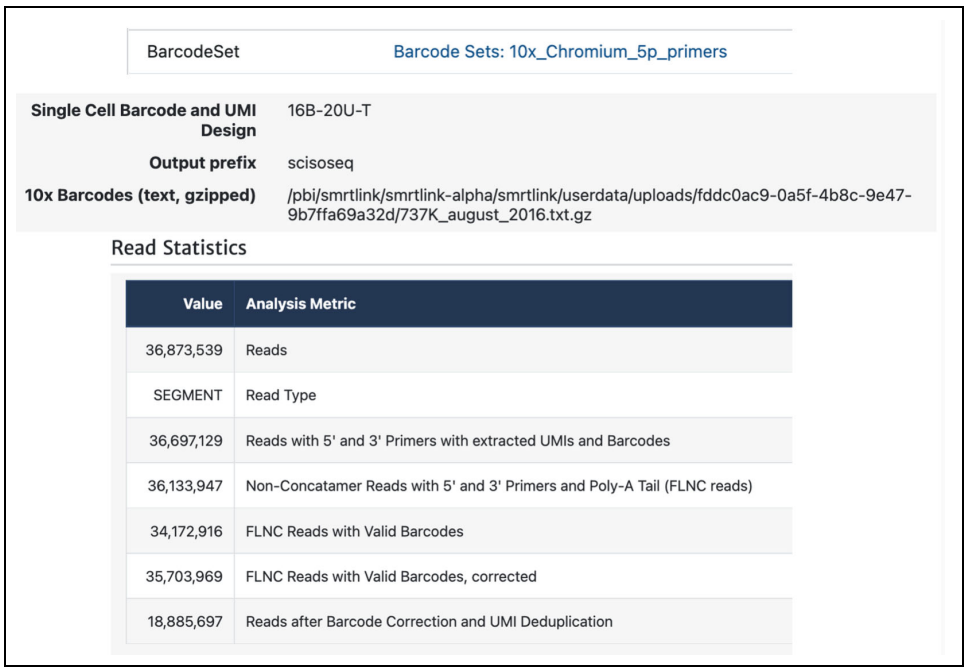
BarcodeSet — Barcode Sets: 10x Chromium single cell 3' cDNA primers

Single Cell Barcode and UMI Design	T-12U-16B
Output prefix	scisoseq
10x Barcodes (text, gzipped)	/pbi/smrtlink/smrtlink-alpha/smrtlink/current/bundles/smrtinub/current/private/pacbio/barcodes/10X_Barcodes/3M-february-2018-REVERSE-COMPLEMENTED.txt.gz

Read Statistics

Value	Analysis Metric
36,873,539	Reads
SEGMENT	Read Type
35,874,218	Reads with 5' and 3' Primers with extracted UMIs and Barcodes
638	Non-Concatamer Reads with 5' and 3' Primers and Poly-A Tail (FLNC reads)
49	FLNC Reads with Valid Barcodes
368	FLNC Reads with Valid Barcodes, corrected
17	Reads after Barcode Correction and UMI Deduplication

Proposed parameters for MAS 5’ unsupported use case

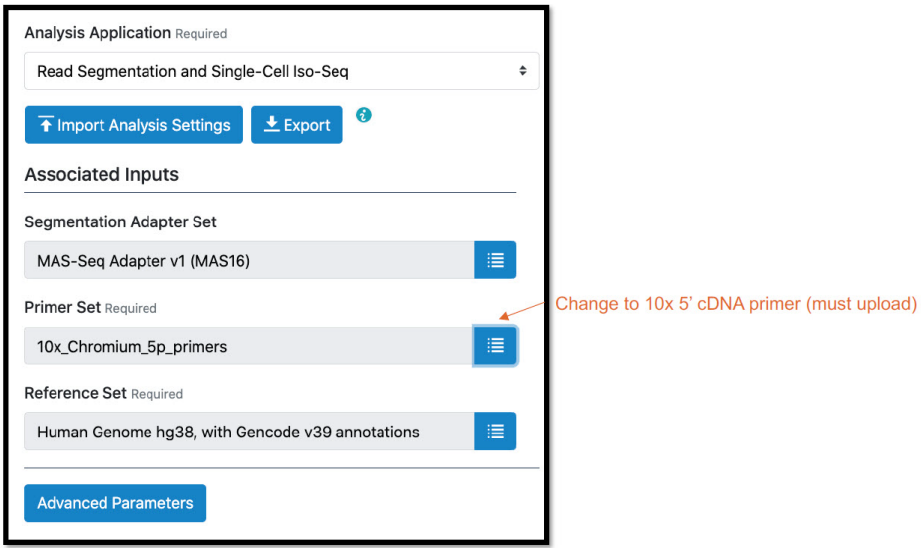


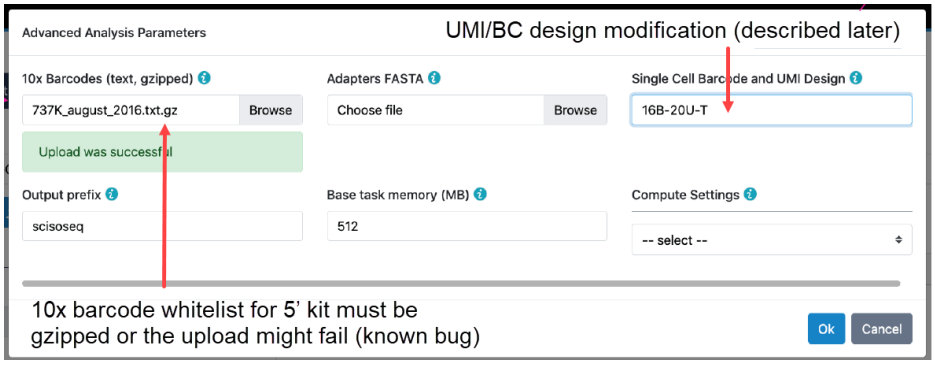
BarcodeSet	Barcode Sets: 10x_Chromium_5p_primers
Single Cell Barcode and UMI Design	16B-20U-T
Output prefix	scisoseq
10x Barcodes (text, gzipped)	/pbi/smrtlink/smrtlink-alpha/smrtlink/userdata/uploads/fddc0ac9-0a5f-4b8c-9e47-9b7ffa69a32d/737K_august_2016.txt.gz

Read Statistics

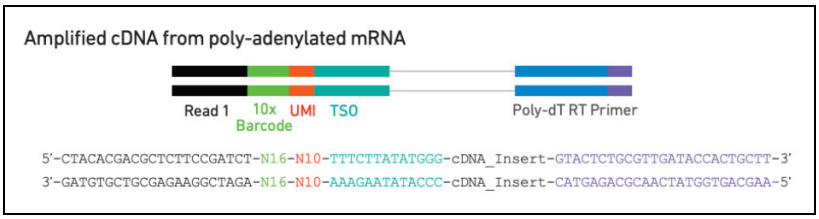
Value	Analysis Metric
36,873,539	Reads
SEGMENT	Read Type
36,697,129	Reads with 5' and 3' Primers with extracted UMIs and Barcodes
36,133,947	Non-Concatamer Reads with 5' and 3' Primers and Poly-A Tail (FLNC reads)
34,172,916	FLNC Reads with Valid Barcodes
35,703,969	FLNC Reads with Valid Barcodes, corrected
18,885,697	Reads after Barcode Correction and UMI Deduplication

Modifying SMRT Link to work with a 10x 5’ kit MAS-Seq run: Unsupported use case





UMI/BC modification for 10x 5' kit

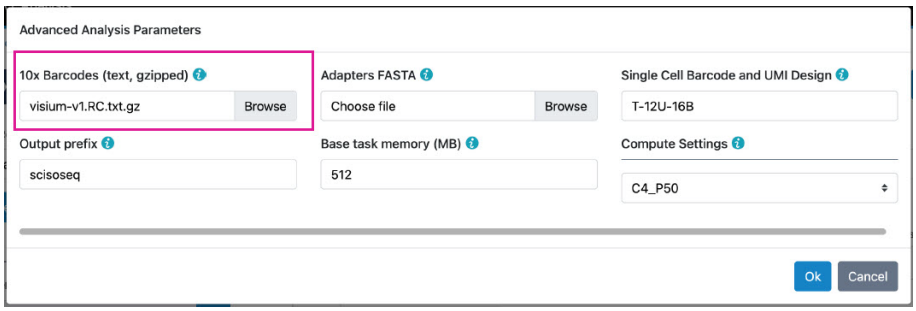


Technically for the 5' kit, it is 16bp BC + 10bp UMI + 10bp TSO. However, the TSO needs to be trimmed away. Here we have the UMI+TSO trimmed together as a 20 bp component with 16B-20U-T design.

Using SMRT Link v12.0 with a Visium sample

Visium samples have the exact same molecular structure as standard 10X 3' kit; the main inputs are identical to 3' analysis.

In the Read Segmentation and Single-Cell Iso-Seq's **Advanced Parameters** dialog, change **10x barcodes** to **Visium** barcodes (~5000 spots). Note that "cells" are basically spots if using SMRT Link to analyze Visium data.



Example unsupported use case: MAS-Seq kit with 10x Visium (spatial) library

The MAS-Seq kit can work directly with Visium libraries **without** modification. Only the SMRT Link parameters require changing.

Incorrect parameters for MAS Visium unsupported use case

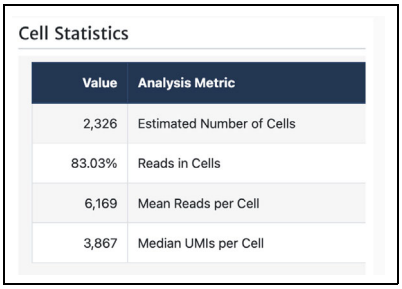
Warning	**Analysis experienced an error, but was able to recover and complete successfully.** `High Barcode errors: [isoseqs] barcode correction ALARM: Missing fraction %99 > threshold 25% (task: pb_sc_isoseq.isoseq_correct-0-a1)`

Analysis Parameters

Adapters FASTA	
Base task memory (MB)	512
Single Cell Barcode and UMI Design	T-12U-16B
Output prefix	scisoseq
10x Barcodes (text, gzipped)	/pbi/smrtlink/smrtlink-alpha/smrtlink/current/bundles/smrtinub/current/private/pacbio/barcodes/10X_Barcodes/3M-february-2018-REVERSE-COMPLEMENTED.txt.gz

When the barcode white list is incorrect, SMRT Link displays a warning in the barcode correction step.

Proposed parameters for MAS Visium unsupported use case



Cell Statistics

Value	Analysis Metric
2,326	Estimated Number of Cells
83.03%	Reads in Cells
6,169	Mean Reads per Cell
3,867	Median UMIs per Cell