Full-Length cDNA Sequencing for Genome Annotation and Analysis of Alternative Splicing

Tyson A. Clark[1], Bo Wang[2], Ting Hon[1], Elizabeth Tseng[1], Michael Regulski[2], and Doreen Ware[2]

[1]Pacific Biosciences, Menlo Park, CA
[2]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

Abstract

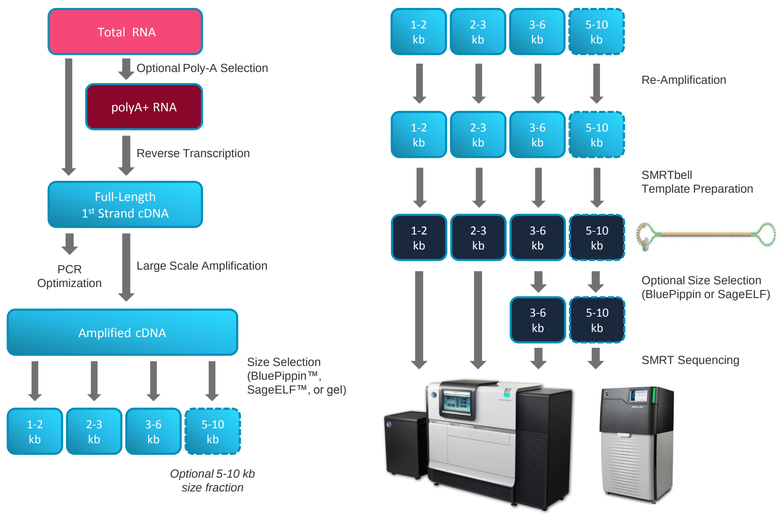
In higher eukaryotic organisms, the majority of multi-exon genes are alternatively spliced. Different mRNA isoforms from the same gene can produce proteins that have distinct properties and functions. Thus, the importance of understanding the full complement of transcript isoforms with potential phenotypic impact cannot be understated. 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.

The Iso-Seq™ protocol developed at PacBio offers the only solution for direct sequencing of full-length, single-molecule cDNA sequences to survey transcriptome isoform diversity useful for gene discovery and annotation. Knowledge of the complete isoform repertoire is also key for accurate quantification of isoform abundance. As most transcripts range from 1 – 10 kb, fully intact RNA molecules can be sequenced using SMRT® Sequencing without requiring fragmentation or post-sequencing assembly. Our open-source computational pipeline delivers high-quality, non-redundant sequences for unambiguous identification of alternative splicing events, alternative transcriptional start sites, polyA tail, and gene fusion events.

We applied the Iso-Seq method to the maize (*Zea mays*) inbred line B73. Full-length cDNAs from six diverse tissues were barcoded and sequenced across multiple size-fractionated SMRTbell libraries. A total of 111,151 unique transcripts were identified. More than half of these transcripts (57%) represented novel, sometimes tissue-specific, isoforms of known genes. In addition to the 2250 novel coding genes and 860 lncRNAs discovered, the Iso-Seq dataset corrected errors in existing gene models, highlighting the value of full-length transcripts for whole gene annotations.

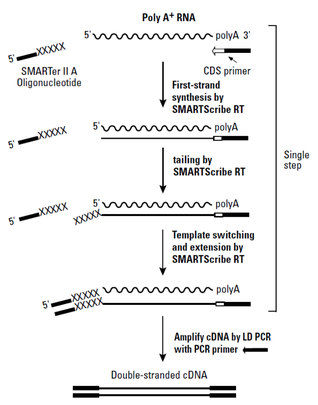
Sample Preparation Methods

Iso-Seq Sample Preparation Workflow

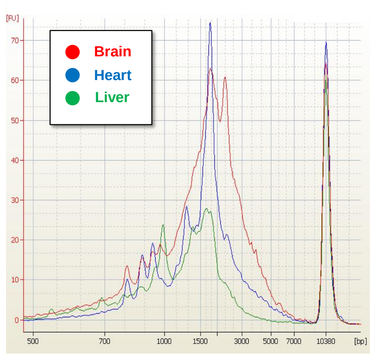


RNA is converted into first strand cDNA using the Clontech SMARTer PCR cDNA Synthesis Kit followed by universal amplification. Amplified cDNA is size fractionated and converted into SMRTbell templates for sequencing on the PacBio RS II or Sequel System.

Clontech SMARTer PCR cDNA Synthesis Kit

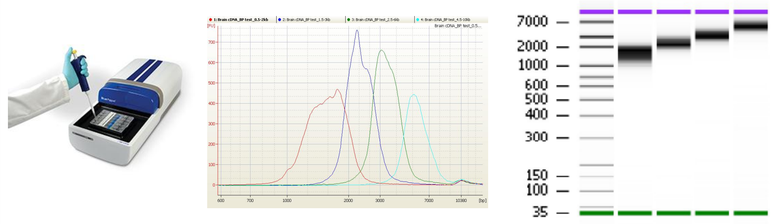


Size Distribution of Amplified cDNA from Multiple Tissues

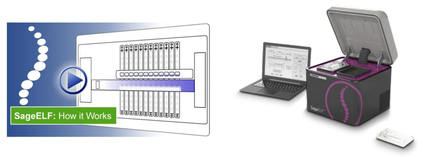


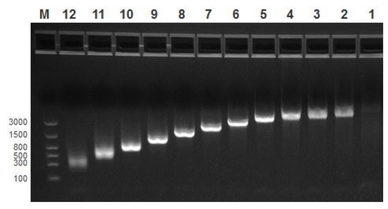
Size Fractionation of Iso-Seq Libraries

Sage Science's BluePippin Size Fractionation



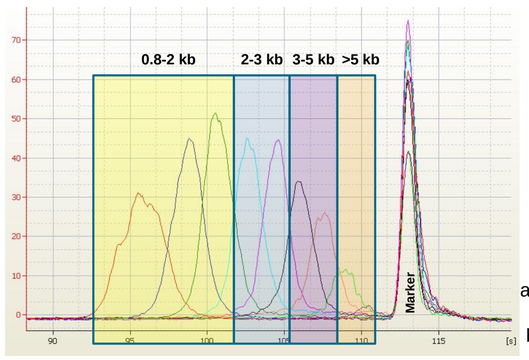
Example Bioanalyzer trace of four size-selected Iso-Seq libraries





Amplified cDNAs after size selection on SageELF system.

Amplified cDNA After Size Fractionation on SageELF System



ELF	SIZE
11	1160
10	1650
9	1900
8	2400
7	2900
6	3700
5	5100
4	6050
3	7000

SageELF system increases size-selection flexibility and allows for isolation of amplified cDNA from several hundred bases up to more than 10 kb.

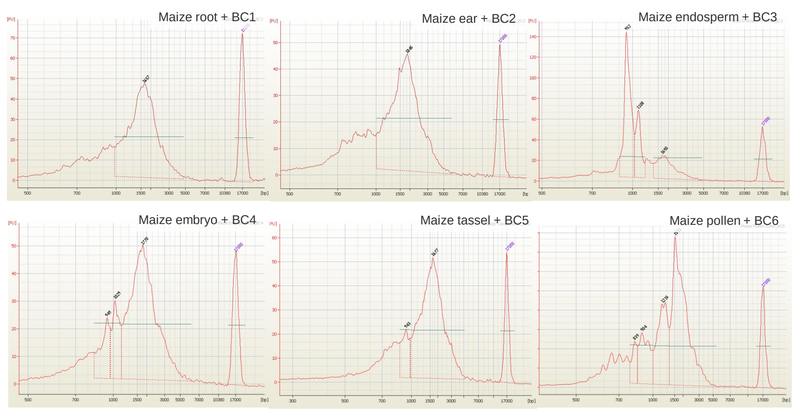
Barcoding cDNA Libraries

Barcoding During Reverse Transcription Step

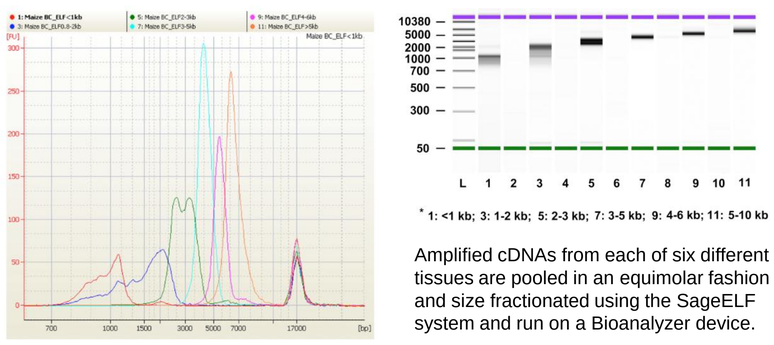
SMARTer_dT_BC1	AAGCAGTGGTATCAACGCAGAGTACtcagacgatgcgtcatTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
SMARTer_dT_BC2	AAGCAGTGGTATCAACGCAGAGTACctatacatgactctgcTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
SMARTer_dT_BC3	AAGCAGTGGTATCAACGCAGAGTACtactagagtagcactcTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
SMARTer_dT_BC4	AAGCAGTGGTATCAACGCAGAGTACtgtgtatcagtacatgTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
SMARTer_dT_BC5	AAGCAGTGGTATCAACGCAGAGTACgatctctactatatgcTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN
SMARTer_dT_BC6	AAGCAGTGGTATCAACGCAGAGTACacagtctatactgctgTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN

Primers containing 16-mer barcodes are used to prime first strand cDNA synthesis. RNA samples are reverse transcribed individually, then pooled prior to size fractionation.

Size Distribution of Individually Amplified cDNAs from Six Diverse Maize Tissues



Size Distribution of Size-Fractionated, Pooled SMRTbell Libraries

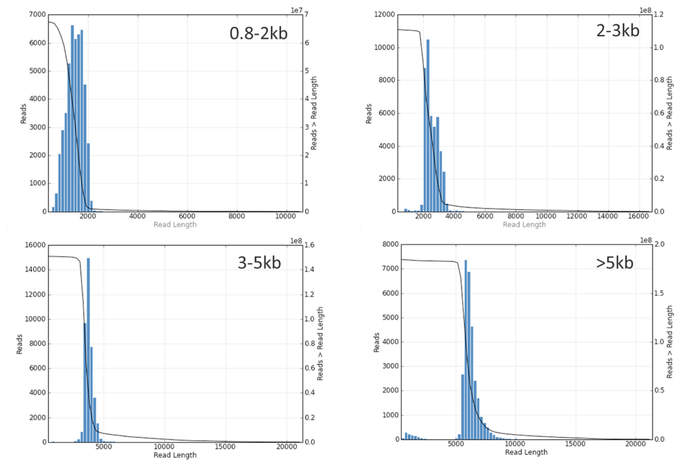


* 1: <1 kb; 3: 1-2 kb; 5: 2-3 kb; 7: 3-5 kb; 9: 4-6 kb; 11: 5-10 kb

Amplified cDNAs from each of six different tissues are pooled in an equimolar fashion and size fractionated using the SageELF system and run on a Bioanalyzer device.

Sequencing Results

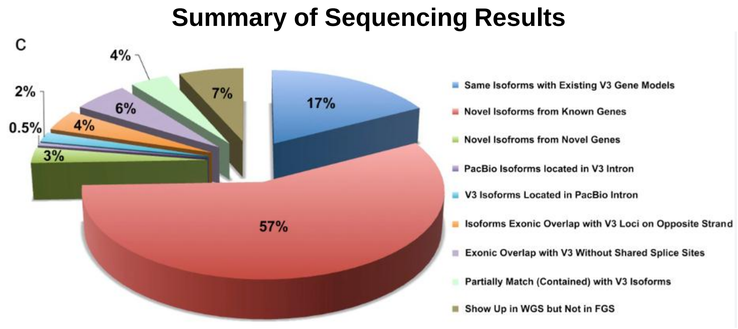
Size Distribution of Full-Length cDNA Sequences



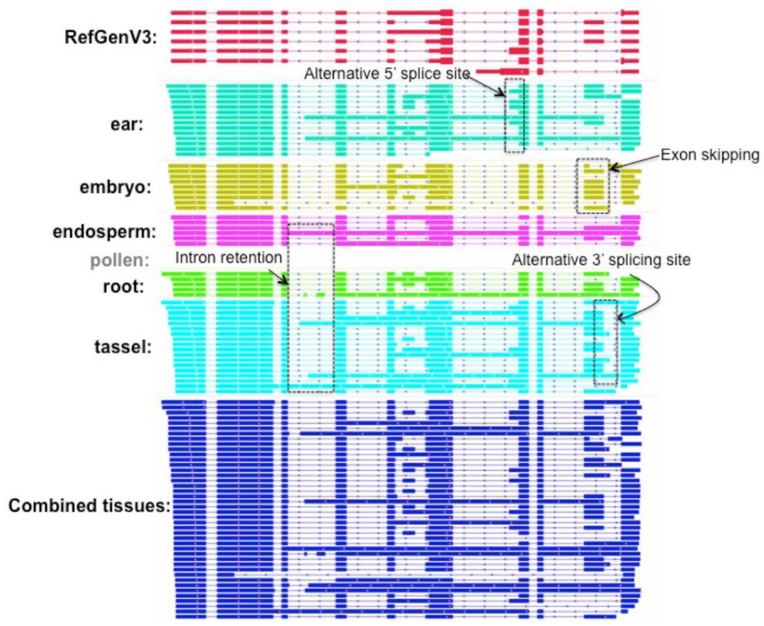
Percentage of Reads from Each Barcoded Tissue

BC	Tissue	1-2 kb	2-3 kb	3-5 kb	>5 kb
1	Root	17%	15%	14%	11%
2	Ear	13%	15%	17%	16%
3	Endosperm	16%	10%	9%	10%
4	Embryo	18%	21%	25%	34%
5	Tassel	15%	16%	17%	15%
6	Pollen	22%	23%	19%	15%

Summary of Sequencing Results



IGV Visualization of Different Splicing Modes in One Gene



Summary and Resources

Summary:

- The Iso-Seq method provides full-length cDNA sequences without the need for assembly.
- Improved sample prep, size-selection, and barcoding methods allows for sequencing of transcripts up to 10 kb from multiple sources.
- Sequencing of full-length transcripts identifies novel isoforms and improves gene annotations

PacBio human three tissue dataset available here:
http://blog.pacificbiosciences.com/2014/10/data-release-whole-human-transcriptome.html

PacBio MCF-7 transcriptome dataset available here:
http://blog.pacificbiosciences.com/2013/12/data-release-human-mcf-7-transcriptome.html

Additional information and Iso-Seq protocols:
http://www.pacb.com/applications/isoseq/index.html

Details on data analysis of Iso-Seq data can be found here:
https://github.com/PacificBiosciences/cDNA_primer/wiki