



PACBIO®



# IsoPhase: Phasing Iso-Seq Data for diploid (and possibly tetraploid) genomes

Elizabeth Tseng, Jan 2018, PAG SMRT Developers Conference

## Online Resources:

 [groups.google.com/forum/#!forum/SMRT\\_isoseq](https://groups.google.com/forum/#!forum/SMRT_isoseq)



[github.com/PacificBiosciences/IsoSeq\\_SA3nUP/](https://github.com/PacificBiosciences/IsoSeq_SA3nUP/)  
(shortened: <http://tinyurl.com/PBisoseq>)



@magdoll

# PLANT AND ANIMAL ISO-SEQ PUBLICATIONS IN 2017

Type	Species	Title
Crop	Coffee	Cheng <i>et al.</i> Long-read sequencing of the coffee bean transcriptome reveals the diversity of full-length transcripts. <i>GigaScience</i> 1–13 (2017).
Medicinal	D. Officinale	He, L. <i>et al.</i> Hybrid Sequencing of Full-Length cDNA Transcripts of Stems and Leaves in Dendrobium officinale. <i>Genes</i> <b>8</b> , 257–13 (2017).
Medicinal	Ginseng	Jo, I.-H. <i>et al.</i> Isoform Sequencing Provides a More Comprehensive View of the Panax ginseng Transcriptome. <i>Genes</i> <b>8</b> , 228–17 (2017).
Medicinal	Huangqi	Li, J. <i>et al.</i> Long read reference genome-free reconstruction of a full- length transcriptome from Astragalus membranaceus reveals transcript variants involved in bioactive compound biosynthesis. <i>Nature Publishing Group</i> 1–13 (2017).
Crop	Cotton	Wang, M. <i>et al.</i> A global survey of alternative splicing in allopolyploid cotton: landscape, complexity and regulation. <i>New Phytol</i> <b>217</b> , 163–178 (2017).
Animal	Rabbit	Chen, S.-Y., Deng, F., Jia, X., Li, C. & Lai, S.-J. A transcriptome atlas of rabbit revealed by PacBio single-molecule long-read sequencing. <i>Sci. Rep.</i> <b>7</b> , 1–10 (2017).
Crop	Quinoa	Jarvis, D. E. <i>et al.</i> The genome of Chenopodium quinoa. <i>Nature</i> <b>542</b> , 307–312 (2017).
Crop	Strawberry	Li, Y., Dai, C., Hu, C., Liu, Z. & Kang, C. Global identification of alternative splicing via comparative analysis of SMRT- and Illumina-based RNA-seq in strawberry. <i>Plant J</i> <b>90</b> , 164–176 (2017).
Crop	Wheat	Clavijo, B. J. <i>et al.</i> An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. <i>Genome Res.</i> <b>27</b> , 885–896 (2017).

# PUBLICLY AVAILABLE SEQUEL ISO-SEQ DATA

## Sequel System Data Release: Iso-Seq Results for Hummingbird and Zebra Finch Brain Tissue

Thursday, August 31, 2017

If you're interested in avian vocal learning or want to explore a PacBio Iso-Seq data set generated with the Sequel System, we have good news. We've just [released data](#) from Iso-Seq interrogations of brain tissue from two avian models of vocal learning, Anna's hummingbird (*Calypte anna*) and zebra finch (*Taeniopygia guttata*), sequenced in collaboration with the Erich Jarvis and Olivier Fedrigo labs at the Rockefeller University.

If you're not familiar with the [Iso-Seq method](#), it's the long-read sequencing answer to short-read RNA-seq studies. By using SMRT Sequencing for a transcriptome project, scientists can generate full-length isoform data, clearly capturing alternative splicing events to see the real diversity of transcripts. Unlike RNA-seq approaches, the Iso-Seq method takes advantage of long-read data to fully span transcript isoforms from the 5' end to their poly-A tails, eliminating the need for error-prone transcript reconstruction and inference processes. With the Sequel System, Iso-Seq projects are low cost and time efficient. Currently we recommend only 1-2 SMRT Cells per tissue type for genome annotation.



Anna's hummingbird photo by Pat Durkin

- 4 Sequel cells
- Barcoded bird brains
- Total: 785k FL reads

	ZEBRAFINCH	HUMMINGBIRD
Runtime	31 hr	28 hr
Unique Genes	7228	7357
Unique Isoforms	17,437	16,898

## ISO-SEQ2: COMING TO YOU IN 2018!

- Faster runtime
- Increased transcript recovery
- Reduced false artifacts
- Available in the next SMRTLink release

SMRT CELLS	CCS READS	FL READS	ISOSEQ2 RUNTIME	UNIQUE TRANSCRIPTS	UNIQUE GENES
1	244,804	212,201	15 hr	13,036	7760
3	796,128	676,905	21 hr	30,956	13,044
6	1,908,507	1,562,039	5 days	63,645	18,227

\* run using all default options except CCS minimum pass changed to 1



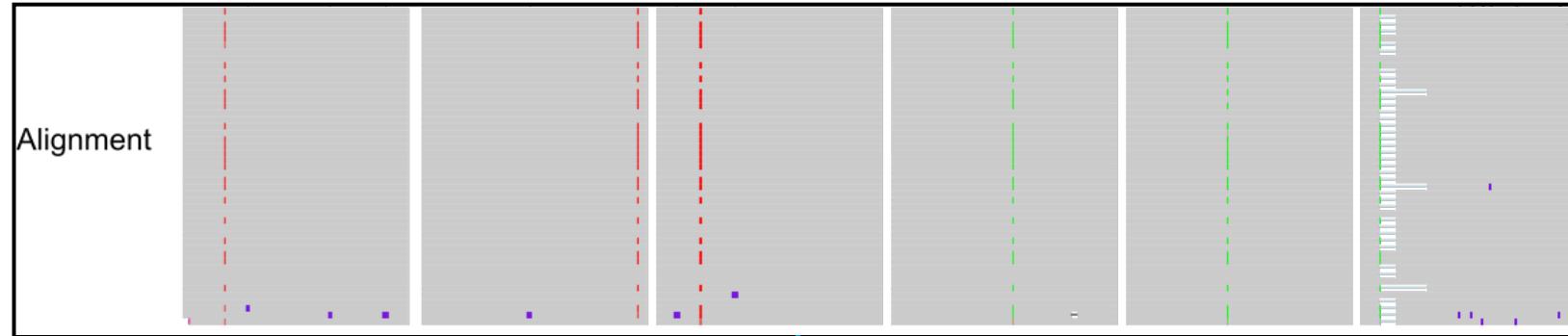
PACBIO®

# Phasing Iso-Seq Data

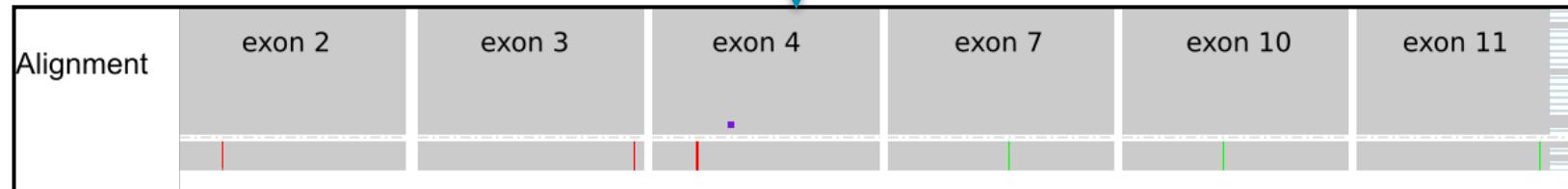
## MOTIVATION

- The Iso-Seq bioinformatics pipeline outputs distinct isoforms (exon skipping, alternative 5' and 3' ends), but collapses SNP-level variations.
- SNP information can be revealed by aligning full-length (FL) CCS reads back to the unique isoforms after Iso-Seq analysis.

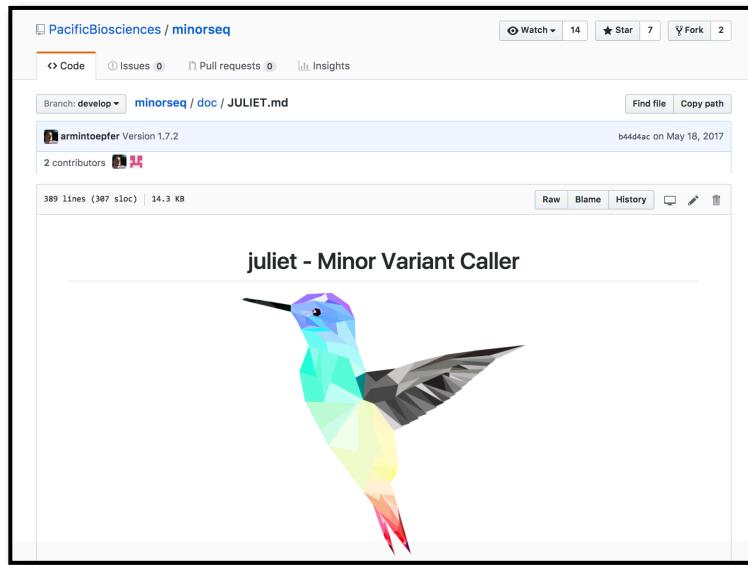
# MOTIVATION



"quickphase" in IGV  
separates aligned reads into two groups



# MOTIVATION



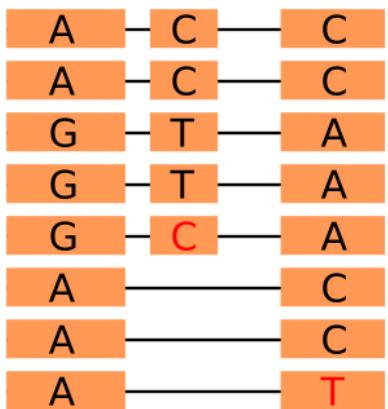
Variant Discovery							
my seq			Sample Variants				
Codon	AA	Pos	AA	Codon	%	Coverage	Affected Drugs*
A G C	S	3	S	A G T	99	2905	
G A G	E	40	E	G A A	100	2828	
A T G	M	41	L	T T G	1	2793	fancy drug
G G G	G	45	G	G G A	100	2596	

\*DrugDB version x.y.z (last updated YYYY-MM-DD)

- Juliet calls minor variants in viral data
- However,
  - It does not handle splicing
  - Performs best with stringent CCS cutoff (> 99%)
  - Performs best with deep coverage (> 250-fold)

# ISOPHASE: ISOFORM PHASING USING ISO-SEQ DATA

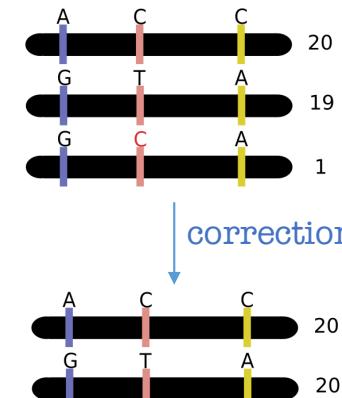
## ALIGNMENT



## SNP CALLING

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

## PHASING



## VCF OUTPUT

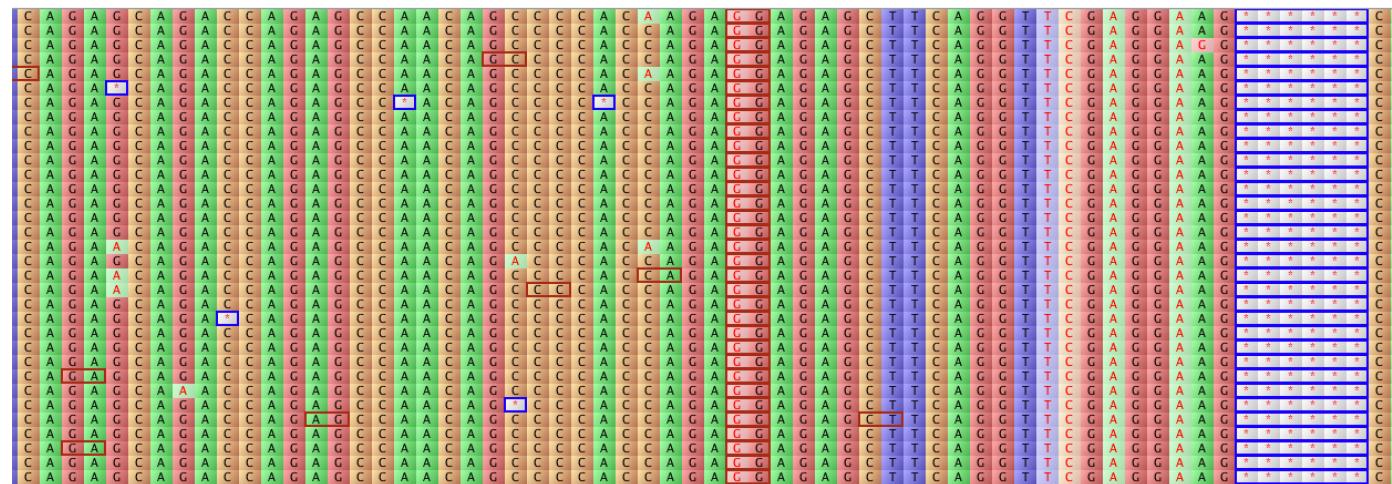
```
##fileformat=VCFv4.2
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT ISOFORM1 ISOFORM2
chr1 105 . A G . PASS DP=40;AF=0.50 GT:HQ 0|1:20,20 0:15
chr1 190 . C T . PASS DP=40;AF=0.50 GT:HQ 0|1:20,20 0:15
chr1 336 . C A . PASS DP=40;AF=0.50 GT:HQ 0|1:20,20 0:15
```

## ISOPHASE METHOD SUMMARY

- Alignment using minimap2, retrieve positions with sufficient coverage
  - If QV is provided, alignment pileup filters out low-quality bases
- SNP calling and phasing using Juliet
- Simple clustering of phased haplotypes to remove errors
- Output VCF denoting SNPs and allele counts for each isoform

# SNP CALLING AND PHASING IN JULIET

- Steps across alignment in codon triples. *not applicable for transcriptome*
- Computes a p-value that the observed bases come from purely noise using a Fisher's exact test.
- If the p-value is significant under a Bonferroni correction, call the variant.
- Phase together variant positions by tallying full-length reads that exhibit different combinations of the variant positions and threshold.
- Limits:
  - CCS reads used to minimize impact of noise. Raw read analysis possible.
  - Currently does not estimate indel variants.



- Reliably identify 1% true variants from sequencing noise.

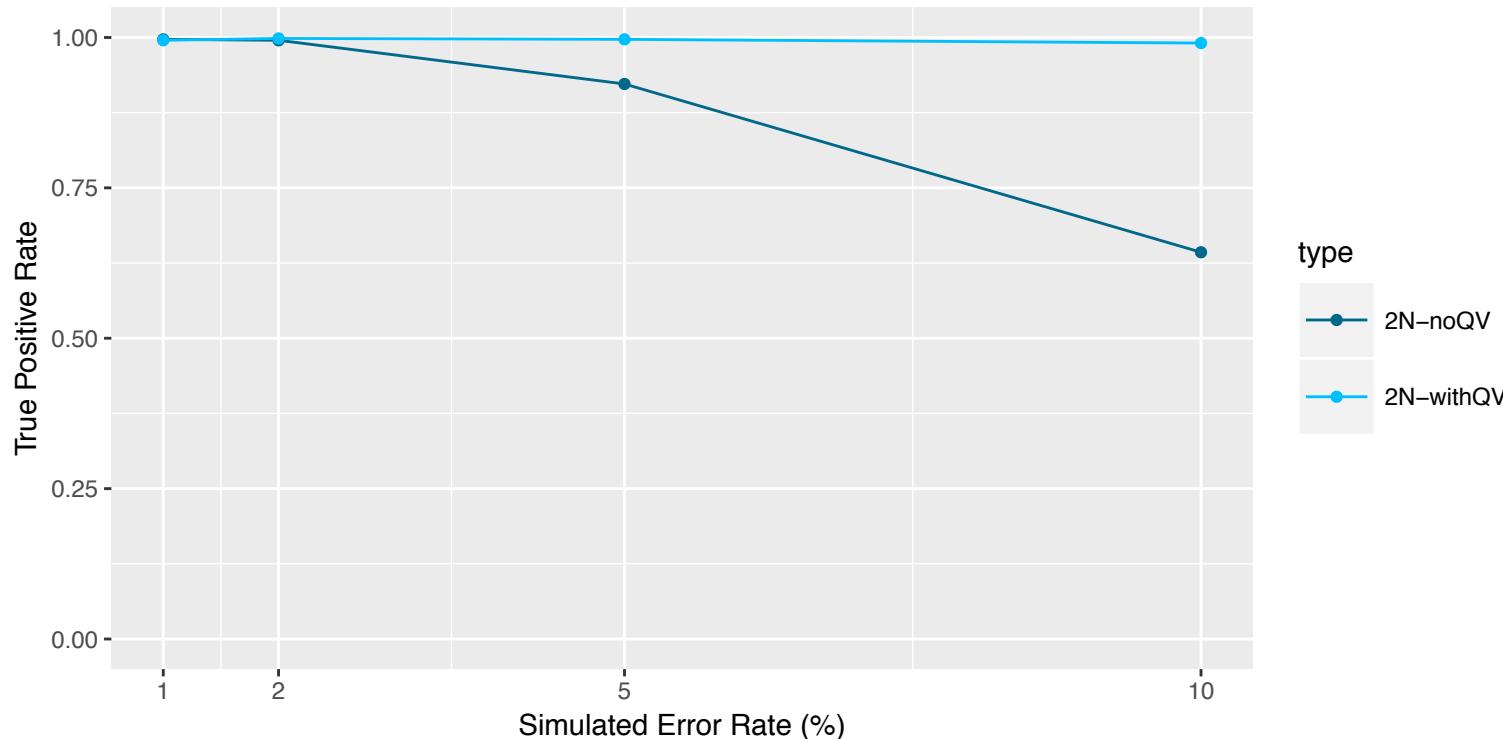


PACBIO®

# Evaluation on Simulated Data

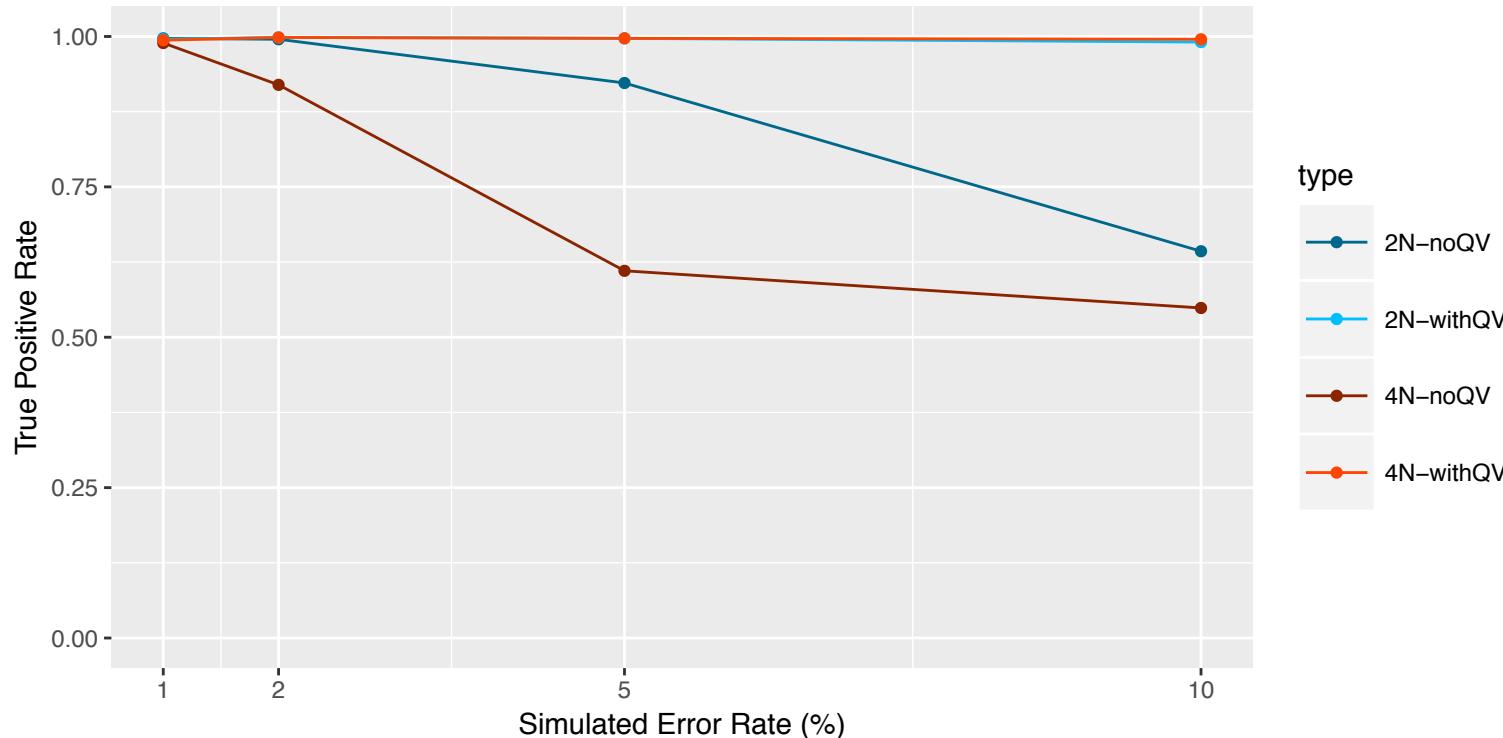
## SNP CALLING ON SIMULATED 100 HUMAN GENES

- Random 100 human genes
- Simulated 1 SNP per 300 bp
- Each allele has 20 copies (20-fold coverage)
- Simulated substitution errors at 1%, 2%, 5%, and 10%



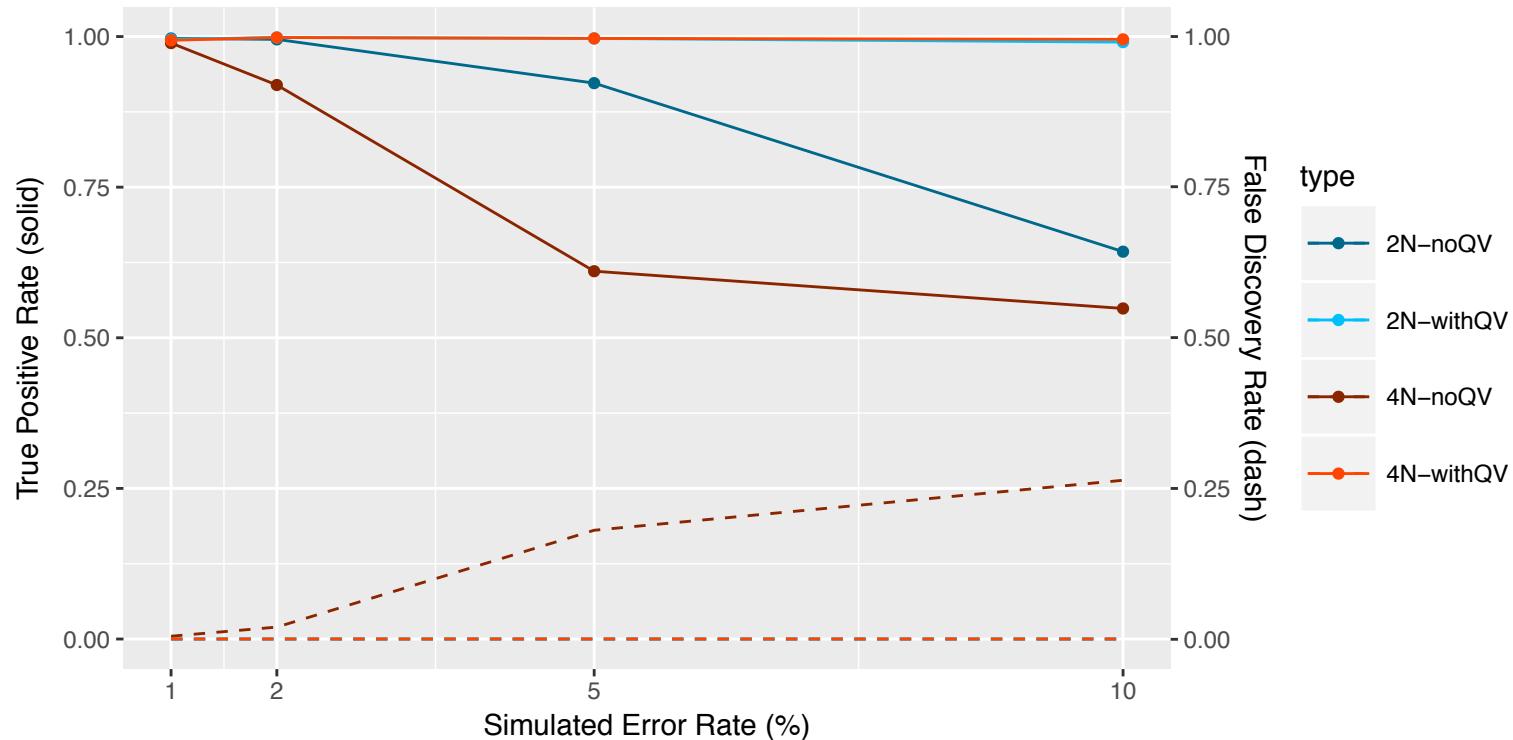
## SNP CALLING ON SIMULATED 100 HUMAN GENES

- Using QV improves SNP recovery
- SNP recovery (TPR) remains high for both diploid and tetraploid even at 10% error

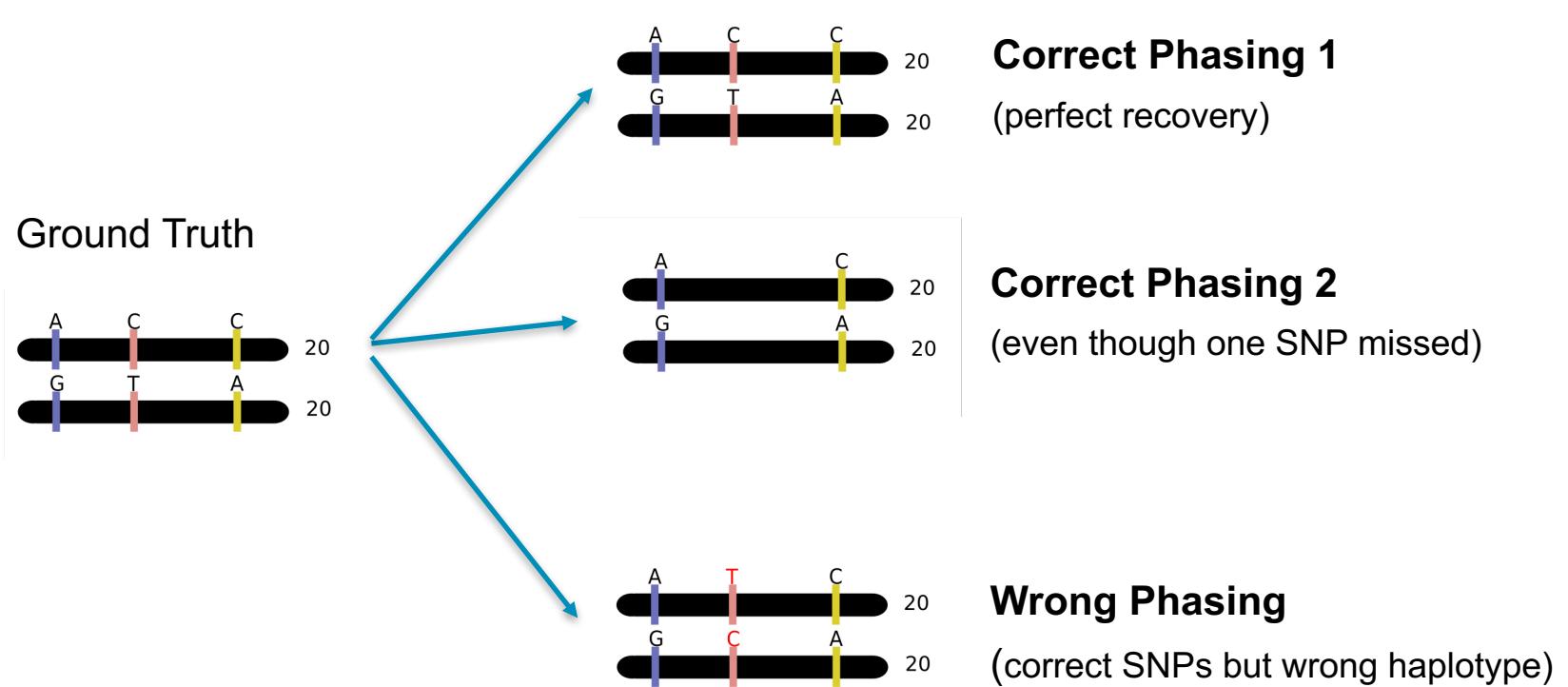


## SNP CALLING ON SIMULATED 100 HUMAN GENES

- False discovery rate remains low until 5% error
- False discovery rate increases with erorr rate for tetraploid-noQV



## PHASING EVALUATION: CRITERION



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.

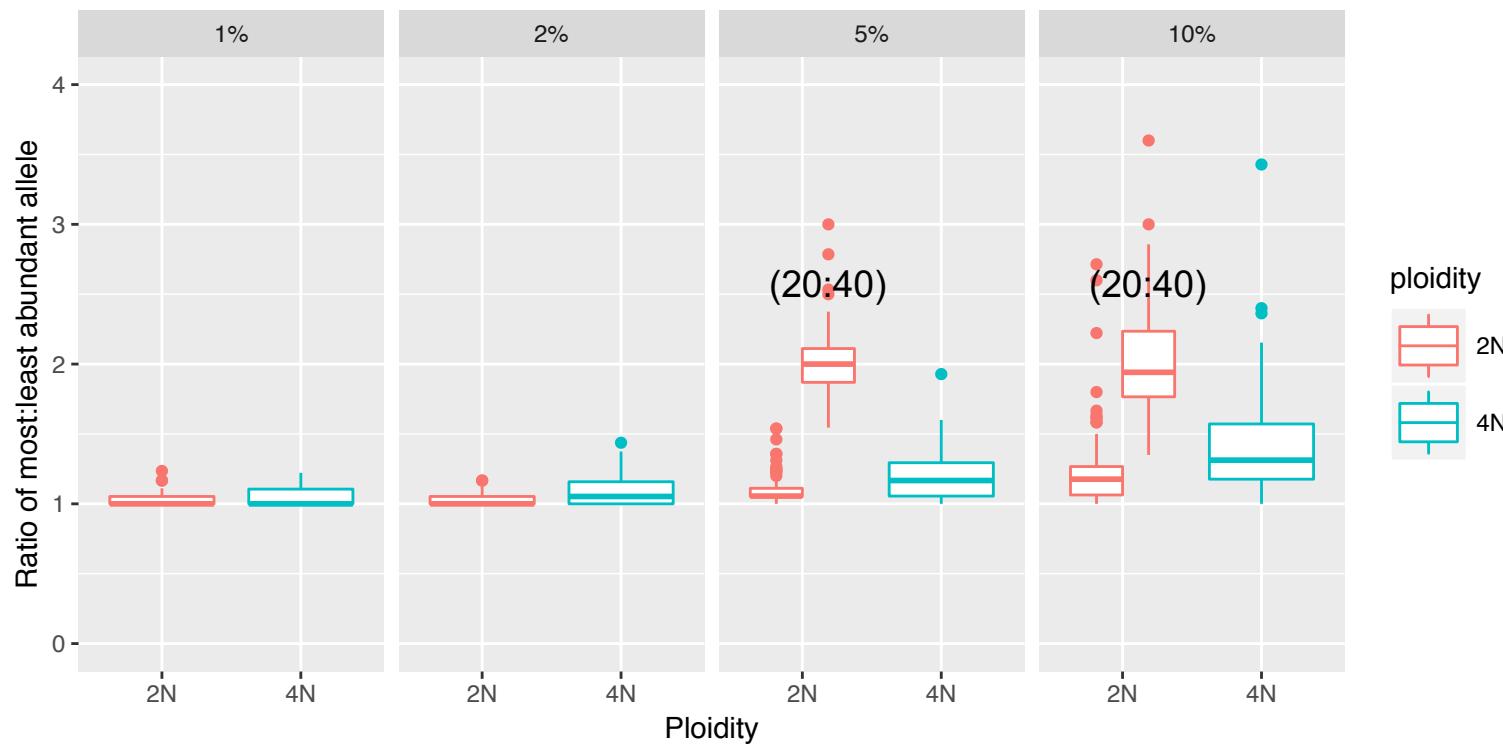
Error Rate	2N no QV	2N with QV	4N no QV	4N with QV
1%	100%	100%	95%	98%
2%	100%	100%	82%	98%
5%	100%	100%	55%	96%
10%	100%	93%	40%	84%



Worse than no QV due to aggressive dropping of reads  
(future work: relax criterion for using reads in haplotyping)

# POTENTIAL TO RECOVER ALLELIC SPECIFIC EXPRESSION

**Ratio of most abundant : least abundant allele.**



Except for (20:40), all 2N simulated with (20:20) read coverage.

All 4N simulated with (20:20:20:20) read coverage.



PACBIO®

# Evaluation on F1 Cattle Data

# BRAHMAN X ANGUS F1 CATTLE

## Genome Assembly

- Brahman x Angus F1 cattle
- 115-fold coverage on PacBio RS II 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

## Iso-Seq Transcriptome Data

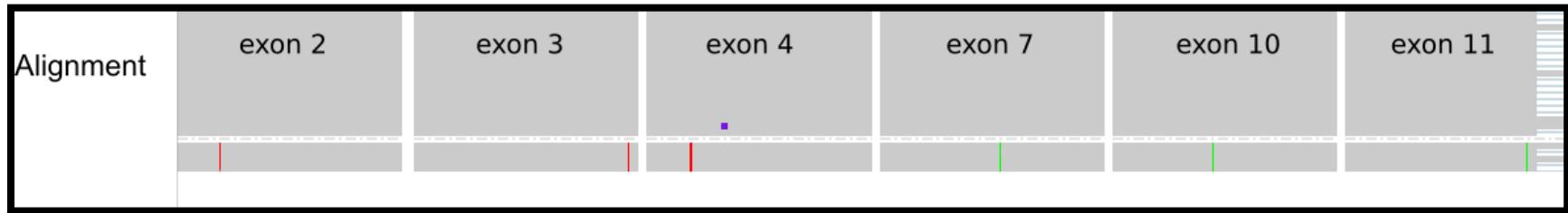
- 8 Sequel cells of tissues from single individual
- Analyzed using IsoSeq2
- Mapped to genome with  $\geq 99\%$  coverage,  $\geq 95\%$  identity
- 30,137 final isoforms (12,101 genes)
- Selected for phasing: 1758 genes with  $\geq 40$  full-length CCS read coverage

## SNP EVALUATION FOR BRAHMAN X ANGUS

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

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

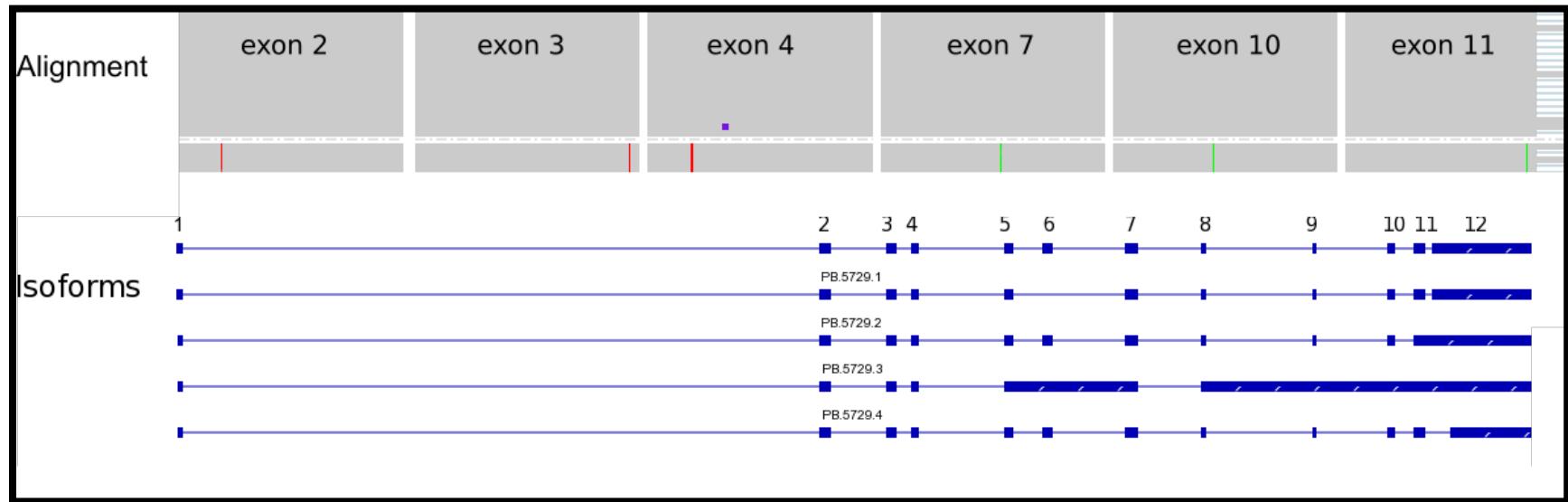
## EXAMPLE OF SNP CALLING VERIFIED BY GENOME



Full-length CCS read alignment, showing only exons with SNPs.

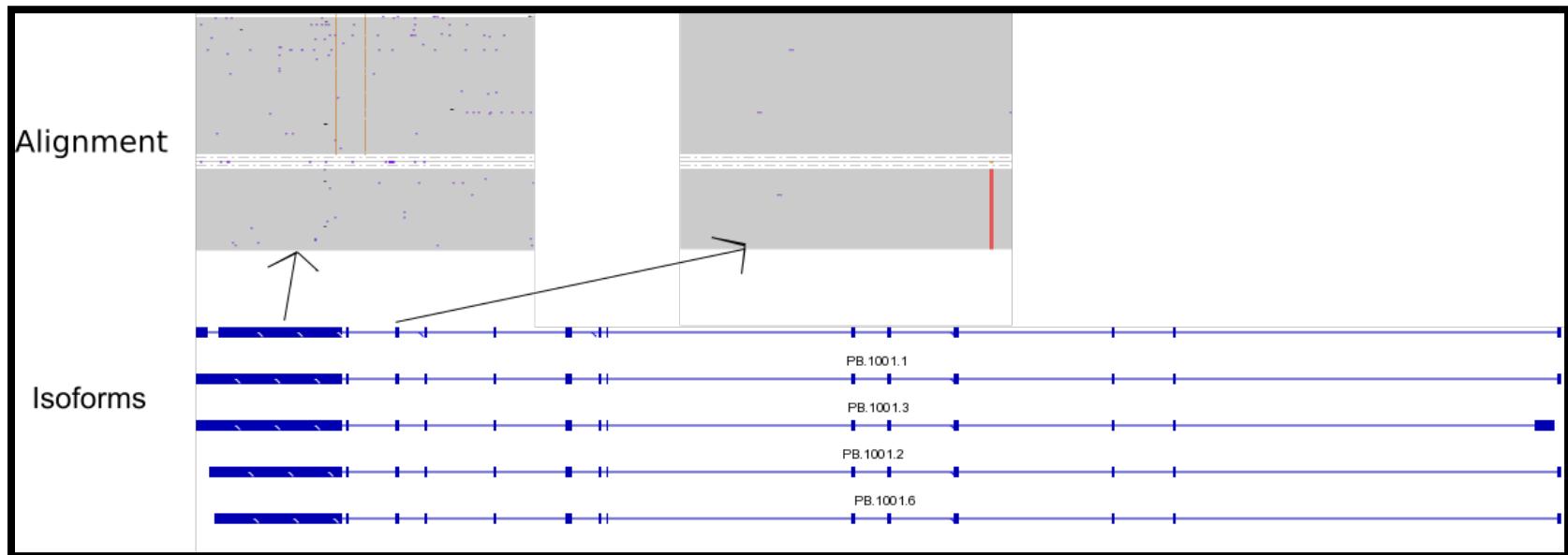
Reads are sorted through “quickphase” in IGV browser showing clear segregation of alleles.  
All 6 SNPs validated by genome assembly Unzip results.

## EXAMPLE OF SNP CALLING VERIFIED BY GENOME



There are 5 different isoforms for this gene. All isoforms cover all 6 SNP sites.

## VPS36 ISOFORMS CALLED SNPS NOT PHASED IN GENOME

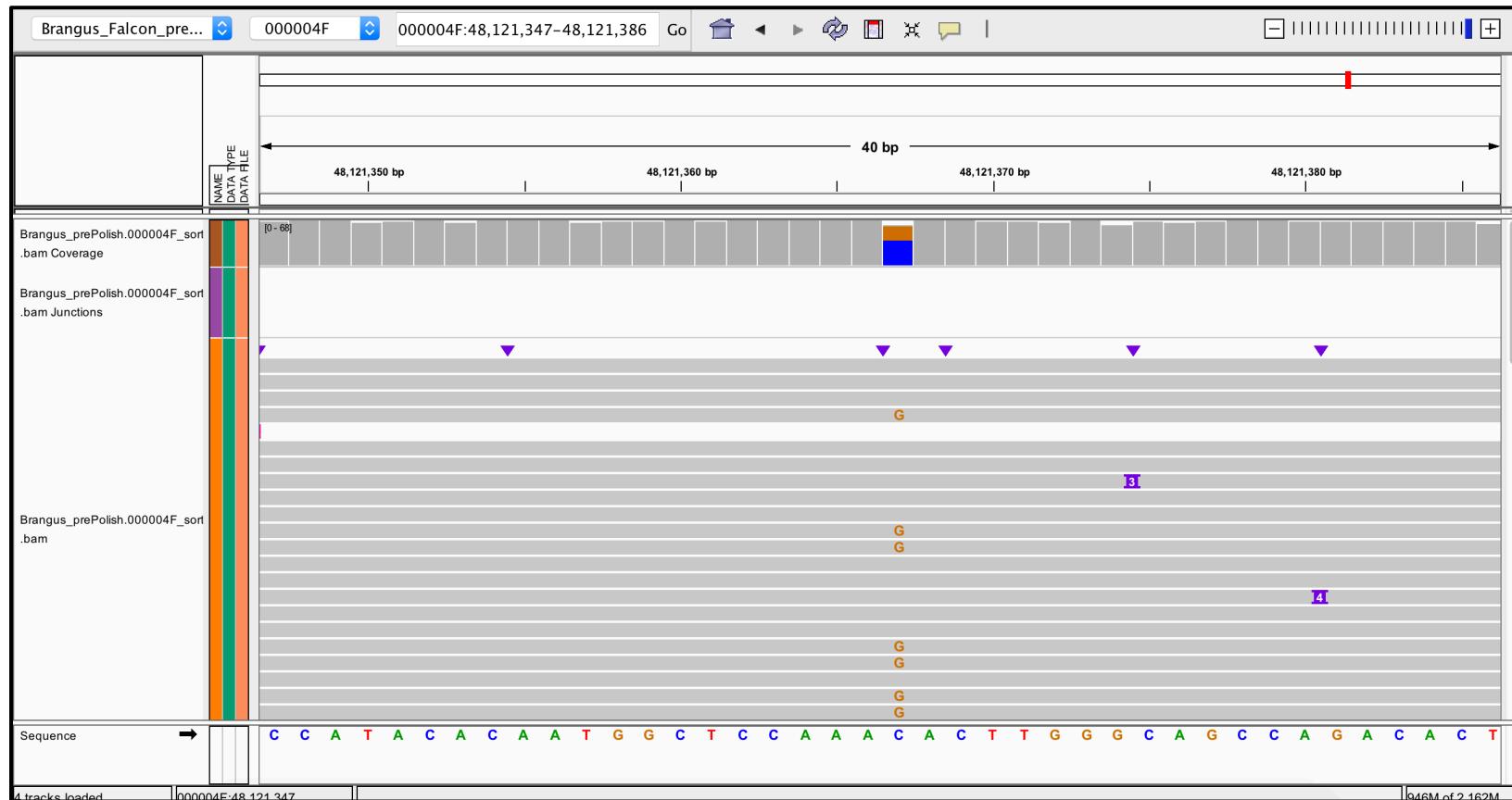


This gene (PB.1001, VPS36) contains 228 FL reads.

- Strong evidence for the 3 SNPs.
- Unzip did not phase this region – so, are the SNPs supported by genome?

# VPS36 ISOFORMS CALLED SNPs NOT PHASED IN GENOME

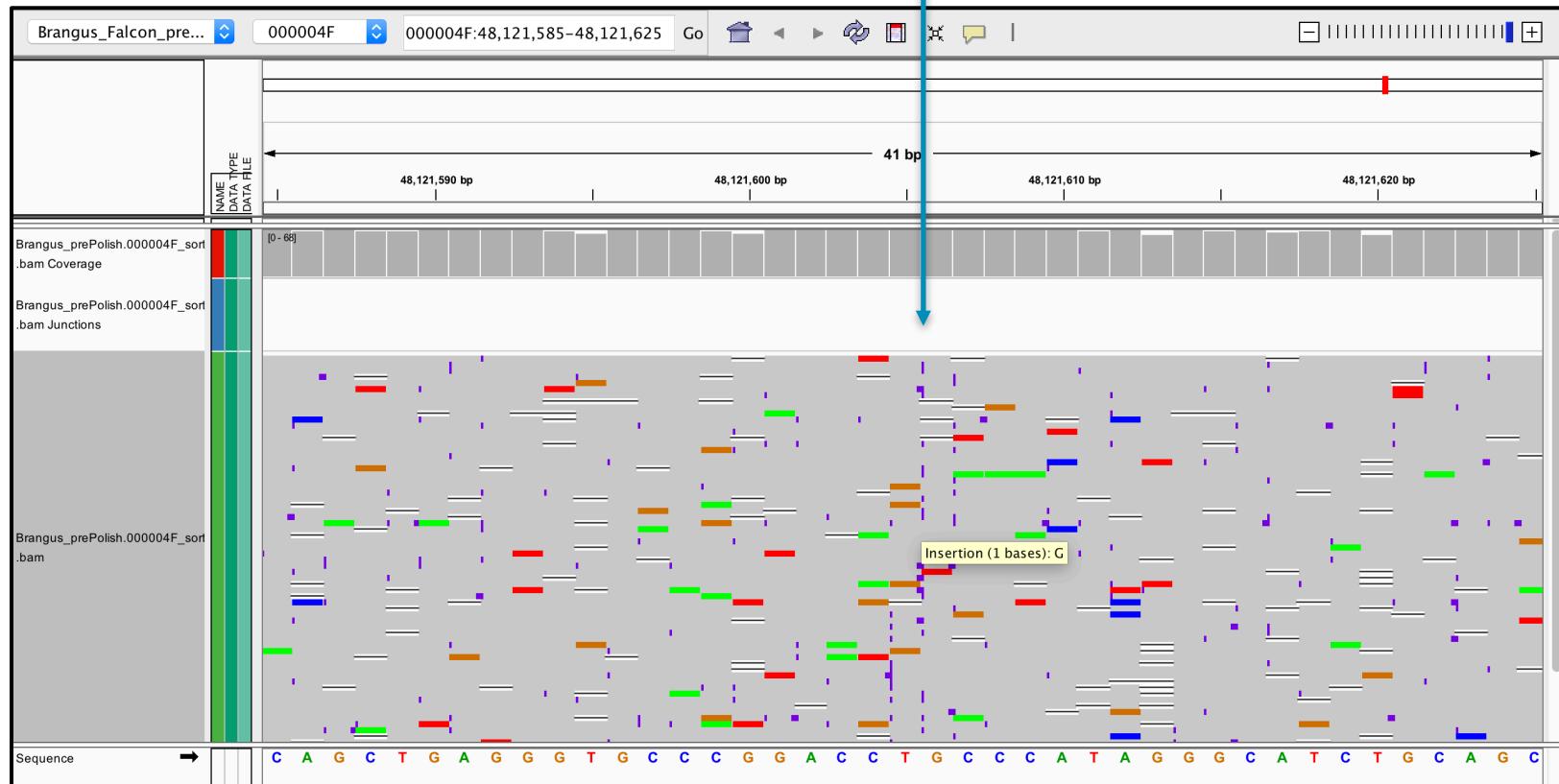
The first SNP 000004F|arrow|arrow:48163477 (C->G) is supported in the pre-polish BAM file.



## VPS36 ISOFORMS CALLED SNPS NOT PHASED IN GENOME

The second SNP 000004F|arrow|arrow:48163716 (T->G) is the second T in the sequence context GCCCGGACCTTGCCCATAGG which in the pre-polish BAM file shows evidence that the second “T” is either a “T” or a “G”.

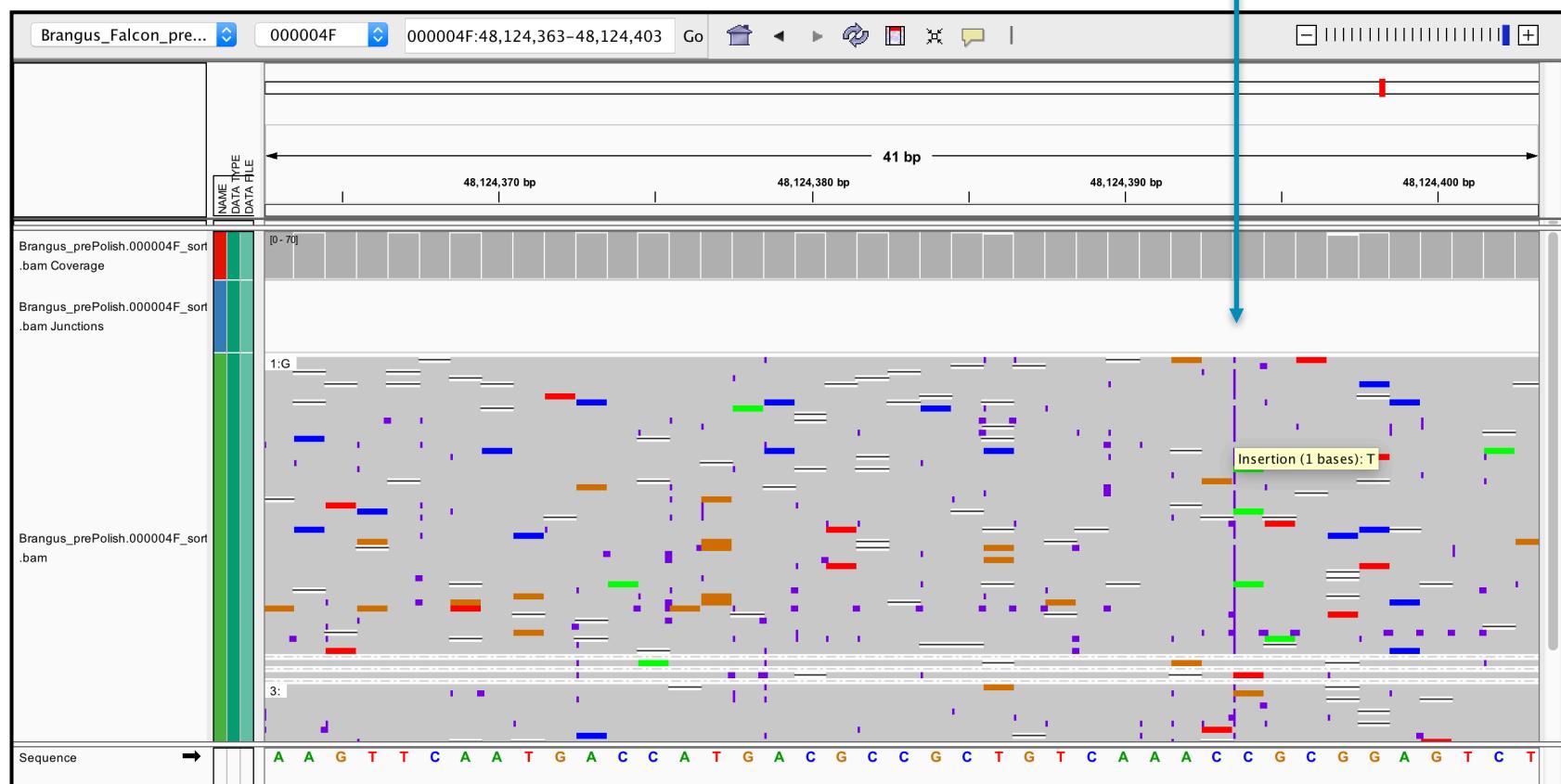
the insertion is either a “T” or a “G” which is the SNP



## VPS36 ISOFORMS CALLED SNPs NOT PHASED IN GENOME

The third SNP 000004F|arrow|arrow:48166508 (A->T) is also an insertion against the pre-polish sequence that is supported by the genome subread data.

the insertion is either a “A” or a “T” which is the SNP

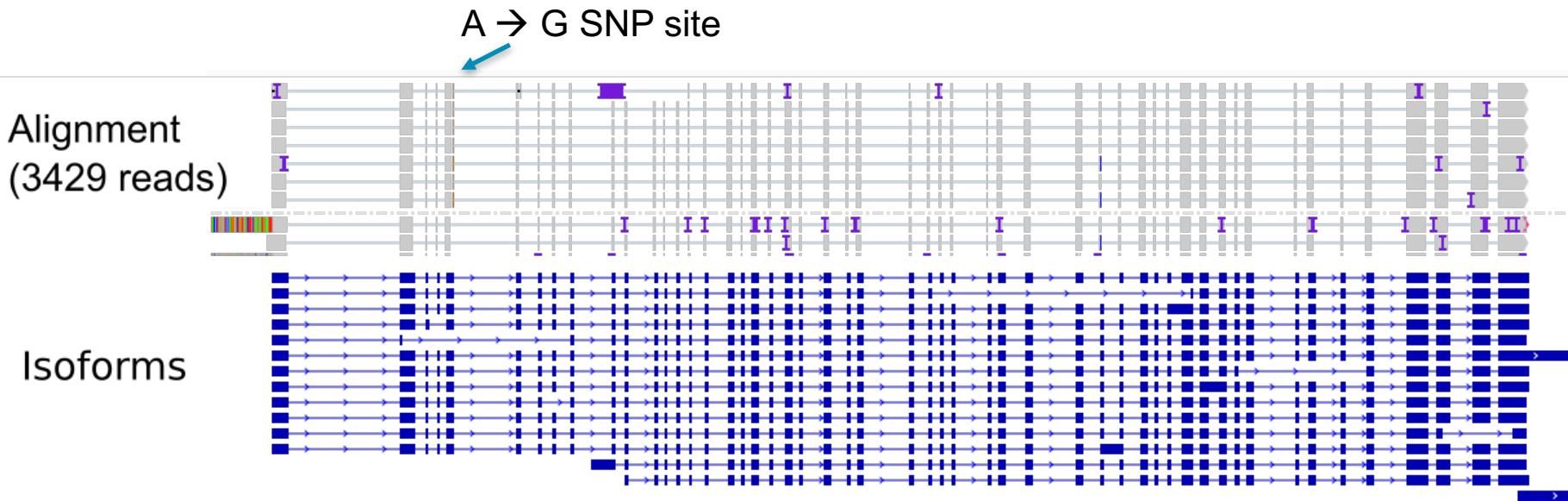


## POTENTIAL A → G RNA EDITING IN COL1A1

CHROM	POS	REF	ALT	SNP IN GENOME?
000071F	7663000	A	G	N
000071F	7671641	T	C	Y

PB.8679 gene (COL1A1) contains a A → G SNP not supported by genome.  
A single alternative contig (000071F\_029) covers the whole region.

## POTENTIAL A → G RNA EDITING IN COL1A1

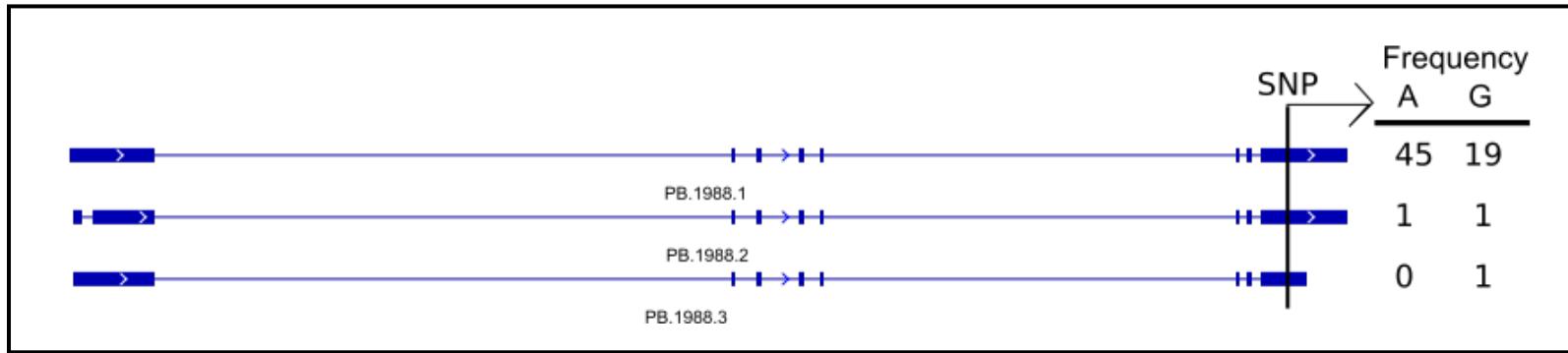


## POTENTIAL A → G RNA EDITING IN COL1A1



**Conclusion:** COL1A1 contains an non-dominant isoform (PB.8679.1) that uses an alternative donor splice site in exon 5 that includes a potential A → G editing site.

## POTENTIAL ALLELE IMBALANCE FOR KIF3C GENE IN BRAIN



- KIF3C is observed in brain only
- The SNP is in the 3' UTR region ( $A \rightarrow G$ ) and is verified by genome
- The major isoform expresses the A allele more dominantly

## SNP EVALUATION FOR BRAHMAN X ANGUS

SNP Type	Count
<b>True Positive</b> (called by both)	8334
<b>False Negative</b> (called by genome only)	259
<b>May Not False Positive</b> (called by transcript only)	1203

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

However, many of the transcript-only SNPs could be true. They could be not phased by the genome due to low heterozygosity, low coverage, or RNA editing.

## ISOPHASE SUMMARY

IsoPhase is a direct extension of the Iso-Seq analysis, utilizing full-length read information to detect SNPs and call haplotypes.

Based on both simulated and real data, it shows **high true discovery rate** and **low false positive rate**.

It has the potential to reveal **allelic specific isoform expressions**.

## FUTURE WORK

### Detect Indels

- Currently, only substitution SNPs are called
- Calling simple (1-3 bp) indels is conceptually possible, but will require work

### Reduce Read Coverage Requirement

- Currently, requires 40-fold per-gene read coverage
- Reducing read coverage may increase detection but also false calls

### Phasing Without a Reference Genome

- Cogent could be used to reconstruct coding “contigs” to map full-length reads back to. However Cogent will require minor modifications to understand unresolved exonic orderings based solely on Iso-Seq data.

### Pipeline for Comparing Genome Phasing Results with IsoPhase

- Automated scripts for showing agreement and disagreement

Please let me know if you have other ideas for making IsoPhase more awesome!



PACBIO®

[www.pacb.com](http://www.pacb.com)

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.