FALCON-Phase: Integrating PacBio and Hi-C data for phased diploid genomes

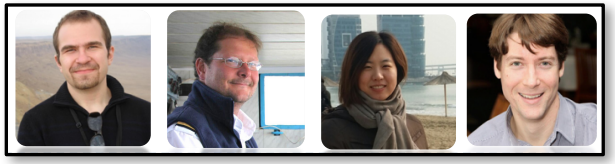
SB Kingan, Staff Scientist, Bioinformatics, PacBio

SMRT Informatics Developers Conference, Leiden, NL, June 14, 2018

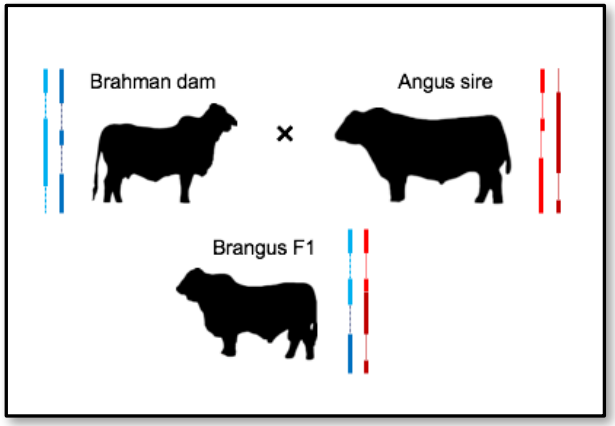
CURRENT APPROACHES TO LONG READ DIPLOID ASSEMBLY

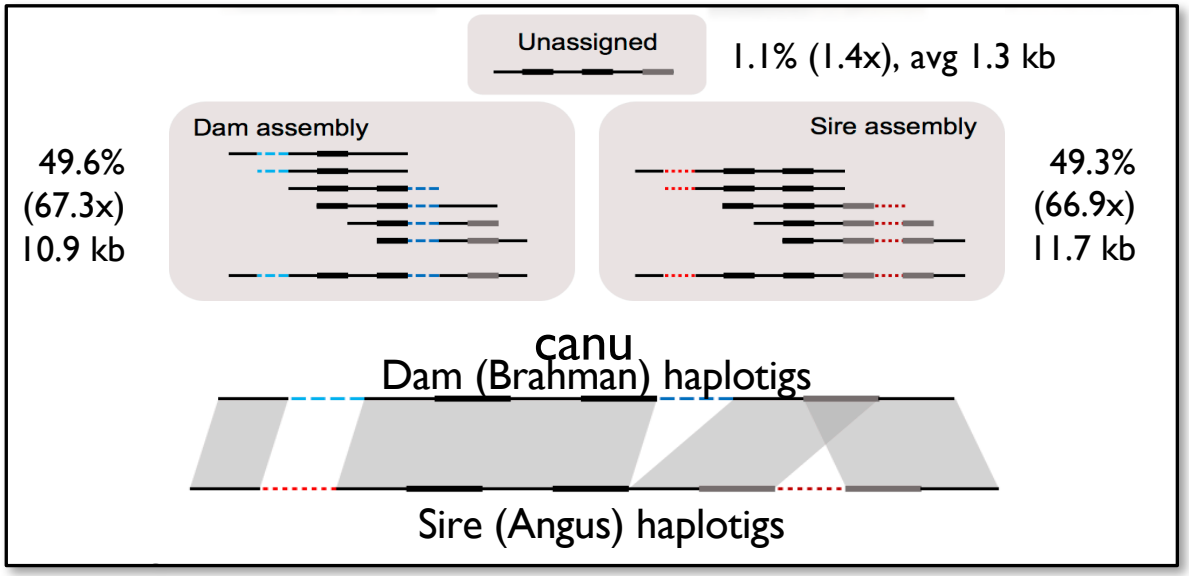
1. Separate Reads with Trio Binning (TrioCanu)

- PacBio data for F1
- ILMN data for parent-specific k-mers
- Bin PacBio reads with k-mers
- Perform two haploid Canu assemblies



EXAMPLE ON F1 BULL



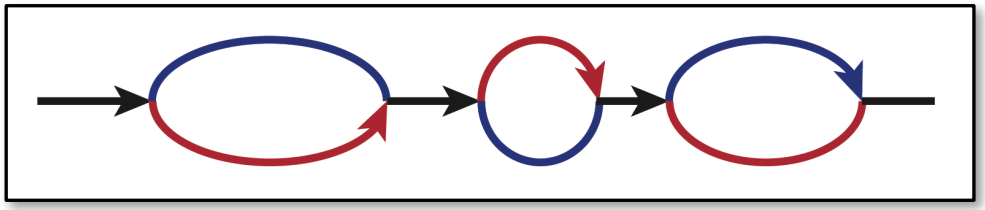


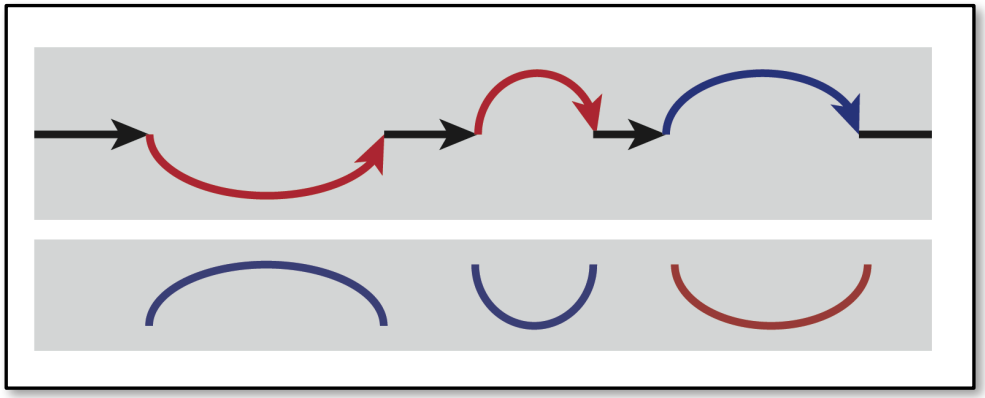
Koren, S. et al. (2018). Complete assembly of parental haplotypes with trio binning. bioRxiv. Available from: https://doi.org/10.1101/271486

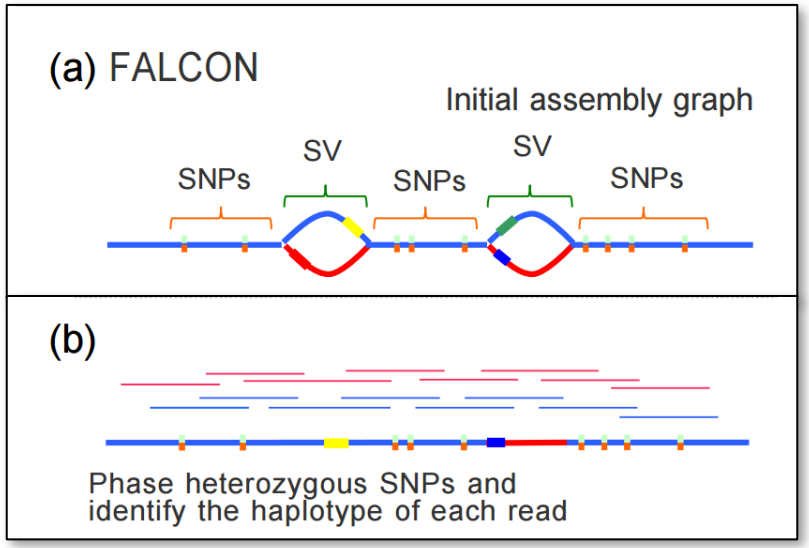
CURRENT APPROACHES TO LONG READ DIPLOID ASSEMBLY

2. Separate Haplotypes During Assembly with FALCON-Unzip

- PacBio data for diploid individual (no trio)
- Phase PacBio reads using SNPs identified in initial assembly graph
- Output phased and collapsed regions in high contiguity contigs



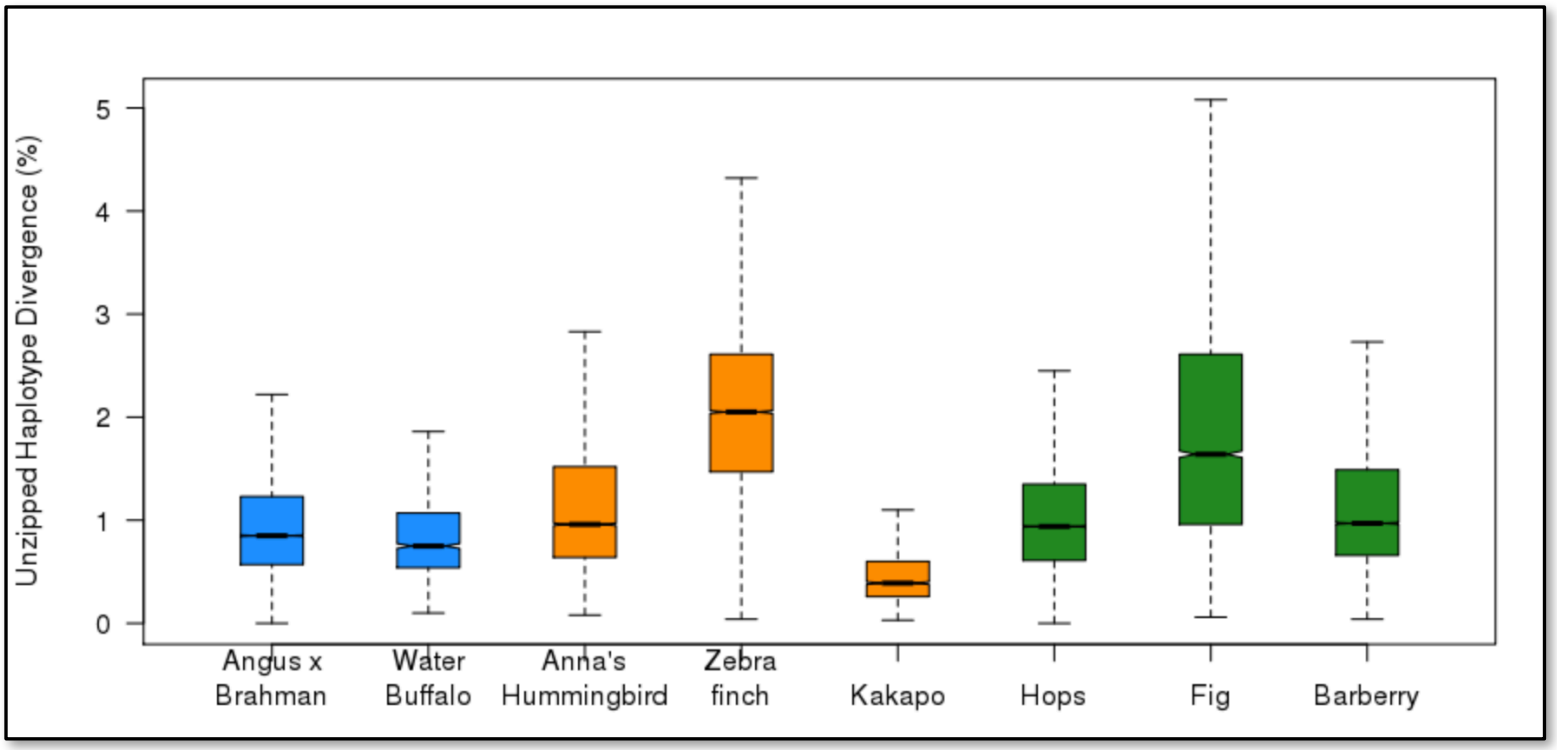




Chin, C.S. et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*. 13(12), 1050.

THE “SWEET SPOT” FOR FALCON-UNZIP

Up to 5% divergence can be unzipped

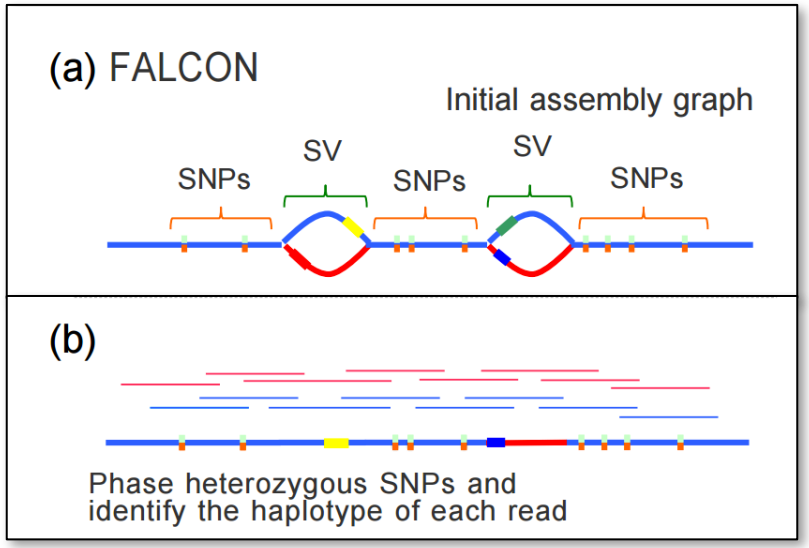


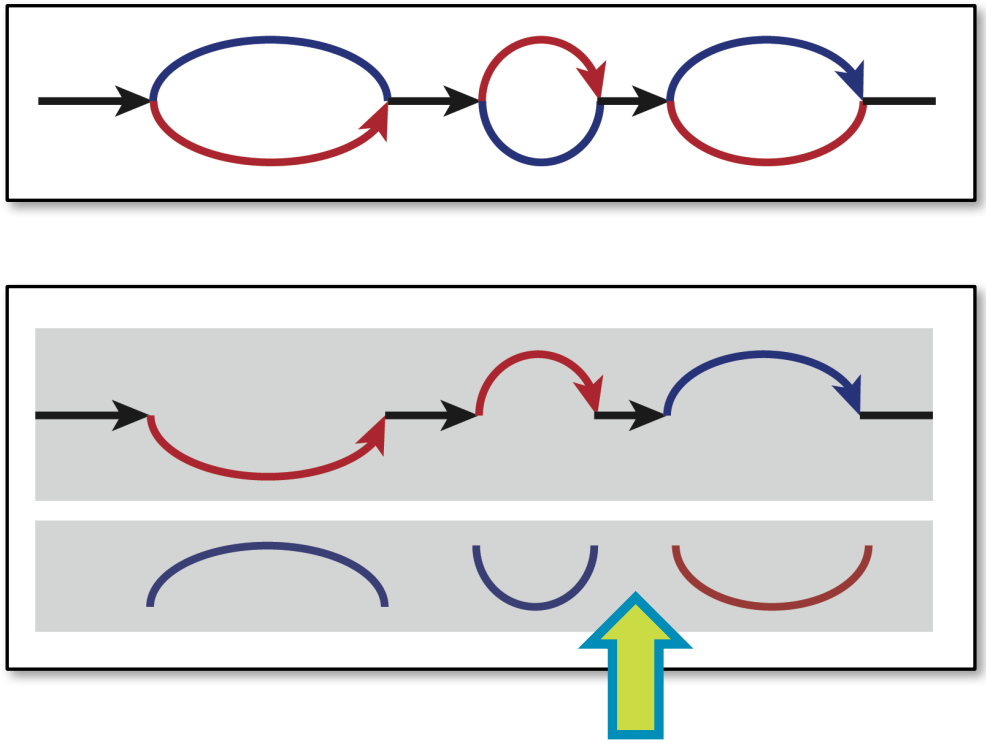
Data Credits: Tim Smith, John Williams, Lloyd Low, Erich Jarvis, John Henning, David Hendrix, Carlos Machado, Iago Hale

CURRENT APPROACHES TO LONG READ DIPLOID ASSEMBLY

2. Separate Haplotypes During Assembly with FALCON-Unzip

- PacBio data for diploid individual (no trio)
- Phase PacBio reads using SNPs identified in initial assembly graph
- Output phased and collapsed regions in high contiguity contigs





“Phase/Haplotype Switch”

Chin, C.S. et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*. 13(12), 1050.

FALCON-PHASE: MOTIVATIONS AND GOALS

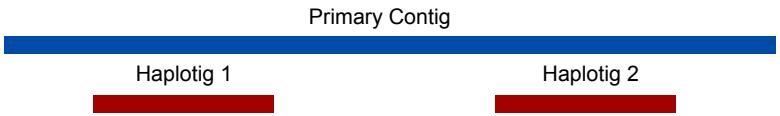
- FALCON-Unzip phase blocks are small
  - Phasing is function of heterozygosity, read depth, read length
  - Phase switches between haplotype blocks are nearly random
- Haplotype/phase switches are problematic
  - “Franken-haplotypes” impact base accuracy, gene prediction
  - Scaffolding errors
- Hi-C contains long-range haplotype information
- FALCON-Phase Tool
  - Open-source snakemake pipeline
  - Co-development project between PacBio and Phase Genomics
  - ***Can be applied at contig and scaffold scale***

Zev Kronenberg

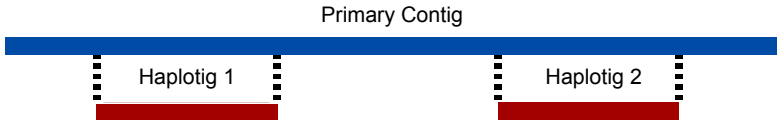
Logo Credit: Kaylee Mueller

FALCON-PHASE WORKFLOW

Input: FALCON-Unzip assembly & HiC data

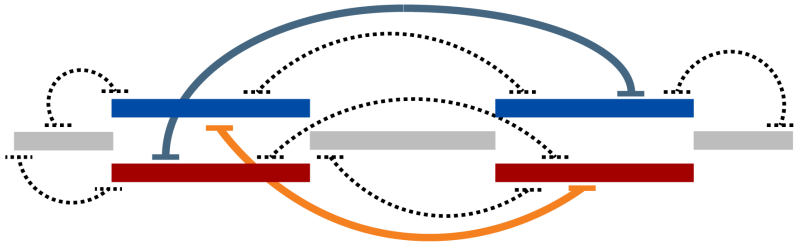


1. Identify **haplotig placement** on primary contigs

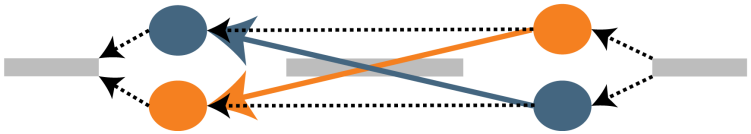


2. **Mince** primary contigs: separate **haplotig pairs** and **collapsed haplotypes**

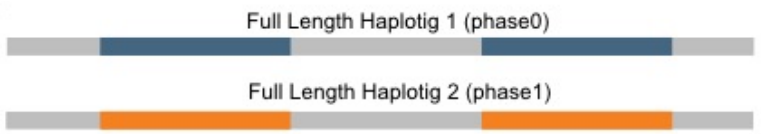
3. **Map** paired HiC reads to minced contigs and generate normalized **contact matrix**



4. **Phase** haplotigs along each primary contig

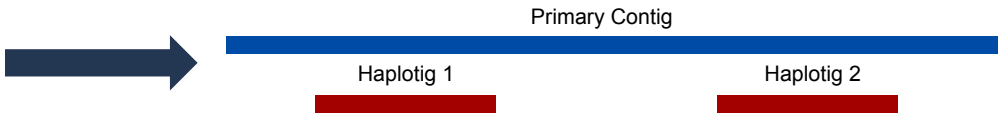


Output: phased full length pseudohaplotypes



FALCON-PHASE WORKFLOW

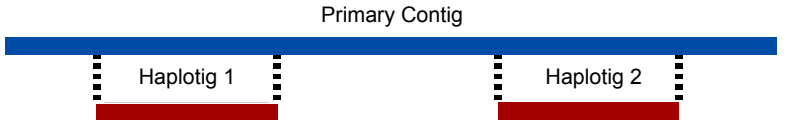
Input: FALCON-Unzip assembly & HiC data



FALCON-PHASE WORKFLOW

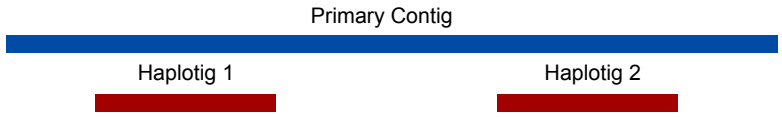
Input: FALCON-Unzip assembly & HiC data

1. Identify **haplotig placement** on primary contigs

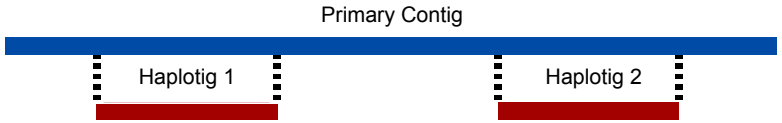


FALCON-PHASE WORKFLOW

Input: FALCON-Unzip assembly & HiC data



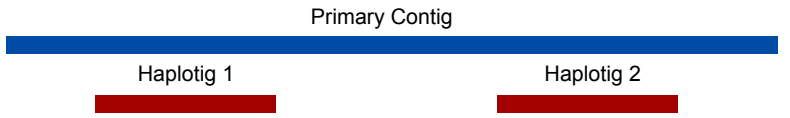
1. Identify **haplotig placement** on primary contigs



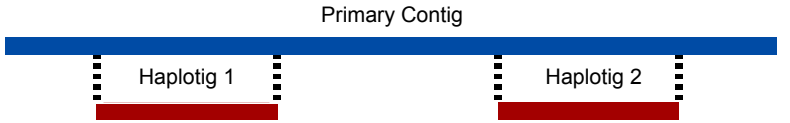
2. **Mince** primary contigs: separate **haplotig pairs** and **collapsed haplotypes**

FALCON-PHASE WORKFLOW

Input: FALCON-Unzip assembly & HiC data

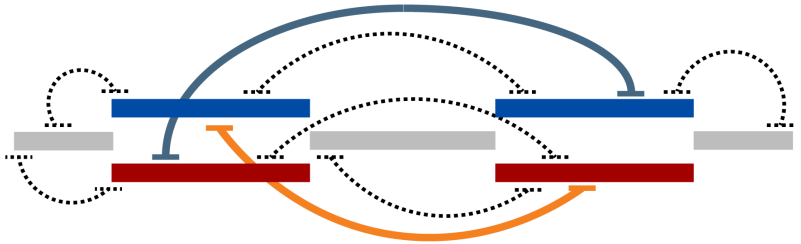


1. Identify **haplotig placement** on primary contigs



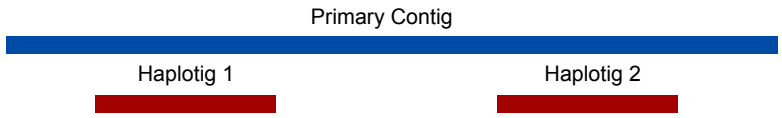
2. **Mince** primary contigs: separate **haplotig pairs** and **collapsed haplotypes**

3. **Map** paired HiC reads to minced contigs and generate normalized **contact matrix**

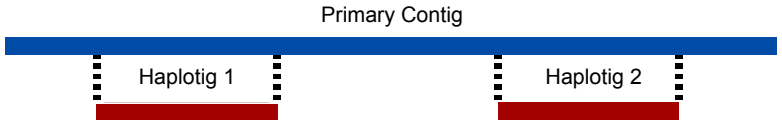


FALCON-PHASE WORKFLOW

Input: FALCON-Unzip assembly & HiC data

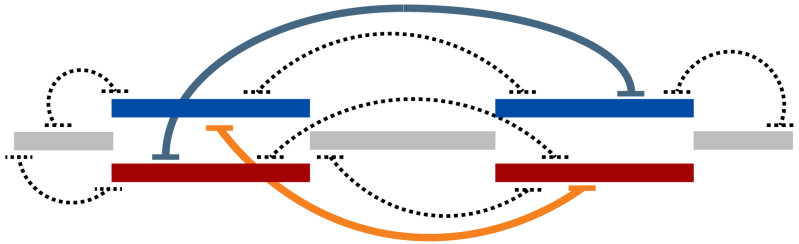


1. Identify **haplotig placement** on primary contigs

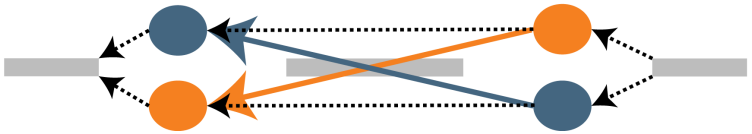


2. **Mince** primary contigs: separate **haplotig pairs** and **collapsed haplotypes**

3. **Map** paired HiC reads to minced contigs and generate normalized **contact matrix**

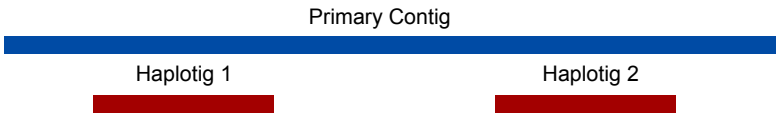


4. **Phase** haplotigs along each primary contig

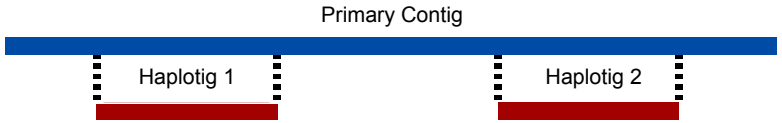


FALCON-PHASE WORKFLOW

Input: FALCON-Unzip assembly & HiC data

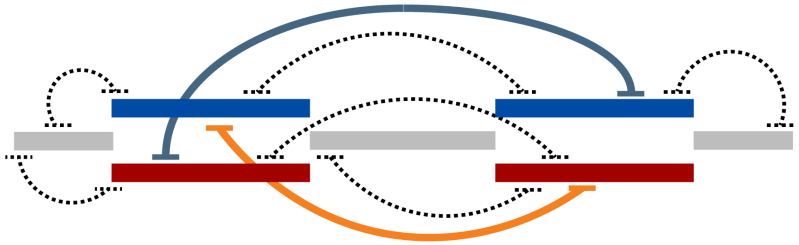


1. Identify **haplotig placement** on primary contigs

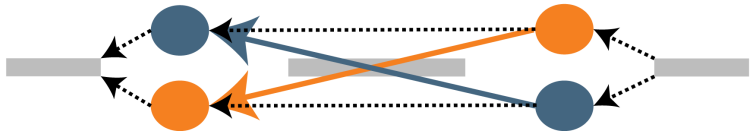


2. **Mince** primary contigs: separate **haplotig pairs** and **collapsed haplotypes**

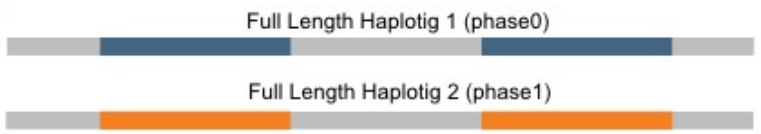
3. **Map** paired HiC reads to minced contigs and generate normalized **contact matrix**



4. **Phase** haplotigs along each primary contig

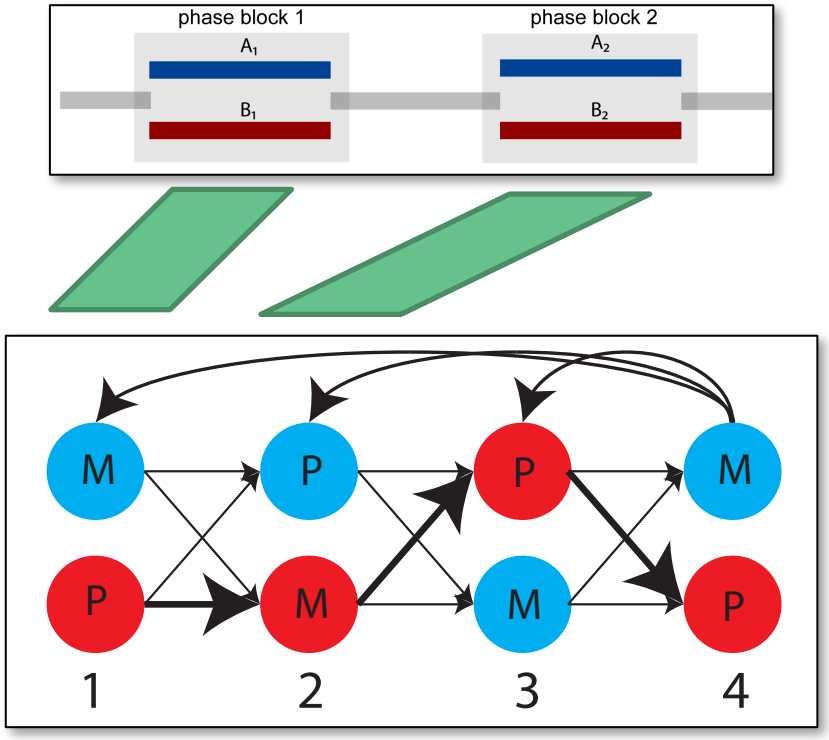


Output: phased full length pseudohaplotypes



PHASING ALGORITHM: INPUTS AND OUTPUTS

- FALCON-Unzip Input
  - Order and pairing of phase blocks along primary contig

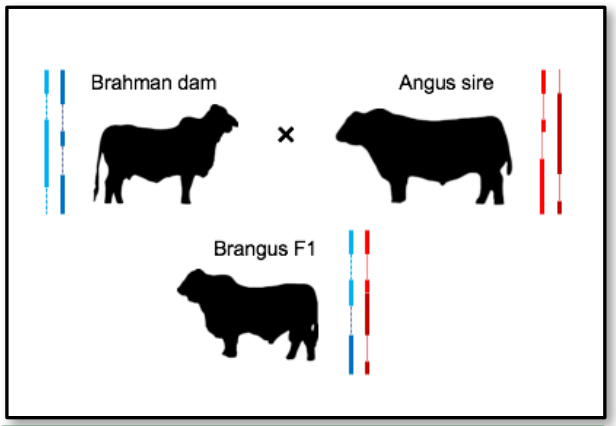


- Hi-C Input
  - Normalized contact matrix between each phase block

	A_1	B_1	A_2	B_2
A_1	18	.	.	.
B_1	1	15	.	.
A_2	2	9	15	.
B_2	7	0	3	12

- Output
  - Majority phase assignment configuration for haplotigs along primary contig

VALIDATION DATASET: ANGUS-BRAHMAN F1 BULL



Data: Tim Smith (USDA), John Williams and Stefan Hiendleder (U Adelaide)
Canu Asms: Adam Phillippy, Sergey Koren, Arang Rhie (NHGRI)

FALCON-Unzip: 90% Unzipped

CONTIGS	NUMBER	LENGTH	N50
PRIMARY	1427	2.71 Gb	31.4 Mb
HAPLOTIGS	5879	2.45 Gb	2.48 Mb

Phase Genomics Hi-C

- 200 million read pairs

TrioCanu Assemblies

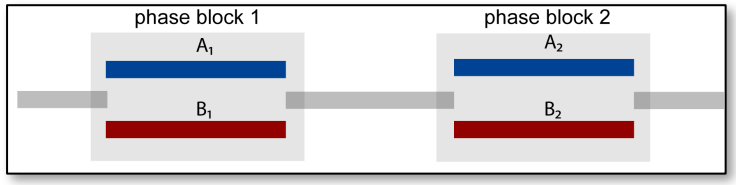
CONTIGS	NUMBER	LENGTH	N50
ANGUS DAM	1747	2.57 Gb	26.7 Mb
BRAHMAN SIRE	1040	2.68 Gb	23.3 Mb

Parental SNP Calls

- 20-25x coverage ILM PE
- read mapping with bwa mem
- SNV calls with freebayes

Koren et al. 2018

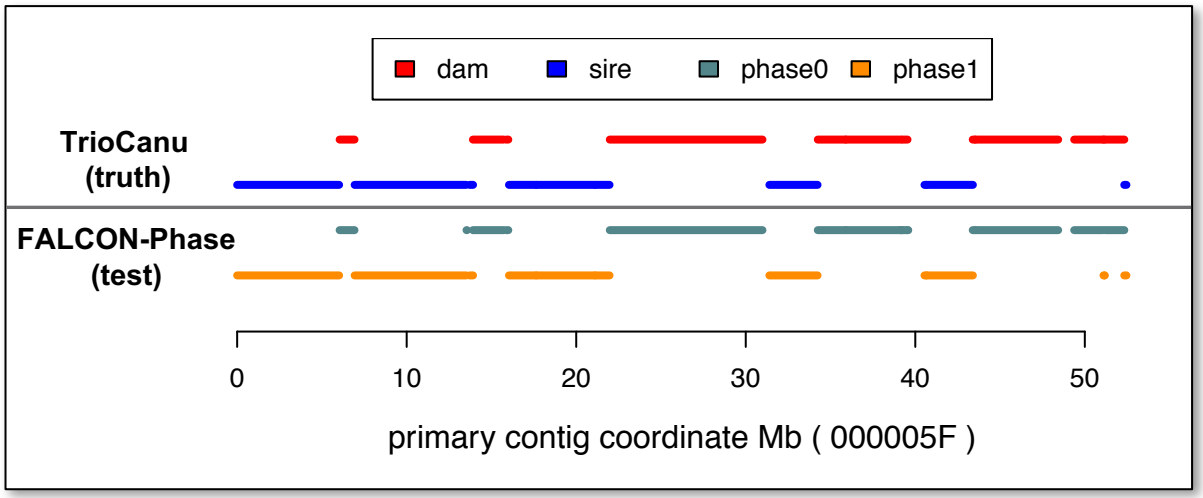
PHASE ASSIGNMENT ACCURACY: PARENTAL ASSIGNMENT



- Minimap to Canu Asms
- Highest PID for longest alignment
- Required concordance between pairs

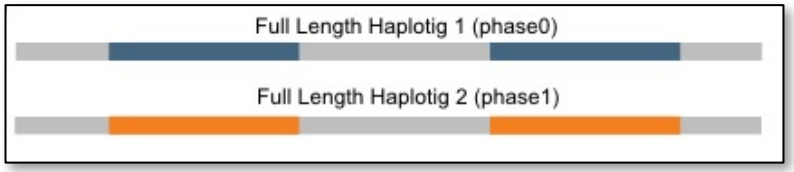
Contig Assignment	Count	Length (%)	Mean Length
Dam	2,305	2.32 Gb (42 %)	1.01 Mb
Sire	2,305	2.32 Gb (42 %)	1.01 Mb
No Parent	1,704	116 Mb (2.1 %)	68.1 kb
Collapsed	3,934	374 Mb (6.8 %)	88.2 kb

RESULTS FOR PRIMARY CONTIG 000005F



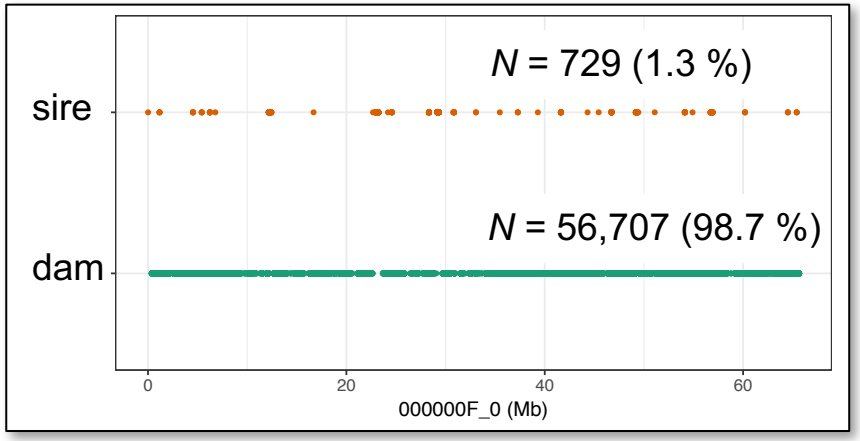
overall accuracy: 96.72%

PHASE ASSIGNMENT ACCURACY: PARENTAL SNV CALLS

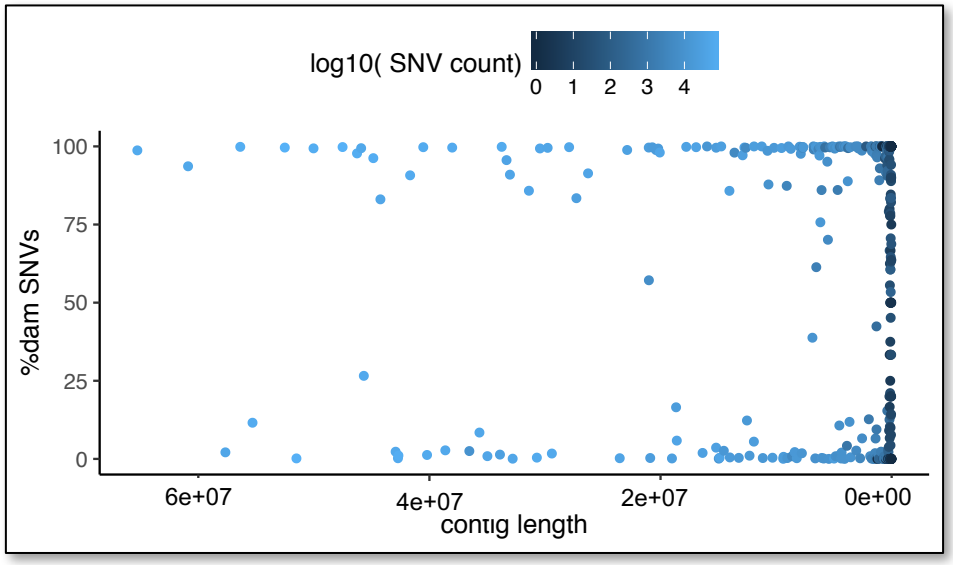


- bwa mem alignment of parental PE ILM data to phase0 haplotigs
- Variant calling with Freebayes
- SNV filtered for homozygous sites that differ between parents

PRIMARY CONTIG 000000F_0 (DAM)



ACCURACY BY PRIMARY CONTIG LENGTH

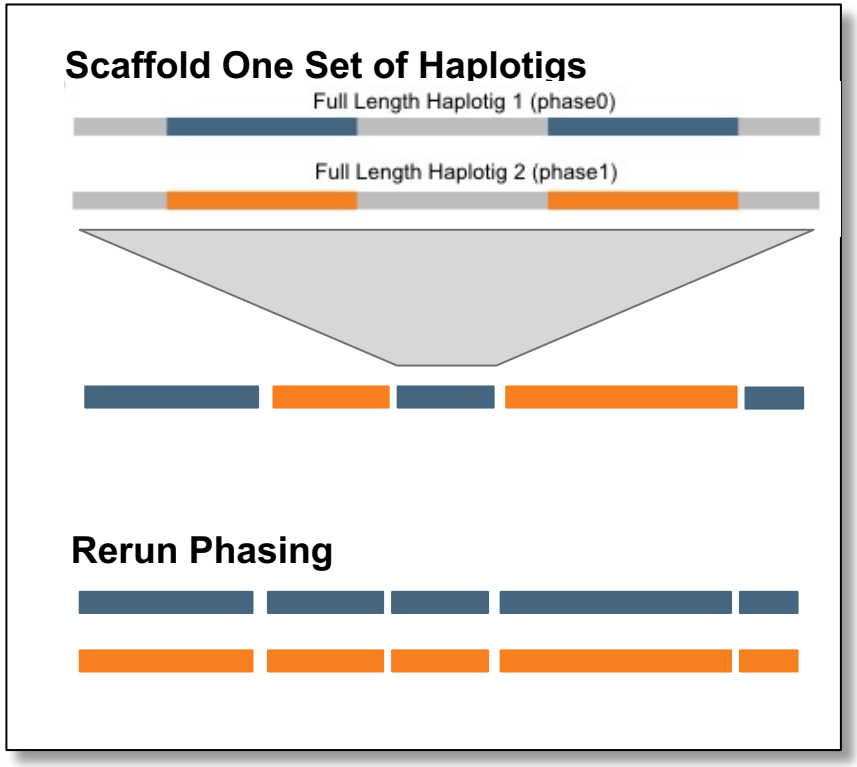


OVERALL PERFORMANCE

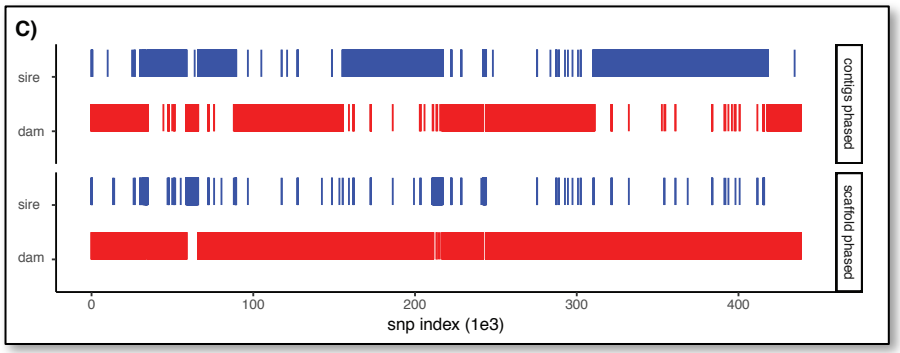
	SNV Count	Correct
dam	2,031,334	97.4 %
sire	1,464,748	95.8 %
total	3,496,082	**96.7 %**

PHASING CHROMOSOME-SCALE SCAFFOLDS

- Scaffold one set of full-length haplotigs with Proximo (Phase Genomics)
- Scaffolds are chromosome-scale
- We know:
  - order of contigs along scaffold
  - pairing of phase 0 and phase 1
- FALCON-Phase Scaffolds



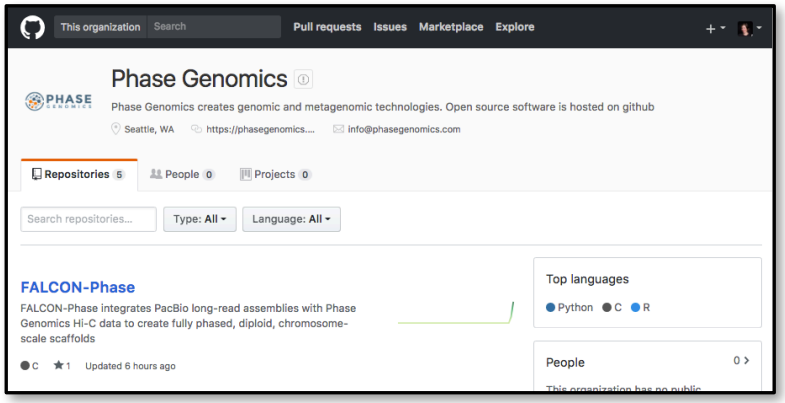
PARENTAL SNVS AFTER SCAFFOLD PHASING

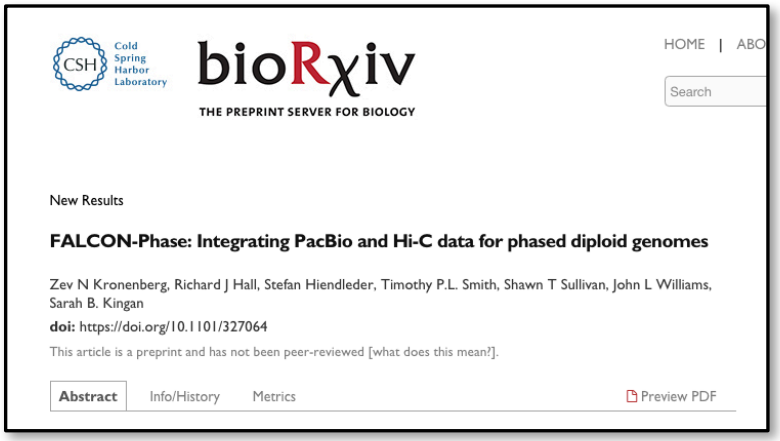


Output: Chromosome-scale, phased, diploid assembly!

SUMMARY

- FALCON-Phase is highly accurate
  - > 96% accuracy when tested against parental assemblies or SNVs
- FALCON-Phase implemented in snakemake pipeline
  - Run locally or on cluster, Open source
- PacBio plus HiC is all you need to produce, phased, chromosome-scale diploid assembly
- More Info:





ACKNOWLEDGEMENTS

- Coauthors
  - Zev Kronenberg (Phase)
  - Richard Hall (PacBio)
  - Stefan Hiendleder (U Adelaide)
  - Timothy Smith (USDA)
  - Shawn Sullivan (Phase)
  - John Williams (U Adelaide)

PACBIO®

Greg Concepcion
Jonas Korlach
Billy Rowell
Ivan Sovic
Liz Tseng
Michelle Vierra

PHASE GENOMICS

Ivan Liachko
Kaylee Mueller
Max Press
Andrew Wiser

Jason Chin
Mark Chaisson
Luke Harmon
Ryan Layer