Scalability and Reliability Improvements to the Iso-Seq Analysis Pipeline Enables Higher Throughput Sequencing of Full-Length Cancer Transcripts

Elizabeth Tseng, Ting Hon, Meredith Ashby
PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

Abstract #7502

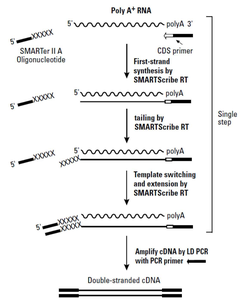
Abstract

The characterization of gene expression profiles via transcriptome sequencing has proven to be an important tool for characterizing how genomic rearrangements in cancer affect the biological pathways involved in cancer progression and treatment response. More recently, better resolution of transcript isoforms has shown that this additional level of information may be useful in stratifying patients into cancer subtypes with different outcomes and responses to treatment.[1] The Iso-Seq protocol developed at PacBio is uniquely able to deliver full-length, high-quality cDNA sequences, allowing the unambiguous determination of splice variants, identifying potential biomarkers and yielding new insights into gene fusion events.

Recent improvements to the Iso-Seq bioinformatics pipeline increases the speed and scalability of data analysis while boosting the reliability of isoform detection and cross-platform usability. Here we report evaluation of Sequel Iso-Seq runs of human UHRR samples with spiked-in synthetic RNA controls and show that the new pipeline is more CPU efficient and recovers more human and synthetic isoforms while reducing the number of false positives. We also share the results of sequencing the well-characterized HCC-1954 breast cancer and normal breast cell lines, which will be made publicly available. Combined with the recent simplification of the Iso-Seq sample preparation[2], the new analysis pipeline completes a streamlined workflow for revealing the most comprehensive picture of transcriptomes at the throughput needed to characterize cancer samples.

Iso-Seq Sample Preparation Methods

Clontech SMARTer PCR cDNA Synthesis Kit



Streamlined Workflow for Iso-Seq Sample Preparation on the Sequel System

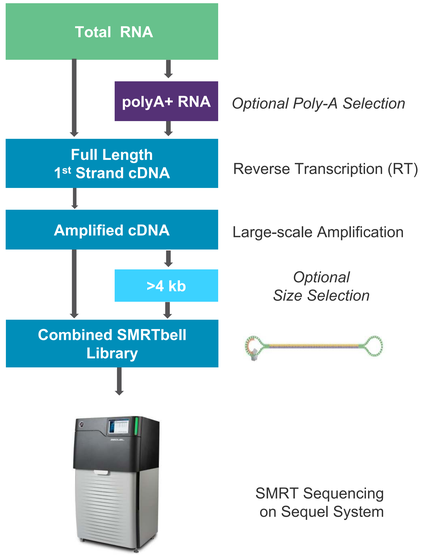


Figure 1. Preparation of HCC-1954 and normal human breast tissue samples. 1ug of total RNA per reaction was reverse transcribed using the Clontech SMARTer cDNA synthesis kit, with four reactions processed in parallel. PCR optimization was used to determine the cycle number for the downstream, large-scale amplification reactions. A single primer (primer IIA from the Clontech SMARTer kit 5' AAG CAG TGG TAT CAA CGC AGA GTA C 3') was used for all post-RT PCR.

Large scale PCR products were purified with either 1X or 0.4X AMPure PB beads (1X or 0.4X cDNA library hereafter) and a Bioanalyzer was used for QC. Taking advantage of the simplified sample prep workflow, equimolar ratios of the 1X and 0.4X cDNA libraries were pooled together to boost representation of long isoforms while avoiding laborious size selection steps. Two SMRT Cells per sample were sequenced on the PacBio Sequel platform using 2.1 chemistry and 10 hour movies.

Optimizing Iso-Seq Analysis for Higher Throughput

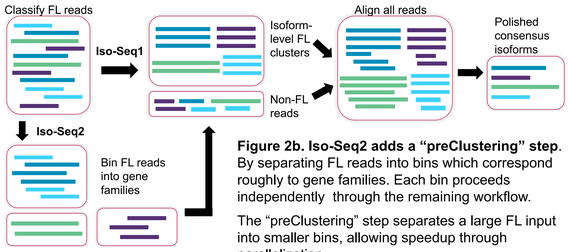


Figure 2a. The complete Iso-Seq analysis workflow involves three steps:

CCS: Getting CCS (circular consensus sequence) reads out of subreads BAM file.

Classify: Identifying full-length (FL) CCS reads based on cDNA primers and polyA tail signal.

Cluster: Isoform-level clustering of full-length reads followed by polishing with all reads to generate high-quality, full-length, consensus isoform sequences.

Figure 2b. Iso-Seq2 adds a "preClustering" step. By separating FL reads into bins which correspond roughly to gene families. Each bin proceeds independently through the remaining workflow.

The "preClustering" step separates a large FL input into smaller bins, allowing speedup through parallelization.

Comparison to Ground Truth

Job	SMRT Cells	FL Reads	Non-FL Reads	Iso-Seq2 Runtime*	Iso-Seq1 Runtime*
Cell 1	1	218,199	128,466	5 hr	27 hr
Cells 1 + 2	2	375,241	439,323	7 hr	20 hr
Cells 1 + 2 + 3	3	623,456	673,209	13 hr	42 hr

* Excludes runtime for CCS and classify

Table 1. Human reference (UHRR) sequel data . Iso-Seq2 was run using 12 CPUs x 20 nodes. Iso-Seq1 was run on SMRT Link job using 24 CPUs x 24 nodes. By speeding up the isoform-level clustering step 3 to 5-fold, the Iso-Seq2 pipeline can now handle the scale of data produced by multi-SMRT Cell Sequel experiments.

Table 2. Recovery of Lexogen SIRV transcripts from UHRR data set. SIRVs are synthetic spike-in variants that enable evaluation of sequencing data against ground truth samples. This SIRV mix consists of 68 isoforms from 7 synthetic gene families, ranging from 100 bp – 2.5 kb. A SIRV counts as a TP recovery if it is seen in FL at least twice with GMAP alignment ≥ 99% coverage and ≥ 95% identity. All internal junctions must match the annotation (max 5 bp diff) in every exon.

Iso-Seq2 recovered more annotated SIRV transcripts (TP; true positive) and produced less artifacts (FP; false positive)

Job	Iso-Seq2 TP	Iso-Seq2 FP	Iso-Seq1 TP	Iso-Seq1 FP
Cell 1	57	10	52	13
Cell 2	54	17	54	15
Cell 3	61	14	58	17

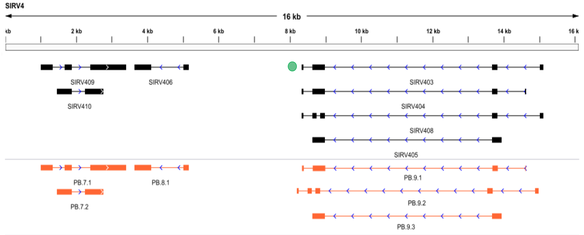


Figure 2. SIRV recovery example. Data is from the 3 cell analysis and visualized with IGV. The SIRV annotation in in black; Iso-Seq2 recovered transcripts in orange. The missed SIRV403 transcript (FN) is marked with a green dot.

Tumor / Normal Sequencing Results

Table 3. Summary of sequencing results. The human breast cancer sample is RNA from HCC-1954 available from ATCC. The control samples is normal human mammary gland pooled from a 27-year-old Caucasian female, available from Clontech.

	Cancer, Breast	Normal, Breast
SMRT Cells	2	2
No. of ZMWs	462,315	470,500
Full-Length Reads	171,444	187,161
Avg. FL Read Lengths	3.5 kb	3.6 kb

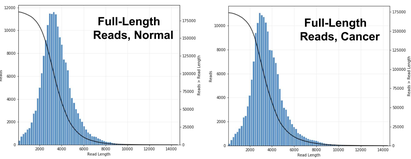
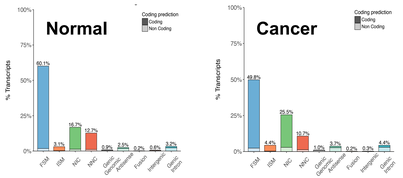


Figure 3 (above). Read length distribution. The simplified Iso-Seq sample preparation workflow produces full-length transcripts from less than 1 kb up to 10 kb, without labor intensive size selection steps.

Figure 4 (right). Isoform Categorization against Gencode v27 annotation using SQANTI. FSM (Full Splice Match) = fully match a transcript for each consecutive junction. ISM (Incomplete Splice Match) = consecutive, but not full, usage of junctions from a known transcript. NIC (Novel In Catalog) = novel isoform using a combination of known junctions. NINC (Novel Not In Catalog) = novel isoform containing novel junctions. Iso-Seq2 faithfully recovers known splice variants while also revealing new combinations of known exons and completely novel isoforms.



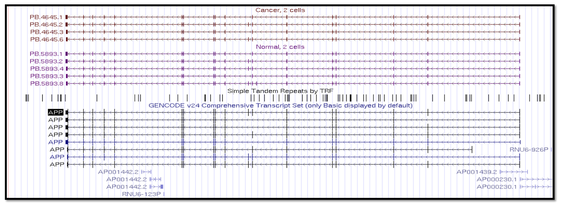


Figure 5. Complex alternative splicing in APP gene. Difference in exon skipping and 3' UTR length are observed in both samples.

Summary and Resources

- The Iso-Seq2 data analysis pipeline speeds analysis to meet the needs of researchers doing full-length transcript sequencing with Sequel while simultaneously improving accuracy, as measured by synthetic spike-in variants.
- The Iso-Seq sample preparation workflow following by sequencing on the Sequel platform yields full-length transcripts across the entire expected size range without laborious gel size selection steps
- PacBio Iso-Seq method generates full-length transcript sequences without the need for assembly of short fragments
- The Iso-Seq method is a powerful tool in the study of cancer providing full-length isoforms, alternative splicing information, and the ability to identify fusion genes.

More information on full-length transcript sequencing (Iso-Seq Application) can be found on the PacBio website: http://pacb.com/isoseq

References

1. Kohli, M, et al. (2017). Kohli, M. et al., 2017. Androgen receptor variant AR-V9 is co-expressed with AR-V7 in prostate cancer metastases and predicts abiraterone resistance. Clin Cancer Res; 1–12.
2. Ashby, et. al. (2017). Simplified Sequencing of Full-Length Isoforms in Cancer on the PacBio Sequel Platform. AACR 2017 General Meeting.