Library Prep and Bioinformatics Improvements for Full-Length Transcript Sequencing on the PacBio Sequel System

Michelle Vierra, Brendan Galvin, Ting Hon, Heather Ferrao, Armin Töpfer, Yuan Li, Elizabeth Tseng
PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

Abstract

The PacBio Iso-Seq method produces high-quality, full-length transcripts of up to 10 kb and longer and has been used to annotate many important plant and animal genomes. Here we describe an improved, simplified library workflow and analysis pipeline that reduces library preparation time, RNA input, and cost.

The Iso-Seq V2 Express workflow is a one day protocol that requires only ~300 ng of total RNA input while also reducing the number of reverse transcription and amplification steps down to single reactions. Compared with the previous workflow, the Iso-Seq V2 Express workflow increases the percentage of full-length (FL) reads while achieving a higher average transcript length.

At the same time, the Iso-Seq 3 analysis recently released in the SMRT Link 6.0 software is a major improvement over previous versions. Iso-Seq 3 is highly accurate at detecting and removing library artifacts (TSO and RT artifacts) as well as differentiating barcodes on multiplexed samples. Iso-Seq 3 achieves the same output performance in high-quality transcript sequences compared to previous versions while reducing the runtime and memory usage dramatically.

Iso-Seq V2 Library Workflow

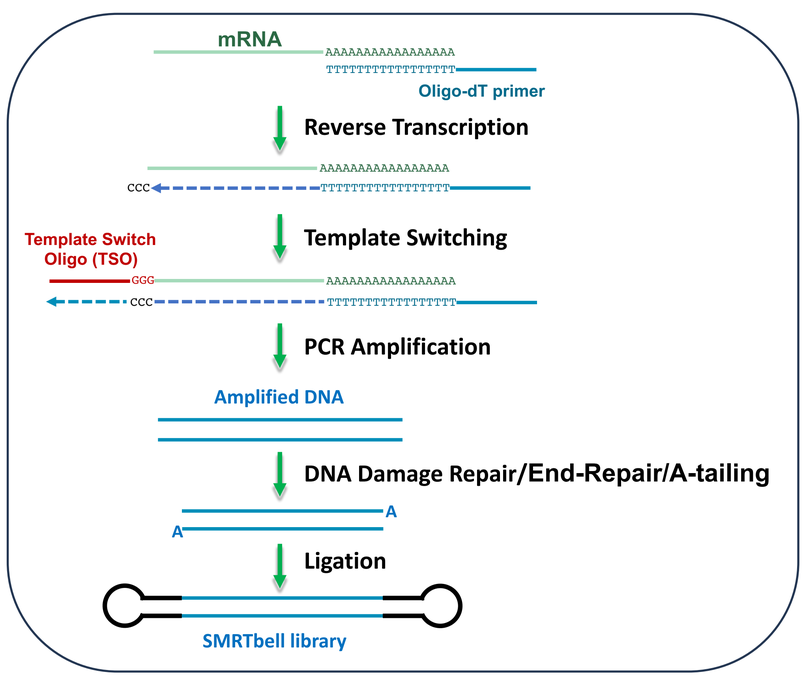

Figure 1. Iso-Seq V2 Workflow. Full-length mRNA is converted into cDNA using the NEBNext Single Cell/Low Input RNA Library Prep Kit followed by PCR amplification. The amplified cDNA is converted into SMRTbell templates using the PacBio SMRTbell Express Template Prep Kit v2 for sequencing on the Sequel System.

Iso-Seq V2 Library Workflow Features:

- RNA to SMRTbell library in 1 day
- ~300 ng total RNA input requirement
- Single RT and amplification reactions
- Increased full-length transcript recovery

	Iso-Seq Current	Iso-Seq V2
Sample	UHRR*	UHRR
Reads	486,997	506,840
FL Reads	361,253 (74%)	433,464 (86%)
FL Read Lengths	2,194 bp	2,901 bp

Table 1. Comparison between current Iso-Seq analysis protocol and Iso-Seq V2. UHHR – Universal Human Reference RNAs; Data for V2 run on Sequel System chemistry 3.0 and software 6.0

Iso-Seq 3 Bioinformatics Workflow

The Iso-Seq analysis workflow begins with the generation of high-fidelity reads using the circular consensus sequencing (CCS) method on a per molecule basis. Then, full-length reads are selected and trimmed of 5' and 3' primers and poly-A tails. The trimmed full-length reads are clustered at the isoform level and consensus is called. Lastly, the consensus isoforms can be optionally mapped back to the reference genome and/or used in downstream analysis.

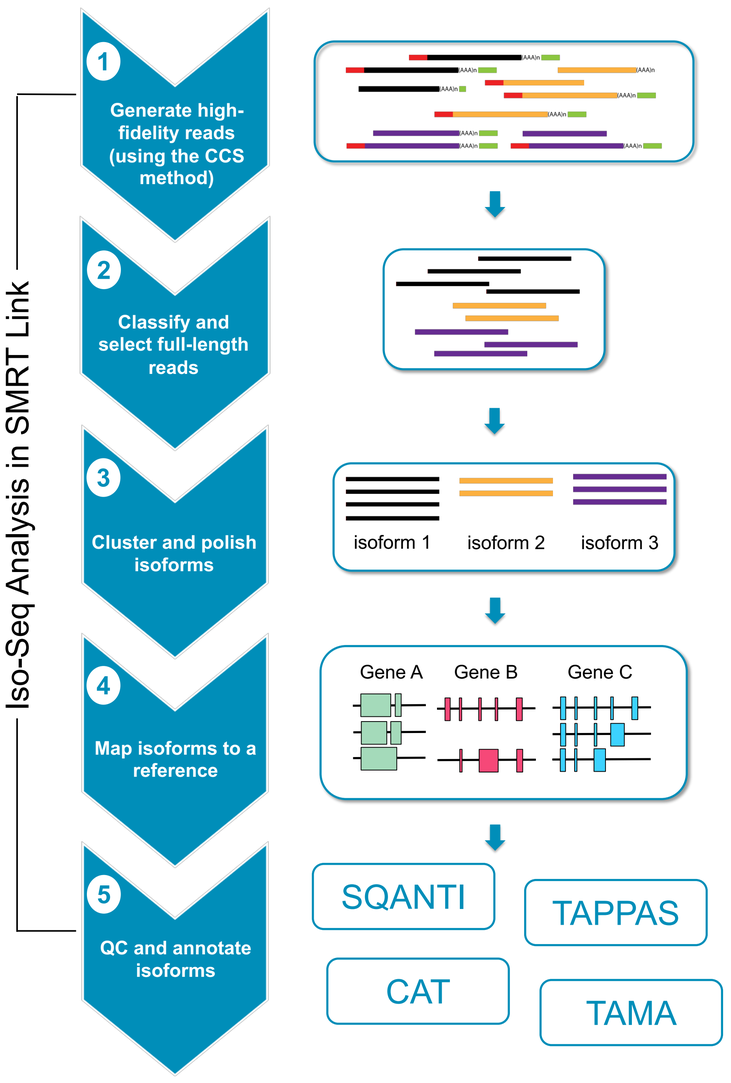

Figure 2. Iso-Seq 3 Bioinformatics Workflow. Analysis pipeline outlined conceptually (left) and graphically (right) to demonstrate FL-read inputs and isoform outcomes of the improved workflow.

Iso-Seq 3 Analysis Pipeline Features:

- Reduced runtime
- Increased de-multiplexing accuracy
- Increased artifact detection
- Compatible with whole and targeted transcriptome experimental designs

	Mouse Liver	Human UHRR
SMRT Cells	1	3
Complexity	Low	High
Runtime	5 hr	13 hr
Speed up (compared to Iso-Seq 1)	5X	6X

Table 2. Runtime comparison between versions of Iso-Seq analysis. For both low and high complexity samples, Iso-Seq 3 analysis results in a ≥5-fold increase in speed compared to previous versions.

Tool Availability and Training:

Iso-Seq 3 and Iso-Seq 3 with mapping are available through SMRT Link 6.0[1] and Bioconda[2]. Additional tutorials on PacBio wesbite[3].

SQANTI for QC

SQANTI is a pipeline for the in-depth characterization of isoforms obtained by full-length transcript sequencing without information about gene/transcript annotation or attribute description. SQANTI provides a wide range of descriptors of transcript quality and generates a graphical report to aid in the interpretation of the sequencing results.[4]

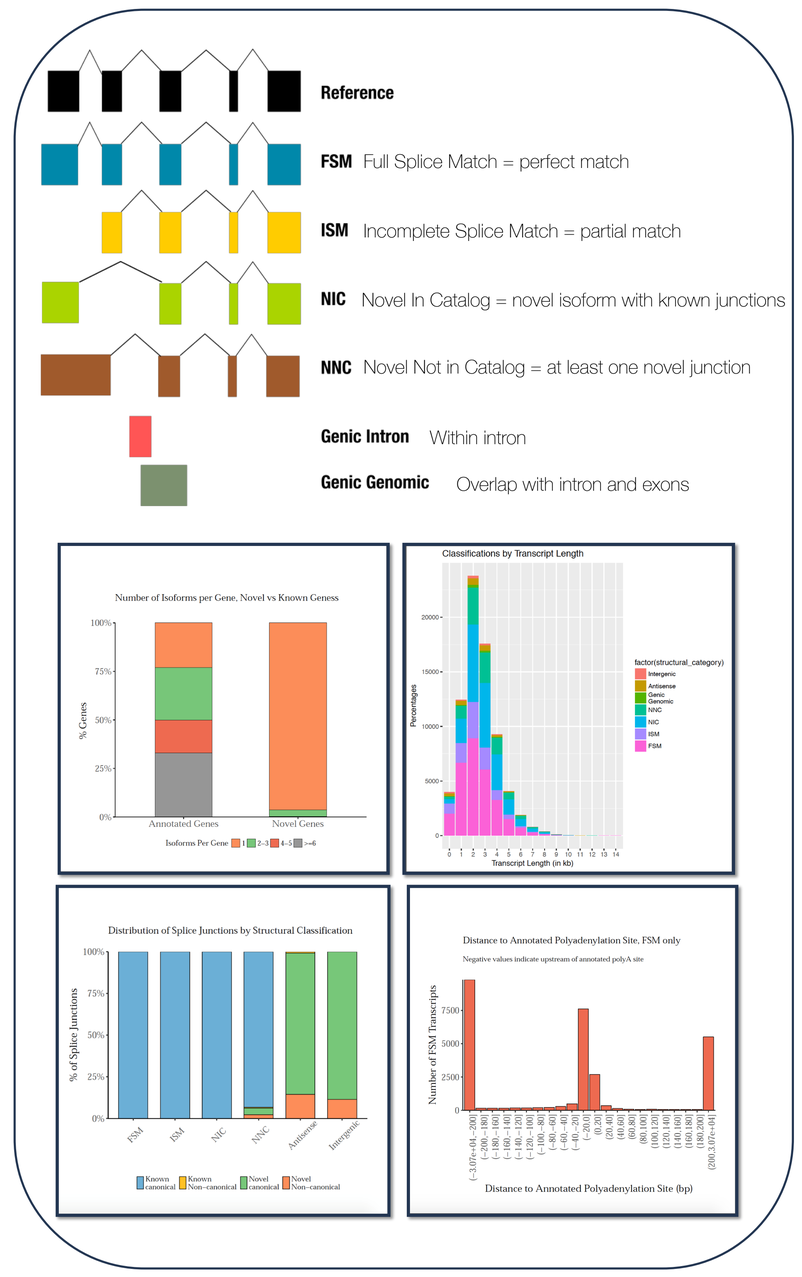

Figure 3. SQANTI Software compares the Iso-Seq analysis output against a reference annotation.

Iso-Seq Community Tools

- **TAMA** for genome annotation (https://github.com/GenomeRIK/tama)
- **Cupcake** for genome annotation (https://github.com/Magdoll/cDNA_Cupcake)
- **Cogent** for genome-free gene clustering (https://github.com/Magdoll/Cogent)
- **SQANTI** for quality control (https://bitbucket.org/ConesaLab/sqanti/)
- **TAPPAS** for functional annotation (http://tappas.org/)
- **CAT** for cross-species annotation (https://github.com/ComparativeGenomicsToolkit/)

References

1. https://www.pacb.com/software
2. https://github.com/PacificBiosciences/pbbioconda
3. https://github.com/PacificBiosciences/IsoSeq_SA3nUP
4. https://bitbucket.org/ConesaLab/sqanti/