Full-length cDNA Sequencing of Alternatively Spliced Isoforms Provides Insight into Human Diseases

Tyson A. Clark, Elizabeth Tseng, Ting Hon, and Jonas Korlach
Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025

Abstract

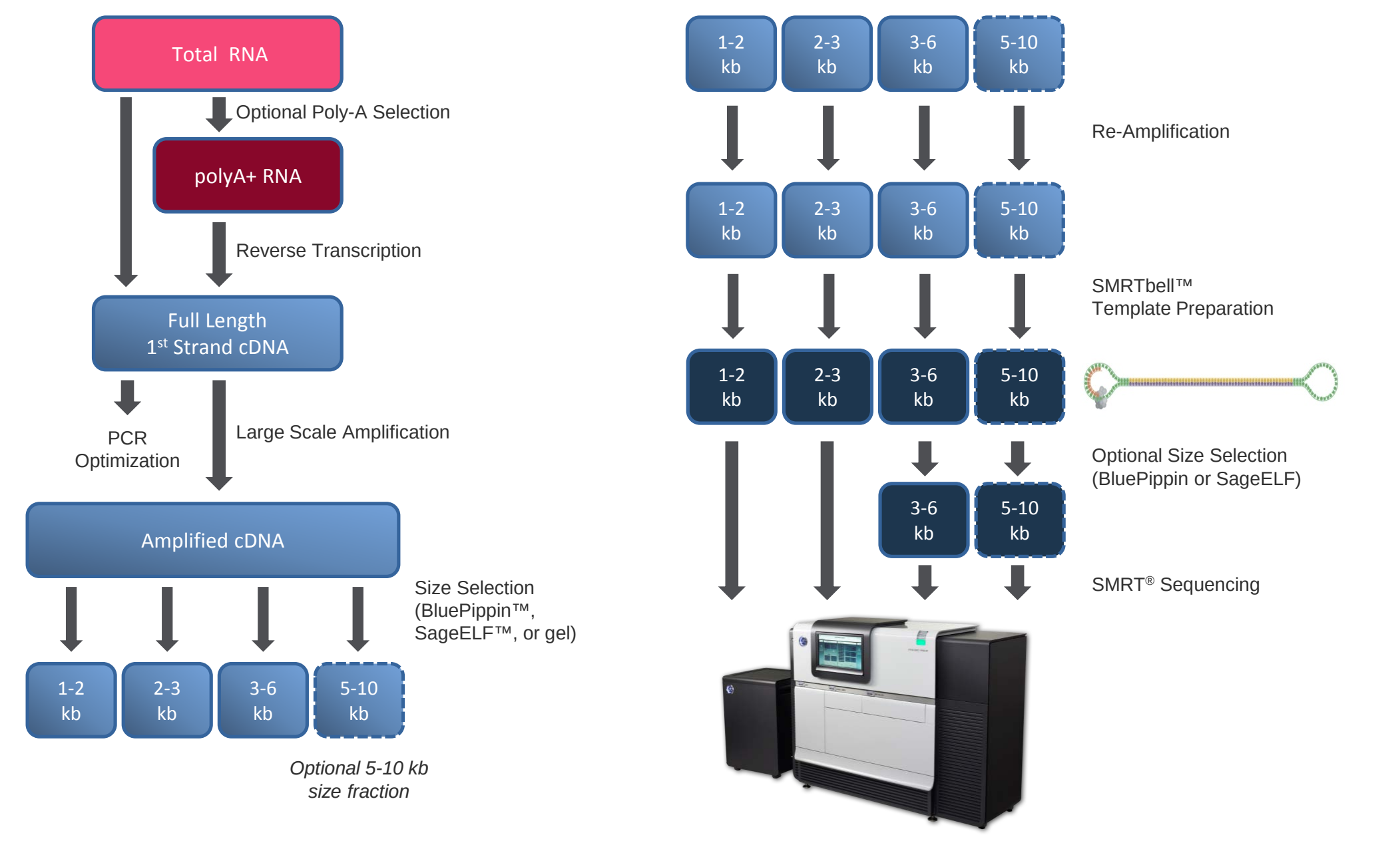
The majority of human genes are alternatively spliced, making it possible for most genes to generate multiple proteins. The process of alternative splicing is highly regulated in a developmental-stage and tissue-specific manner. Perturbations in the regulation of these events can lead to disease in humans. Alternative splicing has been shown to play a role in human cancer, muscular dystrophy, Alzheimer's, and many other diseases. Understanding these diseases requires knowing the full complement of mRNA isoforms. Microarrays and high-throughput cDNA sequencing have become highly successful tools for studying transcriptomes, however these technologies only provide small fragments of transcripts and building complete transcript isoforms has been very challenging.

We have developed the Iso-Seq™ technique, which is capable of sequencing full-length, single-molecule cDNA sequences. The method employs SMRT® Sequencing to generate individual molecules with average read lengths of more than 10 kb and some as long as 40 kb. As most transcripts are from 1 to 10 kb, we can sequence through entire RNA molecules, requiring no fragmentation or post-sequencing assembly. Jointly with the sequencing method, we developed a computational pipeline that polishes these full-length transcript sequences into high-quality, non-redundant transcript consensus sequences. Iso-Seq sequencing enables unambiguous identification of alternative splicing events, alternative transcriptional start and poly-A 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.

Here we characterize the full-length transcriptome of normal human tissues, paired tumor/normal samples from breast cancer, and a brain sample from a patient with Alzheimer's using deep Iso-Seq sequencing. We highlight numerous discoveries of novel alternatively spliced isoforms, gene-fusions events, and previously unannotated genes that will improve our understanding of human diseases.

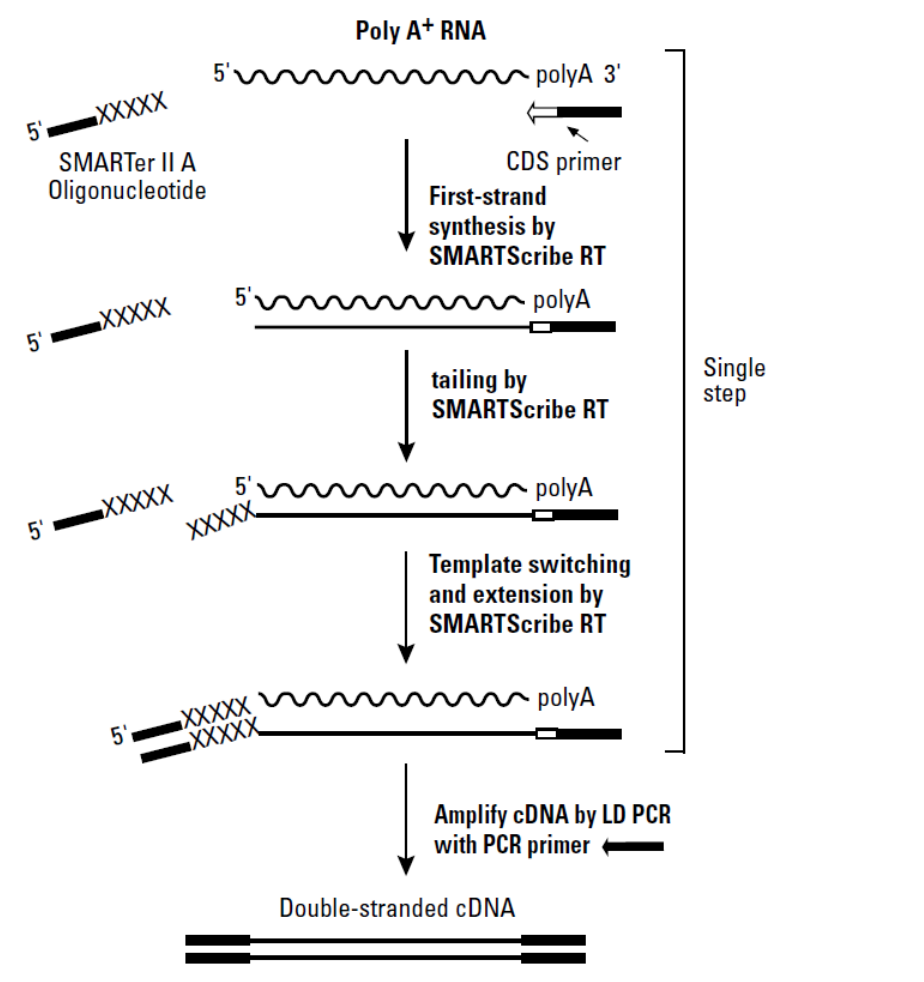
Sample Preparation Methods

Iso-Seq Sample Preparation Workflow

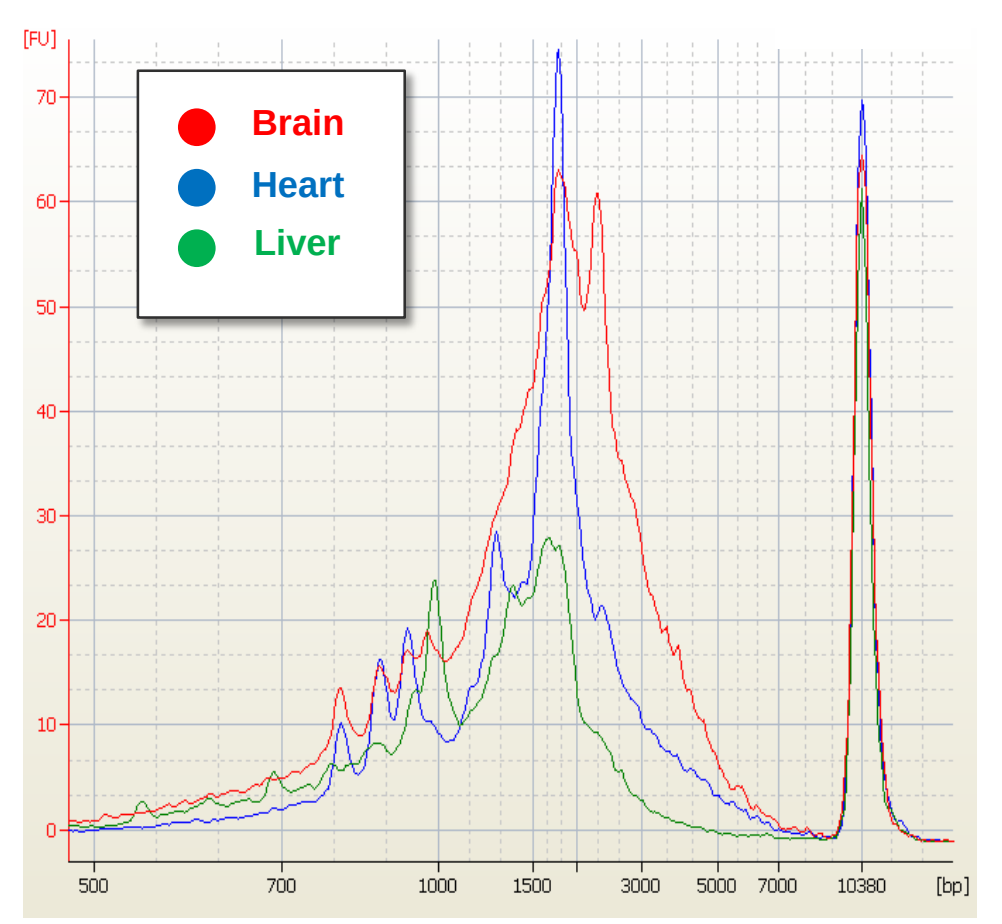


RNA is converted into first strand cDNA using the Clontech SMARTer PCR cDNA Synthesis Kit followed by universal amplification. Amplified cDNA is size fractionated and converted into SMRTbell templates for sequencing on the PacBio® *RS II*.

Clontech® SMARTer® PCR cDNA Synthesis Kit

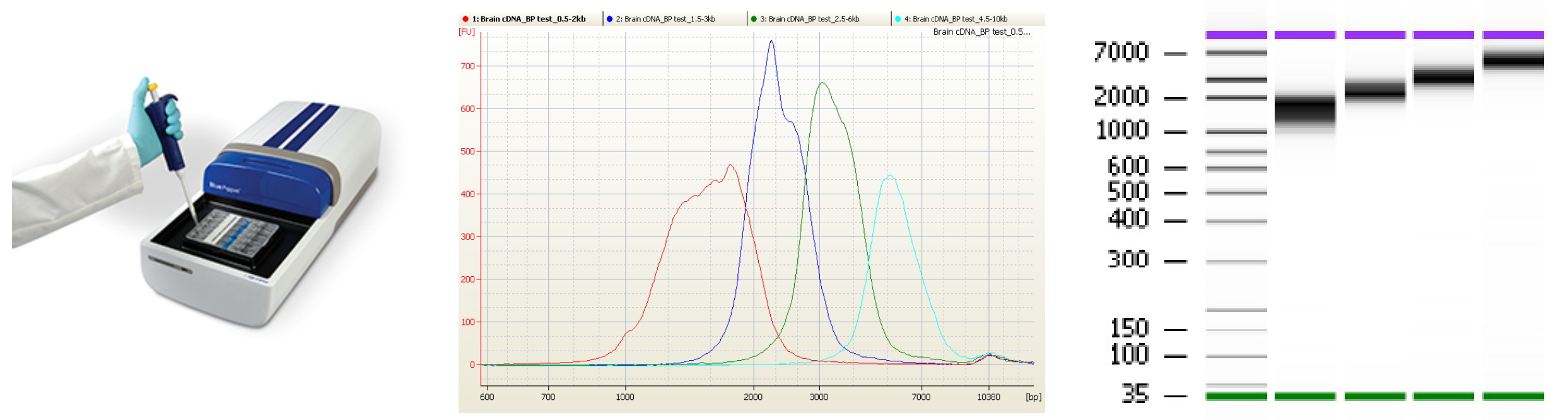


Size Distribution of Amplified cDNA from Multiple Tissues



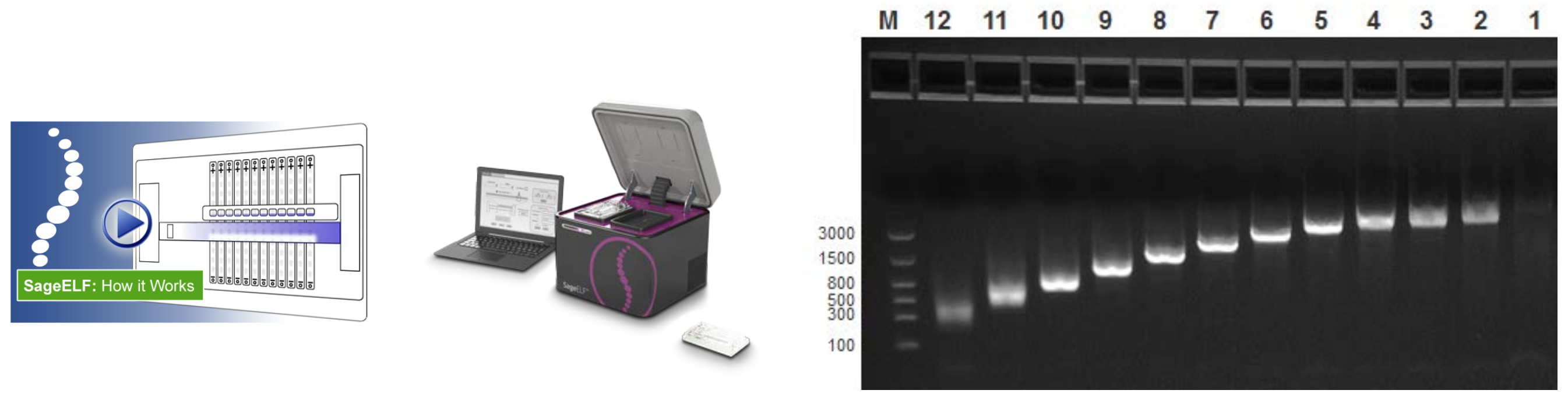
Size Fractionation of Iso-Seq Libraries

Sage Science's BluePippin™ Size Fractionation



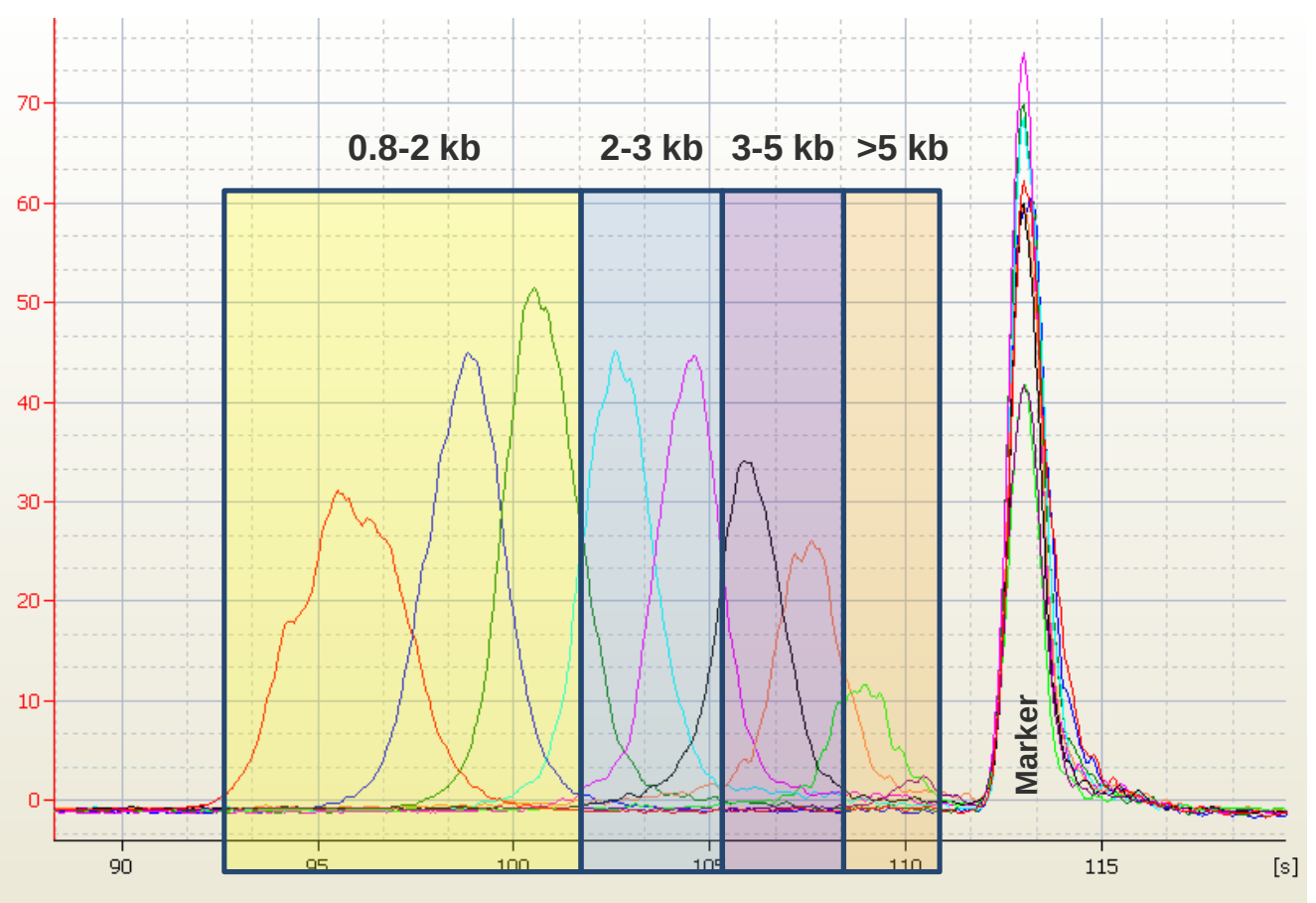
Example Bioanalyzer trace of four size-selected Iso-Seq libraries

SageELF™ Separation Allows for Collection of cDNA Molecules in 12 Fractions Across the Entire Size Distribution



Amplified cDNAs after size selection on Sage ELF system.

Amplified cDNA After Size Fractionation on SageELF System

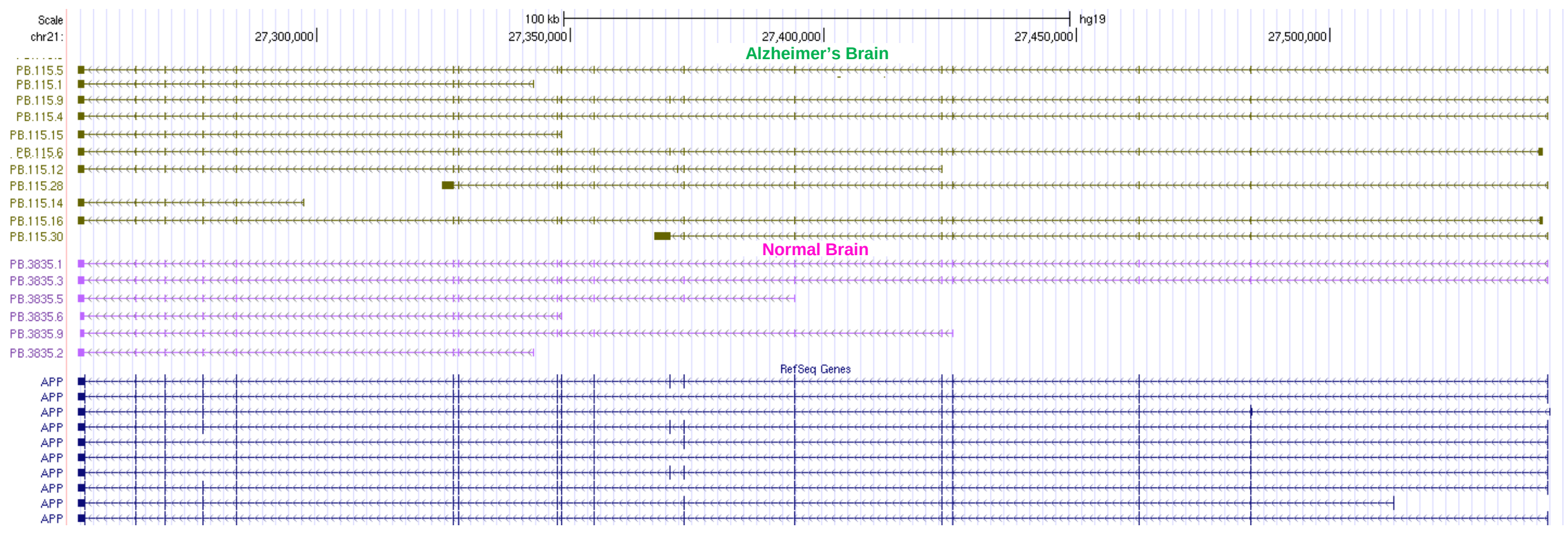


ELF	SIZE
11	1160
10	1650
9	1900
8	2400
7	2900
6	3700
5	5100
4	6050
3	7000

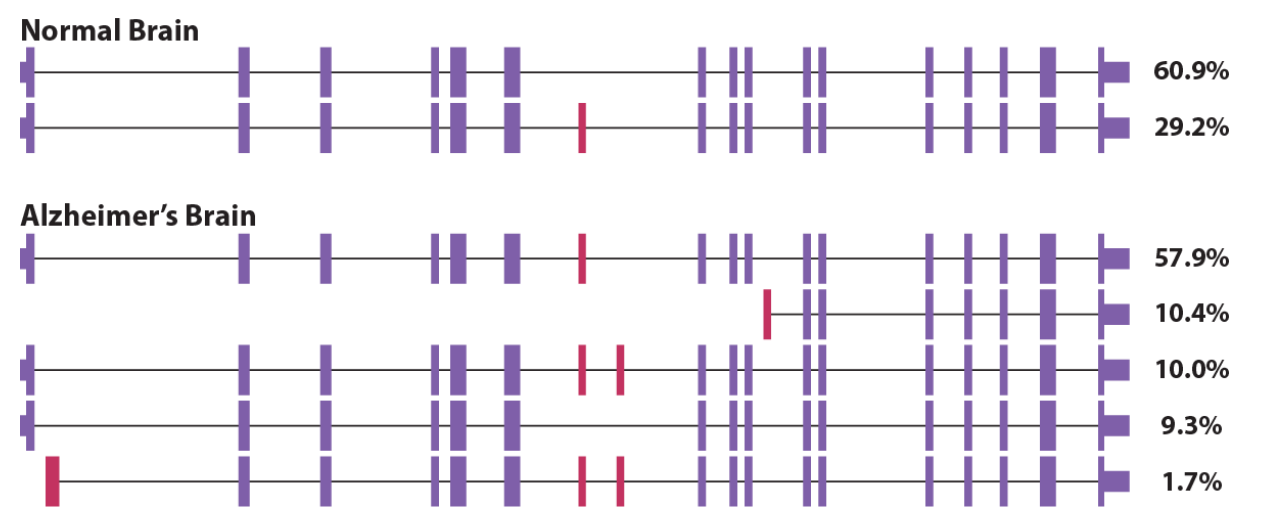
SageELF system increases the flexibility of size selection and allows for isolation of amplified cDNAs from several hundred kb up to more than 10 kb in size.

Alternative Splicing Events in Alzheimer's

Full-Length APP Transcripts From Normal and Alzheimer's Brain Samples



The predominant isoforms from the APP gene (Amyloid Beta A4 Precursor Protein) are different between an RNA sample from normal human brain and RNA from the brain of a patient that suffered from Alzheimer's Disease.

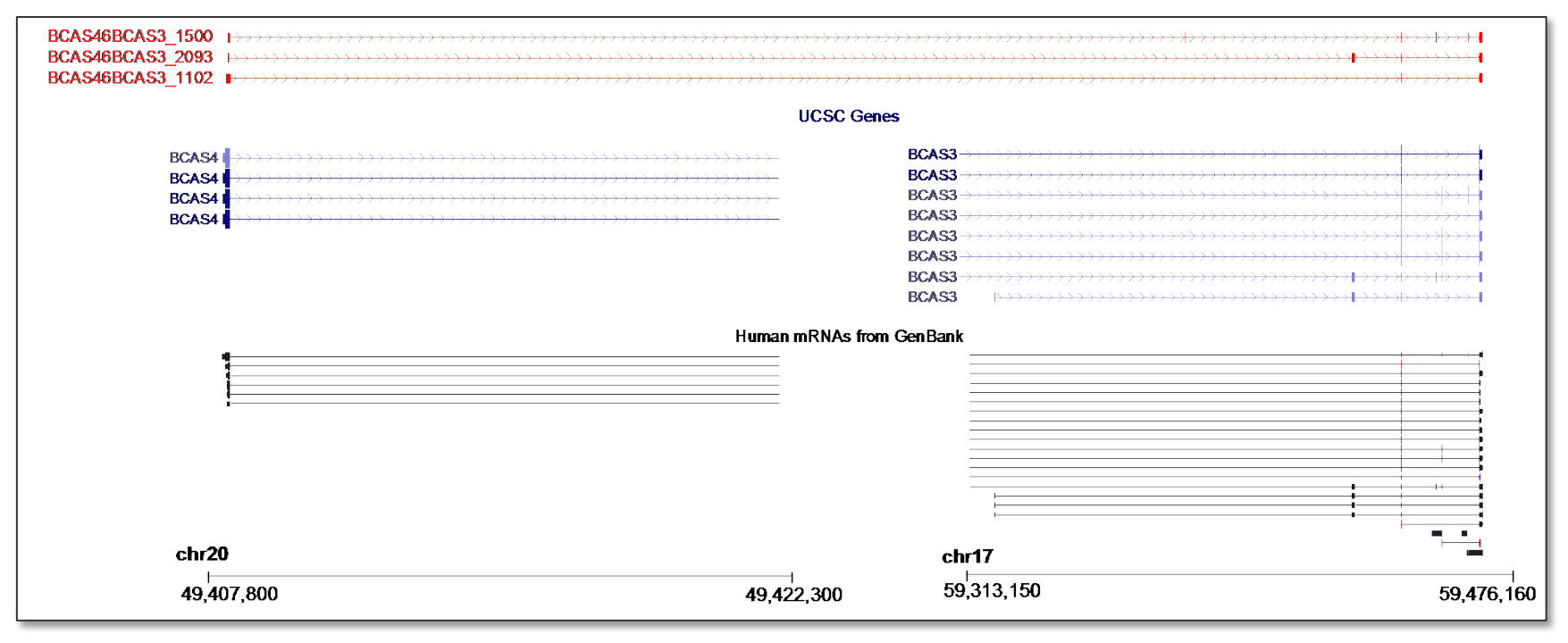


Detection of Fusion Genes in Cancer

93 Gene Fusion Candidates Found in the MCF-7 Cancer Cell Line Iso-Seq Datasets (16 of them are previously known or predicted)

Gene 1	Chrom 1	Gene 2	Chrom 2	Literature Support
ARFGEF2	chr20	SULF2	chr20	experimental
BCAS4	chr20	BCAS3	chr17	experimental
ESR1	chr6	CCDC170	chr6	experimental
FOXA1	chr14	TTC6	chr14	computational
MYH9	chr22	EIF3D	chr22	computational
MYO6	chr6	SENP6	chr6	experimental
PAPOLA	chr14	AK7	chr14	computational
POP1	chr8	MATN2	chr8	experimental
RPS6KB1	chr17	VMP1	chr17	experimental
RPS6KB1	chr17	DIAPH3	chr13	experimental
RSBN1	chr1	AP4B1	chr1	computational
SLC25A24	chr1	NBPF1	chr1	experimental
SYTL2	chr11	PICALM	chr11	experimental
TBL1XR1	chr3	RGS17	chr6	experimental
TXLNG	chrX	SYAP1	chrX	experimental
ZNF217	chr20	SULF2	chr20	computational

Table of known or predicted gene fusions that were detected in the MCF-7 dataset

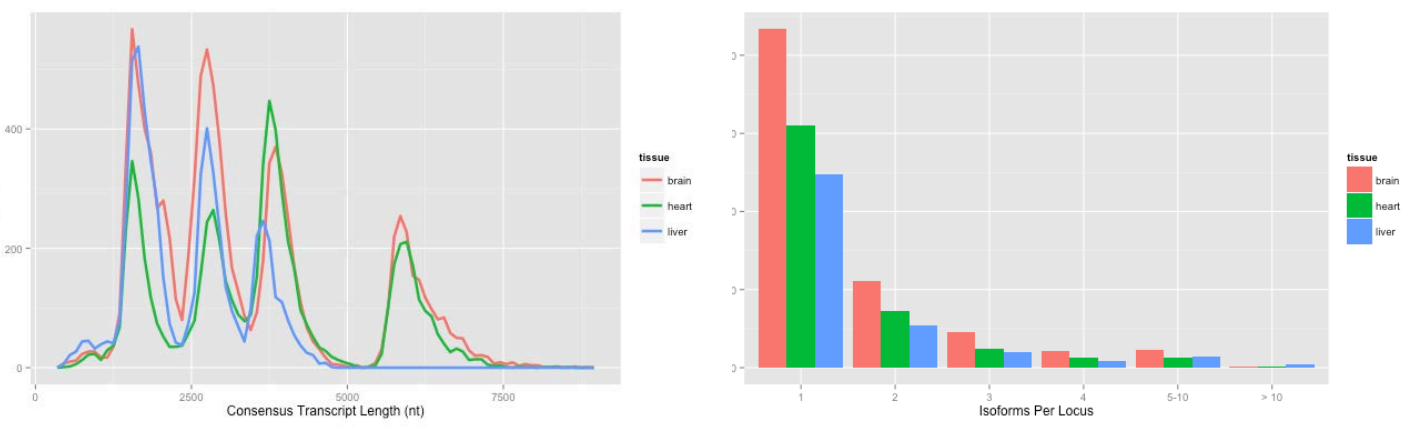


PacBio transcripts (top, red) show three different fusion variants of the BCAS4/BCAS3 genes. All three variants contain a portion of the 5' region of the BCAS4 gene (chr20q13) and a portion of the 3' region of the BCAS3 gene (chr17q23).

Full-Length Human Tissue Transcriptomes

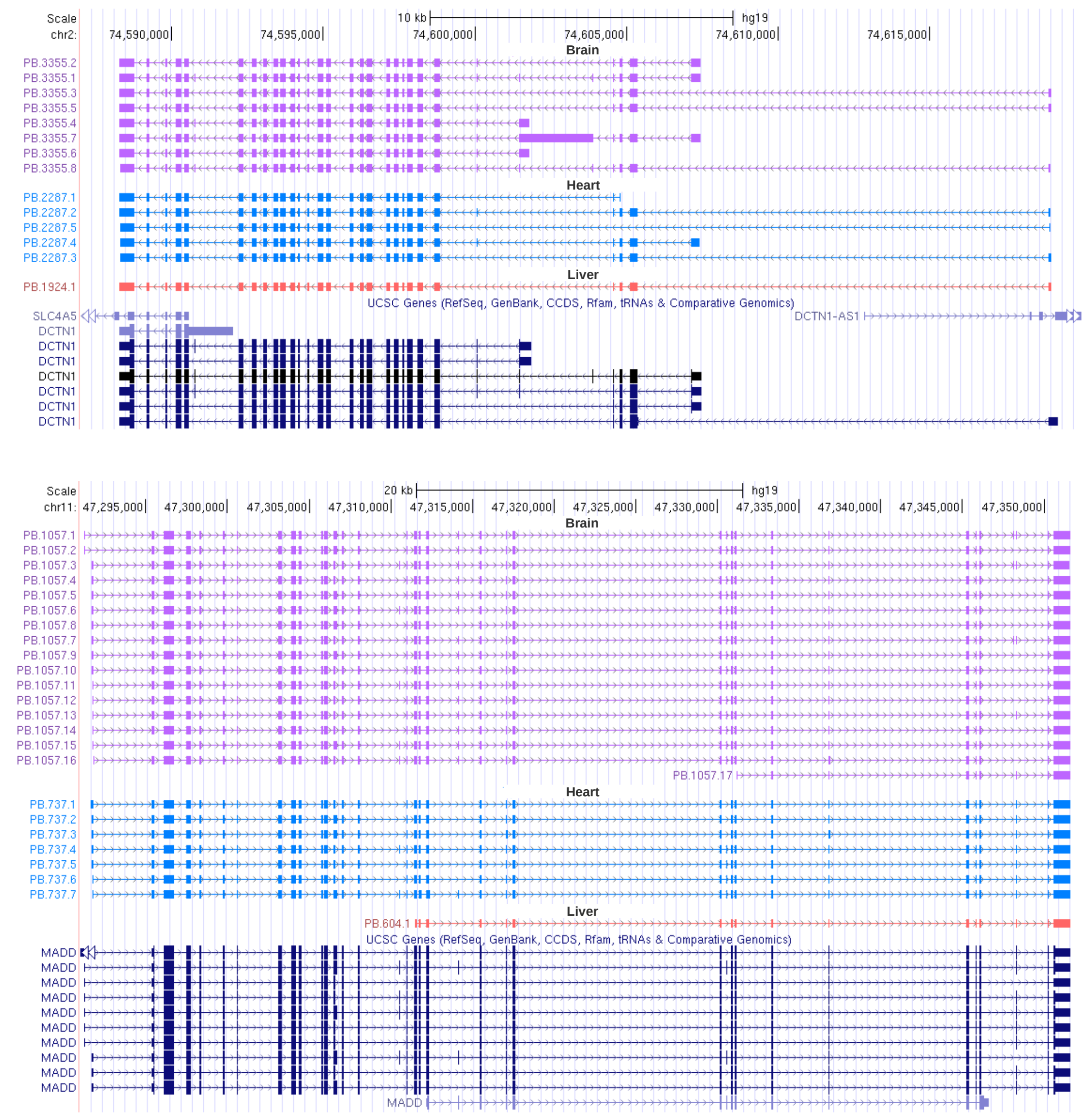
PacBio Sequencing of Iso-Seq Libraries From 3 Human Tissues

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



Overview of the dataset showing numbers of transcripts of various sizes and the number of isoforms per gene

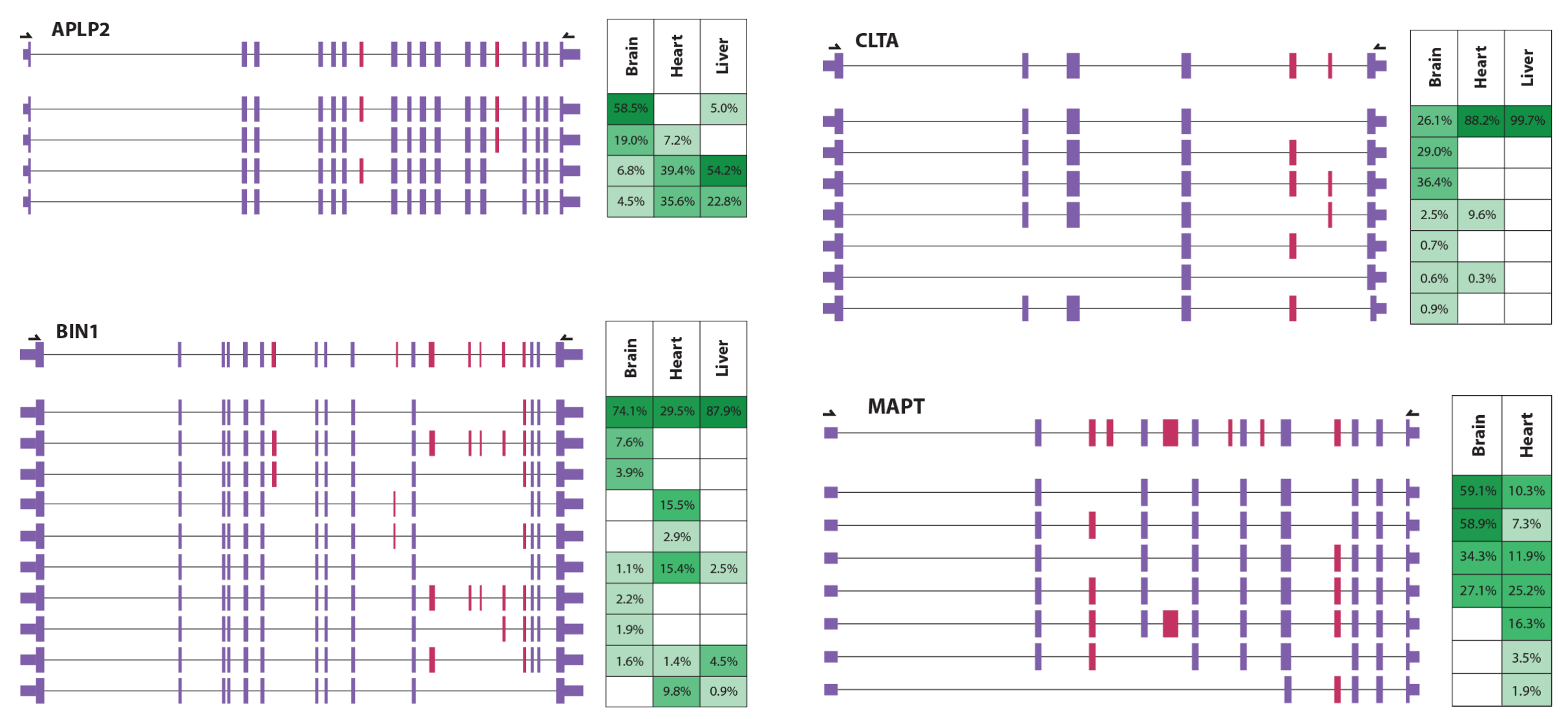
Full-Length Non-Redundant Transcript Sequences



Two examples of genes with differential alternative splicing across the three tissues

Targeted Full-Length cDNA Sequencing

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



PacBio sequencing of full-length RT-PCR products simplifies identification of alternatively spliced isoforms and allows for relative quantification of isoform abundance.

Summary and Resources

Summary:

- The Iso-Seq method provides full-length cDNA sequences without the need for assembly.
- Improved sample prep and size-selection methods allows for sequencing of transcripts up to 10 kb.
- Alternatively spliced transcripts can be easily identified from either whole transcriptome or targeted sequencing.

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

For Research Use Only. Not for use in diagnostic procedures.