A Novel Analytical Pipeline for de novo Haplotype Phasing and Amplicon Analysis using SMRT™ Sequencing Technology

Roberto A Lleras, Brett Bowman, Elizabeth Tseng, Susana Wang, John Harting, Primo Baybayan, Swati Ranade, Jason Chin, Kevin Eng, Patrick Marks

Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025;

Abstract

While the identification of individual SNPs has been readily available for some time, the ability to accurately phase SNPs and structural variation across a haplotype has been a challenge. With individual reads of up to 30kb in length, SMRT® Sequencing technology allows the identification of mutation combinations such as microdeletions, insertions, and substitutions without any predetermined reference sequence. Long amplicon analysis is a novel protocol that identifies and reports the abundance of differing clusters of sequencing reads within a single library. Graphs generated via hierarchical clustering of individual sequencing reads are used to generate Markov models representing the consensus sequence of individual clusters found to be significantly different. Long amplicon analysis is capable of differentiating between underlying sequences that are 99.9% similar, which is suitable for haplotyping and differentiating pseudogenes from coding transcripts. This protocol allowed for the identification of structural variation in the *MUC5AC* gene sequence, despite the presence of a gap in the current genome assembly. Long amplicon analysis allows for the elucidation of complex regions otherwise missed by other sequencing technologies, which may contribute to the diagnosis and understanding of otherwise mysterious diseases.

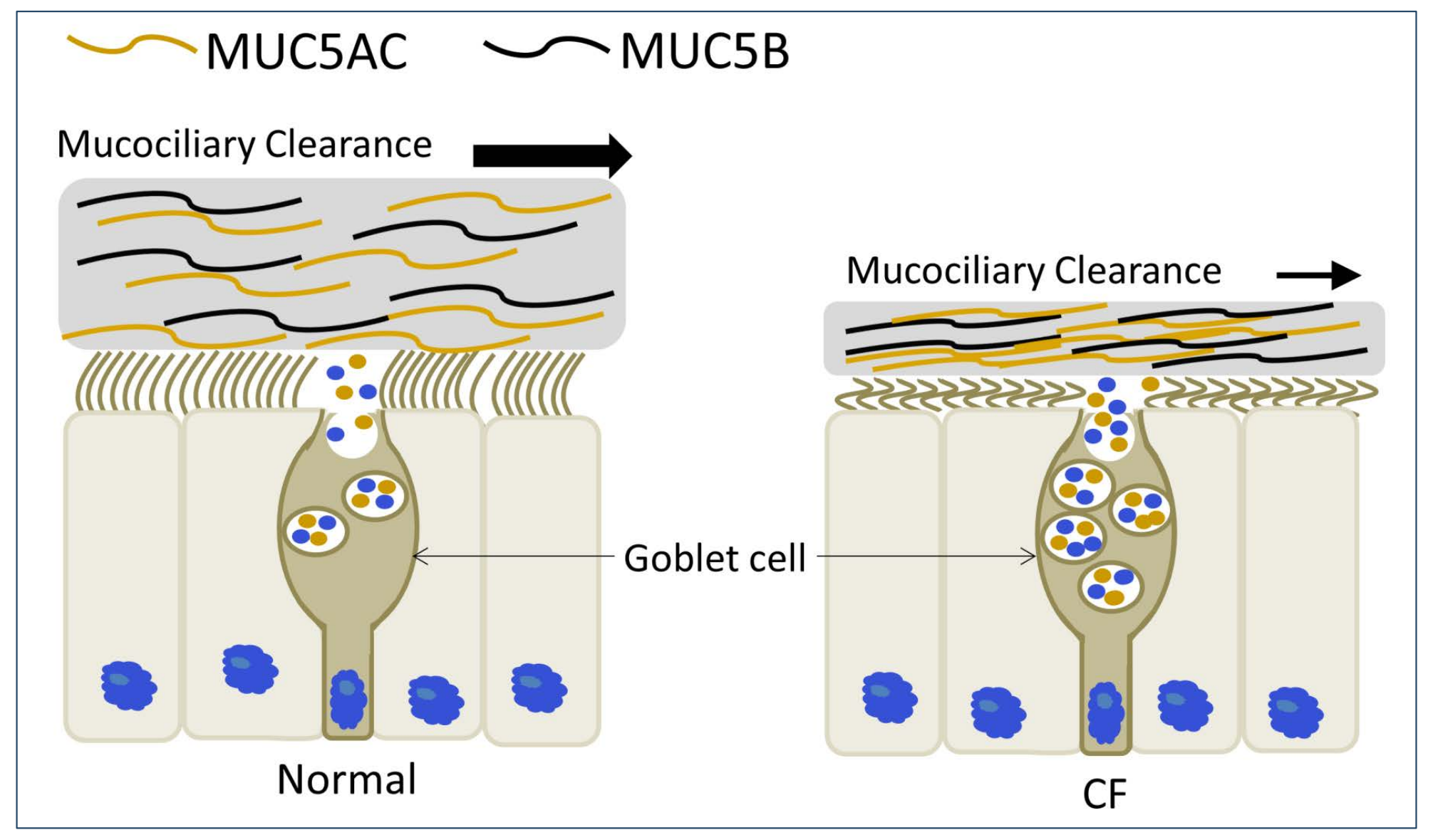

Figure 1. An overview of the function of MUC5AC and MUC5B in mucociliary clearance

Long Amplicon Analysis

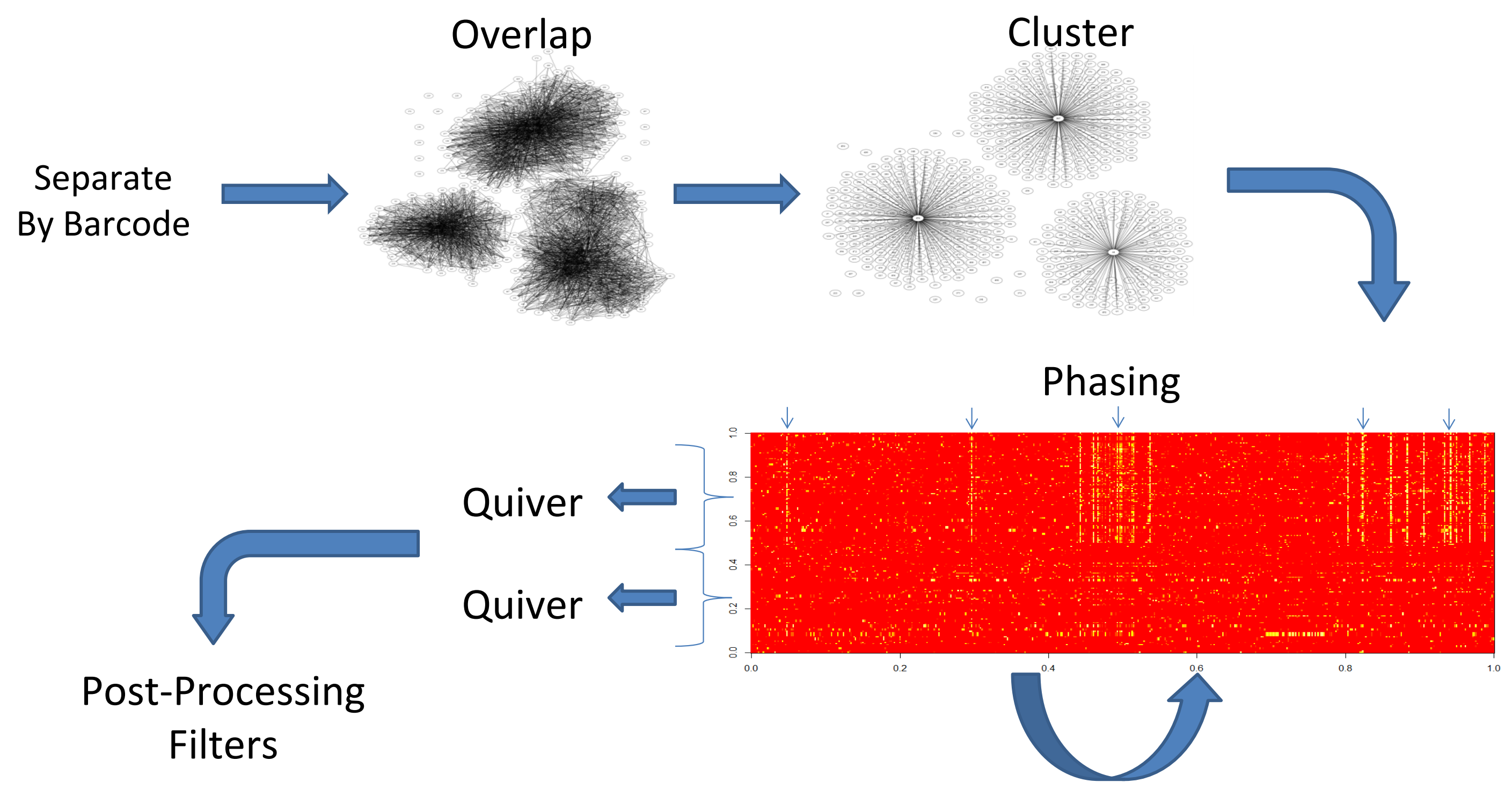

Figure 2: A schematic overview of the long amplicon analysis protocol in SMRTAnalysis v2.2 (to be release April 2014). Reads are identified and binned by barcode. BLASR is utilized to perform rapid pairwise alignments of all reads of a user defined length, and then are clustered with a Markov model based on similarity. Hierarchical clustering is then performed to phase haplotypes together, and Quiver is utilized to create a high quality consensus of individual clusters. Post processing filters remove experimental byproducts. .

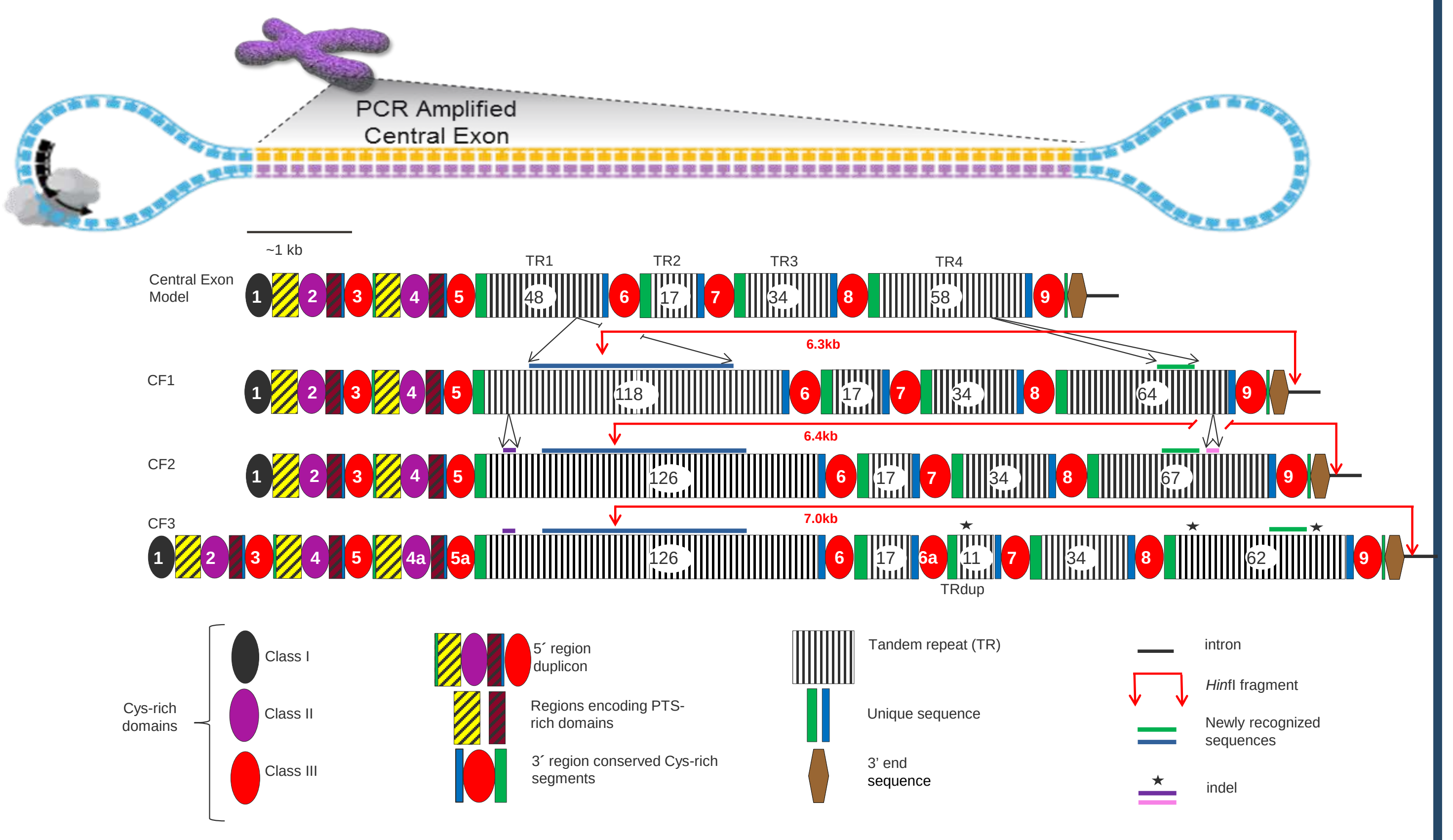

Figure 3: Full length analysis of PCR amplified central exon in MUC5AC of patients reveals structural variation. Patients CF1-CF3 show a 1.9kb expansion of the PTS-TR1 region (blue bar) and a 216 bp expansion of PTS-TR4 (green bar). The increase in the PTS-TR lengths, when compared to the previously known model and more specifically shown in the central exon model of this figure, effectively link the previously available genomic fragments (Figure 1) into one unit and completes the central exon sequence.

Full Length Isoform Sequencing

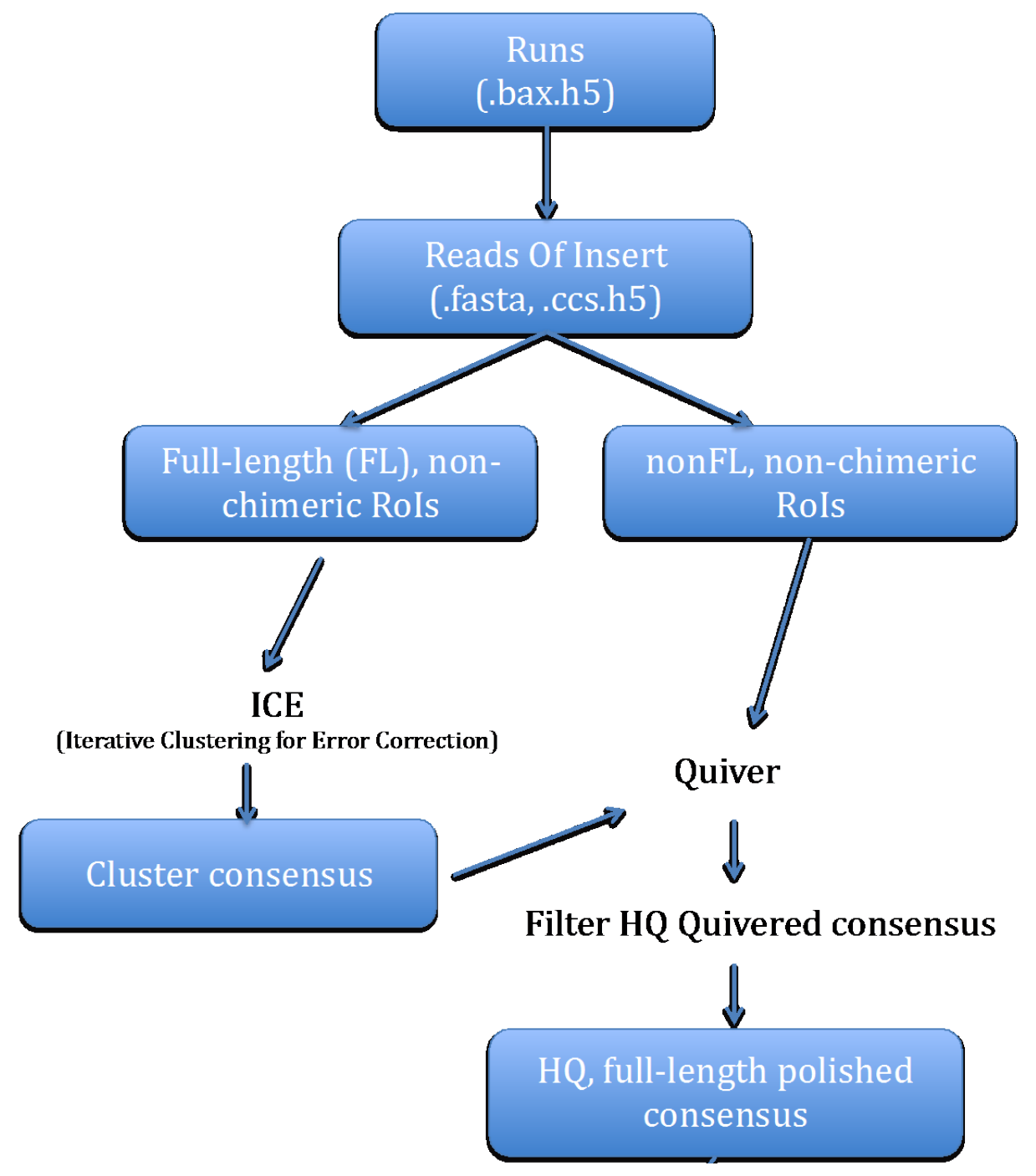
Runs (.bax.h5)
Reads Of Insert (.fasta, .ccs.h5)
Full-length (FL), non-chimeric RoIs
nonFL, non-chimeric RoIs
ICE
(Iterative Clustering for Error Correction)
Cluster consensus
Quiver
Filter HQ Quivered consensus
HQ, full-length polished consensus

Figure 4. 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.

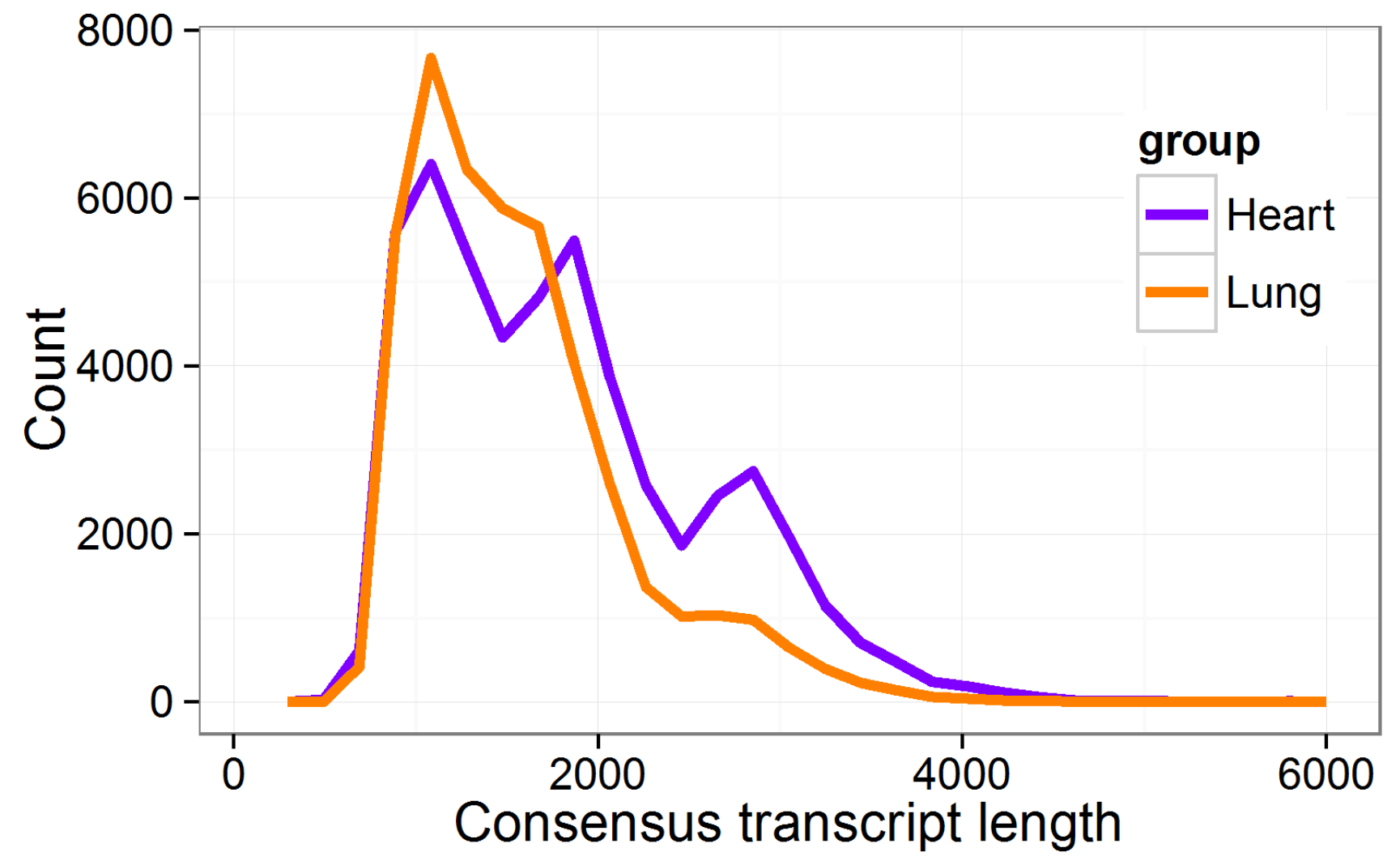

Figure 5. Length distribution of transcript consensus sequences. Each consensus sequence represents the consensus call of a cluster of reads that are considered to be from the same isoform. Since the cDNA 5'/3' primers and polyA tail were used to determine the full-length reads, the consensus transcripts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

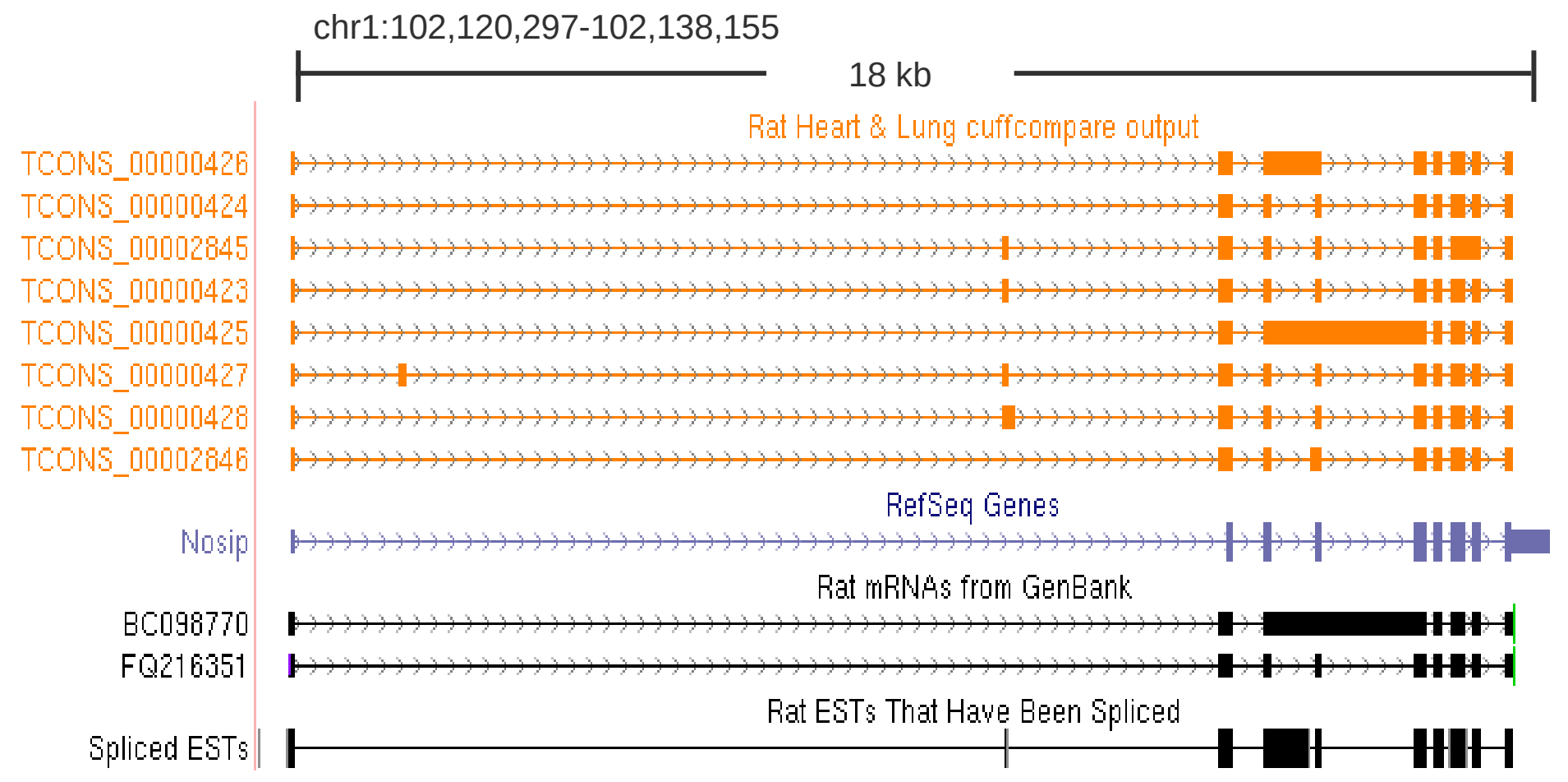

Figure 6. Multiple isoforms observed at a single loci. This UCSC Genome Browser screenshot shows a locus encoding multiple isoforms observed in the PacBio data (top, orange) with alternative splicing and possibly retained introns. Isoforms observed in each sample are marked with (heart) or (lung). See also Poster 44.

Results

Sample ID	HLA-A		HLA-B		HLA-C	
	Allele1	Allele2	Allele1	Allele2	Allele1	Allele2
TU01	*A*02:06:01*	*A*11:01:01*	*B*40:02:01*	*B*55:02:01:02*	*C*01:02:01*	*C*03:03:01*
TU02	*A*02:01:01:01*	*A*31:01:02*	*B*51:02:01*	*B*56:01:01:02*	*C*01:02:01*	*C*03:04:01:02*
TU03	*A*24:02:01:01*	*A*31:01:02*	*B*07:02:01*	*B*35:01:01:02*	*C*03:03:01*	*C*07:02:01:03*
TU04	*A*02:06:01*	*A*02:07:01*	*B*40:02:01*	*B*44:03:01*	*C*03:03:01*	*C*14:03*
TU05	*A*26:01:01*	*A*31:01:02*	*B*15:01:01:01*	*B*35:01:01:02*	*C*03:04:01:02*	*C*07:02:01:04*
TU06	*A*26:03:01*	*A*33:03:01*	*B*15:11:01*	*B*44:03:01*	*C*03:03:01*	*C*14:03*
TU07	*A*02:03:01*	*A*24:02:01:01*	*B*38:02:01*	*B*54:01:01*	*C*01:02:01*	*C*07:02:01:05*
TU08	*A*24:02:01:01*	*A*33:03:01*	*B*44:03:01*	*B*48:01:01*	*C*08:03:01*	*C*14:03*
TU09	*A*02:01:01:01*	*A*02:06:01*	*B*40:06:01:01*	*B*48:01:01*	*C*08:01:01*	*C*15:02:01*
TU10	*A*11:01:01*	*A*31:01:02*	*B*40:01:02*	*B*51:01:01*	*C*07:02:01:01*	*C*15:02:01*
TU21	*A*03:02:01*	*A*24:02:01:01*	*B*07:02:01*	*B*13:02:01*	*C*06:02:01:01*	*C*07:02:01:03*

Sample ID	*HLA-DRB1*	
	Allele1	Allele2
	Allele name	Allele name
TU01	*DRB1*09:01:02:01/02*	*DRB1*15:01:01:03*
TU02	*DRB1*09:01:02:02*	*DRB1*14:05:01:02*
TU03	*DRB1*01:01:01*	*DRB1*14:05:01:02*
TU04	*DRB1*04:10:03:01*	***DRB1*14:54:01:02***
TU05	*DRB1*09:01:02:01*	*DRB1*13:02:01:02*
TU06	*DRB1*04:05:01:01*	*DRB1*13:02:01:02*
TU07	*DRB1*04:03:01:02*	*DRB1*08:03:02:02*
TU08	*DRB1*13:02:01:02*	*DRB1*16:02:01:02*
TU09	*DRB1*14:05:01:02*	–
TU10	*DRB1*09:01:02:01*	*DRB1*12:01:01:02*
TU21	*DRB1*01:01:01*	*DRB1*07:01:01:01*

- 100% concordance with the cDNA references
- One mismatch in intron 2 of TU04 *DRB1*14:54:01:02 compared to SS-SBT generated reference*
- *Alleles were correctly assigned based on PacBio consensus sequences resolving all the ambiguities in the PCR-SSO typing of these samples*

Fig 7. HLA class I (A, B and C) and class II gene (DRB1) typing by comparison to Tokai University reference

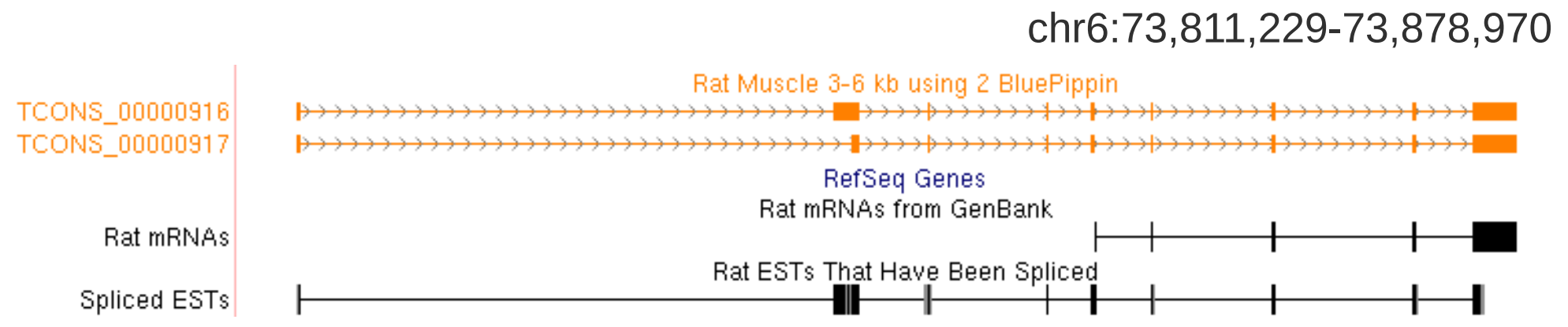
chr6:73,811,229-73,878,970
Rat Muscle 3-6 kb using 2 BluePippin
TCONS_00000916
TCONS_00000917
RefSeq Genes
Rat mRNAs from GenBank
Rat mRNAs
Rat ESTs That Have Been Spliced
Spliced ESTs

Figure 8. (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.

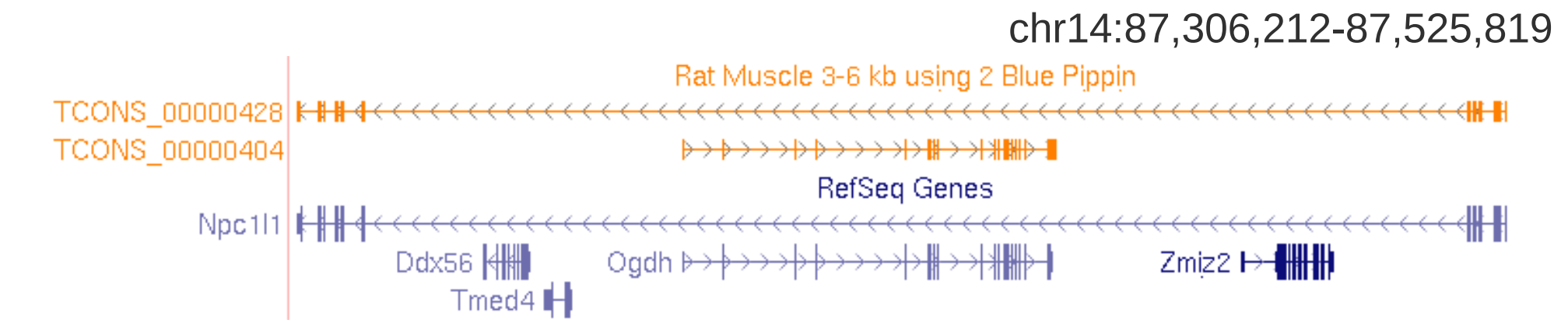

Figure 8. (b) Anti-sense transcription matches known annotation. Orientation of PacBio consensus transcripts are determined by the presence of polyA tails. The transcripts are 4.5 and 4.1 kb long.

Acknowledgements

The authors thank Damian Goodridge and David Sayer, Conexio Genomics, for their contribution to the HLA Haplotyping project. The authors would also like to thank Takashi Shiina, Shingo Suzuki, and Yuki Ozaki for their contributions.

The authors thank Michael R. Knowles, M.D, Xueliang Guo, Ph.D, Rhonda G. Pace, B.S, Jaclyn R. Stonebraker, Ph.D., Wanda K. O'Neal, Ph.D., and Corbin D. Jones, Ph.D for their work on the MUC5AC project.

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.