Image Credit: Iannuzzi, 1996

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

SB Kingan, Staff Scientist, Bioinformatics, PacBio

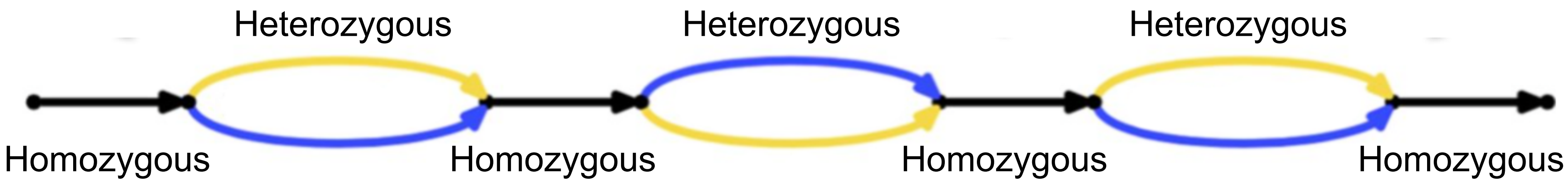
SFAF, May 23rd 2018, Santa Fe, NM

HAPLOTYPE ASSEMBLY IN DIPLOIDS REMAINS A CHALLENGE

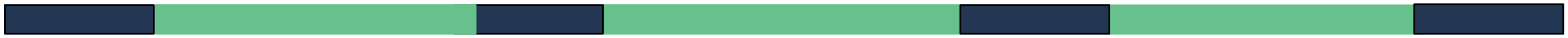
- Traditional assembly paradigm: choose inbred sample
- Non-model genome assembly is increasingly common
- Base accuracy and contiguity suffer if haplotypes collapsed

Korlach et al. 2017, Chin et al. 2016

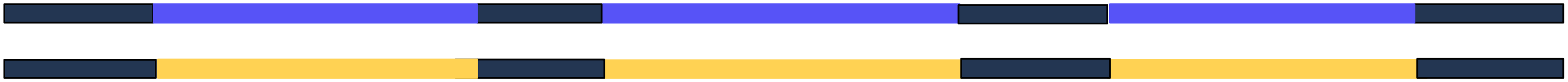
HAPLOTYPE ASSEMBLY IN DIPLOIDS REMAINS A CHALLENGE



“collapsed” haplotypes



contiguous phased haplotigs

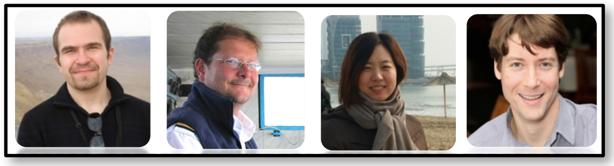


Weisenfeld et al. 2017 , Phillippy 2018

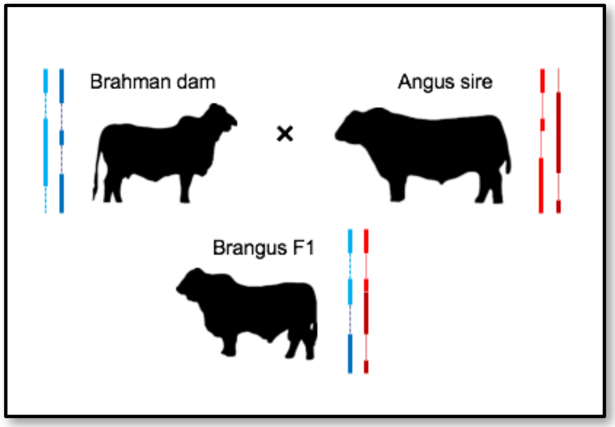
CURRENT APPROACHES TO LONG READ DIPLOID ASSEMBLY

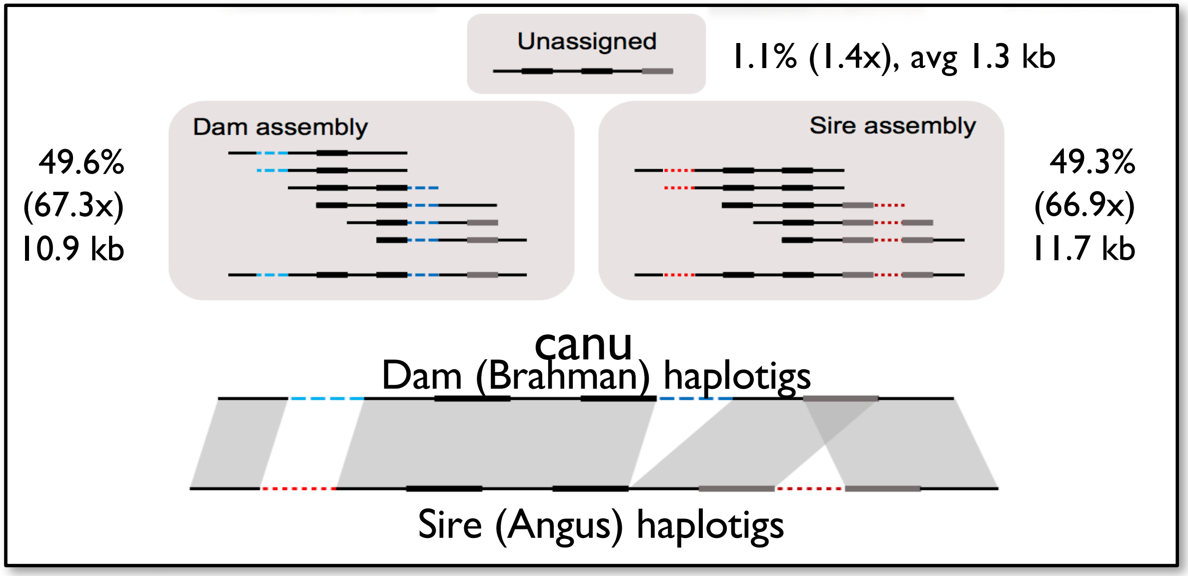
1. Separate Reads with Trio Binning (TrioCanu)

- 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





Unassigned

1.1% (1.4x), avg 1.3 kb

Dam assembly

49.6%
(67.3x)
10.9 kb

Sire assembly

49.3%
(66.9x)
11.7 kb

canu

Dam (Brahman) haplotigs

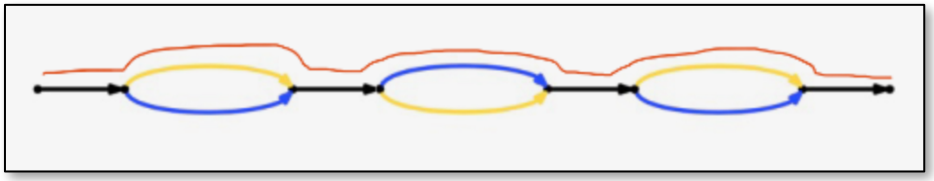
Sire (Angus) haplotigs

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

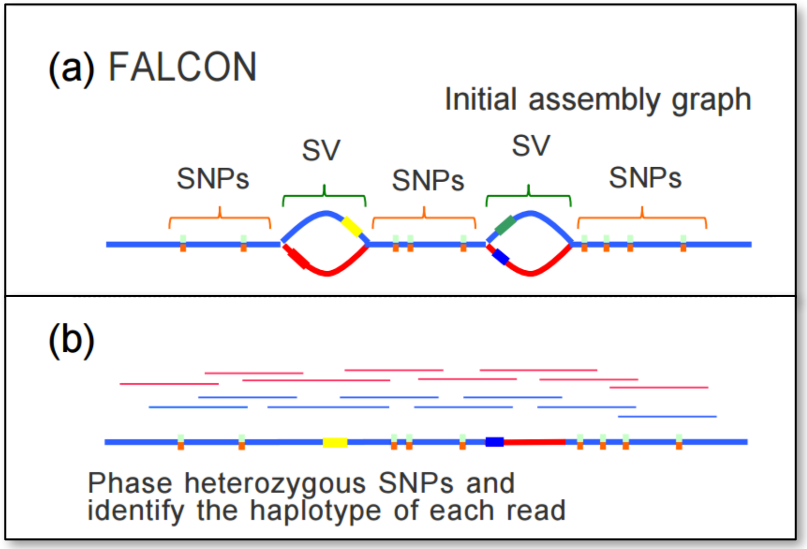
CURRENT APPROACHES TO LONG READ DIPLOID ASSEMBLY

2. Separate Haplotypes During Assembly with FALCON-Unzip

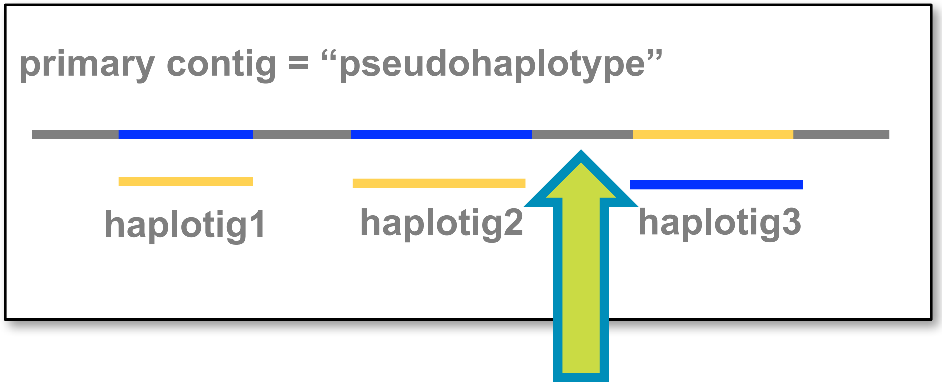
- PacBio data for diploid individual (no trio)
- Phase PacBio reads using SNPs identified in initial assembly graph
- Output phased and collapsed regions in high contiguity contigs



Weisenfeld et al. 2017


(a) FALCON
Initial assembly graph
SNPs
SV
SNPs
SV
SNPs
(b)
Phase heterozygous SNPs and identify the haplotype of each read

PSEUDOHAPLOTYPE AND HAPLOTIGS



"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

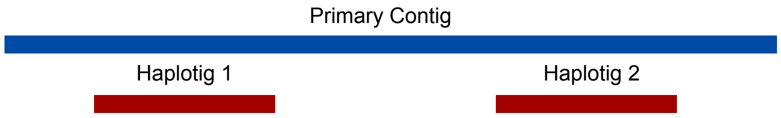
- FALCON-Unzip phase blocks are small
  - Phasing is function of heterozygosity, read depth, read length
  - Phase switches between haplotype blocks are nearly random
- Haplotype/phase switches are problematic
  - "Franken-haplotypes" impact base accuracy, gene prediction
  - Scaffolding errors
- Hi-C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

Zev Kronenberg

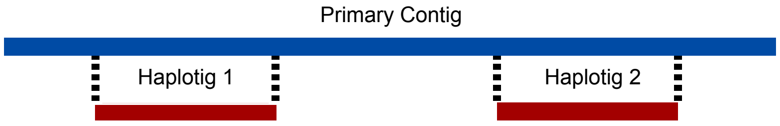
Logo Credit: Kaylee Mueller

FALCON-PHASE WORKFLOW

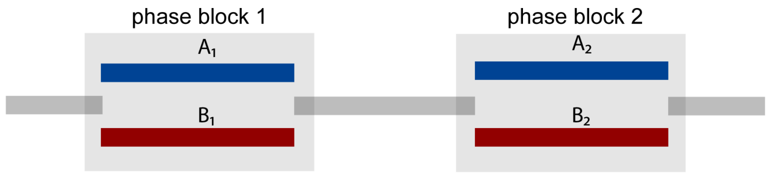
Input: FALCON-Unzip assembly & HiC data



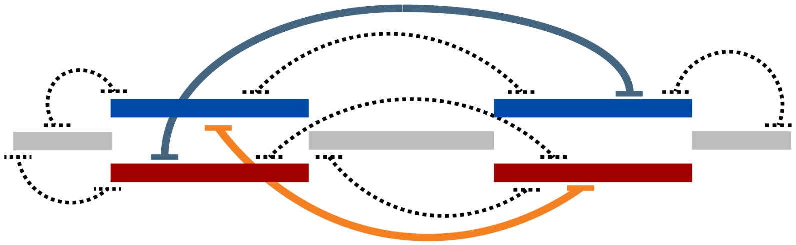
1. Identify **haplotig placement** on primary contigs



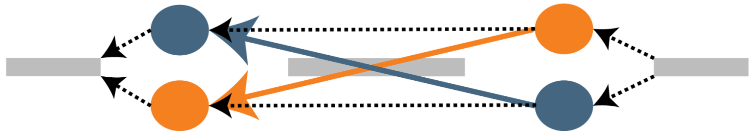
2. **Mince** primary contigs: separate **haplotig pairs** and **collapsed haplotypes**



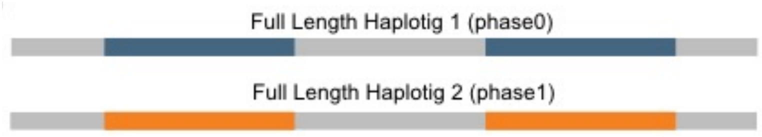
3. **Map** paired HiC reads to minced contigs and generate normalized **contact matrix**



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

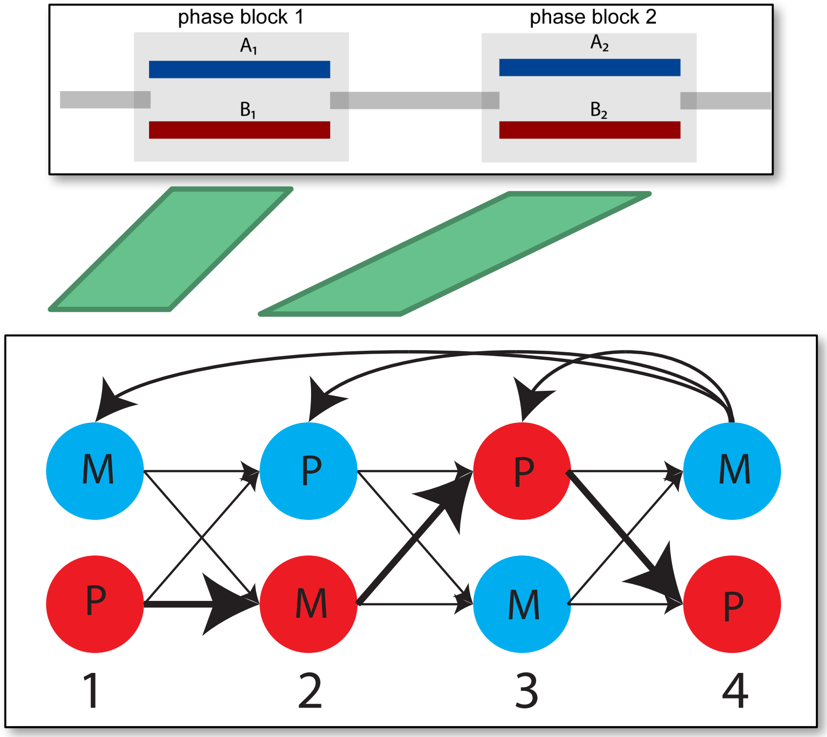


Output: phased full length haplotigs

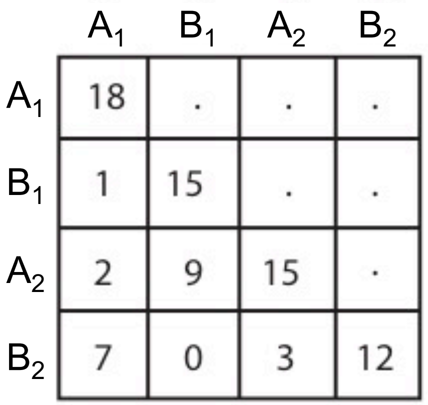

Full Length Haplotig 1 (phase0)
Full Length Haplotig 2 (phase1)

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

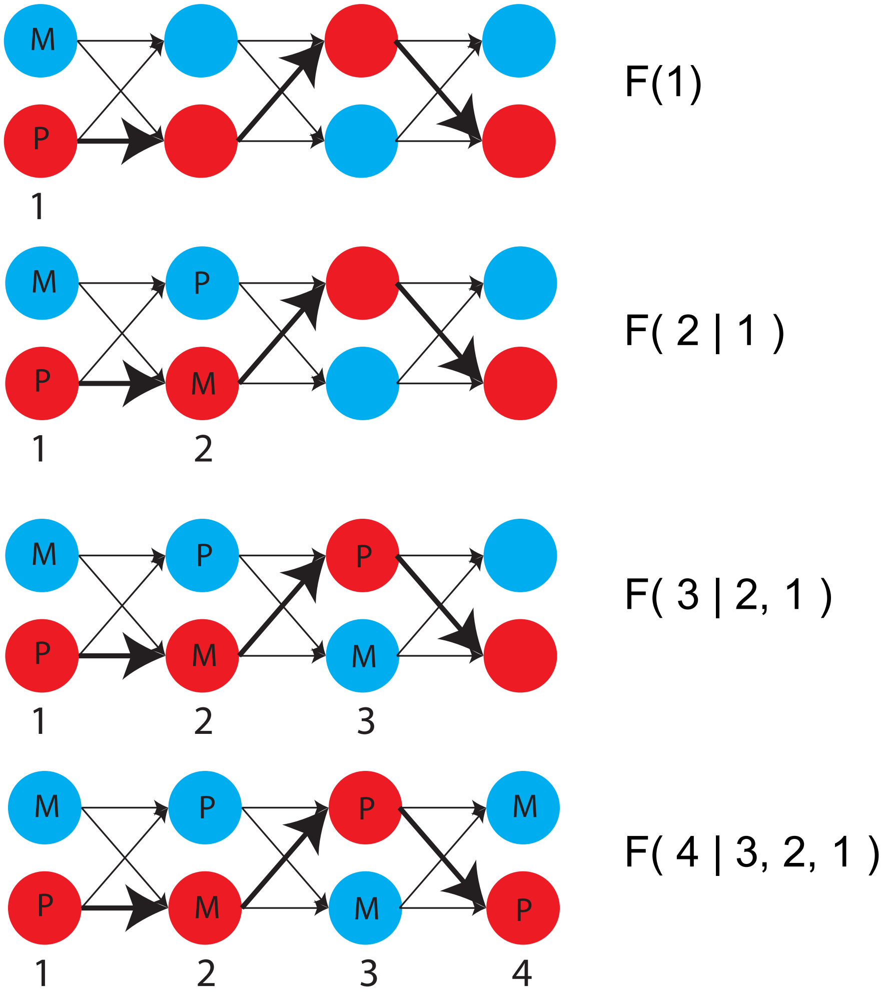


	A_1	B_1	A_2	B_2
A_1	18	.	.	.
B_1	1	15	.	.
A_2	2	9	15	.
B_2	7	0	3	12

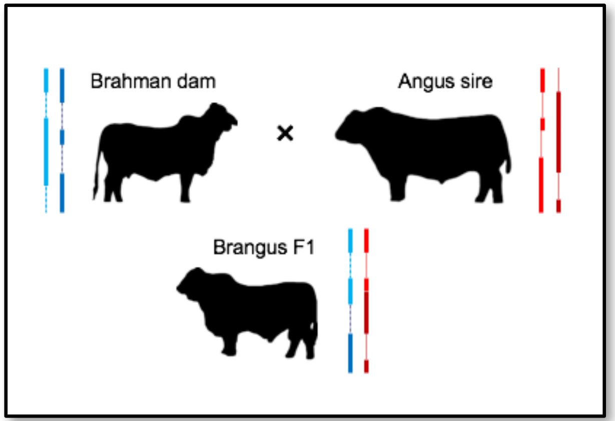
- Output
  - Majority phase assignment configuration for haplotigs along primary contig

PHASING ALGORITHM

- Algorithm sweeps along phase-blocks of primary contig
- Phase assignment is conditioned on ***all*** those before and Hi-C links
- Process repeated for > 10^7 iterations (burn-in = $5x10^6$)
- Complexity ~ $O(n^2)$
  - n phase blocks



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

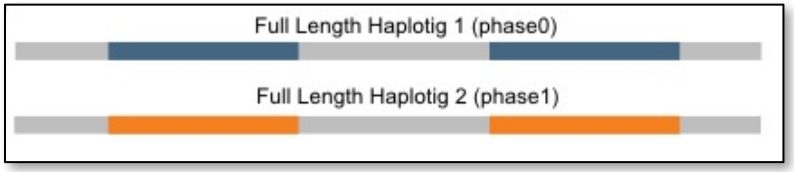
CONTIGS	NUMBER	LENGTH	N50
ANGUS DAM	1747	2.57 Gb	26.7 Mb
BRAHMAN SIRE	1040	2.68 Gb	23.3 Mb

Parental SNP Calls

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

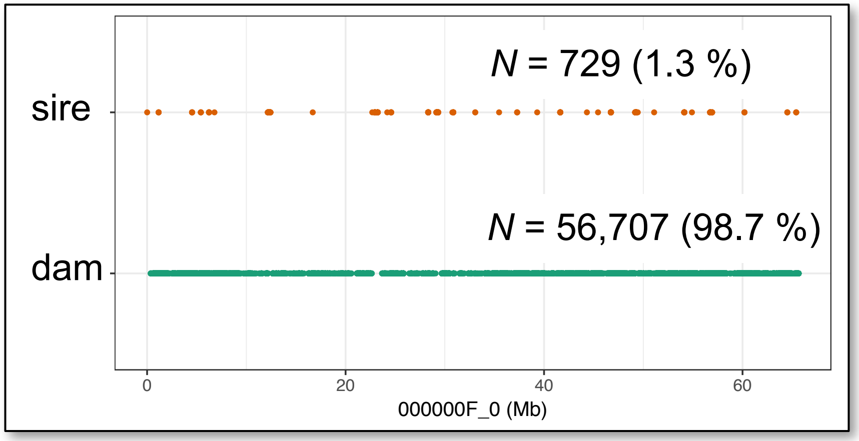
Koren et al. 2018

PHASE ASSIGNMENT ACCURACY: PARENTAL SNV CALLS

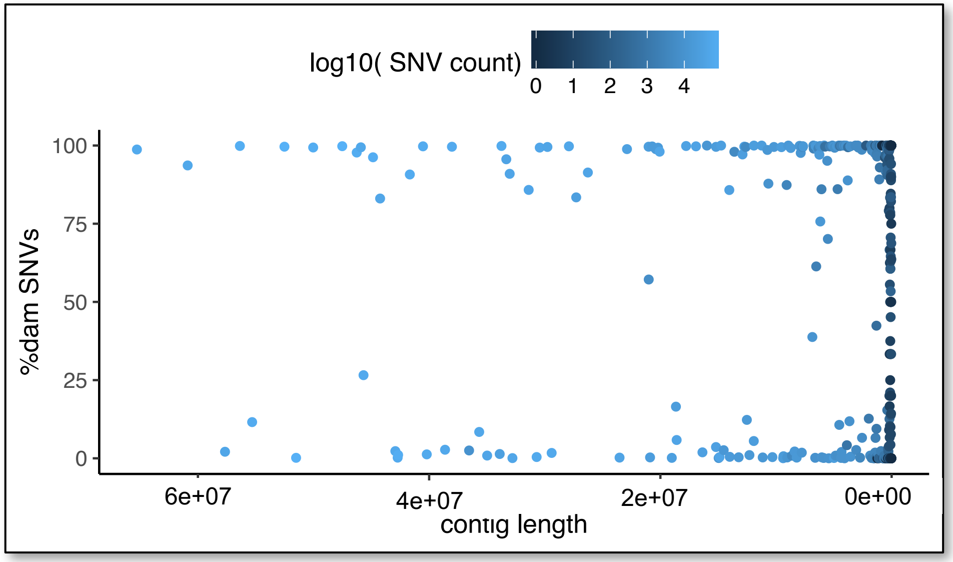

Full Length Haplotig 1 (phase0)
Full Length Haplotig 2 (phase1)

- 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

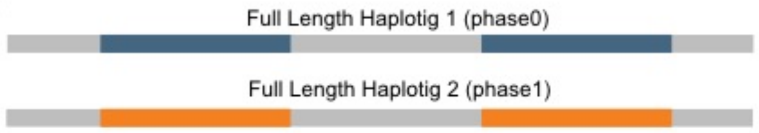


OVERALL PERFORMANCE

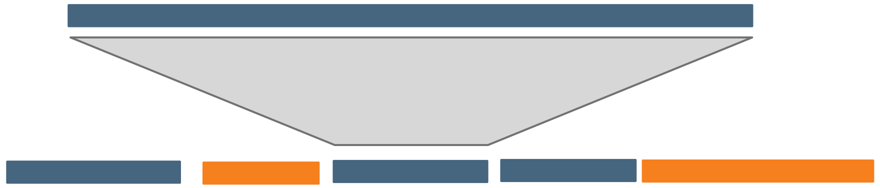
	SNV Count	Correct
dam	2,031,334	97.4 %
sire	1,464,748	95.8 %
total	3,496,082	**96.7 %**

PHASING CHROMOSOME-SCALE SCAFFOLDS

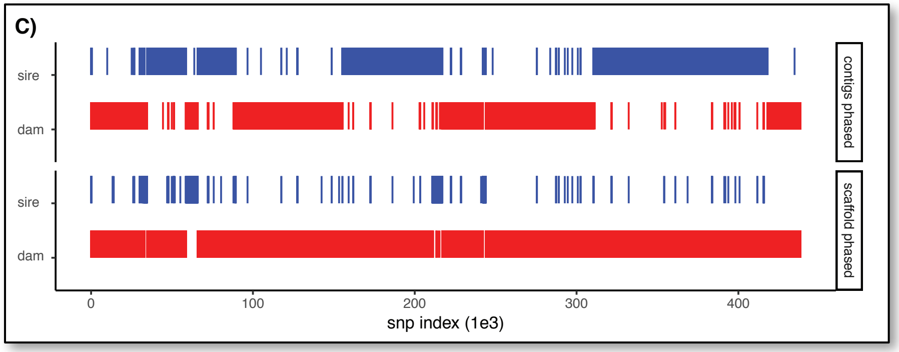
- Scaffold phase0 full-length haplotig with Proximo (Phase Genomics)
- Scaffolds are chromosome-scale
- We know:
  - order of contigs along scaffold
  - pairing of phase 0 and phase 1
- FALCON-Phase Scaffolds


Full Length Haplotig 1 (phase0)
Full Length Haplotig 2 (phase1)

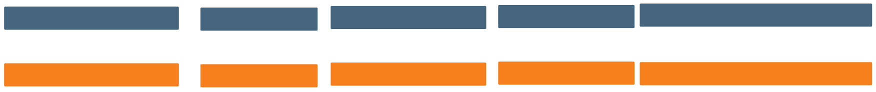
Scaffold Phase 0 Haplotigs



PARENTAL SNVS AFTER SCAFFOLD PHASING



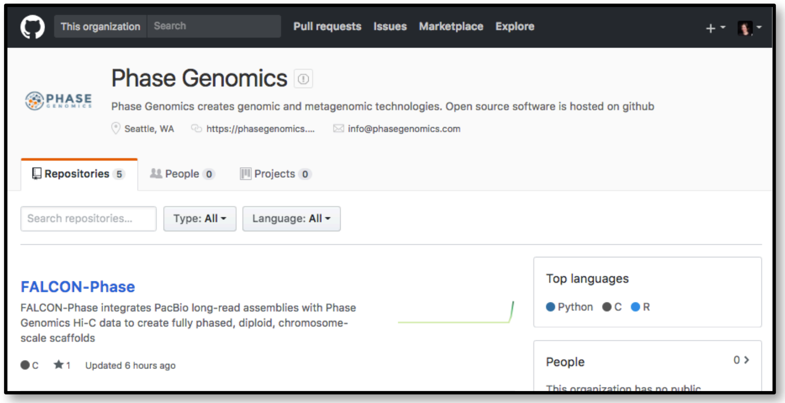
FALCON-Phase Scaffolds



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 gapless, phased, chromosome-scale diploid assembly
- More Info:



Phase Genomics

Phase Genomics creates genomic and metagenomic technologies. Open source software is hosted on github

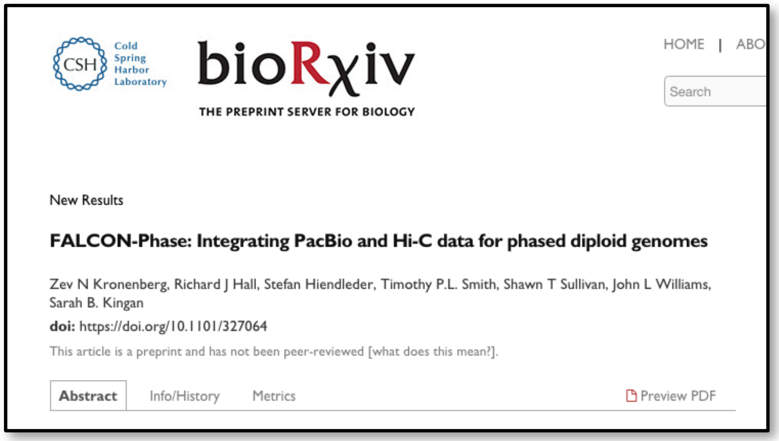
Seattle, WA https://phasegenomics.... info@phasegenomics.com

FALCON-Phase

FALCON-Phase integrates PacBio long-read assemblies with Phase Genomics Hi-C data to create fully phased, diploid, chromosome-scale scaffolds

● C ★ 1 Updated 6 hours ago

This organization has no public


CSH Cold Spring Harbor Laboratory
bioRχiv
THE PREPRINT SERVER FOR BIOLOGY
HOME | ABO
Search
Abstract
Info/History
Metrics
Preview PDF

New Results

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

Zev N Kronenberg, Richard J Hall, Stefan Hiendleder, Timothy P.L. Smith, Shawn T Sullivan, John L Williams, Sarah B. Kingan

doi: https://doi.org/10.1101/327064

This article is a preprint and has not been peer-reviewed [what does this mean?].

ACKNOWLEDGEMENTS

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

PACBIO®

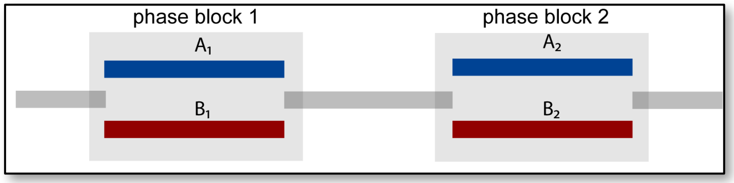
Greg Concepcion
Jonas Korlach
Billy Rowell
Ivan Sovic
Liz Tseng
Michelle Vierra

PHASE GENOMICS

Ivan Liachko
Kaylee Mueller
Max Press
Andrew Wiser

Jason Chin
Mark Chaisson
Luke Harmon
Ryan Layer

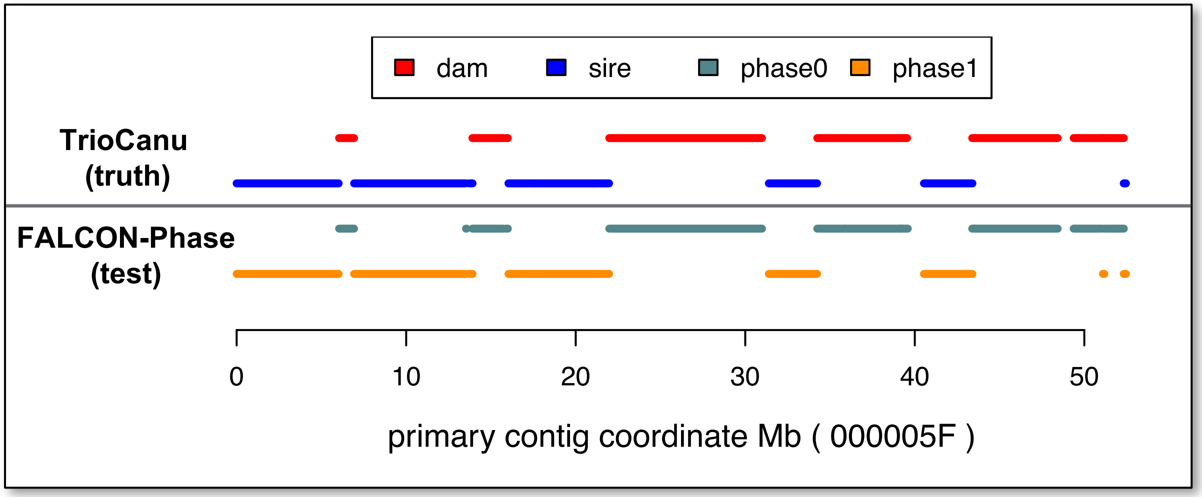
PHASE ASSIGNMENT ACCURACY: PARENTAL ASSIGNMENT



- 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

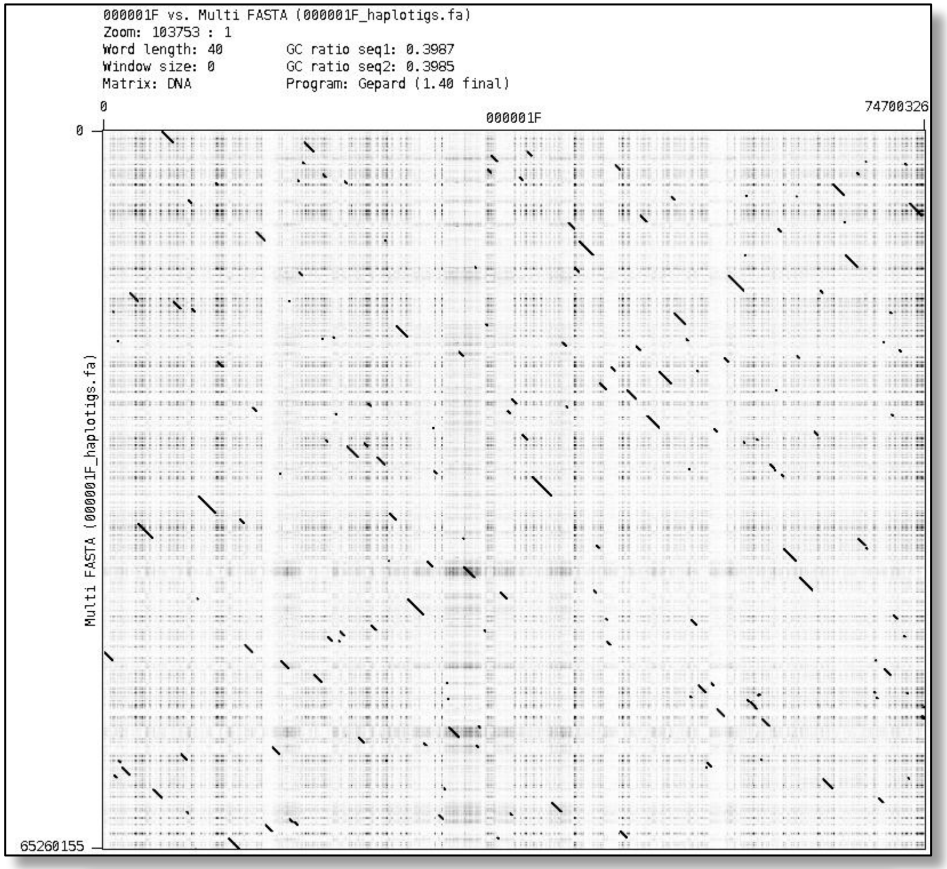
RESULTS FOR PRIMARY CONTIG 000005F



overall accuracy: 96.72%

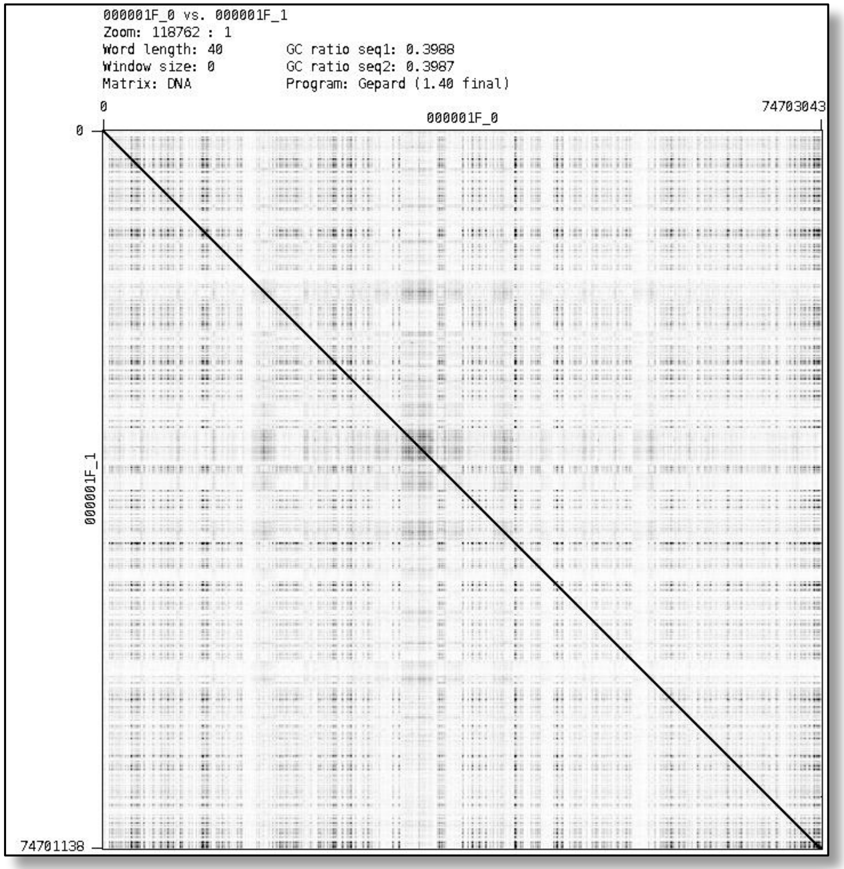
APPLICATION TO HUMAN ASSEMBLY

FALCON-Unzip



- X-axis – 75Mb contig
- Y-axis – Hundreds of haplotigs spanning 87% of the primary contig

FALCON-Phase



- X-axis – 75Mb contig – Phase 0
- Y-axis – 75Mb contig – Phase 1