Isoform Sequencing: Unveiling the Complex Landscape of the Eukaryotic Transcriptome on the PacBio® RS II

Elizabeth Tseng, Susana Wang, Primo Baybayan
Pacific Biosciences, Menlo Park, CA

Introduction

Alternative splicing of RNA is an important mechanism that increases protein diversity and is pervasive in the most complex biological functions. While advances in RNA sequencing methods have accelerated our understanding of the transcriptome, isoform discovery remains computationally challenging due to short read lengths.

Here, we describe the Isoform Sequencing (Iso-Seq) method using long reads generated by the PacBio RS II. We sequenced rat heart and lung RNA using the Clontech® SMARTer® cDNA preparation kit followed by size selection using agarose gel. Additionally, we tested the BluePippin™ device from Sage Science for efficiently extracting longer transcripts ≥ 3 kb. Post-sequencing, we developed a novel isoform-level clustering algorithm to generate high-quality transcript consensus sequences. We show that our method recovered alternative splice forms as well as alternative stop sites, antisense transcription, and retained introns. To conclude, the Iso-Seq method provides a new opportunity for researchers to study the complex eukaryotic transcriptome even in the absence of reference genomes or annotated transcripts.

Full-length Isoform Sequencing

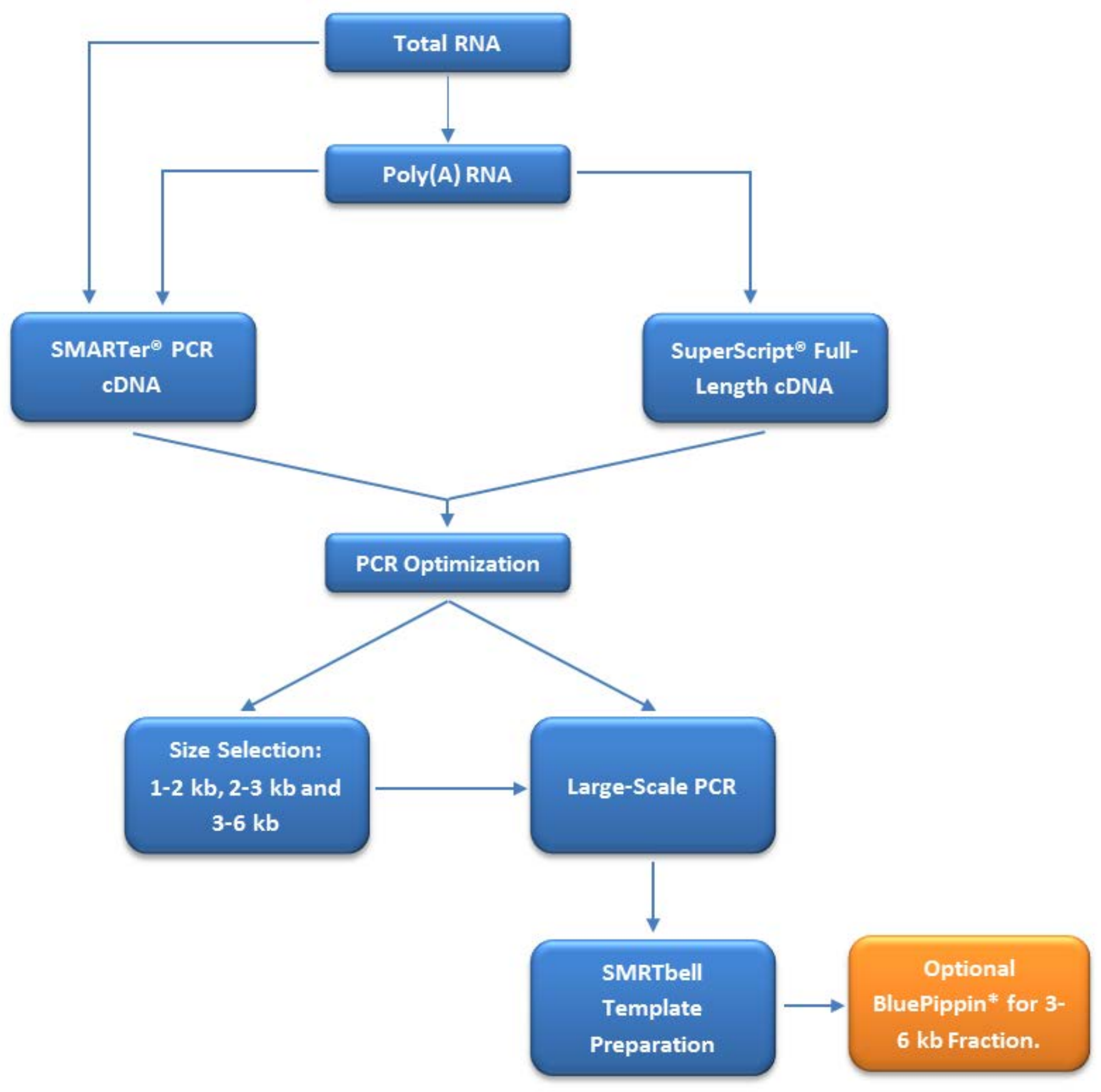

Total RNA
Poly(A) RNA
SMARTer® PCR cDNA
SuperScript® Full-Length cDNA
PCR Optimization
Size Selection: 1-2 kb, 2-3 kb and 3-6 kb
Large-Scale PCR
SMRTbell Template Preparation
Optional BluePippin* for 3-6 kb Fraction.

Figure 1. Isoform Sequencing workflow . Using the Clontech SMARTer kit, rat heart and lung polyA+ RNA were converted to cDNA, then size selected using agarose gel. Additionally, we tested using the BluePippin system for size selection on rat muscle RNA; the BluePippin system was also used for the size selection step (BP Selection 1) and optionally a second time after SMRTbell™ preparation (BP Selection 2).

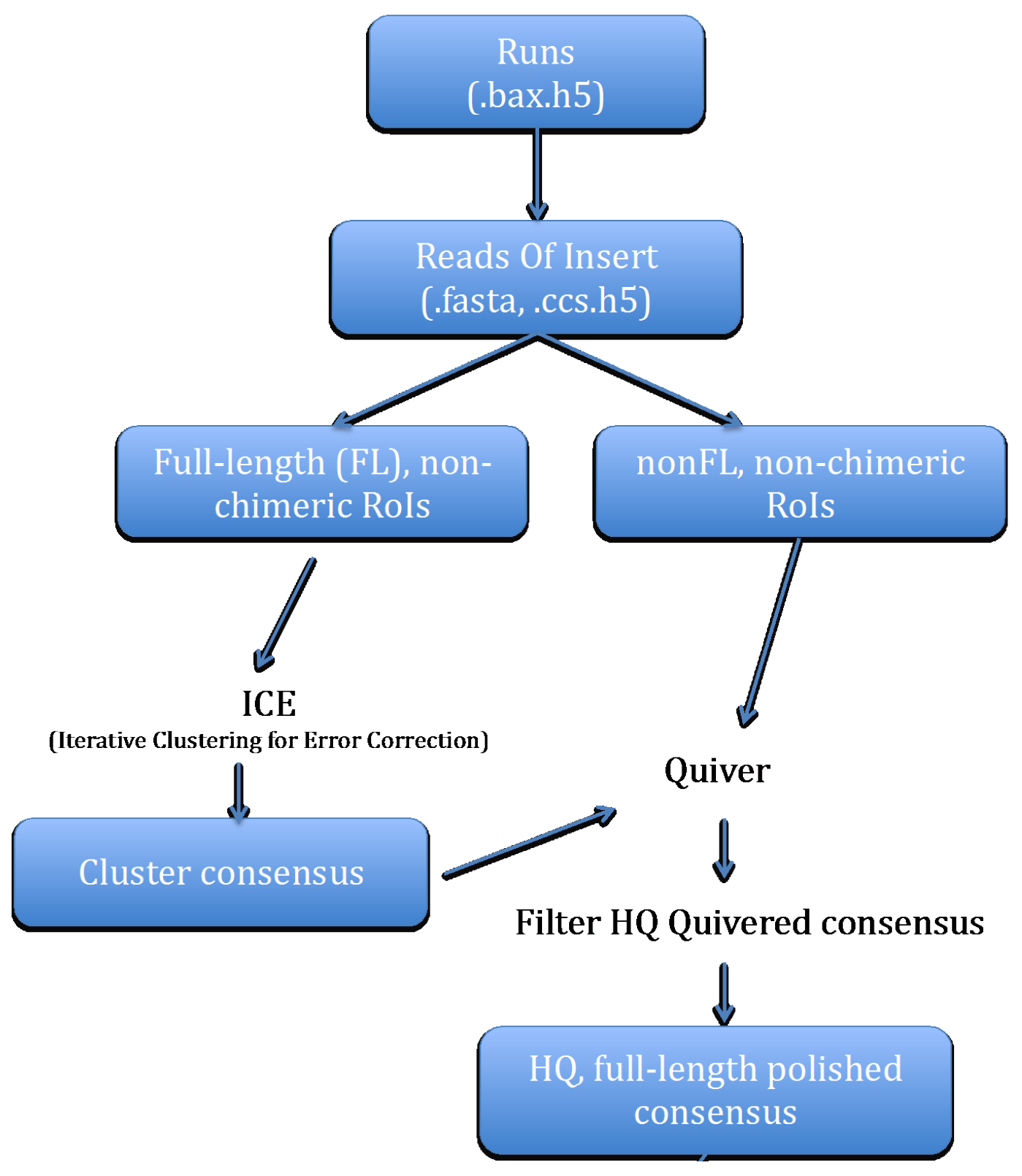

Runs (.bax.h5)
Reads Of Insert (.fasta, .ccs.h5)
Full-length (FL), non-chimeric RoIs
nonFL, non-chimeric RoIs
ICE
(Iterative Clustering for Error Correction)
Cluster consensus
Quiver
Filter HQ Quivered consensus
HQ, full-length polished consensus

Figure 2. Bioinformatics workflow to obtain non-redundant, high-quality transcripts. Reads were defined as ‘full-length’ if both the 5’ and 3’ cDNA primers and polyA tail were observed. Following an iterative clustering algorithm to obtain full-length-only consensus seed sequences, non-full-length reads were aligned to seed consensus for final consensus calling using Quiver. Because the Clontech kit can miss 5’ ends on degraded RNA, transcripts were further collapsed if they have 5’ difference less than 100 bp and are otherwise identical.

Sequencing Rat Transcriptome

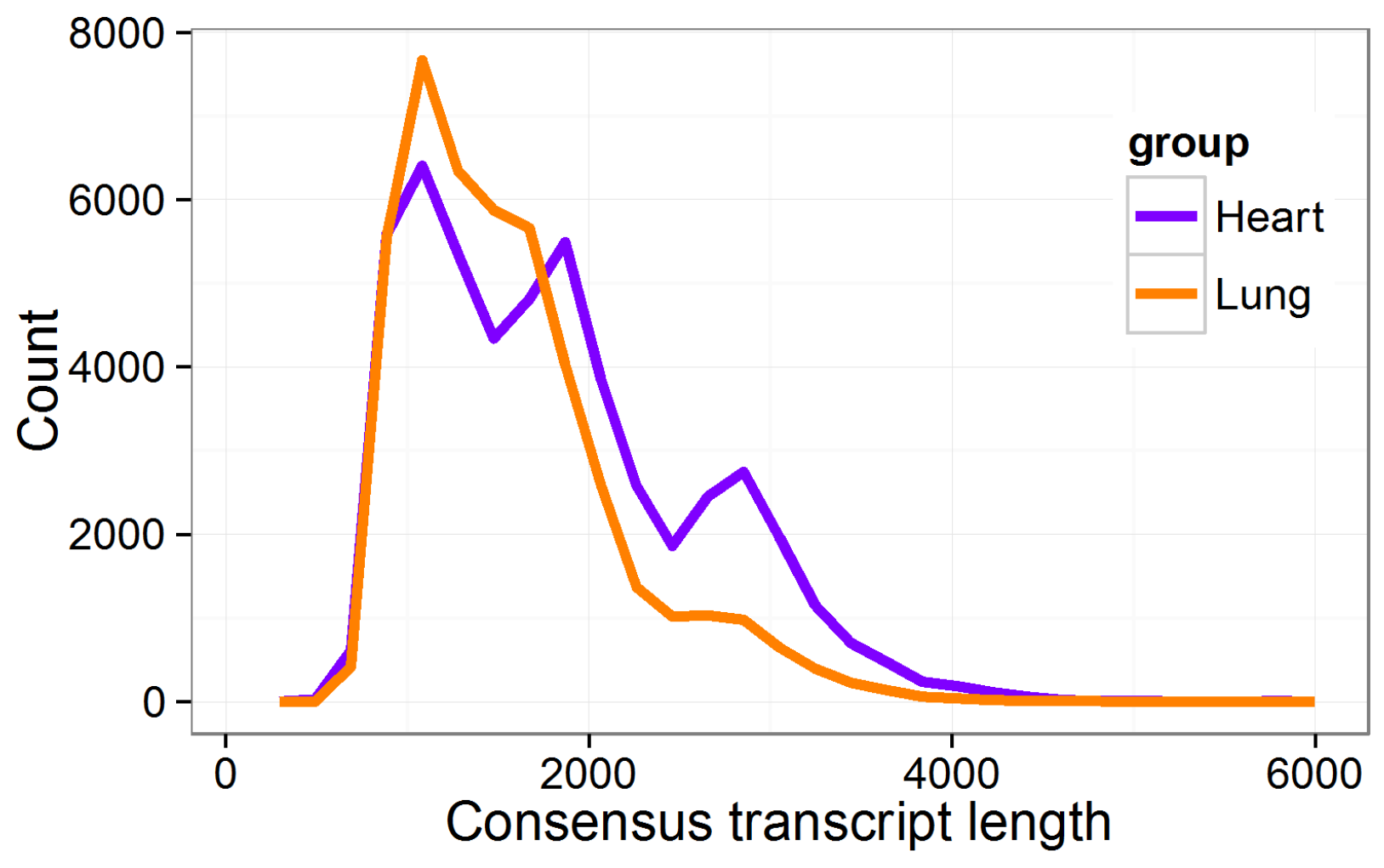
Sequencing rat heart and lung transcriptome

Sample	Number of cells at each size fraction				Number of reads	Number of full-length reads
	1-2 kb	2-3 kb	3-6 kb	Total		
Heart	8	8	16	32	1,849,774	648,997
Lung	8	8	10	26	1,176,609	550,270

- PacBio RS II
- P4-C2 chemistry
- 120 min movies

Rat heart and lung polyA+ RNA were converted to cDNA using the Clontech SMARTer kit, size selected using agarose gel cutting, then sequenced on PacBio RS II.

Isoform-level clustering generates high-quality transcript consensus sequences



min: 138 bp
max: 7,952 bp
median: 1,563 bp

Figure 3. (a) Length distribution of transcript consensus sequences. Each consensus sequence represents the consensus call of a cluster of reads that are considered to be from the same isoform. Since the cDNA 5’/3’ primers and polyA tail were used to determine the full-length reads, the consensus transcripts are putatively full-length.

Sample	Number of transcripts	Aligned transcript coverage		Base differences against reference genome			
		95-99%	100%	Sub	Ins	Del	Total
Heart	51,043	11,262 (22%)	37,105 (73%)	166,359 (0.17%)	91,146 (0.09%)	113,190 (0.12%)	370,695 (0.39%)
Lung	44,083	7,908 (18%)	33,474 (76%)	172,871 (0.25%)	64,189 (0.09%)	101,014 (0.15%)	338,074 (0.49%)

Figure 3. (b) Alignment against rat reference genome rn5. Transcript consensus sequences were aligned against the rat reference genome (rn5) using GMAP. More than ¾ of the transcripts aligned to the genome completely. Base differences against the genome are dominated by substitutions, which could be genuine SNPs, followed by deletions (which are difficult to error correct).

Iso-Seq reveals complex splicing events

cuffcompare was run to compare the rat heart and lung transcripts, which collapsed fully-contained transcripts

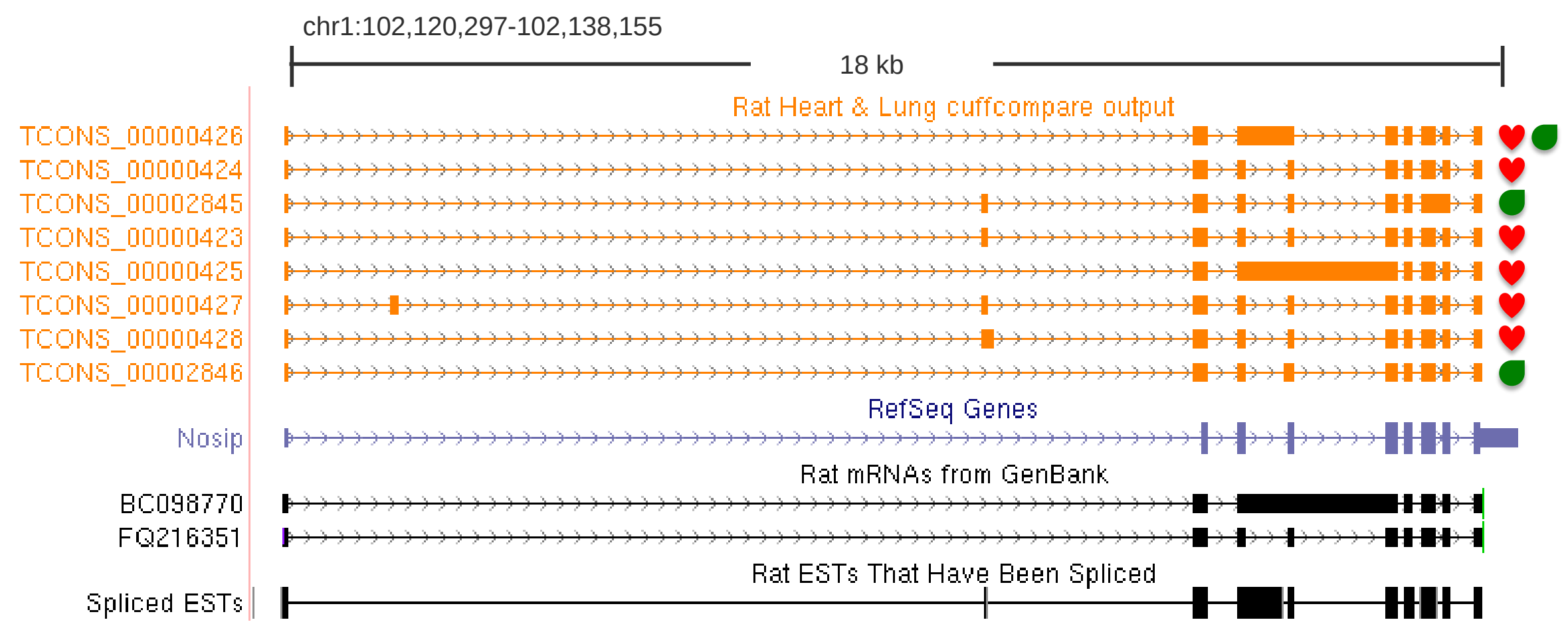


Figure 4. (a) Multiple isoforms observed at a single loci. This UCSC Genome Browser screenshot shows a locus encoding multiple isoforms observed in the PacBio data (top, orange) with alternative splicing and possibly retained introns. Isoforms observed in each sample are marked with ♥ (heart) or ● (lung).

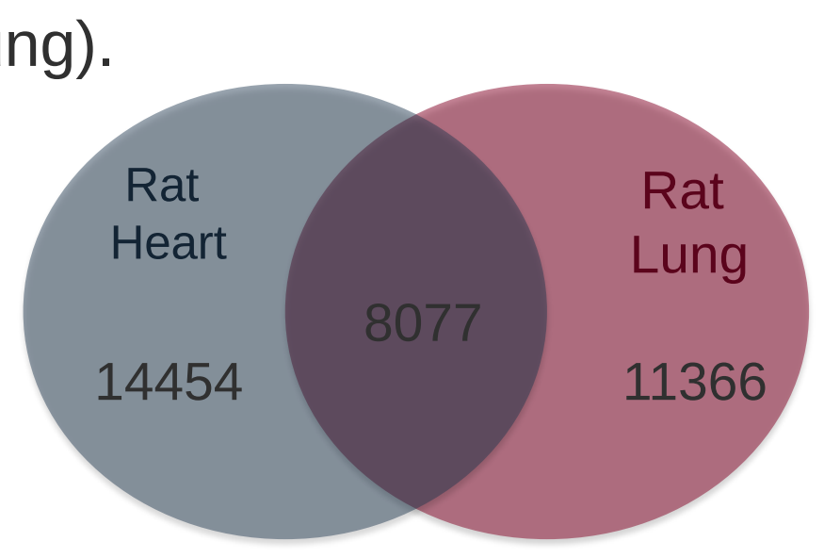


Figure 4. (b) Overlap between rat heart and lung output transcripts using cuffcompare. Note that cuffcompare collapses fully-contained transcripts, which might include genuine alt. start/stops.

Size selection with the BluePippin™ System

Capturing longer transcripts using the BluePippin System

Agarose gel cutting is labor-intensive and less precise above 3 kb.
We used the BluePippin system to size select rat muscle samples to replace the gel cutting step (BP Selection 1) and during an additional selection step after the SMRTbell library preparation step (BP Selection 2).

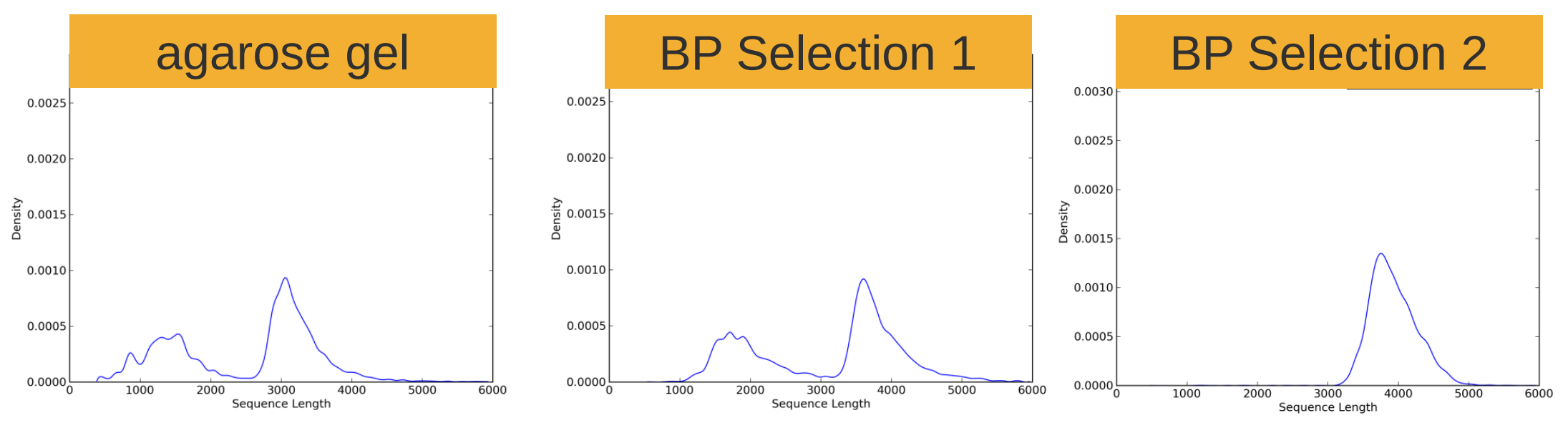


Figure 5. Length distribution of full-length reads using different size selection mechanisms. In all cases, the size-selection target was 3-6 kb. Replacing the gel cutting step with size selection using the BluePippin system resulted in similar contamination of smaller transcripts brought up by the second amplification step. An additional BluePippin size-selection step removed the smaller transcripts.

We sequenced the double-size-selected sample using 2 SMRT® Cells on the PacBio RS II followed by the same bioinformatics analysis.

cuffcompare output:

- 1,118 multi-exonic transcripts (≥ 99% alignment coverage & identity)
- 617 transcribed loci (~1.9 transcript per locus)
- Min: 3,212 bp; Max: 4,732 bp; Median: 3,832 bp

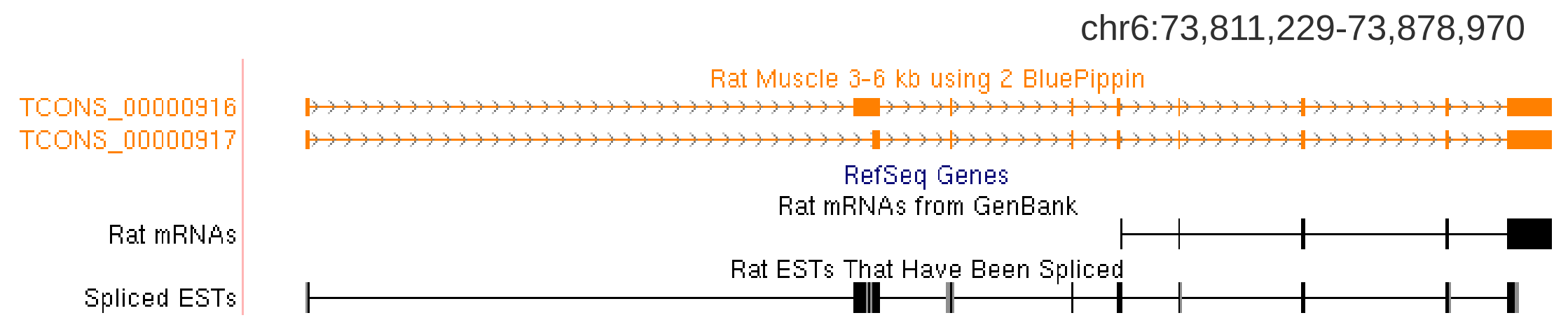

chr6:73,811,229-73,878,970
Rat Muscle 3-6 kb using 2 BluePippin
TCONS_00000916
TCONS_00000917
RefSeq Genes
Rat mRNAs from GenBank
Rat mRNAs
Rat ESTs That Have Been Spliced
Spliced ESTs

Figure 6. (a) BluePippin size-selected rat muscle transcripts matched a predicted protein. Both isoforms had a BLAST hit to a predicted patatin-like phospholipase gene (Pnpla8). The transcripts are 4.6 and 3.6 kb long.

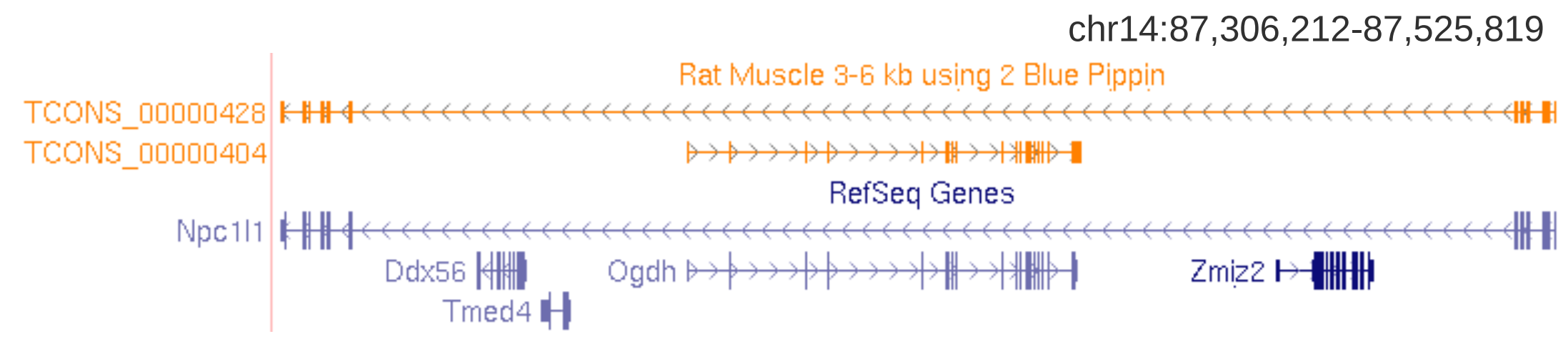


Figure 6. (b) Anti-sense transcription matches known annotation. Orientation of PacBio consensus transcripts are determined by the presence of polyA tails. The transcripts are 4.5 and 4.1 kb long.

Conclusion

Isoform Sequencing

- Construct cDNA libraries enriched in full-length transcripts
- Size selection using agarose gel or the BluePippin system
- Sequence transcripts up to 6 kb in full-length
- Single-molecule observation of each transcript

Bioinformatics Analysis

- Identify putatively full-length transcripts
- Isoform-level clustering generates high-quality transcript consensus sequences

Biological Applications

- Alternative splicing
- Alternative polyadenylation
- Retained introns
- Fusion genes
- Anti-sense transcription

Acknowledgements

The rat heart and lung samples were run by the following trainees during an Iso-Seq training: Miwako Kitazume (Tomy Digital Biology) and PacBio employees Greg Khitrov, Harold Lee and Michael Weiand. We are greatly thankful for their participation and hard work.