PacBio

SMRT® Link Kinnex™ full-length RNA troubleshooting guide

P/N 103-552-100 Version 01 (November 2024)

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

SMRT® Link Kinnex full-length RNA troubleshooting guide (v13.1 and later)

Introduction

This document describes the metrics generated by the **Read Segmentation** and **Iso-Seq®** workflow in SMRT Link v13.1 and later.

- Example data sets 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 Iso-Seq as an Analysis workflow. For Kinnex full-length RNA kit users, **Read Segmentation and 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.)

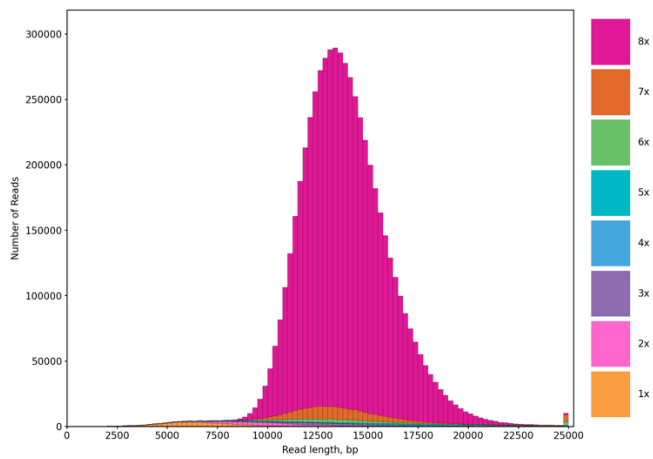
The Kinnex full-length RNA kit enriches for full (8-fold) arrays of cDNA typically generated using the Iso-Seq express 2.0 kit. Therefore, the percentage of full array and concatenation factors should have typical values as shown below.

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	Depends on sample type, but typically ~1.5–2 kb
Percent of Reads with Full Arrays	Percent of HiFi reads with full MAS arrays	85-90%
Mean Array Size	Concatenation factor	~7.xx

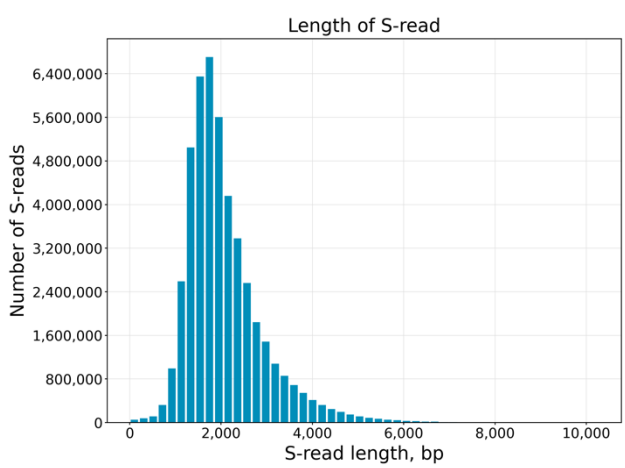
Read Segmentation

Value	Analysis Metric	
6,319,115	Reads	**Depends on sequencing yield**
46,298,636	Segmented reads (S-reads)	**Depends on HiFi yield and concatenation**
2,075	Mean length of S-reads	**Depends on sample type, usually ~2kb**
85.83 %	Percent of reads with full arrays	**More fully arrays indicate good Kinnex concatenation, usuaully ~85-90%**
7.33	Mean array size (concatenation factor)	**Usually ~7.xx for Kinnex full-length RNA kit**

A clean peak between 12,000–15,000 bp indicates good Kinnex array formation and successful enrichment of full arrays:



S-read read length should largely reflect the original cDNA library size, which can be highly sample dependent.



SMRT Link Iso-Seq workflow: Read classification

cDNA primers and polyA tails are removed from S-reads. The resulting full-length, non-concatemer (FLNC) reads are oriented to 5' -> 3' sense strand. It is also at this step that the sample-specific cDNA barcodes (as part of the cDNA primers) are extracted. These steps are performed using the command `lima` and `isoseq refine`.
(See here for the high-level workflow.)

Metric	Explanation	Typical value
Reads	Number of S-reads	Depends on sequencing yield
Reads with 5' and 3' primers	Full-length (FL) reads	>95% of reads should be FL
Non-concatemer reads with 5' and 3' primers and poly-A tail	Full-Length non-concatemer (FLNC) reads	>95% of reads should be FLNC
Mean length of FLNC reads	Mean length of FLNC reads	Dependent on cDNA size
Unique primers	Number of unique barcoded cDNA primers detected	If using Iso-Seq express 2.0 kit, up to 12 unique primers
Mean reads per primer	Mean number of reads per barcoded cDNA primer	Depends on number of multiplexed samples and total read depth
Max reads per primer	Maximum number of reads per barcoded cDNA primer	Depends on number of multiplexed samples and total read depth
Min reads per primer	Minimum number of reads per barcoded cDNA primer	Depends on number of multiplexed samples and total read depth
Reads without primers	Number of reads without a detected barcoded cDNA primer	<3% of reads
Percent bases in reads with primers	Percentage of bases in reads with barcoded cDNA primer	>95% of bases
Percent reads with primers	Percentage of reads with barcoded cDNA primer	>95% of reads

Example of a Read Classification summary for a 12-plex Kinnex full-length RNA dataset:

Value	Analysis Metric
59,799,238	Reads
57,734,211	Reads with 5' and 3' Primers
57,667,867	Non-Concatamer Reads with 5' and 3' Primers
57,517,651	Non-Concatamer Reads with 5' and 3' Primers and Poly-A Tail (FLNC Reads)
1,668	Mean Length of FLNC Reads
12	Unique Primers
4,811,184	Mean Reads per Primer
8,757,463	Max. Reads per Primer
393,810	Min. Reads per Primer
2,065,027	Reads without Primers
96.12%	Percent Bases in Reads with Primers
96.54%	Percent Reads with Primers

SMRT Link Iso-Seq workflow: Transcript clustering

FLNC reads from the previous step are clustered at the isoform level to generate high-quality (HQ) transcript sequences. This is performed using the `isoseq cluster2` command. (See here for information.)

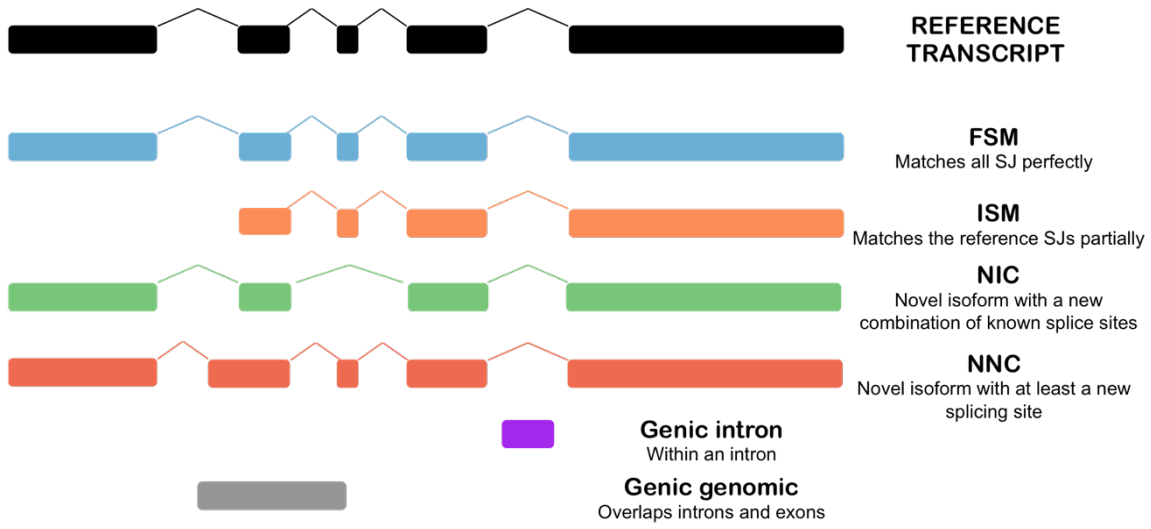
Note that the number of HQ transcripts are not necessarily a good indication of the number of unique genes or isoforms as library artifacts have not yet been filtered, mapped to genome, and a comparison against existing annotations have not been done. These steps are done in the subsequent part of the workflow, Transcript mapping and classification.

Metric	Explanation	Typical value
Number of high-quality isoforms	Number of high-quality isoforms based on isoform-level clustering of FLNC reads	Depends on sample complexity. Not a good indicator of final number of genes and isoforms.

SMRT Link Iso-Seq workflow: Transcript mapping and classification

HQ transcripts are aligned to the reference genome then classified and filtered against a reference annotation using `pigeon` software (which is the PacBio implementation of SQANTI3). This is performed using the command `pbmm2/isoseq3 collapse/pigeon`. (See here for information.)

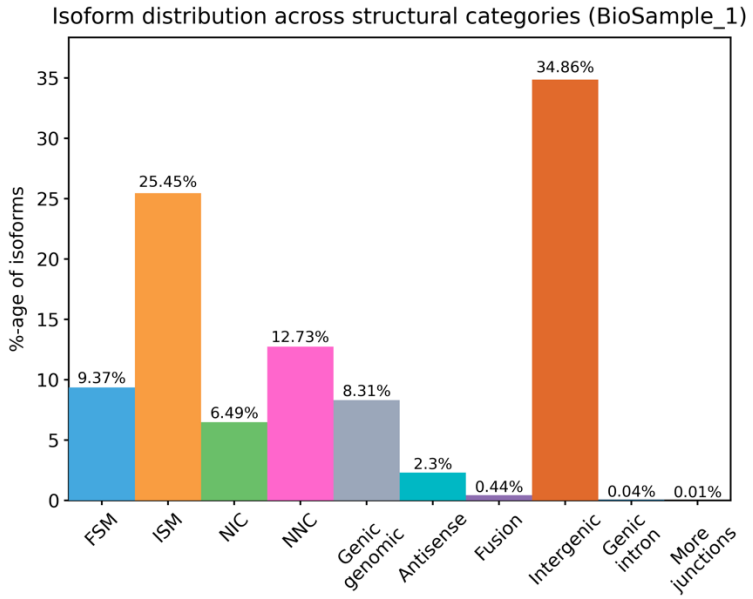
A full description of the SQANTI3 (pigeon) classification types can be found in the SQANTI3 GitHub wiki as well as the publication. They are described briefly below.



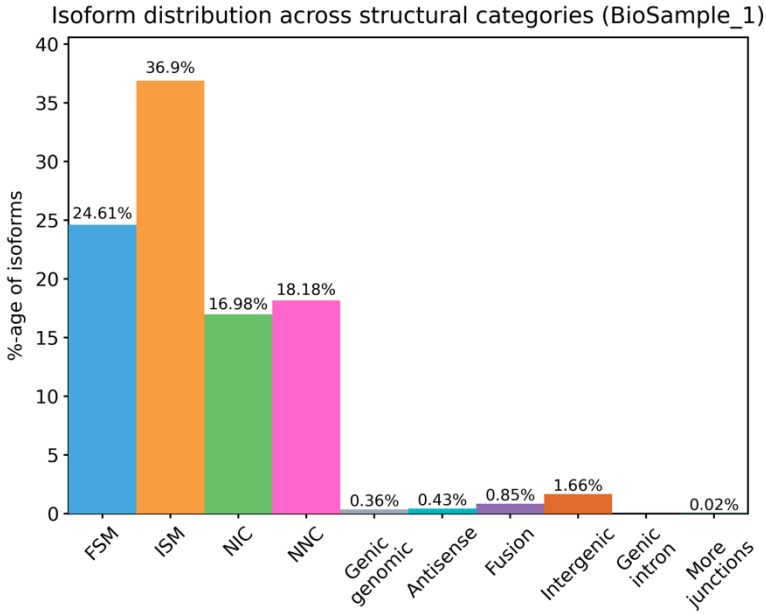
Isoform classification	Explanation	Typical value in human sample (after *pigeon filter*)
FSM: full-splice match	Matches a reference transcript for all junctions. May differ in start or end with reference.	Sample dependent, but typically >15% after filtering
ISM: Incomplete splice match	Matches a reference transcript for a subset of internal junctions. May differ in start or end with reference.	Sample dependent, but typically <40% after filtering
NIC: Novel In Catalog	No match to reference transcript but uses a combination of known donor and acceptor sites	Sample dependent, but often range from 5–30% after filtering
NNC: Novel Not In Catalog	No match to reference transcript and at least one donor or acceptor site is novel	Sample dependent, but often range from 5–30% after filtering
Genic intron	Isoform is completely contained within intron of a reference transcript	Sample dependent, but typically <1% after filtering
Genic genomic	Isoform overlaps some intronic and exonic region of a reference transcript	Sample dependent, but typically <1% after filtering
Antisense	No match to reference transcript but overlaps on the opposite strand	Sample dependent, but typically <1% after filtering
Fusion	Matches multiple reference transcripts usually on adjacent genes (readthrough)	Sample dependent, but typically <1% after filtering
More junctions	Similar to fusion, but one or more junctions are shared by multiple genes	Sample dependent, but typically <1% after filtering
Intergenic	No match to reference transcripts	Sample dependent, but typically <1% after filtering

Pigeon filtering will often remove many library artifacts (see Appendix 1). Therefore, isoform classification percentages can change before and after filtering.

As an example, the figure below is the isoform classification for a human cerebellum sample before filtering:



Below are the percentages after filtering, where most of the intergenic and genic genomic artifacts have been removed.



The post-filtering percentages for different isoform categories can vary greatly based on RNA quality, sequencing depth, sample type, and annotation. Here are some of the considerations:

- At lower read depths (e.g., high sample multiplexing), it is likely to see more known isoforms (FSM, ISM) and fewer novel isoforms (NIC, NNC) since highly abundant isoforms tend to be annotated.
- A high percentage of ISM (>40%) might indicate either lower sample quality (5' RNA degradation) or sample specificity.
- After filtering, the percentage of non-annotated genes (e.g., intergenic, genic intron) tends to be <1%.

Note that the Iso-Seq workflow will exclude true fusion (inter-chromosomal) transcripts as the default parameters exclude reads that map to multiple loci. To identify fusion transcripts, refer to long-read fusion tools described in the tertiary analysis guide.

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

A full description of the files available for download can be found in the SMRT Link User Guide. Briefly, the post-pigeon filtered GFF and classification files are the most common starting place for downstream analysis.

File	Size	Type
High-Quality Isoforms (All Samples)	3 GB	Fasta
Isoform Counts by Barcode (All Samples)	35 MB	csv
Mapped High-Quality Isoforms (All Samples)	1 GB	bam
Mapped High-Quality Isoforms (BAM Index) (All Samples)	3 MB	bam_bai
SMRT Link Log	35 KB	log
Unique mapped transcripts, GFF (All Samples)	643 MB	gff
Unique mapped transcripts, classification TXT (All Samples)	361 MB	txt
Unique mapped transcripts, filtered, GFF (All Samples)	373 MB	gff
Unique mapped transcripts, filtered, classification TXT (All Samples)	124 MB	txt
Unique mapped transcripts, filtered, junctions TXT (All Samples)	383 MB	txt
Unique mapped transcripts, junctions TXT (All Samples)	514 MB	txt

Post pigeon-filter GFF and files

Tertiary analysis recommendations

For a list of community tools supporting annotation, quantification, differential analysis, fusion detection, variant calling, phasing, visualization, etc., please consult the full-length isoform sequencing bioinformatics Application note.

Appendix 1: Common cDNA library artifacts and pigeon filtering

There are many different kinds of cDNA library artifacts as part of the cDNA synthesis process, which are extensively described in Verwilt, Mestdagh, and Vandesompele (2023). The two most common types of cDNA artifacts that can be readily detected and bioinformatically removed are **intrapriming** and **RT switching artifacts**.

Intrapriming artifacts arise from mis-priming off stretches of 'A's that are part of the transcript but not the poly(A) tail. As a result, the cDNA sequence would appear to represent a novel transcript that has a novel 3' end. This can be detected by mapping an Iso-Seq transcript to the genome and searching for stretches of 'A's immediately downstream of the transcript end. Pigeon labels a transcript as an intrapriming artifact if there are >60% of the 'A's or 6 consecutive 'A's found in the 20 bp downstream of the reported TTS at the genomic level. These can be adjusted with the `--poly-a-percent` and `--poly-a-runlength` parameter for `pigeon filter`.

RT artifacts, or template switching artifacts, happen as part of the intrinsic property of reverse transcriptase (RT) that allows them to jump within or across template positions without terminating DNA synthesis. Secondary structures in RNA templates can enhance RT switching activity and cause gaps during cDNA synthesis. As a result, the cDNA sequence would appear to have a new splicing event, which most commonly will represent a novel intron with noncanonical junctions (that is, not GT-AG or GC-AG). The original SQANTI1 paper (Tardaquila et al. 2018) referenced Cocquet et al. (2006) and Houseley and Tollervey (2010) which described this phenomenon. SQANTI, and therefore pigeon, both identify the hallmark of RT switching based on a specific sequence signature (the presence of a direct repeat between the upstream mRNA boundary of the noncanonical intron and the intron region adjacent to the downstream exon boundary) and label a transcript as an RT switching artifact if one or more of the splicing junctions has this sequence signature.

More sophisticated machine learning-based filtering approaches are employed by the academic SQANTI3 software but are not implemented in pigeon. Note that the `pigeon filter` uses the following rules:

A transcript is filtered if:

(1) It is an intrapriming artifact, or
(2) The parameter `--mono-exon` is used and it is mono-exonic, or
(3) It is not a full-splice match (FSM) and is (i) an RT switching artifact; and/or (ii) was given a matching orthogonal junction BED file (via `pigeon classify --coverage`) and one or more of the junctions is non-canonical and has less than the minimum junction support as specified by the `--min-cov` parameter in `pigeon filter`.