Simplified Sequencing of Full-Length Isoforms in Cancer on the PacBio Sequel Platform

Meredith Ashby, Ting Hon, Elizabeth Tseng, and Tyson A. Clark

PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

Abstract

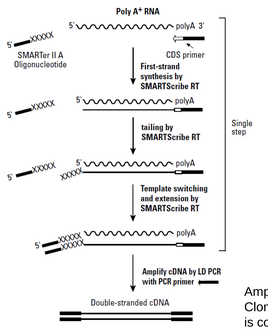
The aberrant transcription and expression of alternative RNA isoforms has been observed in multiple types of cancer, and is hypothesized to contribute to oncogenesis in certain cancer subtypes. Identification and annotation of cancer-specific mRNA isoforms is critical to understanding how mutations in the genome affect the biology of cancer cells. While microarrays and other NGS-based methods have become useful for studying transcriptomes, these technologies yield short, fragmented transcripts that remain a challenge for accurate, complete reconstruction of splice variants. In cancer proteomics studies, the identification of biomarkers from mass spectroscopy data is often limited by incomplete gene isoform expression information to support protein to transcript mapping

The Iso-Seq protocol developed at PacBio offers the only solution for direct sequencing of full-length, single-molecule cDNA sequences needed to discover biomarkers for early detection and cancer stratification, to fully characterize gene fusion events, and to elucidate drug resistance mechanisms. Knowledge of the complete isoform repertoire is key for accurate quantification of isoform abundance. As most transcript sizes range from 1 – 10 kb, fully intact RNA molecules can be sequenced using SMRT Sequencing without requiring fragmentation or post-sequencing assembly. However, some cancer research applications have presented a challenge for the Iso-Seq protocol, due to the combination of limited sample input and the need to deeply sequence heterogenous samples.

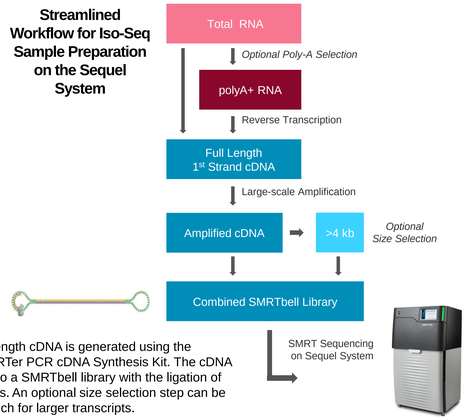
Here, we report the optimization of the Iso-Seq library preparation protocol for the PacBio Sequel platform and its application to cancer cell lines and tumor samples. We demonstrate how loading enhancements on the higher-throughput Sequel instrument have decreased the need for size-fractionation steps, reducing sample input requirements while simultaneously simplifying the sample preparation workflow and increasing the number of full-length transcripts per SMRT Cell. The results highlight the potential for broader application of the Iso-Seq method to more comprehensively characterize alternative splicing in cancer.

Iso-Seq Sample Preparation Methods

Clontech SMARTer PCR cDNA Synthesis Kit



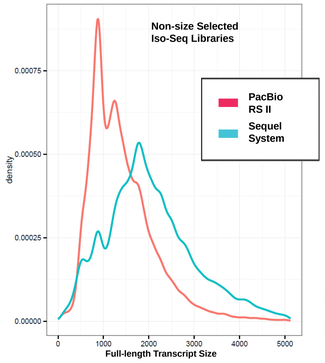
Streamlined Workflow for Iso-Seq Sample Preparation on the Sequel System

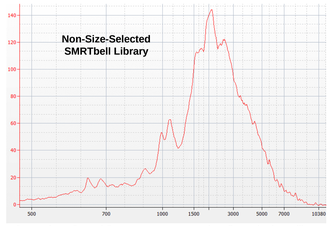


Amplified full-length cDNA is generated using the Clontech SMARTer PCR cDNA Synthesis Kit. The cDNA is converted into a SMRTbell library with the ligation of hairpin adapters. An optional size selection step can be included to enrich for larger transcripts.

Optimizing the Iso-Seq Application on the Sequel Platform

Magbead-loaded Samples on Sequel System Have Decreased Loading Bias

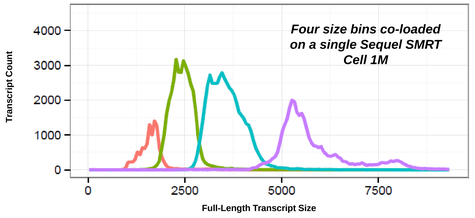




Output from Sequel System more accurately reflects the input library

A non-size selected Iso-Seq library was run on both the PacBio RS II and Sequel System. A histogram of number of full-length sequences by transcript length is plotted on the left. The full-length cDNA sequences run on the Sequel System closely resembles the size distribution of the input SMRTbell library, as shown in the bioanalyzer trace of the non-size selected library above.

Multiple Size-selected Libraries Can Be Co-loaded on the Sequel System



Size-selected Iso-Seq libraries were combined and run together on a single Sequel SMRT Cell 1M. Amplified cDNA had been size fractionated into four size bins (1-2 kb, 2-3 kb, 3-6 kb, and 5-10 kb) using the Sage Science BluePippin system, per the RSII Iso-Seq protocol. Barcoded SMRTbell libraries were made from each of the four size bins, to enable deconvolution post sequencing. A histogram of the full-length transcript lengths from each of the four size bins is shown at left, demonstrating that all four size bins are well represented when co-loaded. To boost representation of the longest isoforms, customers may optionally size select cDNA > 4 kb and pool at a non 1:1 ratio.

Sequencing of NALM6 Precursor B-ALL Cell Line on Sequel

Typical Sequencing Statistics From a Single Sequel SMRT Cell 1M

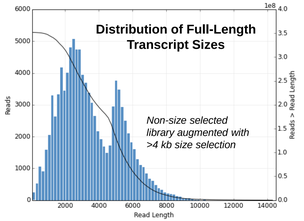
Transcript Classification

Value	Analysis Metric
595,533	Number of consensus reads
294,535	Number of five prime reads
338,803	Number of three prime reads
323,415	Number of poly-A reads
230	Number of filtered short reads
348,362	Number of non-full-length reads
246,941	Number of full-length reads
244,521	Number of full-length non-chimeric reads
738,399,613	Number of full-length non-chimeric bases
3,019	Mean full-length non-chimeric read length

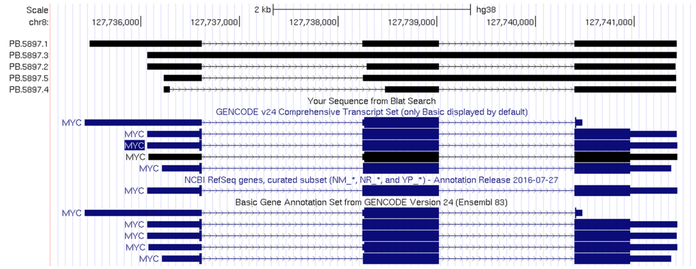
Transcript Clustering

Value	Analysis Metric
143,392	Number of unpolished consensus isoforms
17,765	Number of polished high-quality isoforms
125,467	Number of polished low-quality isoforms
3,196	Mean unpolished consensus isoforms read length

Data from a single, optimally loaded SMRT Cell processed in SMRT Analysis using the Iso-Seq analysis pipeline.

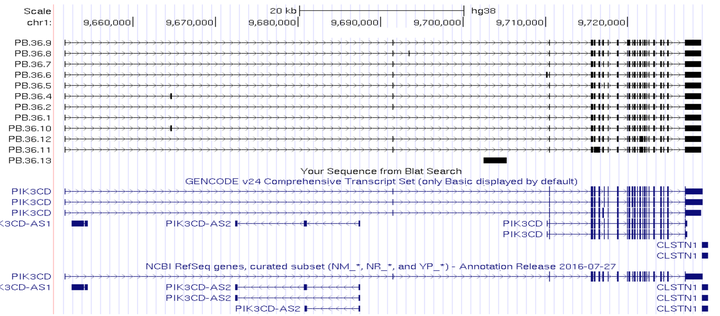


Alternative Splicing in Cancer



Example 1. MYC isoforms from NALM6 cell line.

Example 2. PIK3CD isoforms from NALM6 cell line.



Summary and Resources

- PacBio Iso-Seq method generates full-length transcript sequences without the need for assembly of short fragments
- Decreased loading bias on the Sequel System allows for a simplified Iso-Seq sample prep workflow that does not require multiple size-selection steps
- Barcoding during the cDNA generation allows multiplexing samples in a single SMRTbell library
- Improved throughput on the Sequel instrument increases the number of full-length transcripts per SMRT Cell.
- The Iso-Seq method is a powerful tool in the study of cancer providing full-length isoforms, alternative splicing information, and the capability 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