Haplotyping Using Full-Length Transcript Sequencing Reveals Allele-Specific Expression

Elizabeth Tseng[1], Bo Wang[2], Kevin Eng[1], Primo Baybayan[1], Doreen Ware[2]
1 PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025
2 Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724

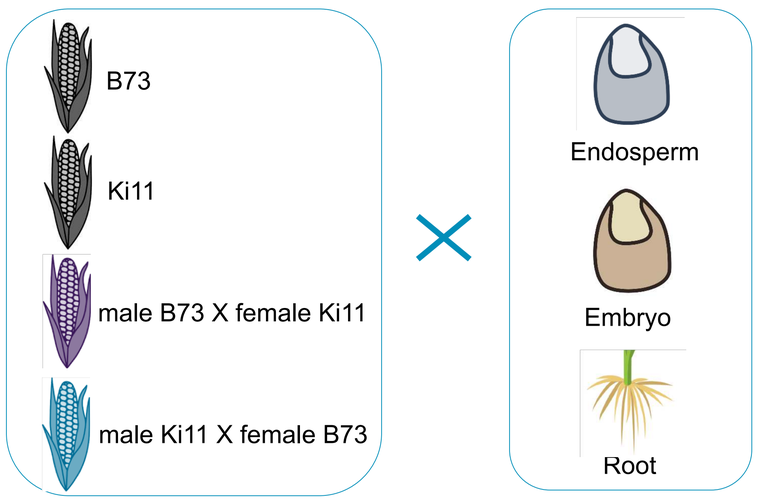
Abstract

An important need in analyzing complex genomes is the ability to separate and phase haplotypes. While whole genome assembly can deliver this information, it cannot reveal whether there is allele-specific gene or isoform expression.

The PacBio Iso-Seq method, which can produce high-quality transcript sequences of 10 kb and longer, has been used to annotate many important plant and animal genomes. We present an algorithm called IsoPhase that post-processes Iso-Seq data for transcript-based haplotyping.

We applied IsoPhase to a maize Iso-Seq dataset consisting of two homozygous parents and two F1 cross hybrids. We validated the majority of the SNPs called with IsoPhase against matching short read data and identified cases of allele-specific, gene-level and isoform-level expression.

Experimental Design



- 12 samples total (4 individuals x 3 tissues)
- Multiplexed using PacBio barcodes
- Sequenced 15 SMRT Cells total on the Sequel System

INDIVIDUAL	TISSUE	FULL-LENGTH READS
B73	Embryo	339,048
Ki11	Embryo	254,342
(P)B73x(M)Ki11	Embryo	305,307
(P)Ki11x(M)B73	Embryo	444,580
B73	Endosperm	284,678
Ki11	Endosperm	290,122
(P)B73x(M)Ki11	Endosperm	232,168
(P)Ki11x(M)B73	Endosperm	288,205
B73	Root	362,431
Ki11	Root	225,208
(P)B73x(M)Ki11	Root	287,485
(P)Ki11x(M)B73	Root	426,238
	TOTAL	3,739,812

Table 1. Number of Full-Length Reads per sample-tissue after demultiplexing as part of the Iso-Seq 3 analysis in SMRT Link 6.0.

Iso-Seq for Genome Annotation

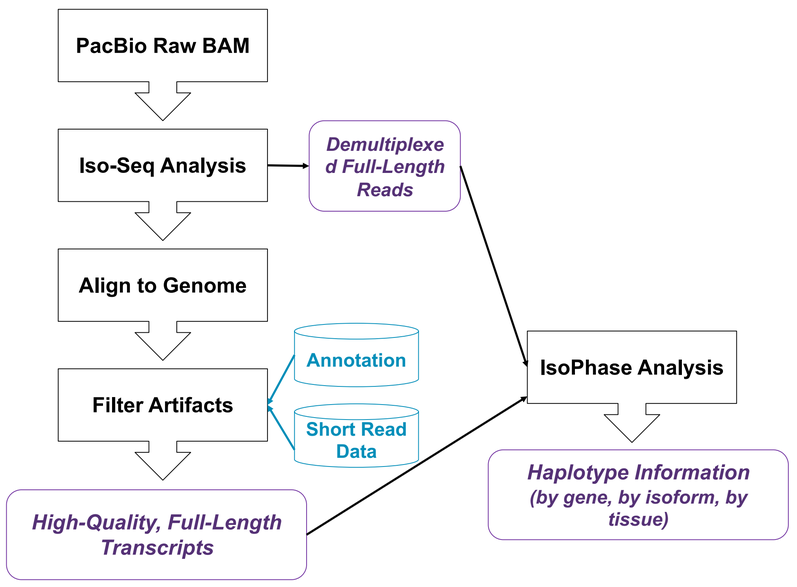


Figure 1. Bioinformatics Analysis Workflow. The Iso-Seq 3 analysis workflow in SMRT Link 6.0 [1, 2] generates full-length, high-quality transcript sequences, which are then mapped to a B73 reference genome using minimap2. Artifacts are filtered using supporting short read junction data and reference annotation. IsoPhase post-processes mapped transcripts to extract isoform-level phasing information [3].

	SQANTI report
# Known Genes	20,068
# Novel Genes	3,344
# Isoforms	75,118

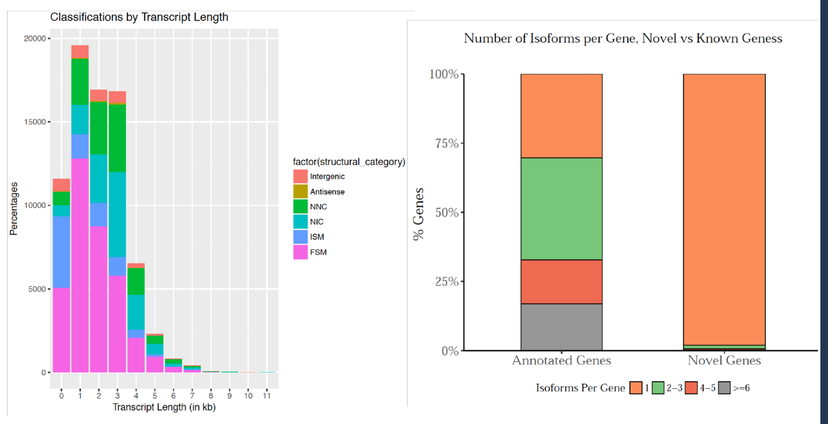


Figure 2. Comparison of Iso-Seq transcripts against B73 v4 annotation using the SQANTI software [4]. SQANTI compares each aligned transcript against a reference annotation and categorizes each transcript as follows:
FSM = perfect match;
ISM = incomplete match/5' degradation;
NIC = novel isoform using all known junctions;
NNC = novel isoform using at least one novel junction

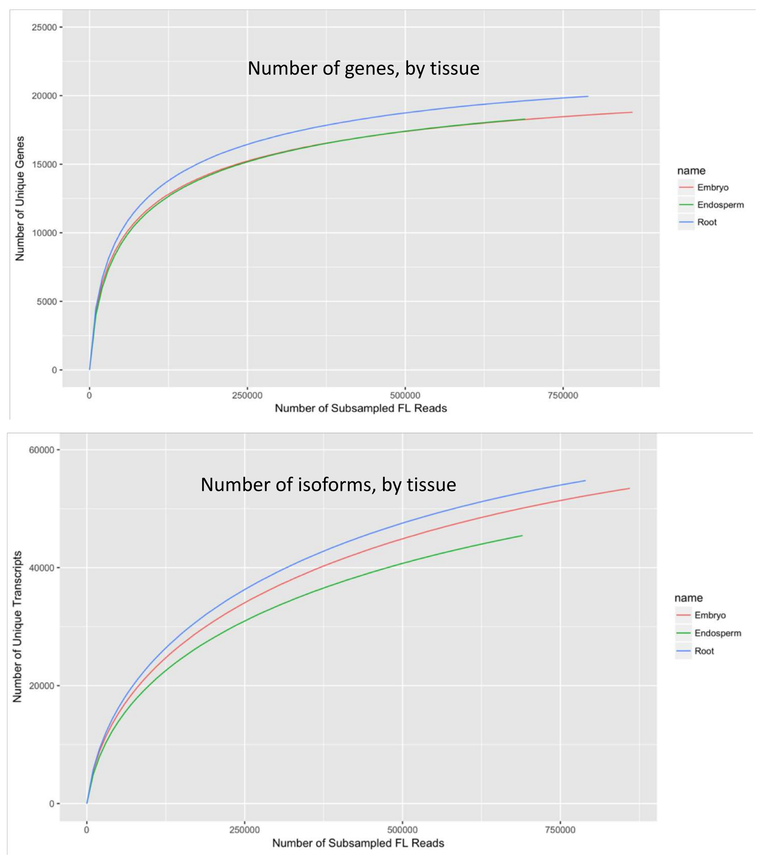


Figure 3. Rarefaction analysis by tissue. Rarefaction curves were drawn by subsampling full-length reads at different depths for 100 iterations and plotting the average number of observed unique genes or transcripts per tissue.

IsoPhase: Isoform-Level Phasing

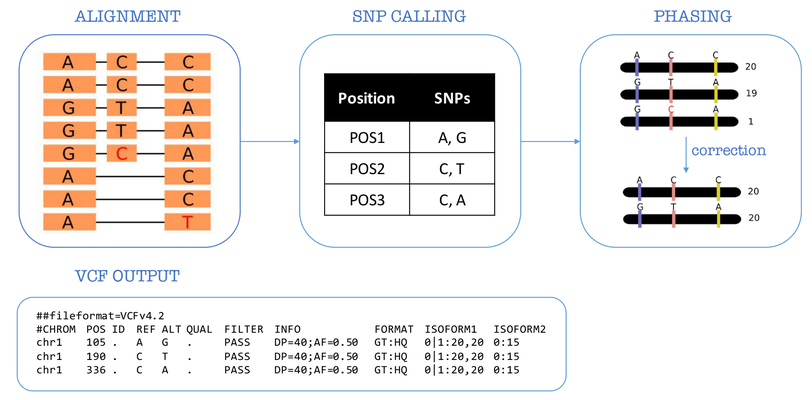


Figure 4. IsoPhase workflow. Full-length reads are mapped to the genome to create a pileup. After substitution SNPs are called, the reads are used to tally haplotype counts followed by error correction.

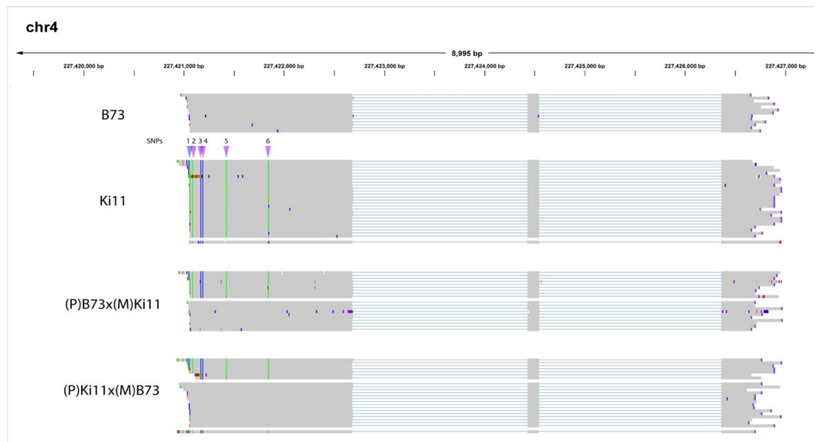


Figure 5. IsoPhase phasing example. Full-length reads are aligned against B73 reference genome. There are 6 SNPs detected by either short read or Iso-Seq data. SNP 2-6 are detected by both. SNP 1 is missed by Iso-Seq due to reduced coverage. The two F1 hybrids express both alleles.

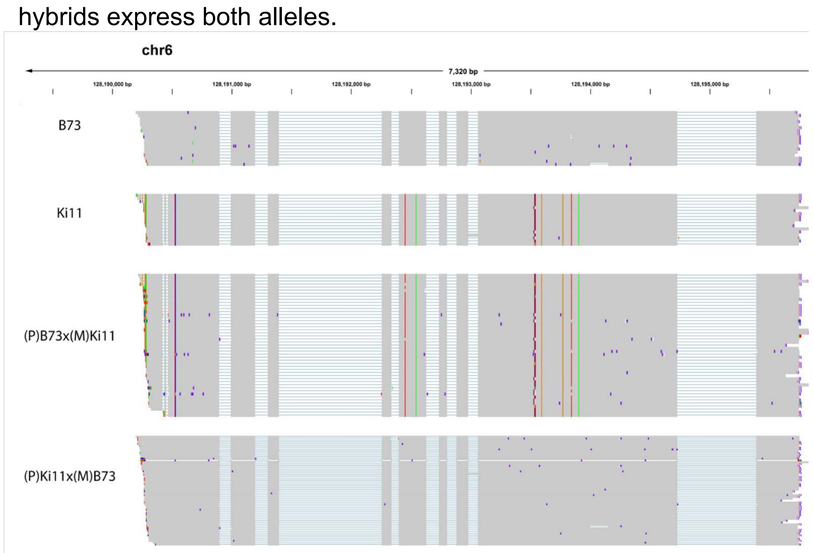


Figure 6. IsoPhase phasing example in which F1 hybrids express only the maternal allele. Full-Length reads are aligned against B73 reference genome. The two F1 hybrids express only the (M)aternal allele.

Conclusions

- The Iso-Seq method generates high-quality, full-length transcripts for genome annotation
- Iso-Seq 3 bioinformatics analysis in SMRT Link 6.0 is fast and works for multiplexed data
- IsoPhase reveals allelic-specific isoform expression

References

1. https://www.pacb.com/software
2. https://github.com/PacificBiosciences/IsoSeq_SA3nUP
3. https://github.com/Magdoll/cDNA_Cupcake
4. Tardaguila, M. *et al.* SQANTI: extensive characterization of long read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Research* (2018)