Single Molecule, Real-Time Sequencing of Full-length cDNA Transcripts Uncovers Novel Alternatively Spliced Isoforms

Tyson A. Clark[1], Elizabeth Tseng[1], Susana Wang[1], Jason G. Underwood[2], and Jonas Korlach[1]
[1]Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025
[2]University of Washington Genome Sciences, Seattle, WA

Abstract

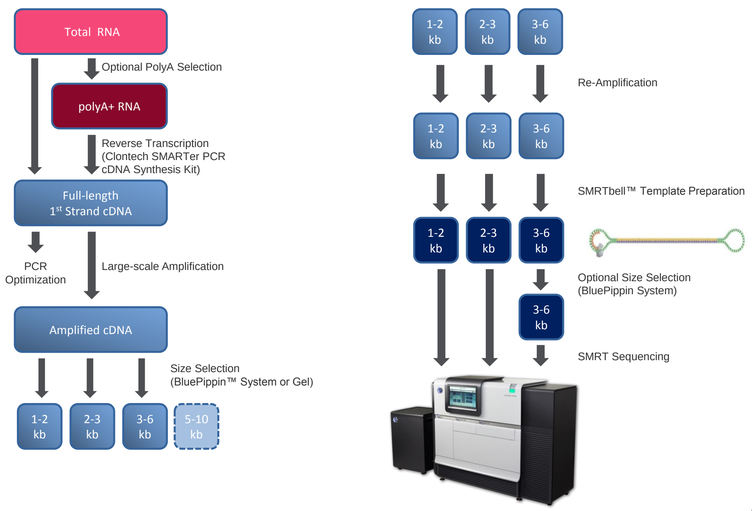
Alternative splicing of mRNA molecules is a tool used by eukaryotic organisms to expand the protein-coding potential of their genomes. In humans, for example, nearly all multi-exon genes are alternatively spliced. Different mRNA isoforms from the same gene can generate changes in RNA stability as well as produce proteins that can have distinct properties such as structure, function or subcellular localization. Thus, understanding the biology of an organism requires knowing the full complement of isoforms. Microarrays and high-throughput cDNA sequencing have become incredibly useful tools for studying transcriptomes, yet these technologies provide small fragments of transcripts and building complete transcripts has been challenging (1).

We have developed a technique that is capable of sequencing full-length, single-molecule cDNA sequences. The method employs PacBio's SMRT® Sequencing, which has the capability to sequence individual molecules with read lengths that average 8 kb and can reach as long as 40 kb. Thus, we are able to generate sequence for complete individual transcripts from the polyA-tail to the 5' end of the RNA molecule. Full-length cDNA sequencing allows for unambiguous identification of alternative splicing events, alternative transcriptional start and polyA sites, and transcripts from gene fusion events. Knowledge of the complete set of isoforms from a sample of interest is key for accurate quantification of isoform abundance when using any technology for transcriptome studies (2).

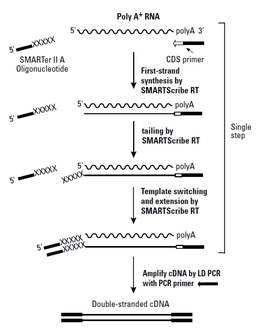
Using a deep dataset of full-length cDNA sequences from the MCF-7 human breast cancer cell line and a comparative study of human brain, heart, and liver, we demonstrate the ability to obtain full-length cDNA sequences from transcripts longer than 10 kb. Even in extensively profiled sample types, the method has been able to uncover large numbers of novel alternatively spliced isoforms and previously unannotated genes.

Sample Preparation Methods

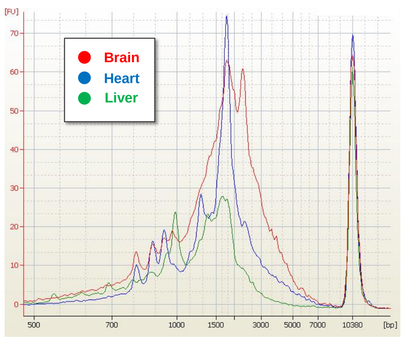
Iso-Seq Sample Preparation Workflow



Clontech® SMARTer® PCR cDNA Synthesis Kit

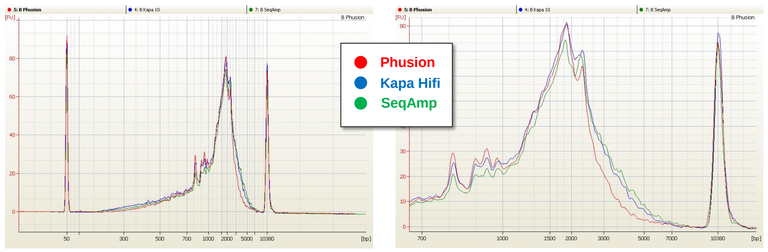


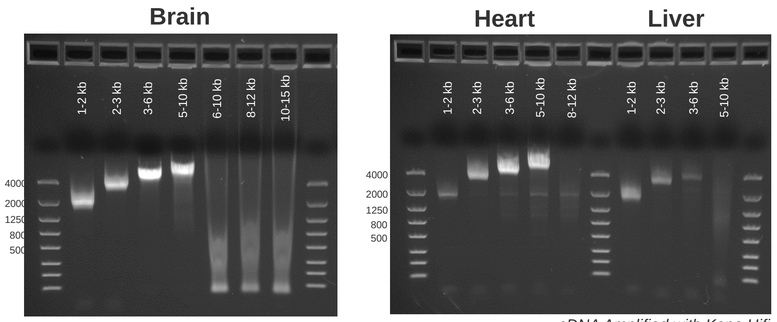
Size Distribution of Amplified cDNA From Multiple Tissues



Sample Prep Improvements

Protocol Adjustments Improve Representation of Longer Transcripts



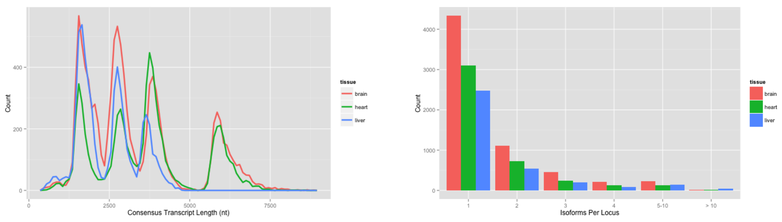


cDNA Amplified with Kapa Hifi

Full-Length Human Tissue Transcriptomes

PacBio® Sequencing of Iso-Seq Libraries From 3 Human Tissues

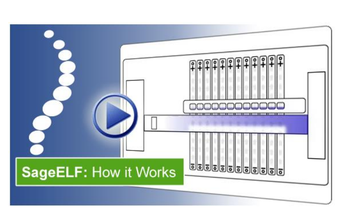
Tissue	Size Fractions Sequenced				Number of Isoforms	Number of Genes	Transcript Lengths
	1-2 kb	2-3 kb	3-6 kb	5-10 kb			
Brain	■	■	■	■	10289	6356	418 – 8823 nt
Heart	■	■	■	■	6896	4351	467 – 8528 nt
Liver	■	■	■		6124	3497	419 – 4754 nt

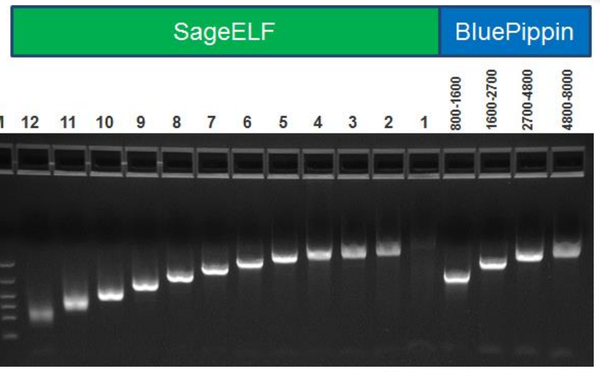


Full-Length Non-Redundant Transcript Sequences

SageELF™ Size Fractionation

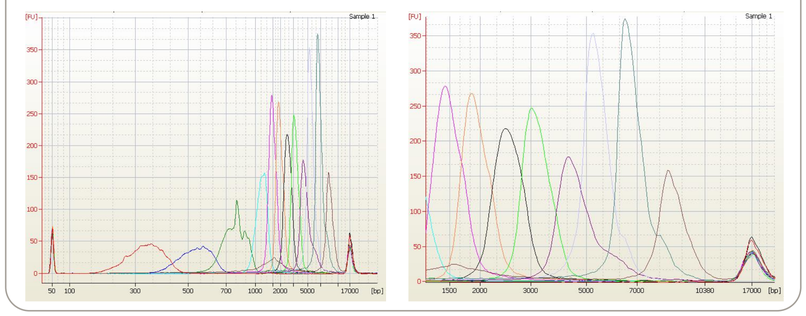
SageELF Allows For Collection of cDNA Molecules in 12 Fractions Across the Entire Size Distribution





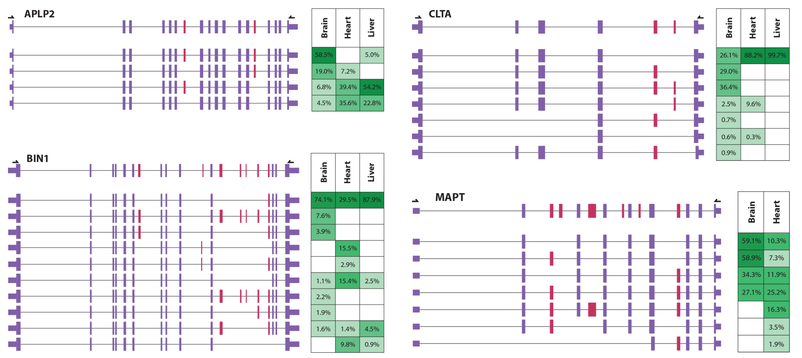
Amplified cDNA from Human Brain

Bioanalyzer® Traces of SageELF Size-Selected cDNA from Human Brain



Targeted Full-Length cDNA Sequencing

Sequencing of Full-Length RT-PCR Products Shows Differential Alternative Splicing Across Three Human Tissues



References

(1) Steijger T, et al. *Nat Methods*. 2013 Dec;10(12):1177-84.
(2) Au KF, et al. *Proc Natl Acad Sci USA*. 2013 Dec 10; 110(50):E4821-30.

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

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