Haplotyping of full-length transcript reads from long-read sequencing can reveal allelic imbalances in isoform expression

Elizabeth Tseng[1], Timothy P.L. Smith[2], Sarah B. Kingan[1], Stefan Hiendleder[3], Cynthia Liu[3], John L. Williams[3]

1 PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

2 USDA-ARS, U.S. Meat Animal Research Center, Clay Center, Nebraska

3 Davies Research Centre, School of Animal and Veterinary Sciences, University of Adelaide, Australia

Abstract

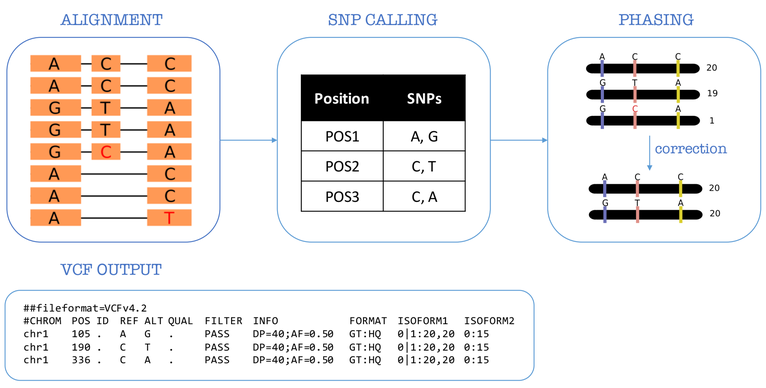
The Pacific Biosciences Iso-Seq method, which can produce high-quality isoform sequences of 10 kb and longer, has been used to annotate many important plant and animal genomes. Here, we develop an algorithm called IsoPhase that post-processes Iso-Seq data to retrieve allele-specific isoform information.

Using simulated data, we show that for both diploid and tetraploid genomes, IsoPhase results in good SNP recovery with low FDR at error rates consistent with CCS reads.

We apply IsoPhase to a haplotype-resolved genome assembly and multiple fetal tissue Iso-Seq dataset from a F1 cross of Angus x Brahman cattle subspecies. IsoPhase-called haplotypes were validated by the phased assembly and demonstrate the potential for revealing allelic imbalances in isoform expression.

Workflow

IsoPhase: Isoform Phasing with or without a reference genome



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.

F1 Cattle Genome

Genome Assembly

- Brahman x Angus F1 cattle
- 115-fold coverage on PacBio RS2 and Sequel systems
- Assembled using Falcon
- ~90% of genome phased using Unzip

CONTIG	NUMBER	LENGTH	N50	LONGEST
PRIMARY	1427	2.71 Gb	31.4 Mb	65.3 Mb
HAPLOTIGS	5879	2.45 Gb	2.48 Mb	14.0 Mb

Phasing Simulated Data

Data simulated from 100 randomly selected human genes from Gencode. SNPs were simulated at a rate of 1 per 300 bp. Errors are substitution only. Each allele was simulated for 20 copies. If QVs are used, low quality bases are filtered out by `samtools mpileup --min-BQ=13`.

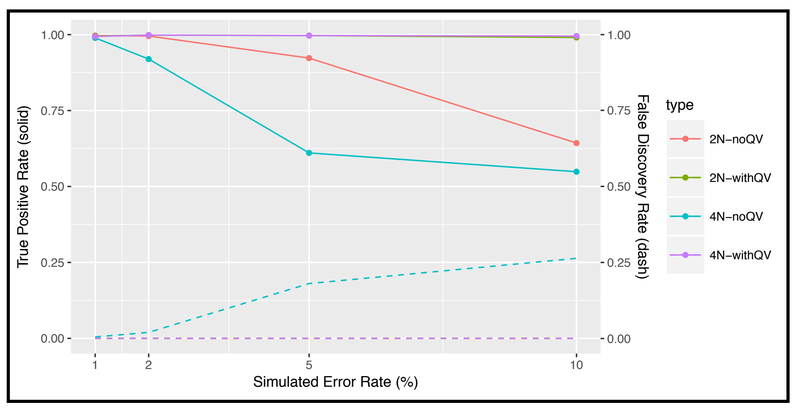


Figure 2. Evaluation of SNP recovery on simulated data at different error rates and ploidy.
(solid line) Fraction of simulated SNPs that were recovered; true positive rate. (dashed line) Fraction of called SNPs that are false; false discovery rate.

Error Rate	2N no QV	2N with QV	4N no QV	4N with QV
1%	100%	100%	95%	98%
2%	100%	100%	82%	98%
5%	100%	100%	55%	96%
10%	100%	93%	40%	84%

Table 1. Fraction of genes with correctly recovered alleles. For each of the 100 simulated genes, correct recovery means each recovered allele (2 for 2N, 4 for 4N) contains the correct phase of the simulated SNPs, disregarding SNPs that were not recovered in Figure 1.

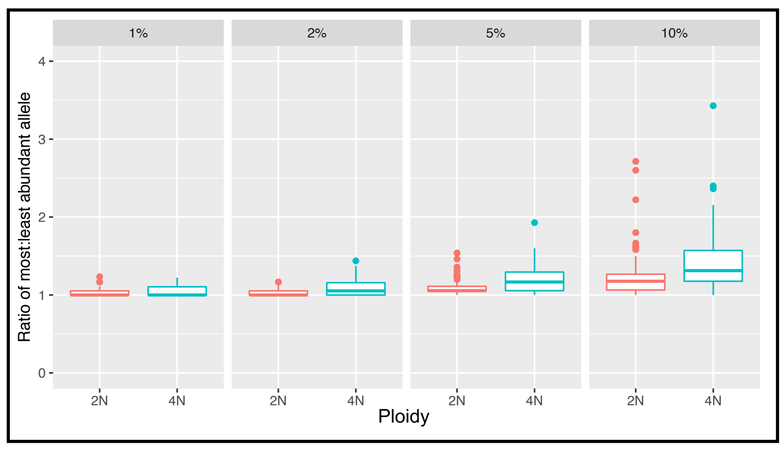


Figure 3. Distribution of recovered allelic ratio. The abundance ratio of the most abundant vs least abundant allele based on the number of supporting reads. Since each allele is simulated with 20 copies, the expected ratio is 1.

Summary from Simulated Data

- Using QV reduces number of falsely predicted SNPs
- For diploid species, correct haplotype construction is possible for up to 5% error (substitutions only)
- Can estimate allelic ratio based on number of supporting full-length reads

Phasing F1 Cattle Data

Iso-Seq Transcriptome Data

- 8 Sequel cells of tissues from single individual
- Analyzed using IsoSeq2 then mapped to genome with ≥99% coverage, ≥95% identity
- 30,137 final isoforms (12,101 genes)
- Selected for phasing: 1758 genes with ≥40 full-length CCS read coverage

SNP Evaluation based on Genome

*Only SNPs with ≥40-fold read coverage are considered.

SNP Type	Count
True Positive (called by both)	8334
False Negative (called by genome only)	259
False Positive (called by transcript only)	1203

Using genome phasing results as truth, IsoPhase SNP calling achieves 97% sensitivity and 87% specificity.

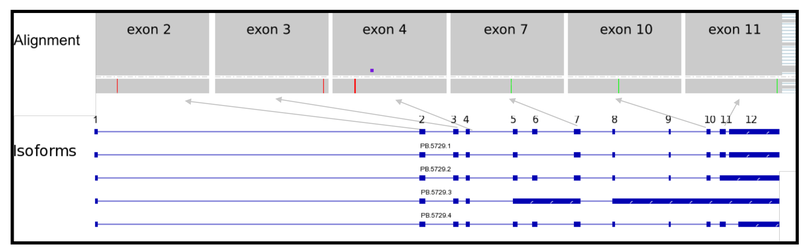


Figure 4. Example of genome-validated SNPs.
(top) Full-length CCS read alignment, showing only exons with SNPs.
(bottom) Isoforms for this gene. All six called SNPs are validated by genome.

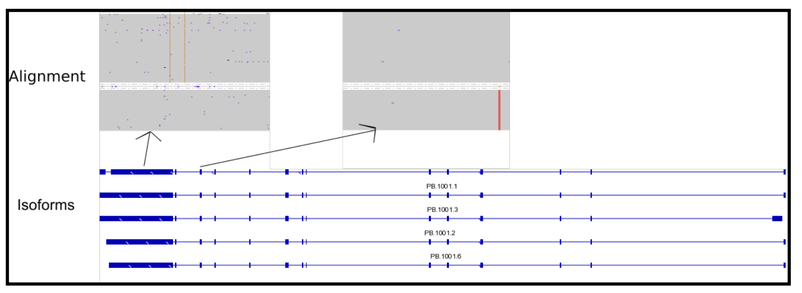


Figure 5. Example of possible novel SNPs.
(top) Full-length CCS read alignment for exon 2 and exon 4
(bottom) Isoforms for this gene. The three SNPs called are not validated by the genome, but are well supported by Iso-Seq data.

Conclusions

- IsoPhase is a tool for calling SNPs and phasing Iso-Seq transcriptome data
- Simulated data shows it is possible to phase diploid and tetraploid species
- Applying IsoPhase to F1 cattle Iso-Seq data shows concordance with genome phasing information

For Research Use Only. Not for use in diagnostics procedures. © Copyright 2018 by Pacific Biosciences of California, Inc. All rights reserved. Pacific Biosciences, the Pacific Biosciences logo, PacBio, SMRT, SMRTbell, Iso-Seq, and Sequel are trademarks of Pacific Biosciences. BluePippin and SageELF are trademarks of Sage Science. NGS-go and NGSengine are trademarks of GenDx. FEMTO Pulse and Fragment Analyzer are trademarks of Advanced Analytical Technologies. All other trademarks are the sole property of their respective owners.