Full-length Transcript Profiling with the Iso-Seq Method for Improved Genome Annotations

Michelle Vierra[1], Sarah B. Kingan[1], Elizabeth Tseng[1], Ting Hon[1], William J. Rowell[1], Jacquelyn Mountcastle[2], Olivier Fedrigo[2], Erich D. Jarvis[2], Jonas Korlach[1]
1.PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025
2.The Rockefeller University, New York, NY & Howard Hughes Medical Institute

Abstract

Incomplete annotation of genomes represents a major impediment to understanding biological processes, functional differences between species, and evolutionary mechanisms. Often, genes that are large, embedded within duplicated genomic regions, or associated with repeats are difficult to study by short-read expression profiling and assembly. In addition, most genes in eukaryotic organisms produce alternatively spliced isoforms, broadening the diversity of proteins encoded by the genome, which are difficult to resolve with short-read methods.

Short-read RNA sequencing (RNA-seq) works by physically shearing transcript isoforms into smaller pieces and bioinformatically reassembling them, leaving opportunity for misassembly or incomplete capture of the full diversity of isoforms from genes of interest. In contrast, Single Molecule, Real-Time (SMRT) Sequencing directly sequences full-length transcripts without the need for assembly and imputation.

Here we apply the Iso-Seq method (long-read RNA sequencing) to detect full-length isoforms and the new IsoPhase algorithm to retrieve allele-specific isoform information for two avian models of vocal learning, Anna's hummingbird (*Calypte anna*) and zebra finch (*Taeniopygia guttata*).

Sample Preparation & Sequencing

Data Generation:

- RNA isolated from tissue stored at -80°C
- A portion of cDNA from each sample was enriched for >5 kb transcripts and pooled with non size-selected cDNA
- cDNA libraries were sequenced on 7 SMRT Cells 1M on the Sequel System

SMRT Cell ID	Reads	Yield	Polymerase Read N50	Subread N50
1	507,715	7.7 Gb	35 kb	5.8 kb
2	517,278	7.4 Gb	32 kb	5.8 kb
3	464,945	7.2 Gb	35 kb	5.3 kb
4	347,004	6.1 Gb	38 kb	5.3 kb
5	401,826	5.5 Gb	31 kb	5.3 kb
6	348,472	4.8 Gb	31 kb	5.3 kb
7	252,380	4.6 Gb	38 kb	4.3 kb

Table 1. Performance Statistics for Sequencing. Barcoded libraries were sequenced on seven SMRT Cells 1M. The first four cells were used for annotation, and all seven were used as input for IsoPhase analysis.

Genome Annotation

Iso-Seq transcriptome data from 4 SMRT Cells 1M were used to annotate the most recent hummingbird and zebra finch genome assemblies[1]. Iso-Seq data were analyzed with Iso-Seq2, an advanced version of Iso-Seq analysisthat is currently available in beta release[2,3] in SMRT Link 5.1.

	Zebra Finch	Hummingbird
HQ isoforms	17,451	16,944
Mean length	4,112 bp	3,938 bp
Number of loci	7,258	7,418
Mean isoforms per locus	2.40	2.28
Novel isoforms	901 (5.2%)	1,517 (9.0%)
Mapped back to PacBio reference[1]	17,450 (99%)	16,944 (100%)
Mapped to Previous Reference[4,5]	16,183 (93%)	14,608 (86%)

Table 2. Isoform Characterization and Mapping with ToFU2 and Cupcake. High Quality (HQ): ≥ 2 full-length reads, >99% accuracy. Isoforms belong to same locus if they overlap by at least 1 bp on same strand. Novel isoforms lack BLAST hit to current transcriptome (GCF_000151805.1, GCF_000699085.1) with e-value < 1^{-10}. HQ Isoforms mapped to references with GMAP filters: min. coverage 90%, min. identity 95%.

Example of Full-Length Transcript Capture: New Exons and UTRs

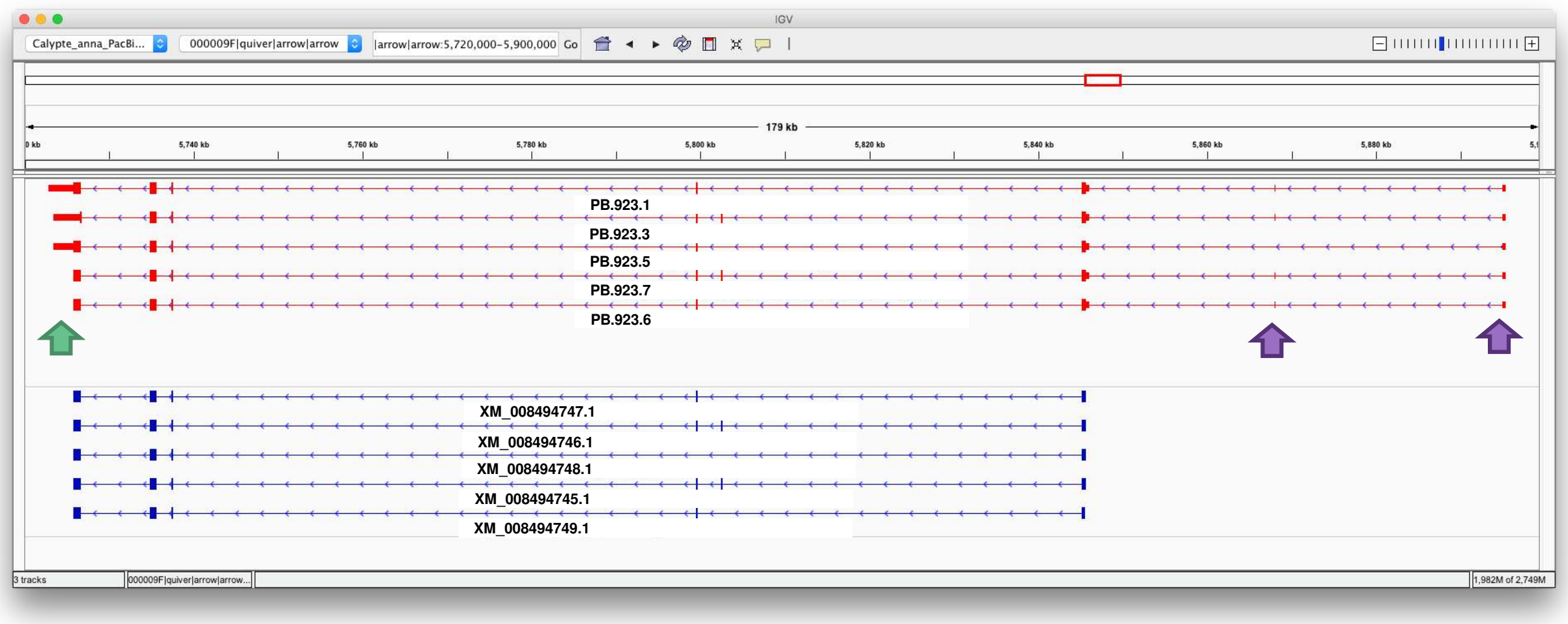
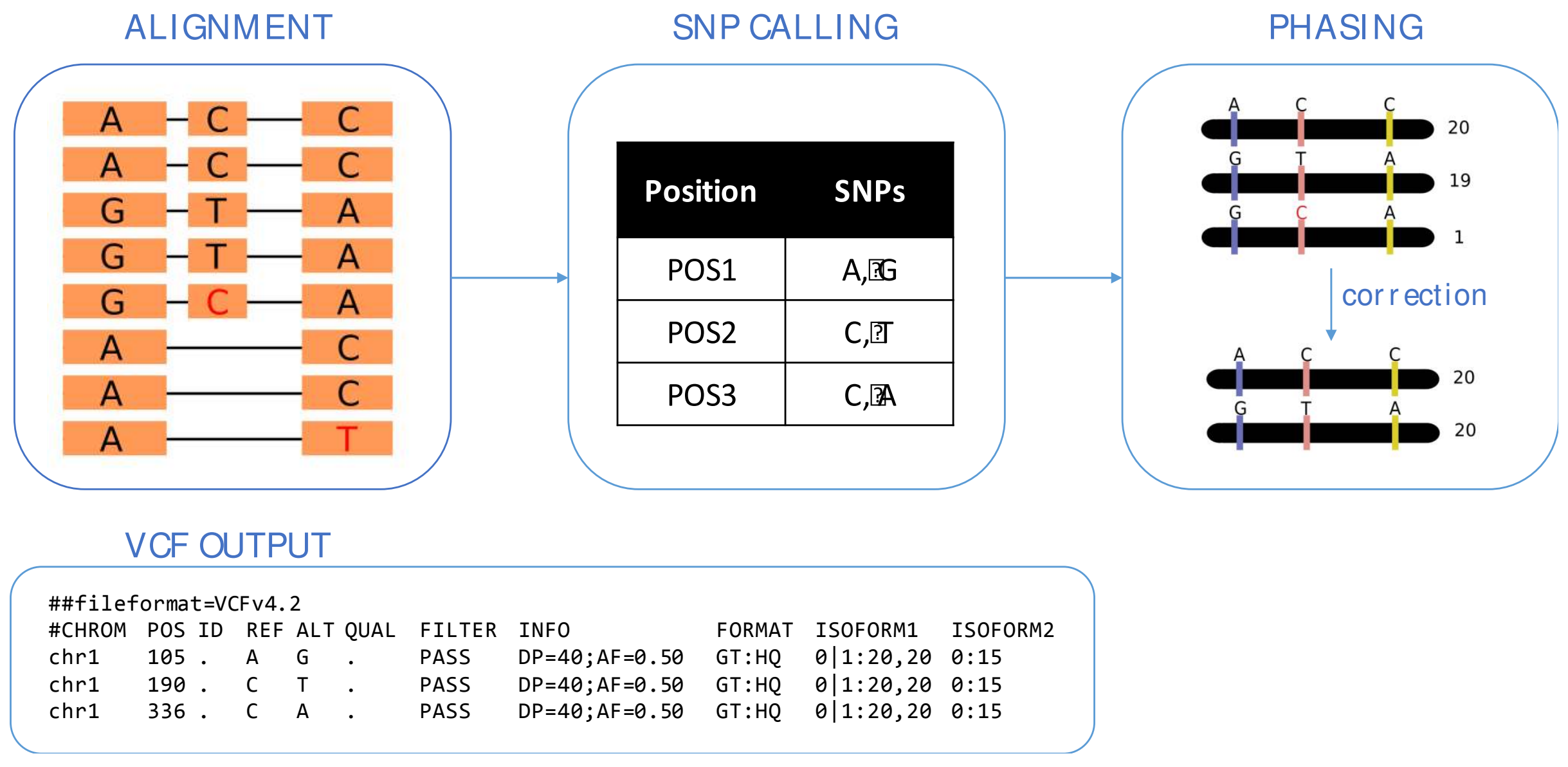


Figure 1. Multiple isoforms for neuroligin-4 in hummingbird. Full-length isoform sequences (red transcript models) identify two additional non-coding 5' exons (purple arrows) and extended 3' UTRs (green arrow), while also capturing all five known splice variants (blue transcript models). Neuroligins are a class of postsynaptic cell-adhesion molecules essential for normal synapse function; neuroligin-4 is implicated in autism spectrum disorder[6]. Isoforms visualized in IGV v3.0_beta.

Isoform Phasing Using Iso-Seq Data



Position	SNPs
POS1	A,G
POS2	C,T
POS3	C,A

```
##fileformat=VCFv4.2
#CHROM  POS ID  REF ALT QUAL  FILTER  INFO           FORMAT  ISOFORM1   ISOFORM2
chr1    105 .   A   G   .     PASS    DP=40;AF=0.50  GT:HQ   0|1:20,20  0:15
chr1    190 .   C   T   .     PASS    DP=40;AF=0.50  GT:HQ   0|1:20,20  0:15
chr1    336 .   C   A   .     PASS    DP=40;AF=0.50  GT:HQ   0|1:20,20  0:15
```

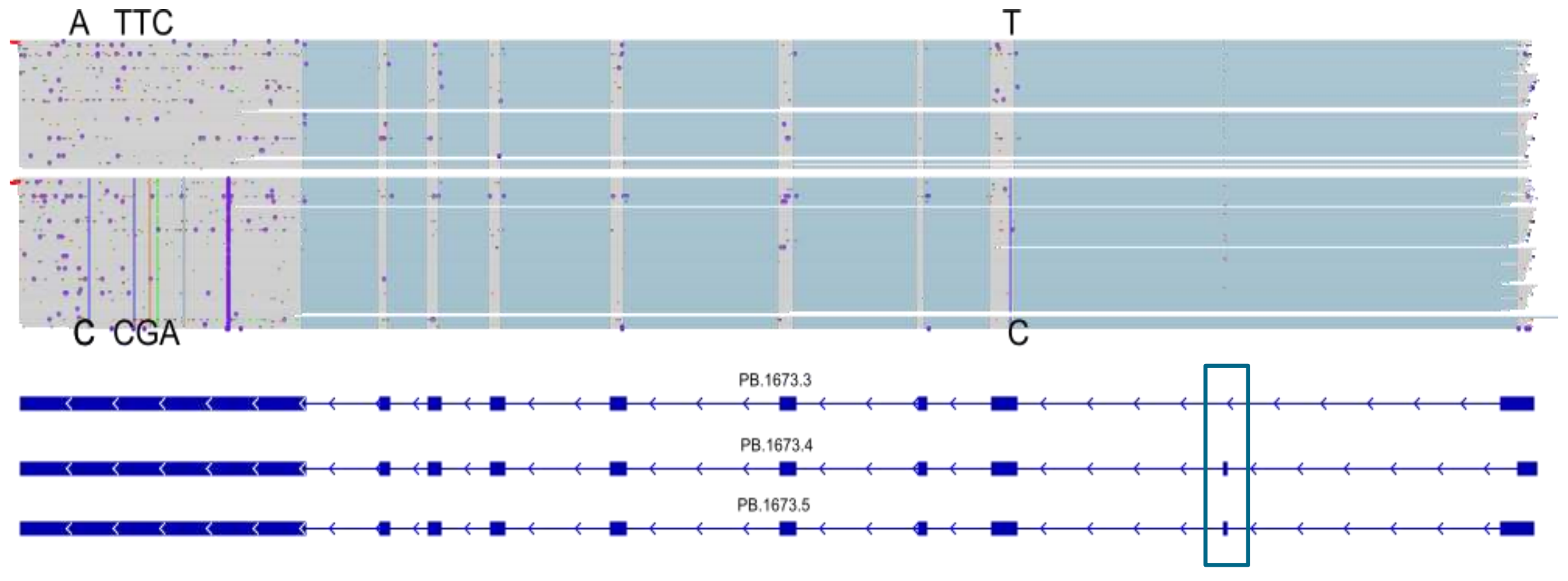
Figure 2. IsoPhase algorithm. IsoPhase takes full-length CCS reads and aligns them to a reference genome to get per-gene coverage. First, individual SNPs are called. Then, full-length reads are used to infer haplotypes. Residual sequencing errors are corrected to get to the number of expected alleles.

SNP Evaluation with IsoPhase

SNP Type	Count
SNPs called by Genome & IsoPhase	7462
SNPs called by IsoPhase Only (in regions not phased by genome)	1883
SNPs called by Genome Only	979

Table 3. SNP evaluation in hummingbird. Genome SNPs were called by mapping FALCON-Unzip alternative haplotigs back to primary contigs using mummer 4.0.0[7]. Iso-Seq SNPs were called using IsoPhase on 1,714 gene families that had ≥40 Iso-Seq full-length read coverage.

Example of Allele-Specific Isoform Expression



Isoform ID	Allele 1 (ATTCT) [# transcripts]	Allele 2 (CCGAC) [# transcripts]
PB.1673.3	37	7
PB.1673.4	23	61
PB.1673.5	3	13

Figure 3. Allele-specific isoform expression of HAGH gene in Hummingbird brain. The hydroxyacylglutathione hydrolase (HAGH) gene expressed three dominant isoforms where one allele was expressed dominantly in the isoform that skipped exon 2 (PB.1673.3) and the other allele was expressed dominantly in isoforms that included exon 2 (PB.1673.4, PB.1673.5).

Conclusion

Long-read RNA SMRT Sequencing has demonstrated utility for:

- Surveying transcript diversity for genome annotation without RNA-seq assembly
- Characterizing novel isoforms
- Detecting new exon and UTR sequences
- Phasing Iso-Seq transcriptome data and identifying allele-specific expression
- Phasing SNPs in low-heterozygosity regions of the genome

References

1. Korlach, J. et al. (2017). De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. *Gigascience* 6(10), 1-16.
2. Tseng E. (2017). https://github.com/PacificBiosciences/IsoSeq_SA3nUP/
3. Tseng E. (2017) https://github.com/Magdoll/cDNA_Cupcake/
4. Warren WC, et al. (2010). The genome of a songbird. *Nature* 464(7289), 757-762.
5. Zhang G, et al. (2014). Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* 346(6215),1311-1320.
6. Südhof TC (2008). Neuroligins and neurexins link synaptic function to cognitive disease. *Nature* 455(7215), 903–911.
7. Mummer4 https://github.com/mummer4/mummer