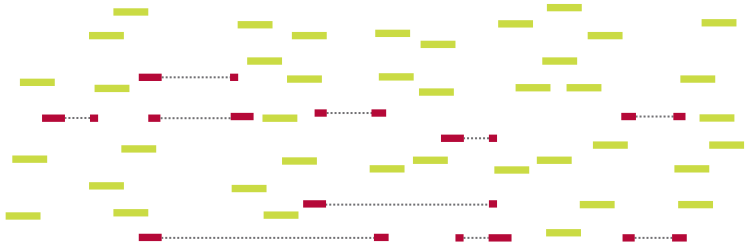
Haplotyping Using Full-Length RNA-seq Reveals Allele-Specific Expression

Elizabeth Tseng, Bo Wang, Kevin Eng, Primo Baybayan, Doreen Ware

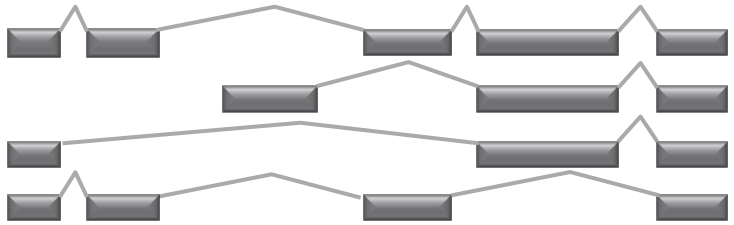
RNA-SEQ METHODS

SHORT READS



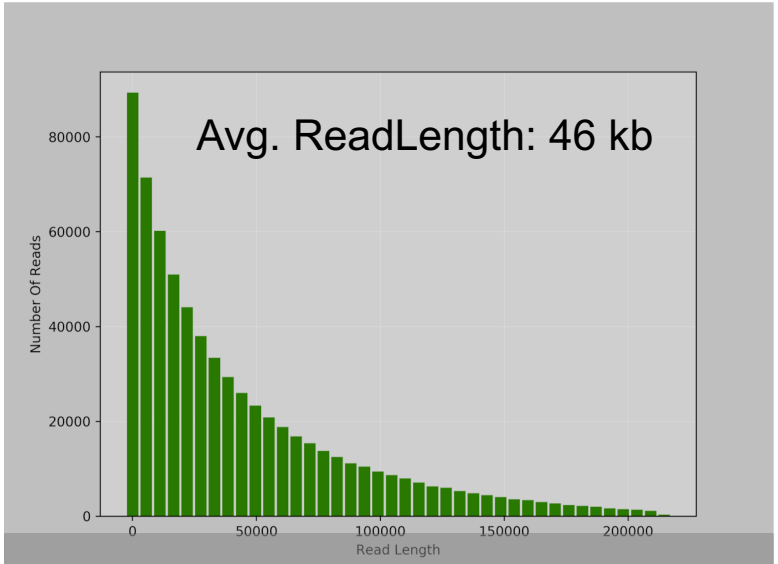
- Fragmented reads
- Challenge with complex splicing
- Deep coverage

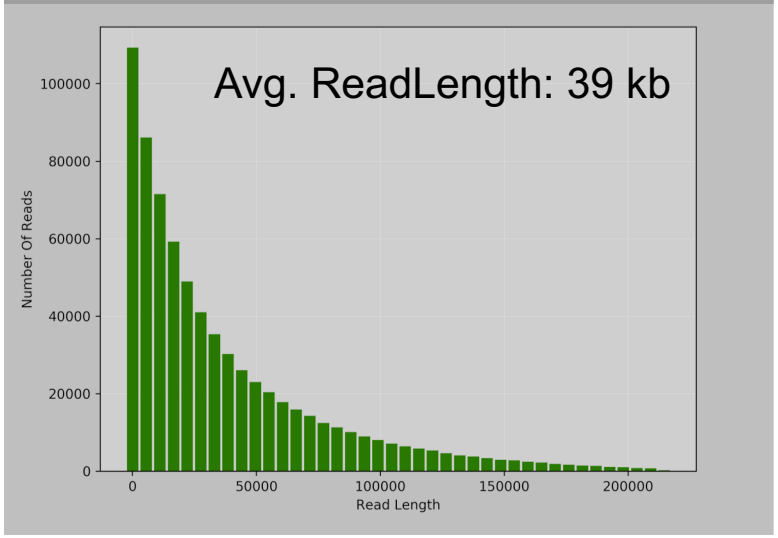
LONG READS



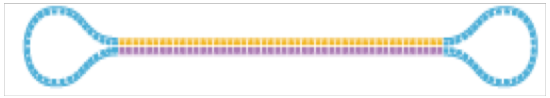
- Full-length transcript
- No assembly required
- Lower throughput

LONG READ LENGTHS ENABLE FULL-LENGTH SEQUENCING



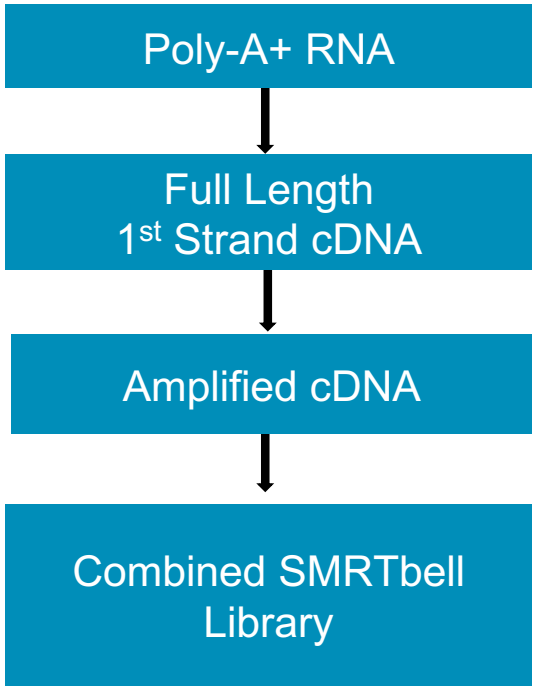


PacBio Sequencing



- Typical transcripts are 0 - 10 kb long

ISO-SEQ: FULL-LENGTH RNA-SEQ ON THE PACBIO PLATFORM



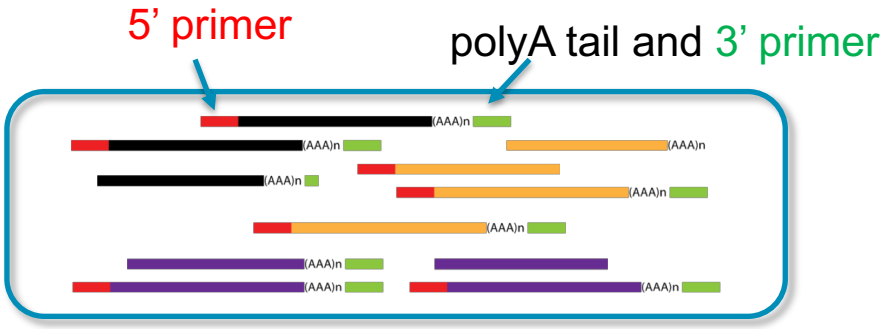
Sample Prep:

- Full-Length cDNA
- One read = one molecule

Analysis:

- Group full-length reads at transcript-level

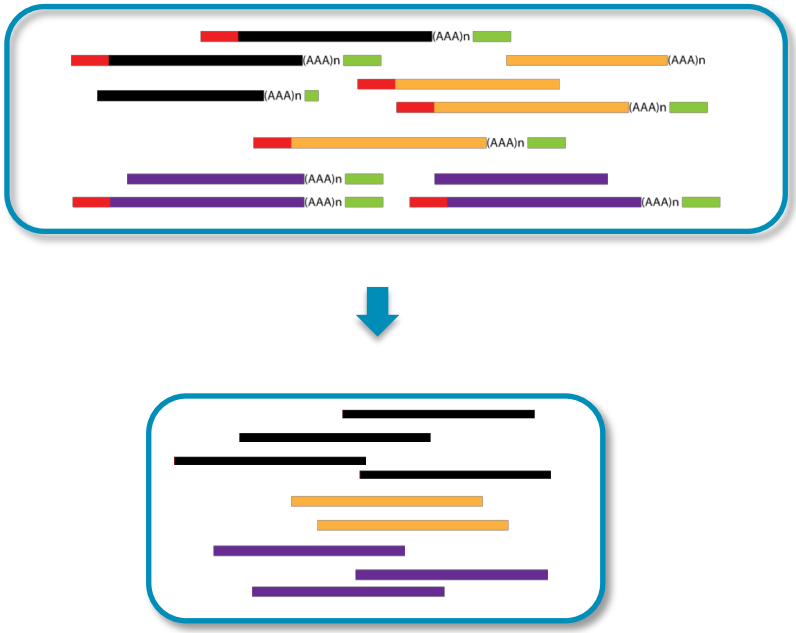
BIOINFORMATICS ANALYSIS FOR FULL-LENGTH RNA-SEQ



Iso-Seq Analysis is available through PacBio’s SMRT Link software

BIOINFORMATICS ANALYSIS FOR FULL-LENGTH RNA-SEQ

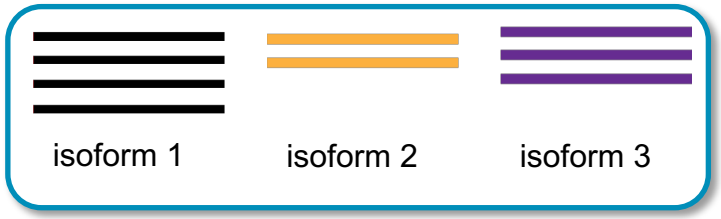
1. Generate Reads
2. Classify Full-Length Reads



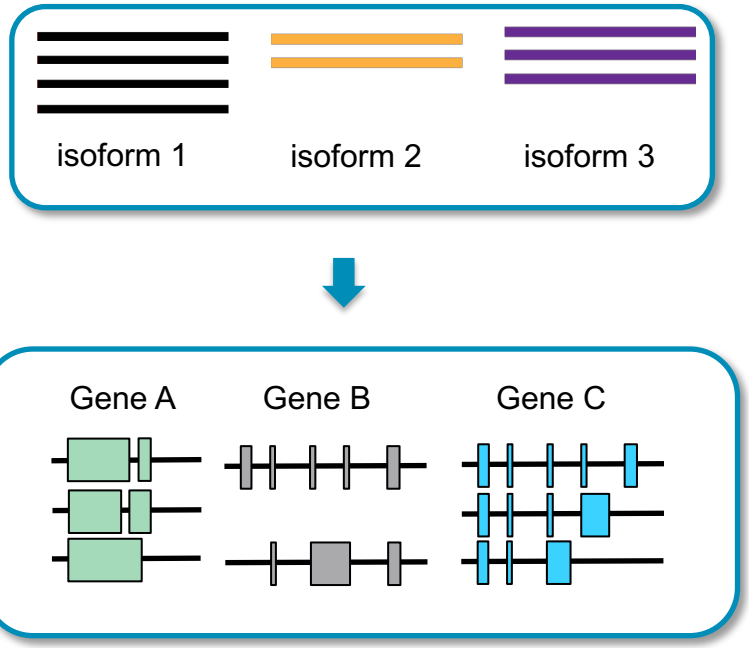
BIOINFORMATICS ANALYSIS FOR FULL-LENGTH RNA-SEQ

3

Cluster and Polish Isoforms



BIOINFORMATICS ANALYSIS FOR FULL-LENGTH RNA-SEQ

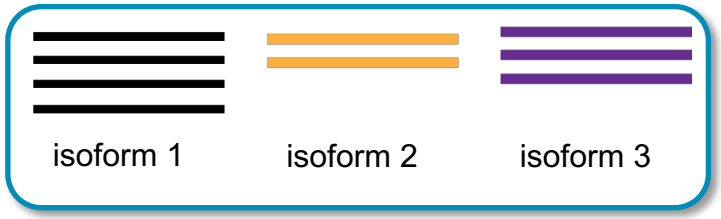


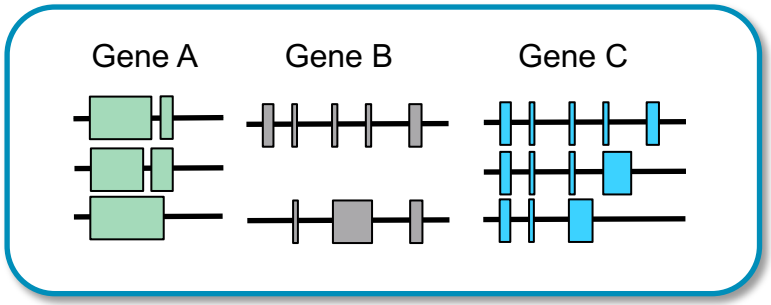
BIOINFORMATICS ANALYSIS FOR FULL-LENGTH RNA-SEQ

3 **Cluster and Polish Isoforms**

4 **Map to Genome**

5 **QC and Annotate**





SQANTI

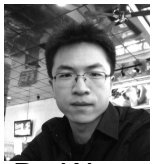
TAPPAS

CAT

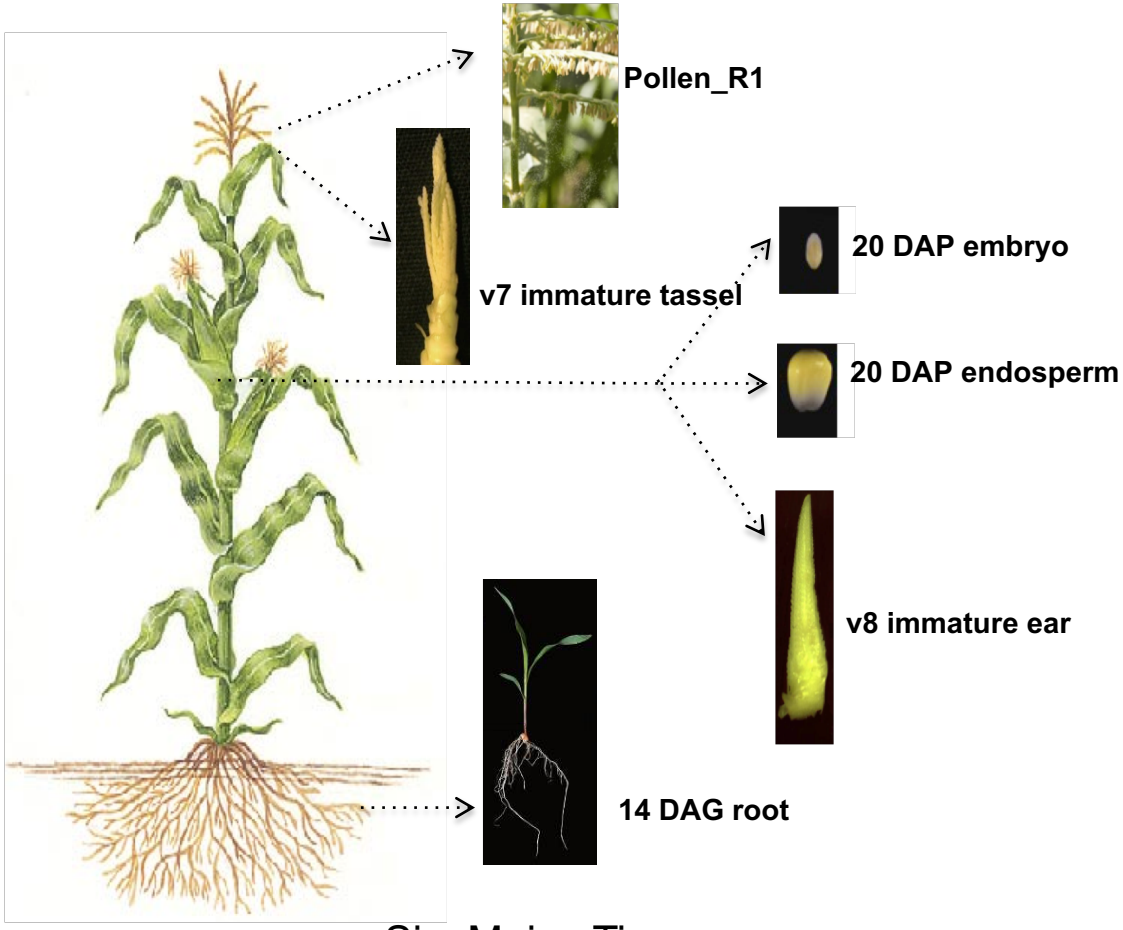
TAMA

History of Maize Iso-Seq

B73 GENOME ANNOTATION (2016)


Bo Wang

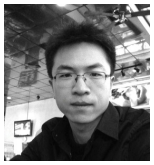

Doreen Ware
(CSHL)

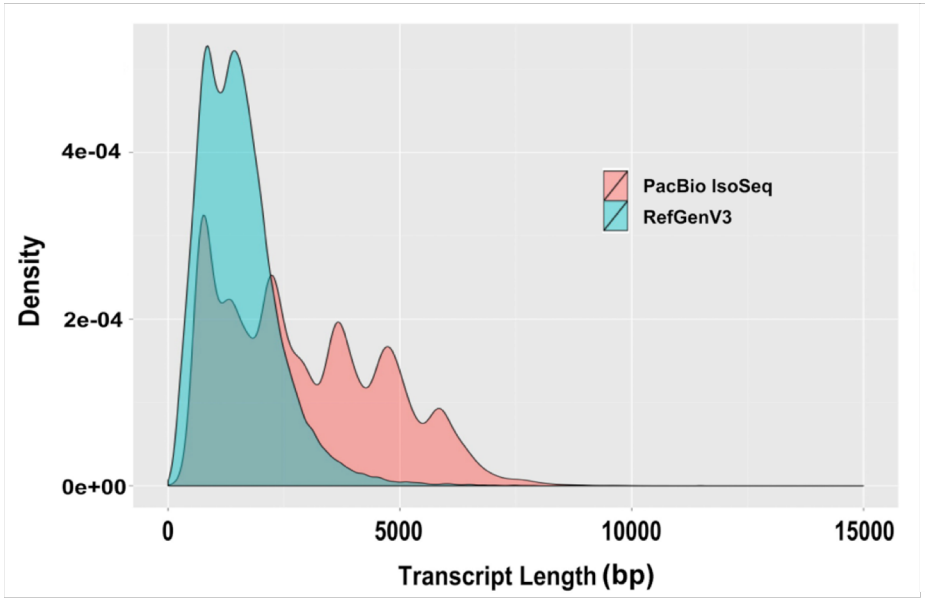


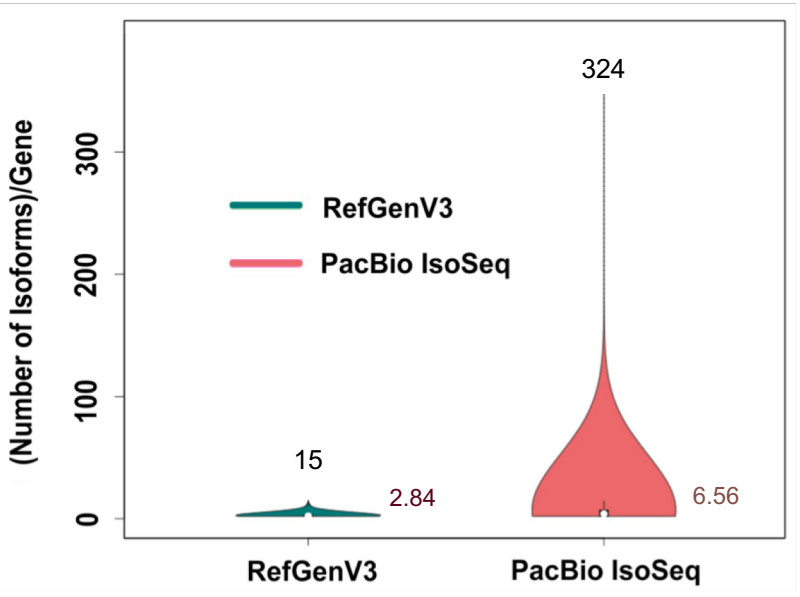
Six Maize Tissues

Wang, B. *et al.* Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat Comms* **7,** 11708 (2016).

B73 GENOME ANNOTATION (2016)

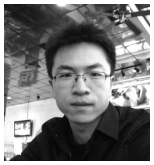


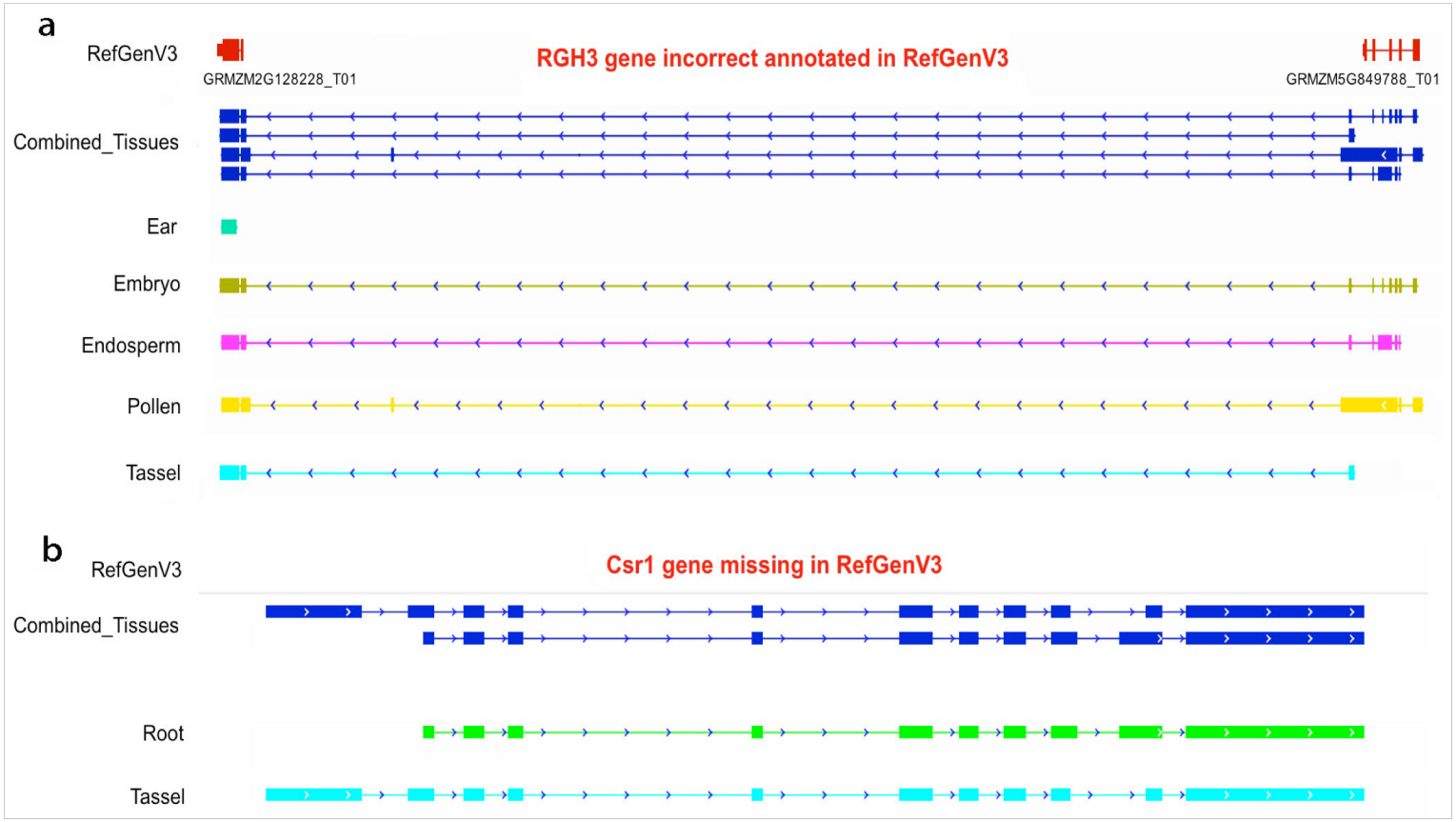




Wang, B. *et al.* Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat Comms* **7,** 11708 (2016).

B73 GENOME ANNOTATION (2016)





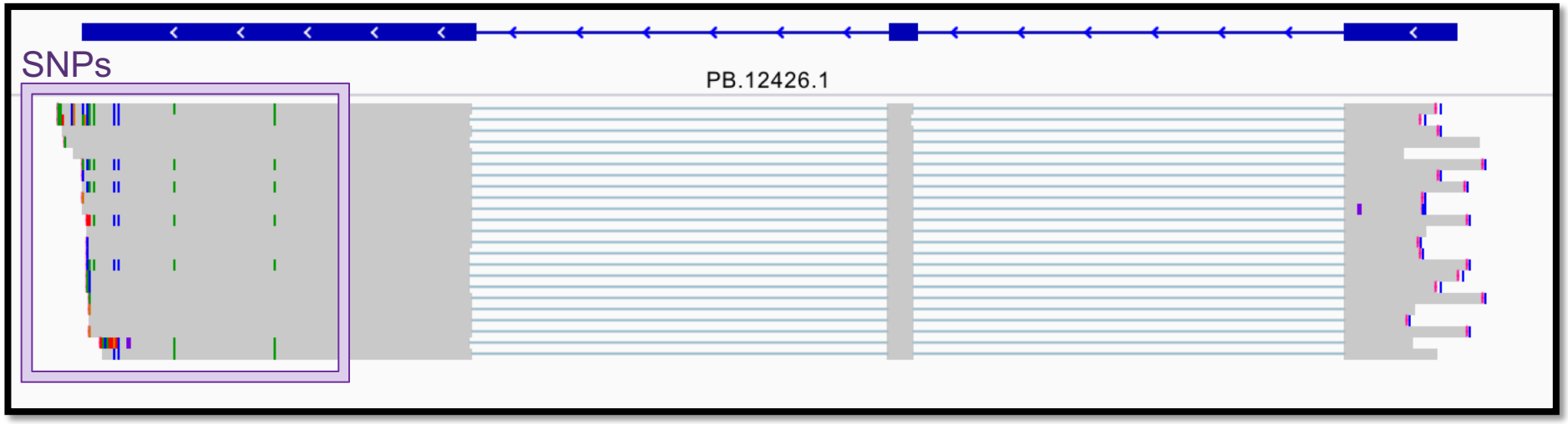
Wang, B. *et al.* Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat Comms* **7,** 11708 (2016).

PACBIO®

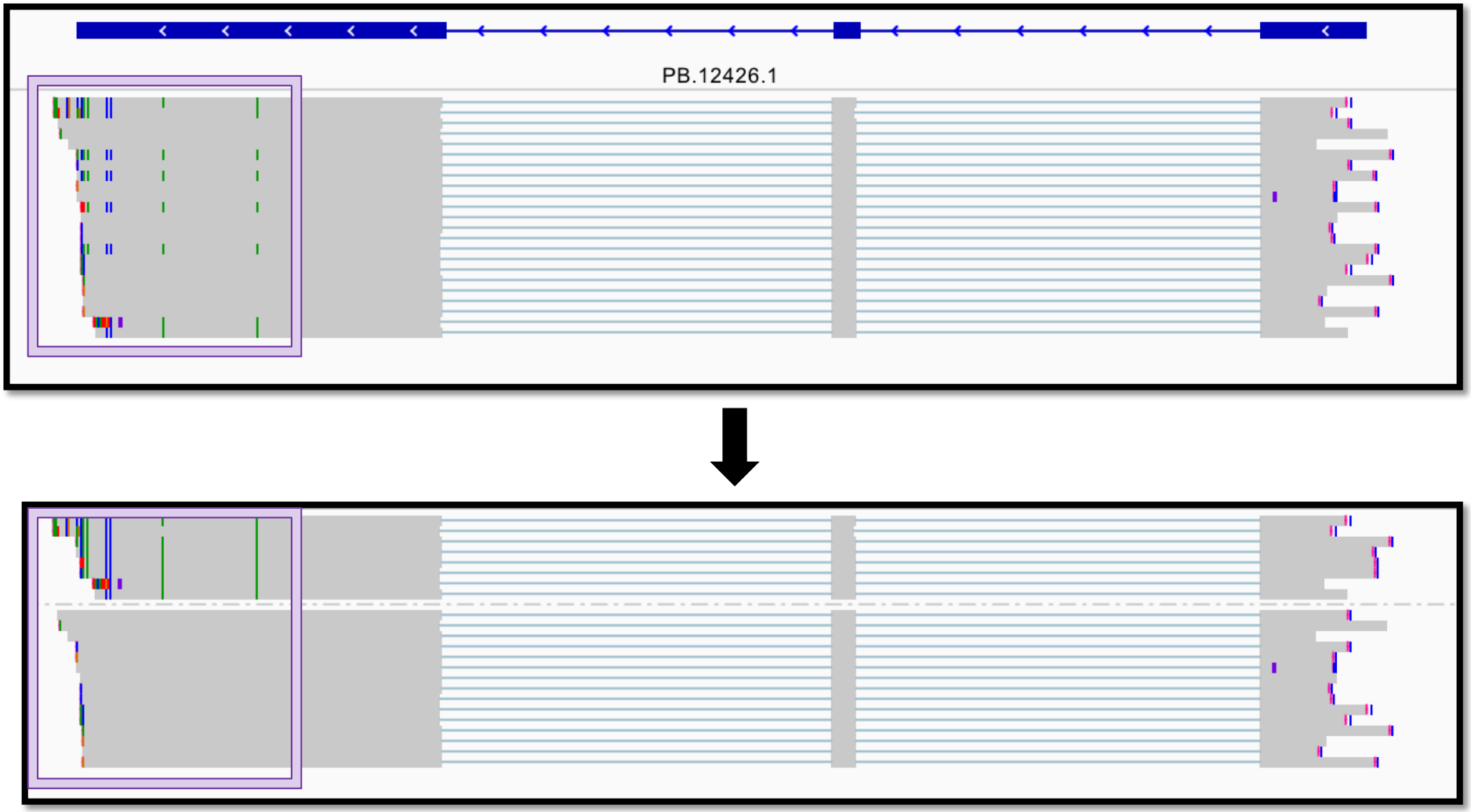
Iso-Seq Phasing on F1 Maize

ISOFORM-LEVEL PHASING IN ISO-SEQ

The full-length and single-molecule nature of Iso-Seq data enables isoform-level phasing.

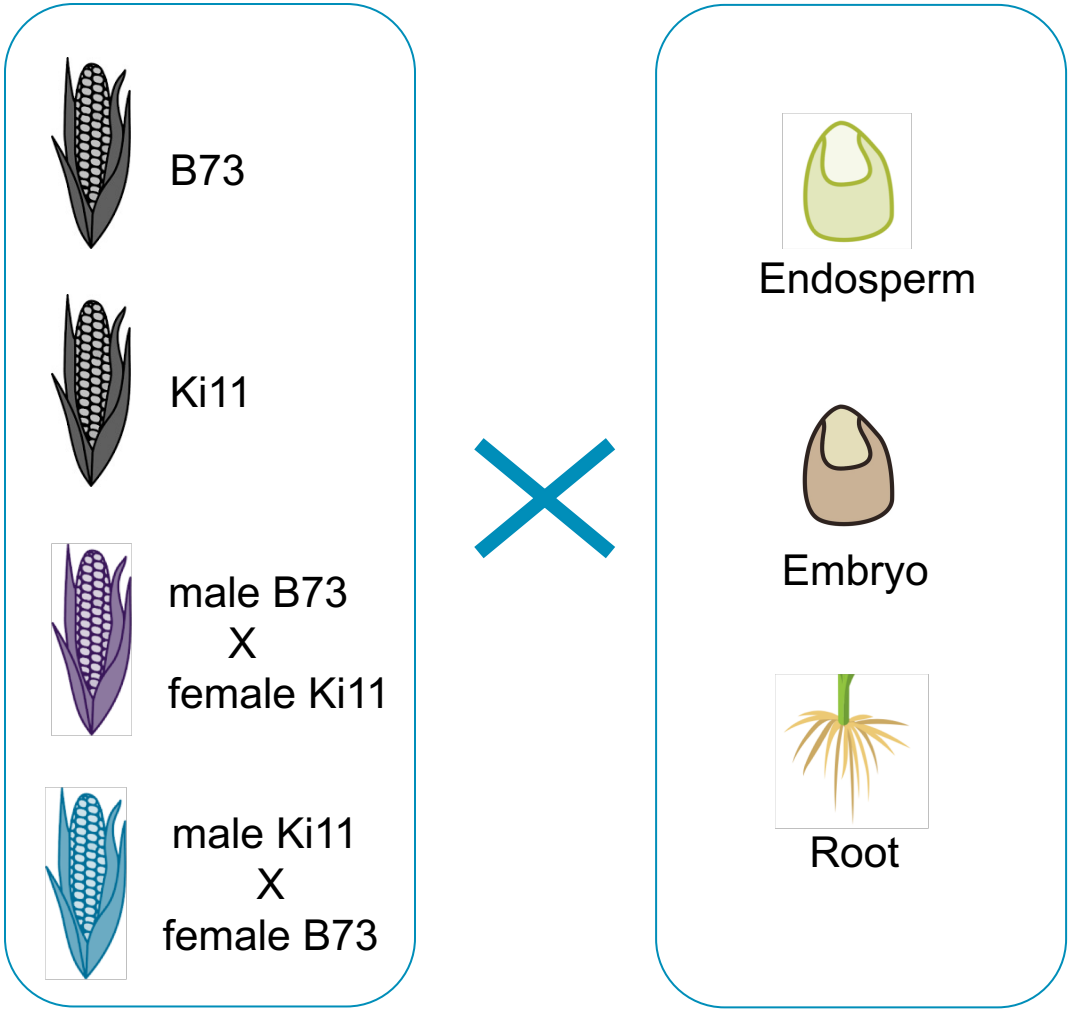


ISOFORM-LEVEL PHASING IN ISO-SEQ



SAMPLE SETUP

Wang et al. (in preparation)



12 barcoded cDNA samples

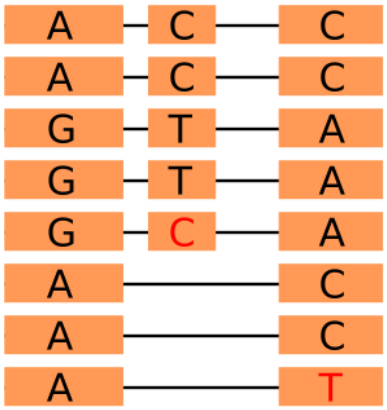
Pooled into 4 libraries

Sequenced for each lib

Photo Credit: www.vecteezy.com

ISO-PHASE: ISOFORM PHASING USING ISO-SEQ DATA

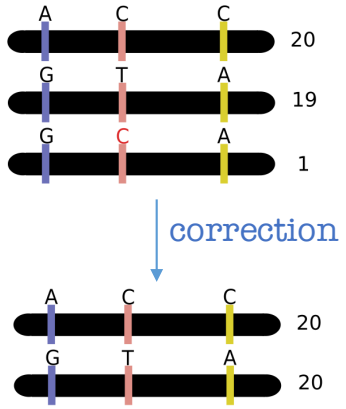
ALIGNMENT



SNP CALLING

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

PHASING



Can optionally include RNA-seq as input for SNP calling

VCF OUTPUT

```
##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
```

ISO-PHASE SNP CALLS CONCORDANT WITH RNA-SEQ

Iso-Phase applied only to 6907 genes with ≥ 40 long read coverage

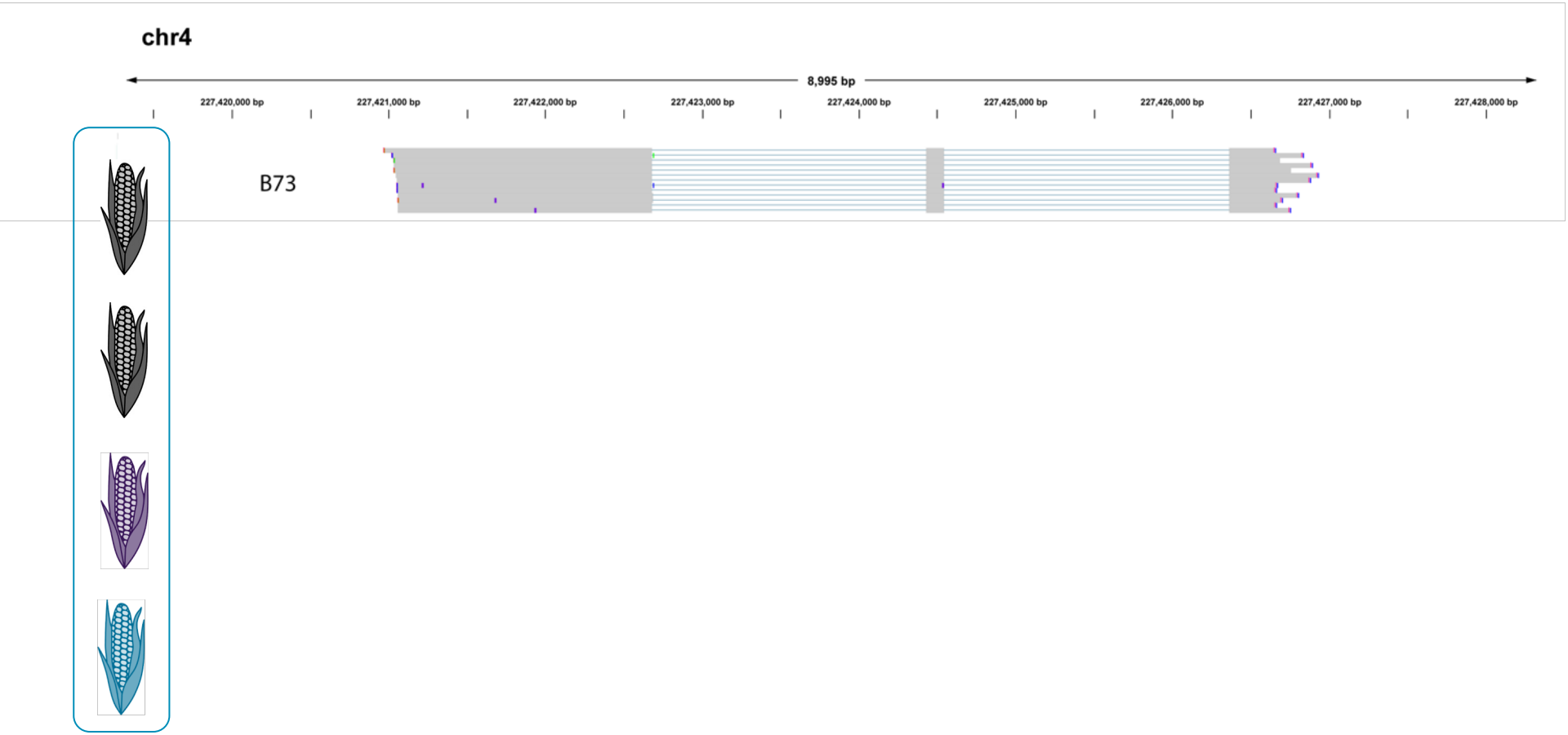
TYPE	NUMBER OF SNPS
Iso-Phase & RNA-seq	74,280
RNA-seq only	26,774
Iso-Seq only	3260

Possible Reasons for difference:

1. Low coverage (of long reads)
2. Short read alignment artifact
3. Ki11 indels (currently Iso-Phase calls only substitution SNPs)

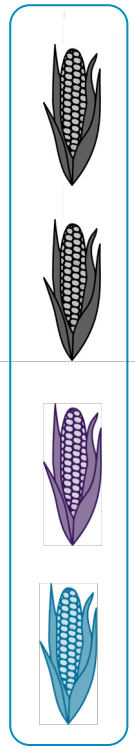
PHASING EXAMPLE: ISO-PHASE DETECTS SNPS

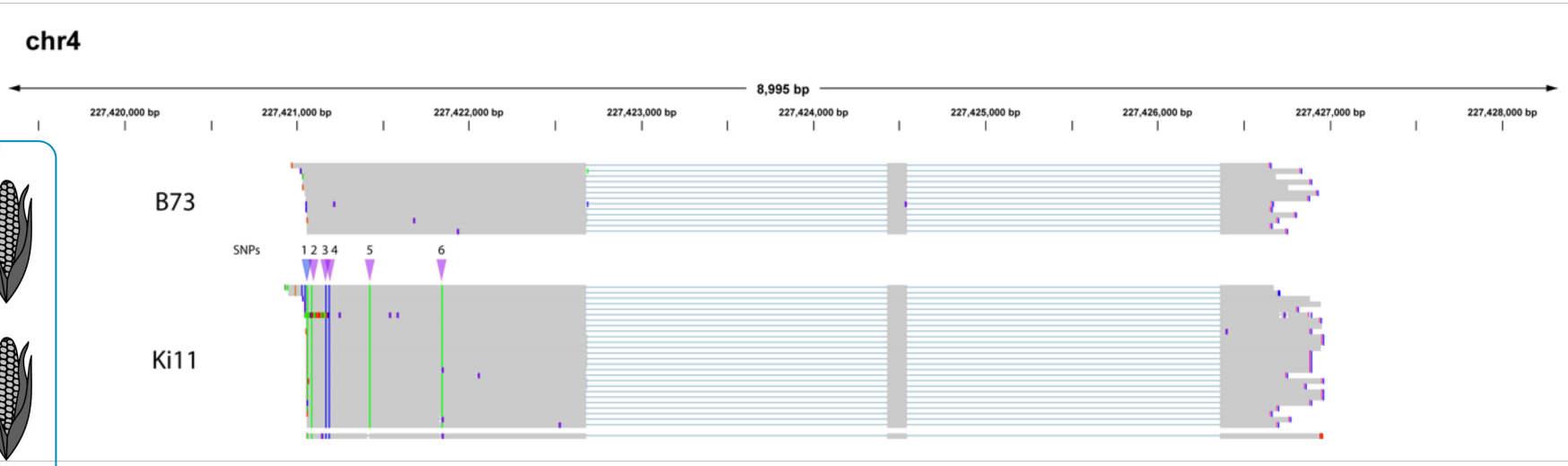
- B73 Iso-Seq data aligns to B73 reference genome → no SNPs expected



PHASING EXAMPLE: ISO-PHASE DETECTS SNPS

- Ki11 Iso-Seq data aligns to B73 identifies SNPs #2 through #6
- SNP #1 found by short read only



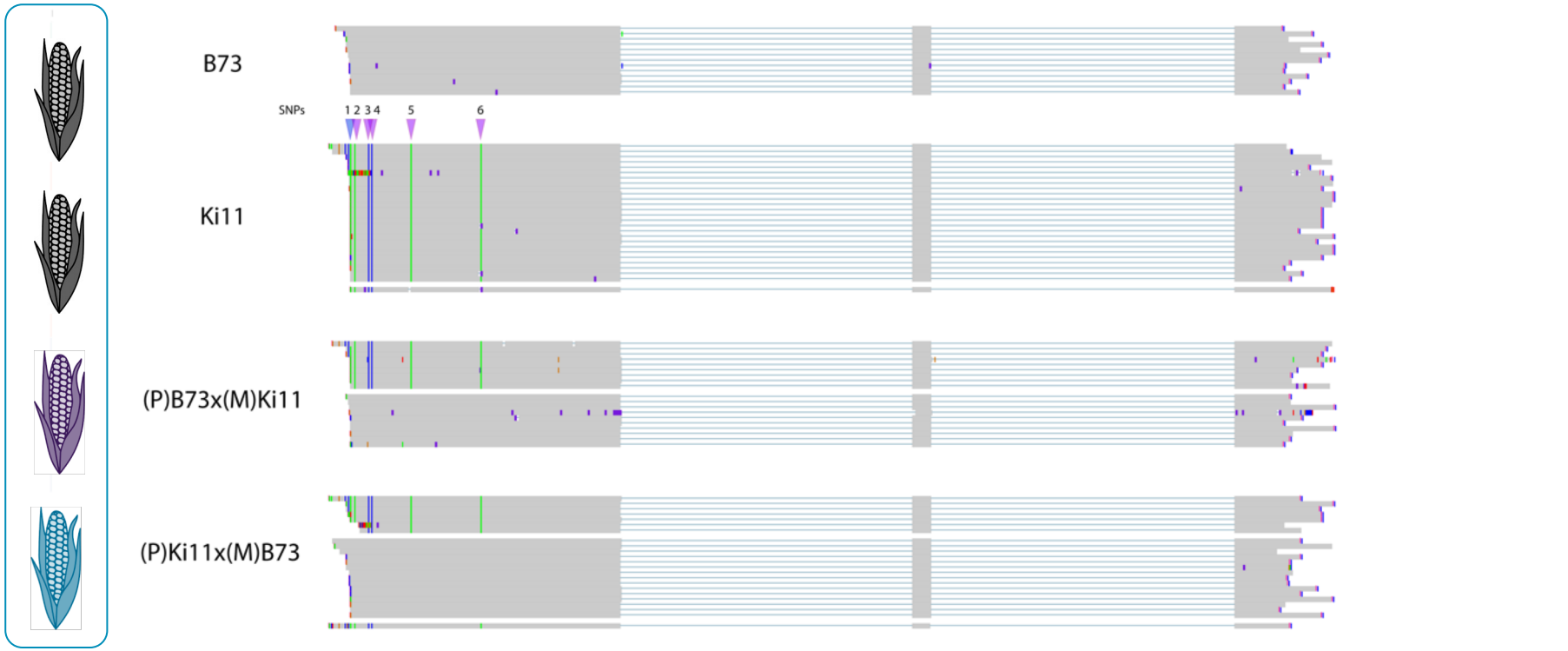


PHASING EXAMPLE: ISO-PHASE DETECTS SNPS

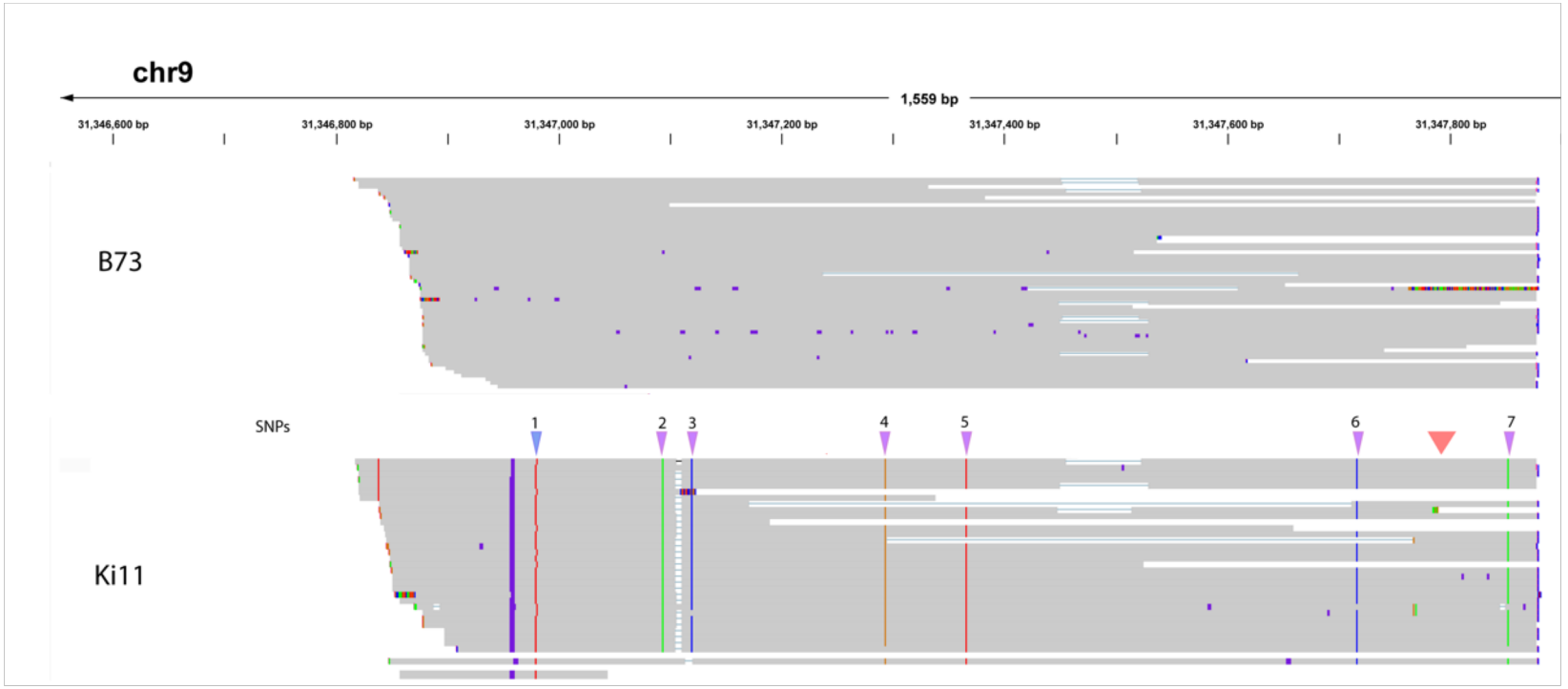
- F1 Iso-Seq data shows expression of both alleles

PHASING EXAMPLE: ISO-PHASE DETECTS SNPS

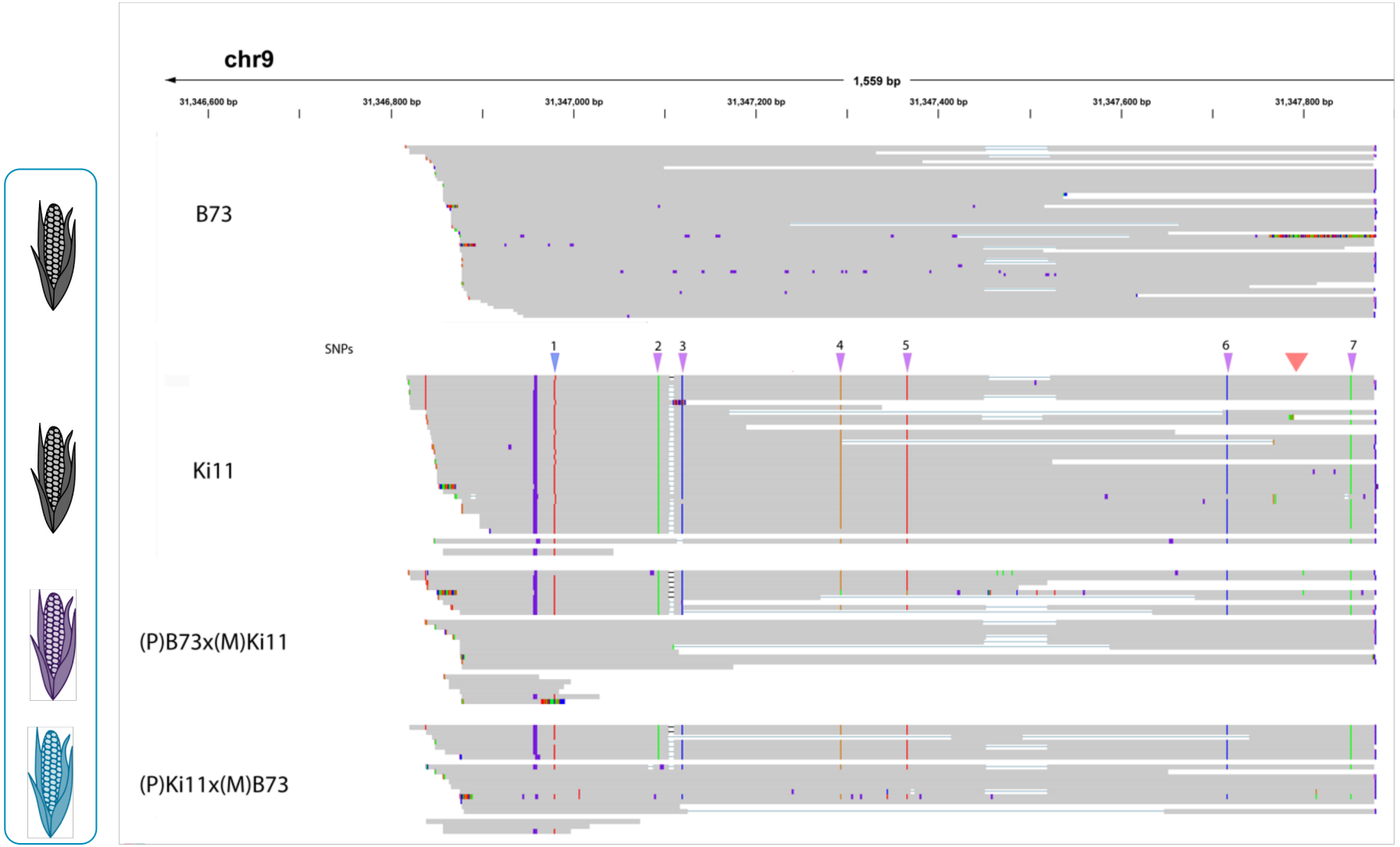
- F1 Iso-Seq data shows expression of both alleles



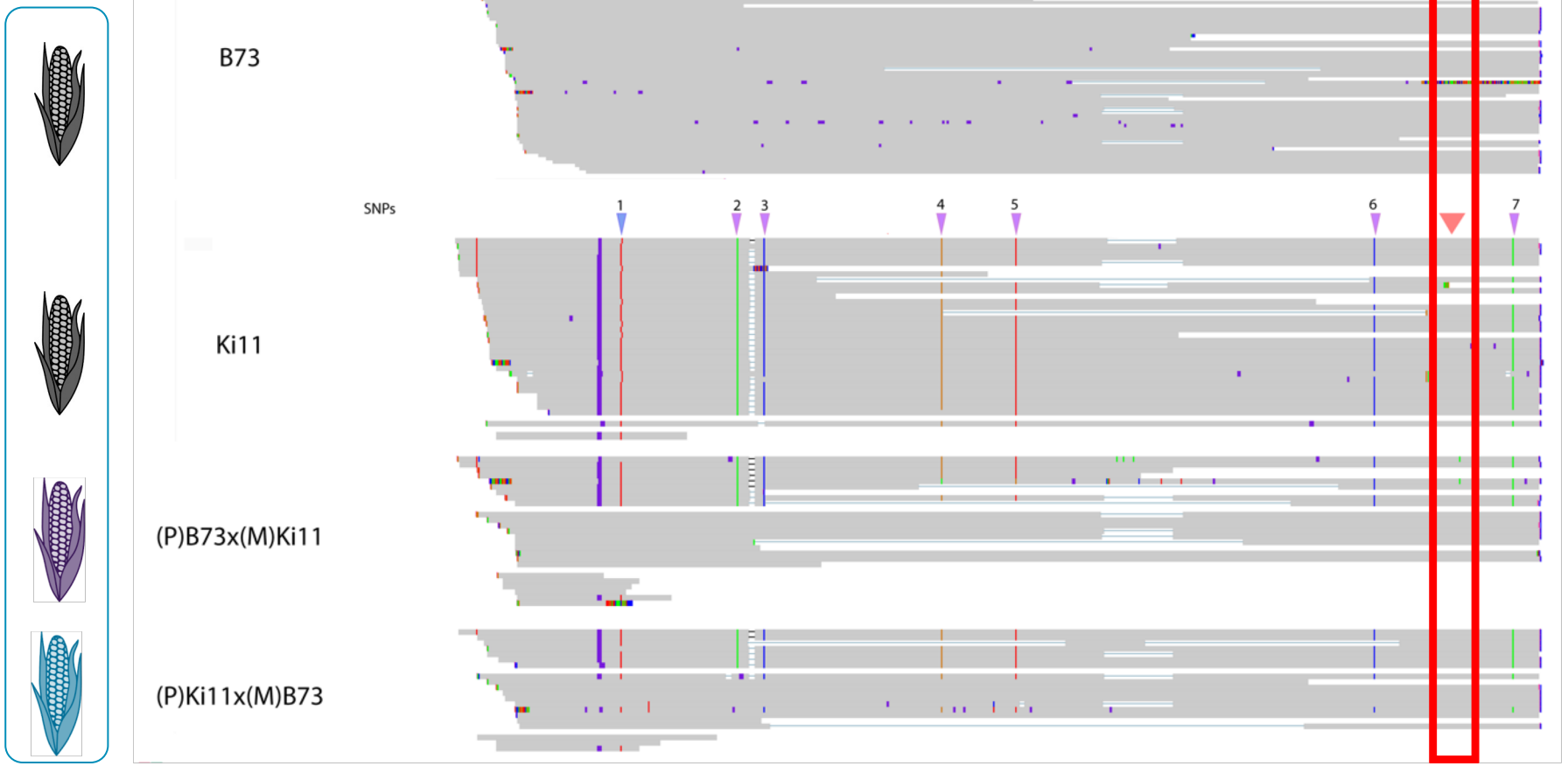
POSSIBLE SHORT READ SNP MIS-CALLS: PB.21897



POSSIBLE SHORT READ SNP MIS-CALLS: PB.21897

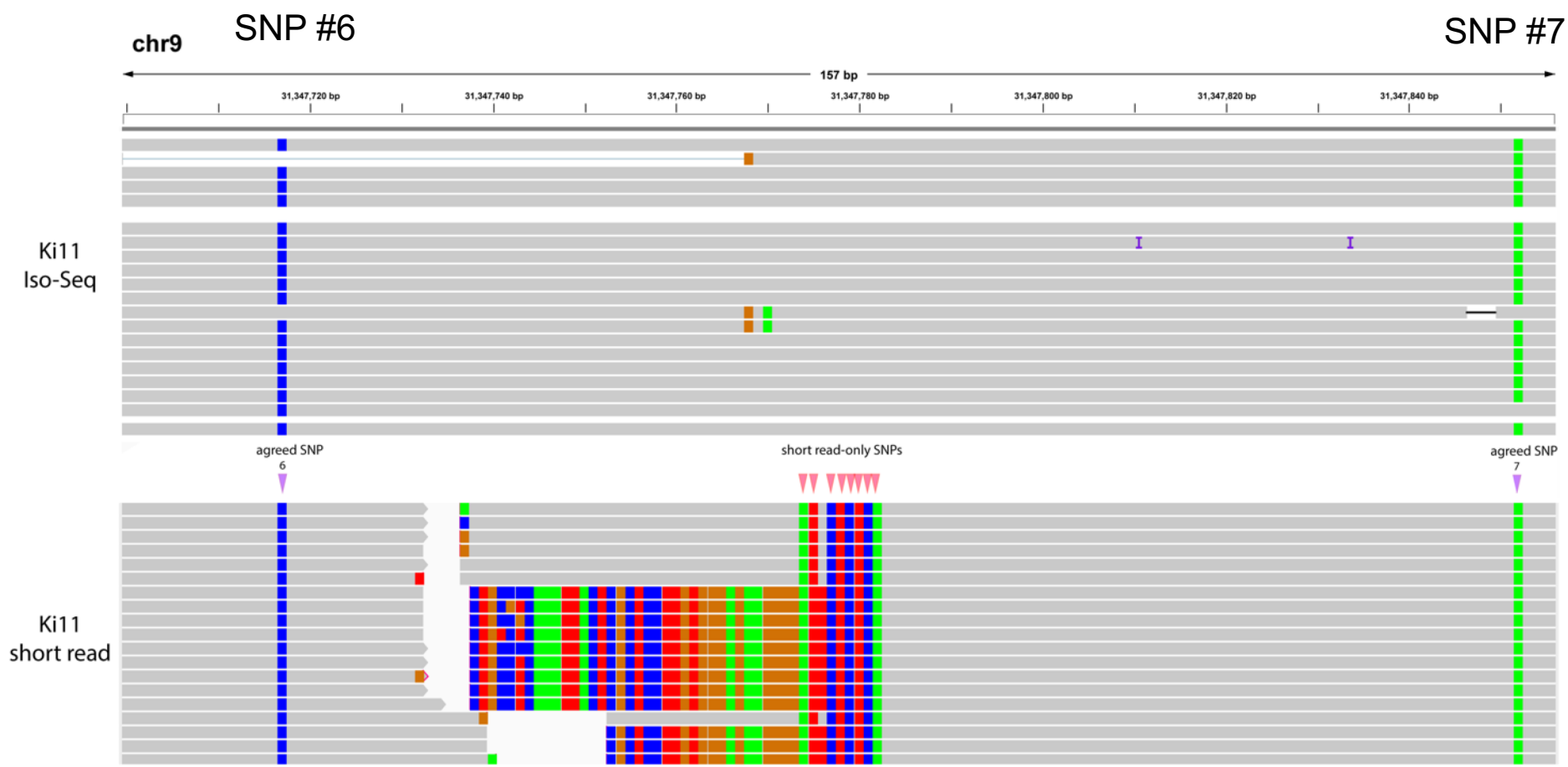


POSSIBLE SHORT READ SNP MIS-CALLS: PB.21897



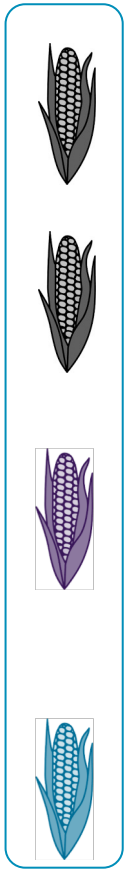
SNPs called by short reads only

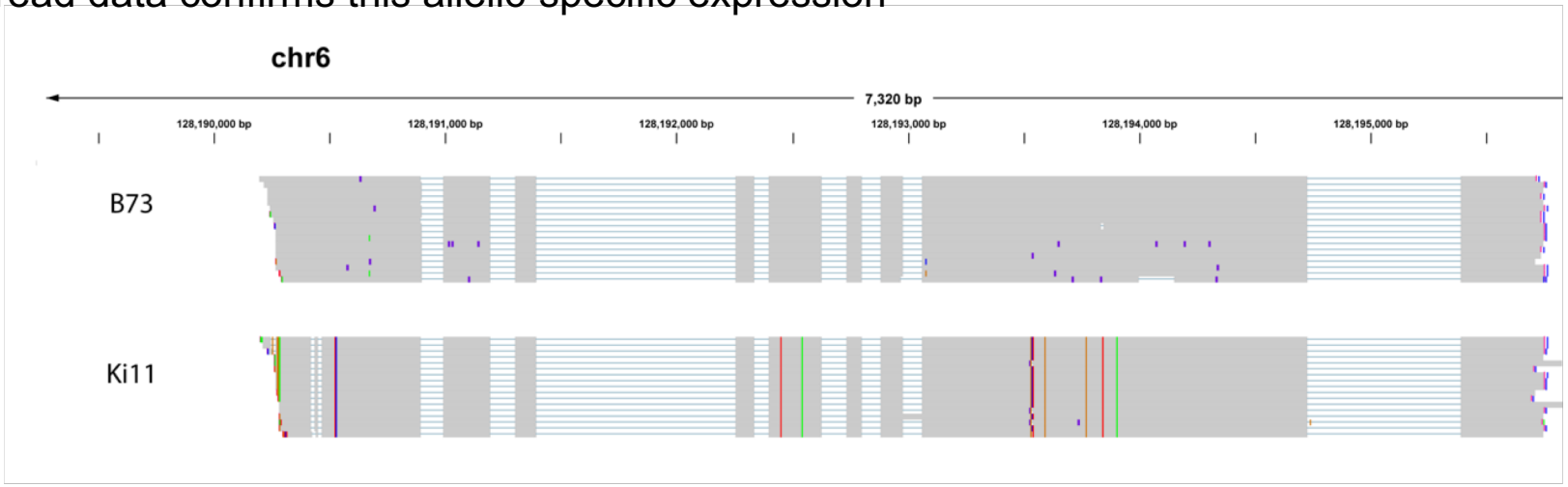
POSSIBLE SHORT READ SNP MIS-CALLS: PB.21897



ALLELE-SPECIFIC EXPRESSION

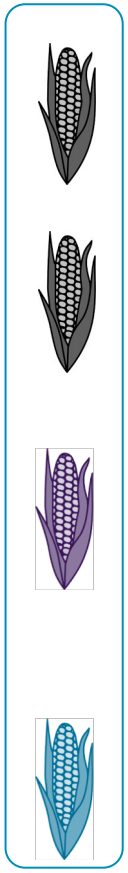
- B73xKi11 only expresses the Ki11 (female) allele
- Ki11xB73 only expresses the B73 (female) allele
- Short read data confirms this allelic-specific expression

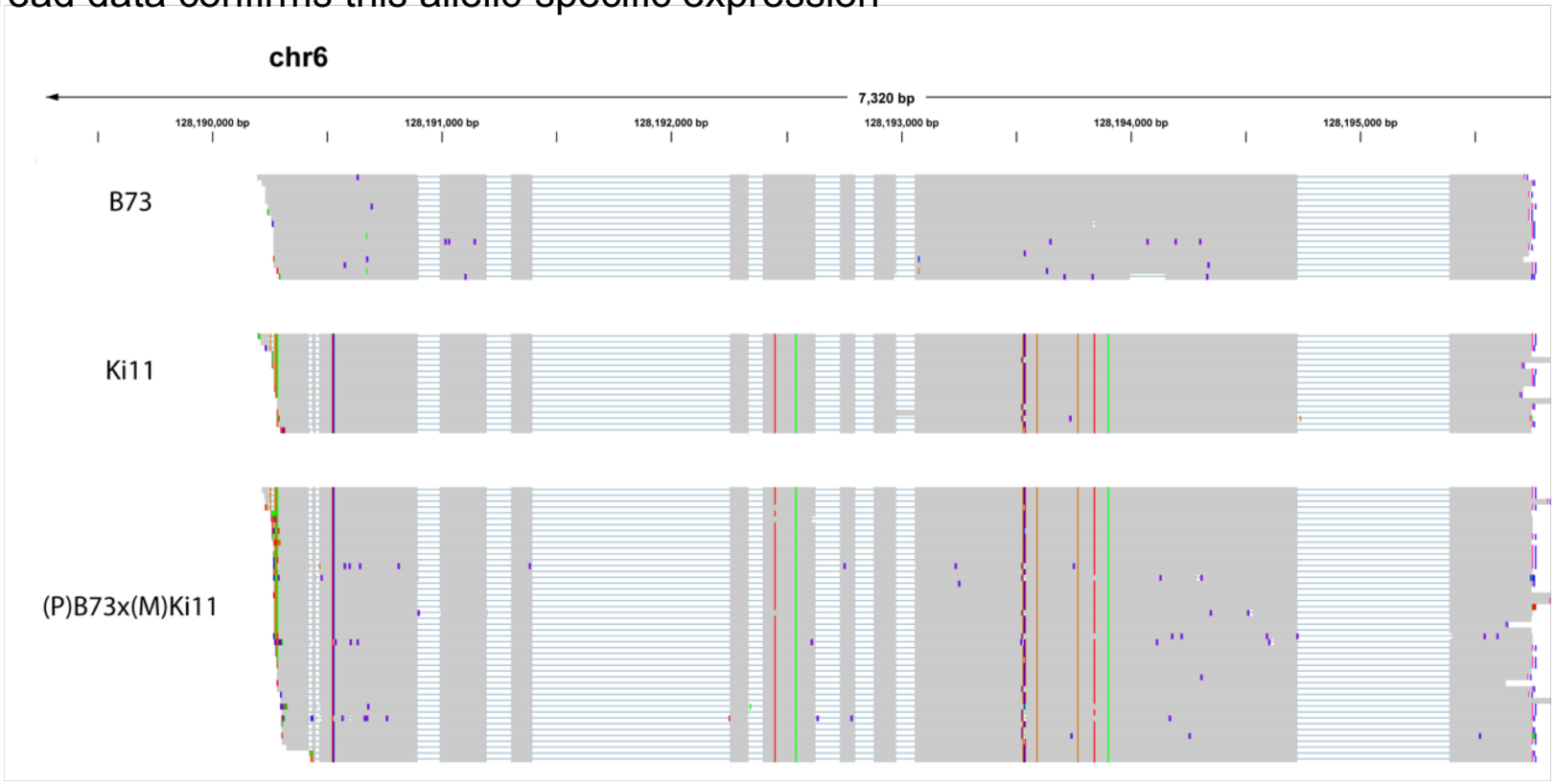




ALLELE-SPECIFIC EXPRESSION

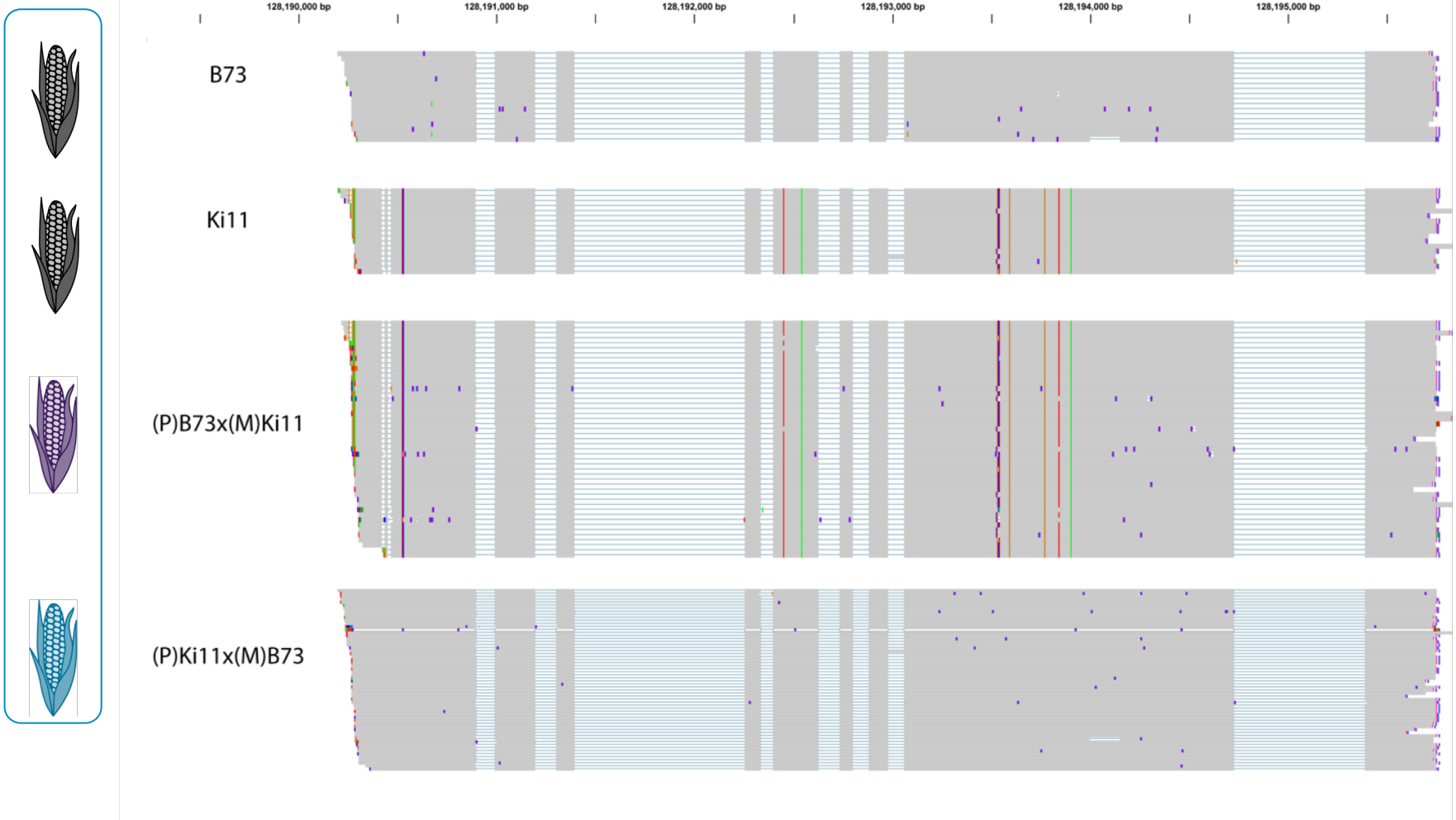
- B73xKi11 only expresses the Ki11 (female) allele
- Ki11xB73 only expresses the B73 (female) allele
- Short read data confirms this allelic-specific expression



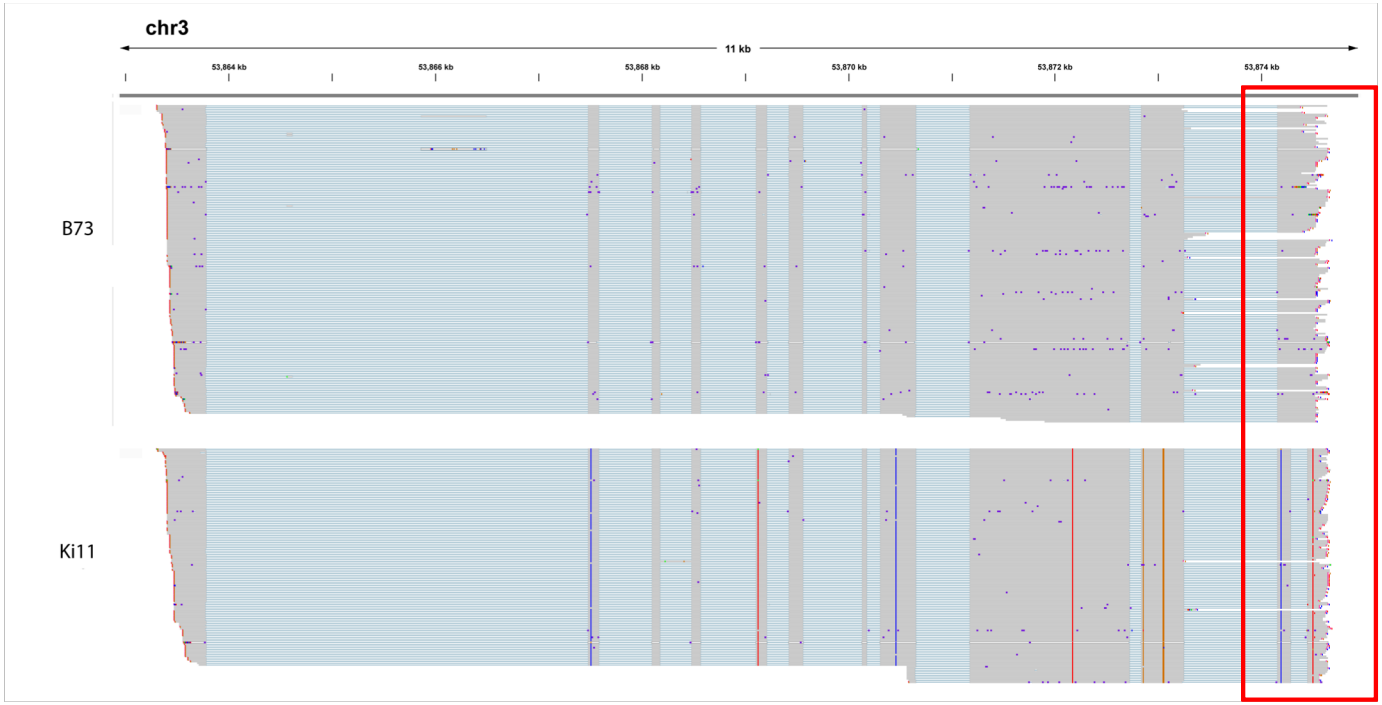


ALLELE-SPECIFIC EXPRESSION

- B73xKi11 only expresses the Ki11 (maternal) allele
- Ki11xB73 only expresses the B73 (maternal) allele
- Short read data confirms this allele-specific expression (not shown)

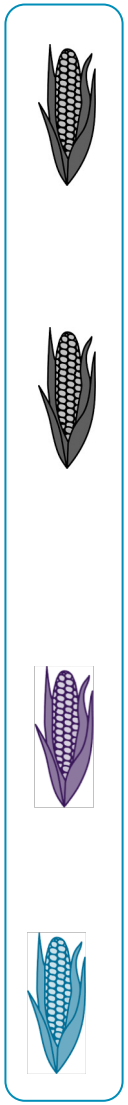


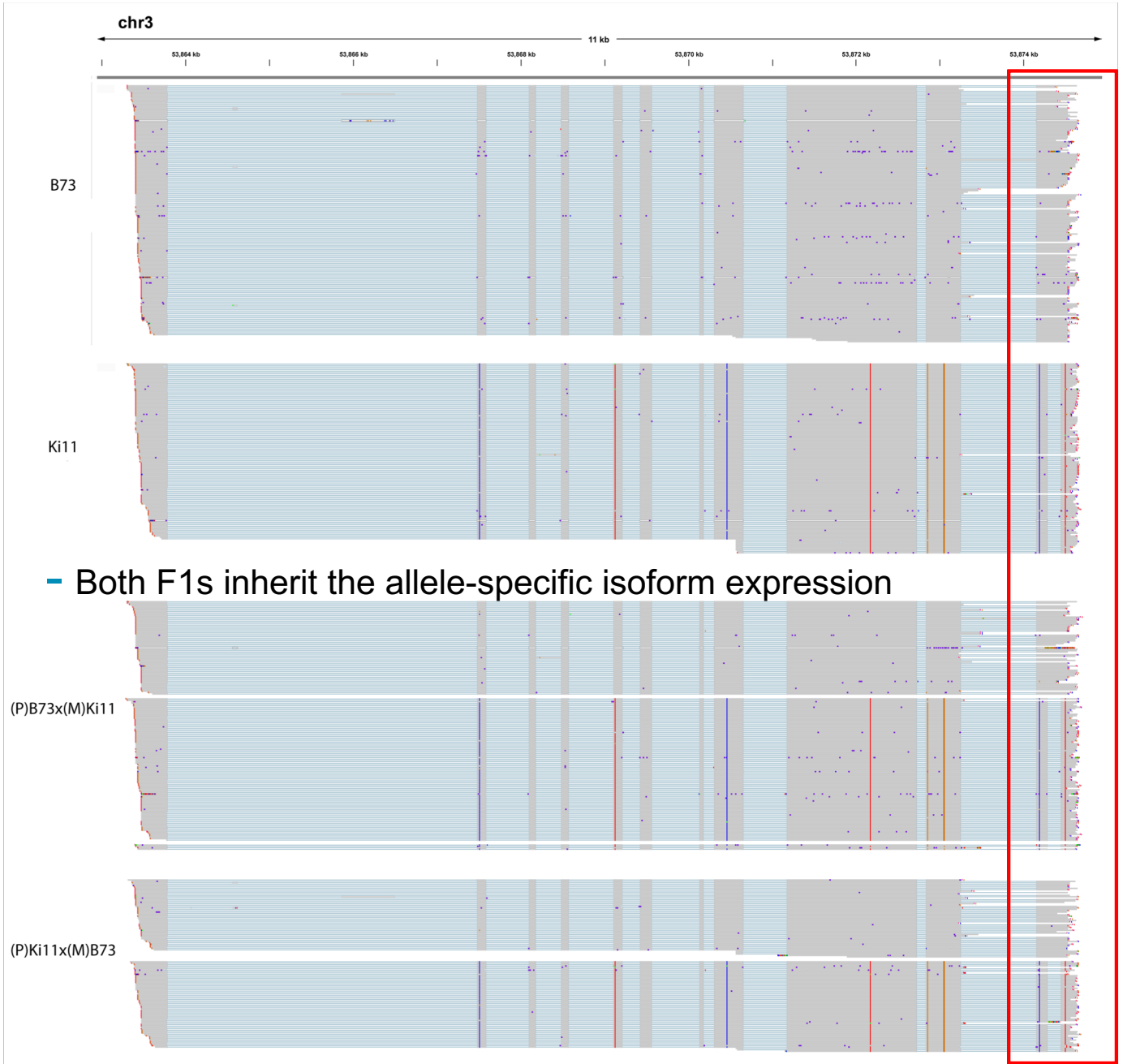
ALLELE-SPECIFIC ISOFORM EXPRESSION



- Two parents express different isoforms (3’ exon difference)

ALLELE-SPECIFIC ISOFORM EXPRESSION



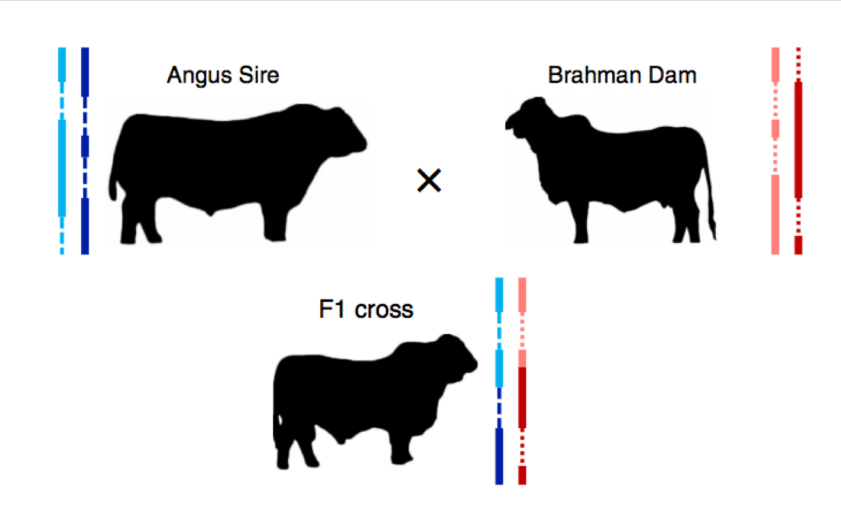


- Both F1s inherit the allele-specific isoform expression

Iso-Seq Phasing on a F1 cattle

PHASING F1 CATTLE

- Genome info available for parental & F1
- Can use genome SNPs for validation

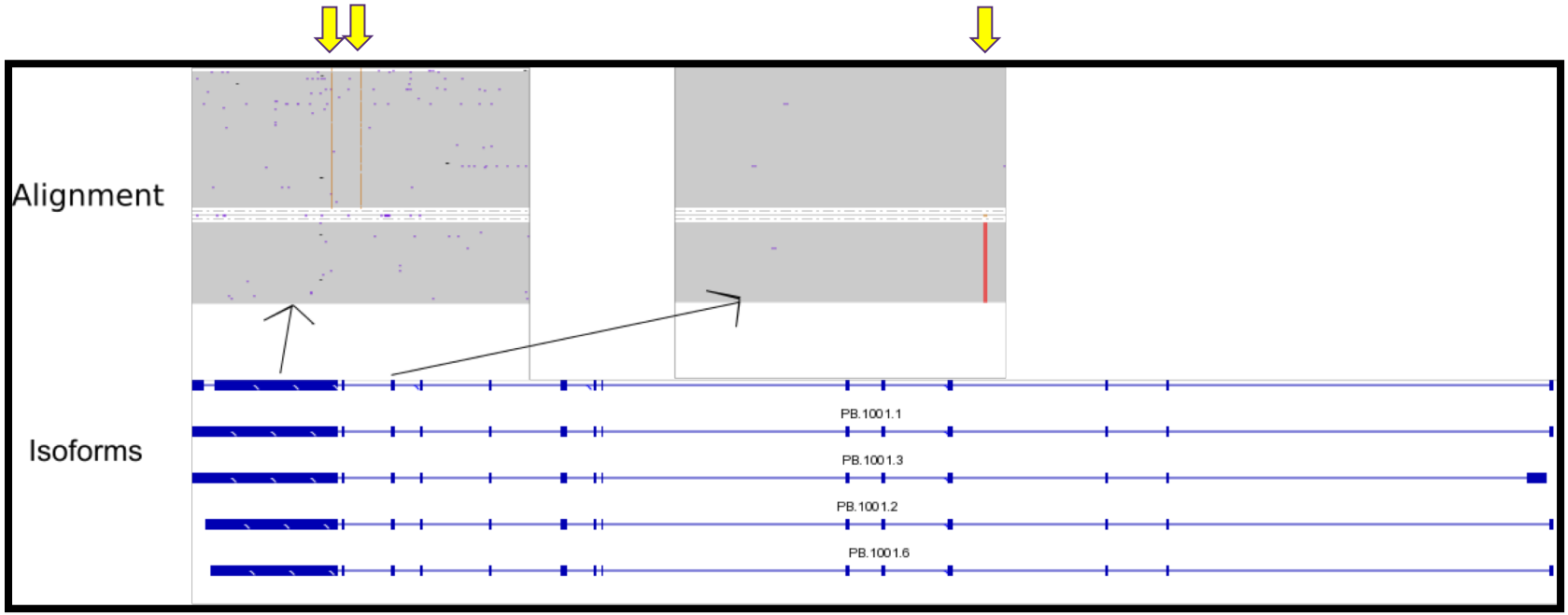


SNP Type	Count
True Positive (called by both genome & Iso-Seq)	8334
False Negative (called by genome only)	259
Unphased by Genome (called by Iso-Seq only)	1203

Collaboration with John Williams (U of Adelaide)

Koren, S. *et al.* De novo assembly of haplotype-resolved genomes with trio binning. *Nat Biotechnol* 1–14 (2018). doi:10.1038/nbt.4277

ISO-SEQ CALLED SNPS NOT PHASED IN GENOME

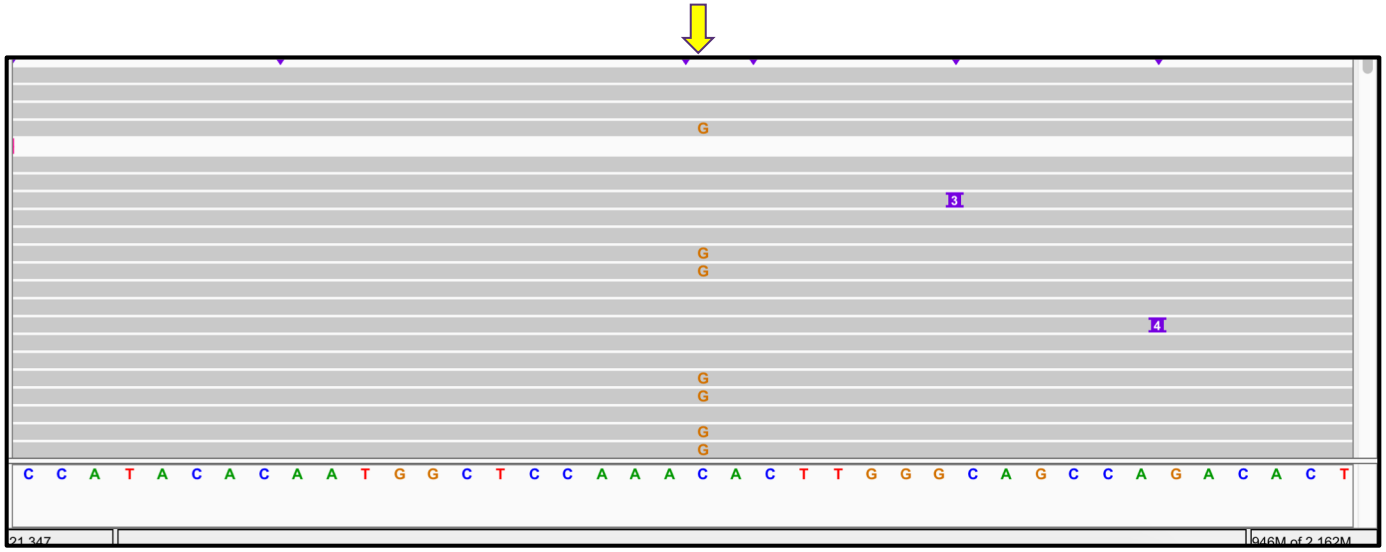


This gene contains 228 FL reads.

- Strong evidence for the 3 SNPs.
- Genome assembler (Falcon UNZIP) did not phase this region - WHY?

ISO-SEQ CALLED SNPS NOT PHASED IN GENOME

SNP #1 showed C/G SNP support in genome raw pileup.

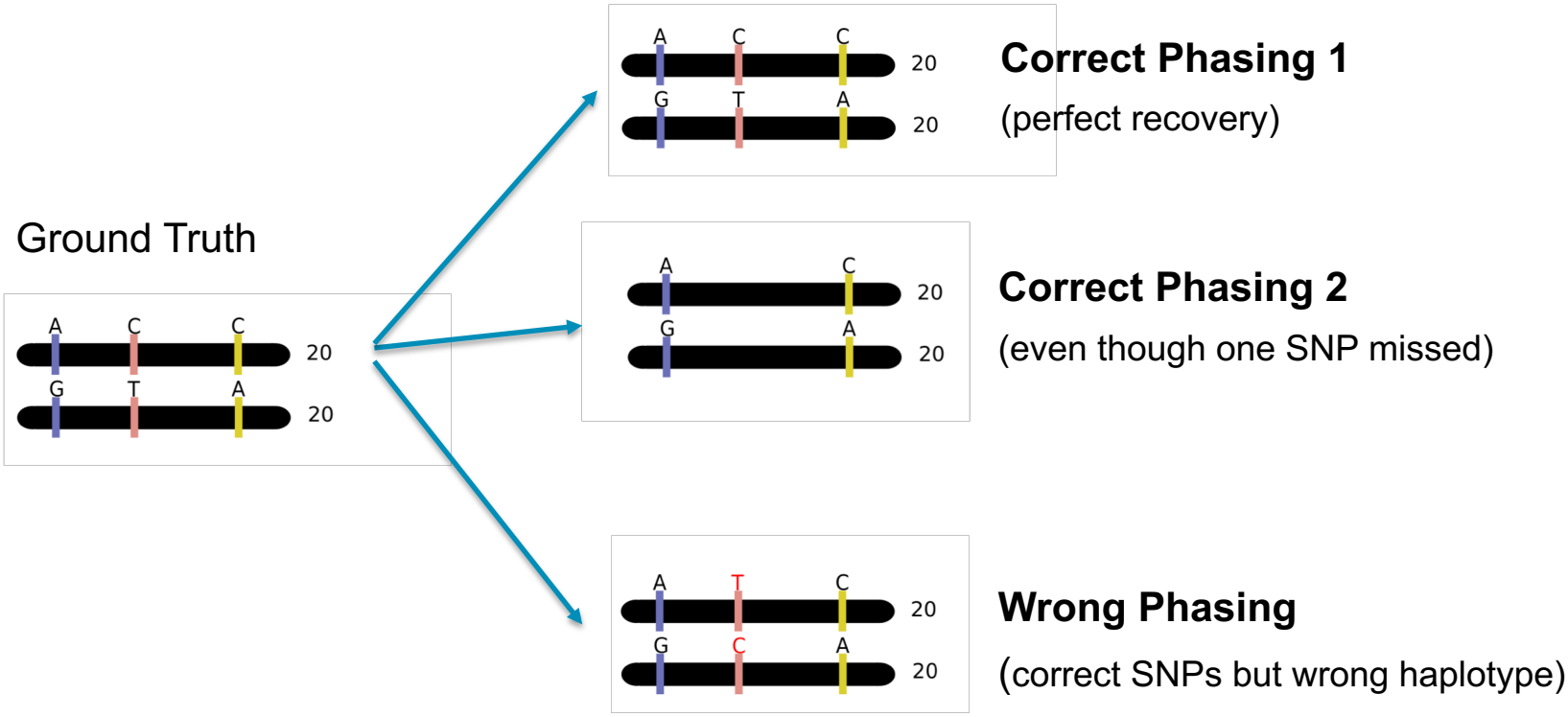


- SNP #2 and #3 were also supported by genome raw data
- This region was low heterozygosity

PACBIO®

Tetraploid Phasing

PHASING EVALUATION USING SIMULATED DATA



- Simulated 100 human genes. Each allele with 20-fold coverage with varying error rate.
- For 4N, “correct phasing” means getting all 4 alleles correct. Getting 3 → still wrong.

PHASING EVALUATION ON SIMULATED 100 HUMAN GENES

Percentage of 100 genes that were correctly phased.

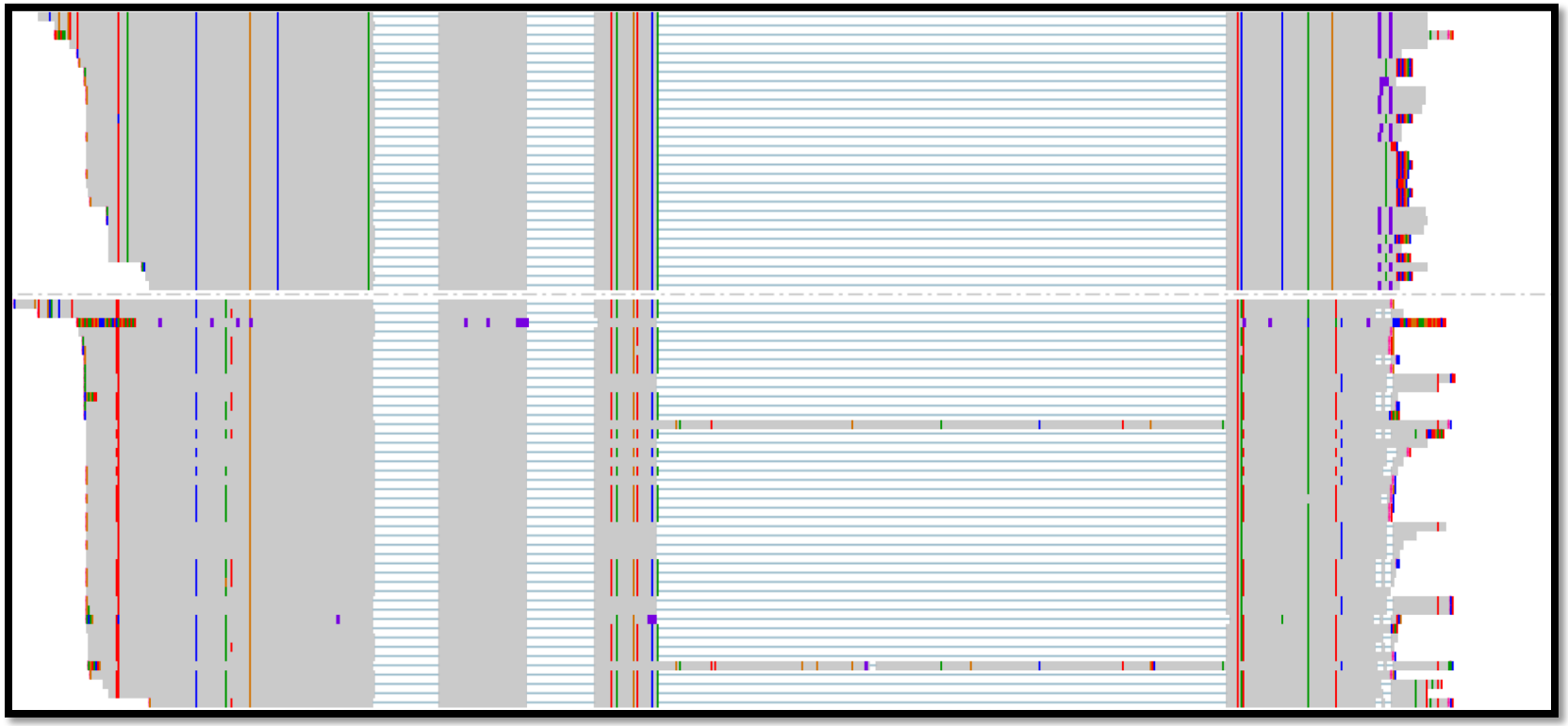
Error Rate	2N no QV	2N with QV	4N no QV	4N with QV
1%	100%	100%	95%	98%
2%	100%	100%	82%	98%
5%	100%	100%	55%	96%

- Simulated with 20-fold coverage per allele
- 2N = 40-fold coverage
- 4N = 80-fold coverage

TETRAPLOID POTATO ISO-SEQ PHASING

Collaboration w/ Marko Petek & Kristina Gruden (NIB, Slovenia)

- Each variant position has two SNPs
- Iso-Phase estimated three alleles, not four



SUMMARY

- Iso-Phase takes advantage of full-length transcript reads from the PacBio for haplotyping
- Diploid phasing on F1 maize
  - Identifies allelic specific gene and isoform expressions
- Diploid phasing on F1 cattle
  - Identifies SNPs not called by genome due to low heterozygosity
- Tetraploid phasing challenging due to (a) higher depth requirement (80-fold); (b) non-uniform allelic expression

ADDITIONAL POSTER & TALK INFORMATION

PO0087 Isophase: Haplotyping Using Full-Length Transcript Sequencing in a F1 Maize Hybrid Reveals Allele-Specific Expression

PO0185 Library Prep and Bioinformatics Improvements for Full-Length Transcript Sequencing on the PacBio Sequel System

W463 Single-Molecule Sequencing Reveals Increased Complexity of the Transcriptome Landscape in Maize and Sorghum

Sunday, January 13, 2019, Pacific Salon 1, by Bo Wang (CSHL)
3:01 PM - 3:14 PM

Sequence with Confidence – How SMRT Sequencing is Accelerating Plant and Animal Genomics

Monday, January 14, 2019 Golden West Ballroom
12:50 - 3:00 PM

PacBio SMRT Developers Conference

Wednesday, January 16, 2019 Hampton and Sheffield Ballrooms
12:00 - 4:30 PM

ACKNOWLEDGEMENTS

Bo Wang

Doreen Ware

Kevin Eng

Primo Baybayan