The pleasures and perils of assembling insect genomes

Adam M. Phillippy
Head, Genome Informatics Section, NHGRI

NIH National Human Genome Research Institute
Advancing human health through genomics research

@aphillippy

The assembly problem

Genome assembly with short reads

Bigger pieces are better

“It”	>1,000	SSR
“It was”	320	TE
“It was the best”	2	SegDup
“It was the best of times”	1	Unique
“With his hands in his pockets”	3	Meta

Genome assembly with long reads

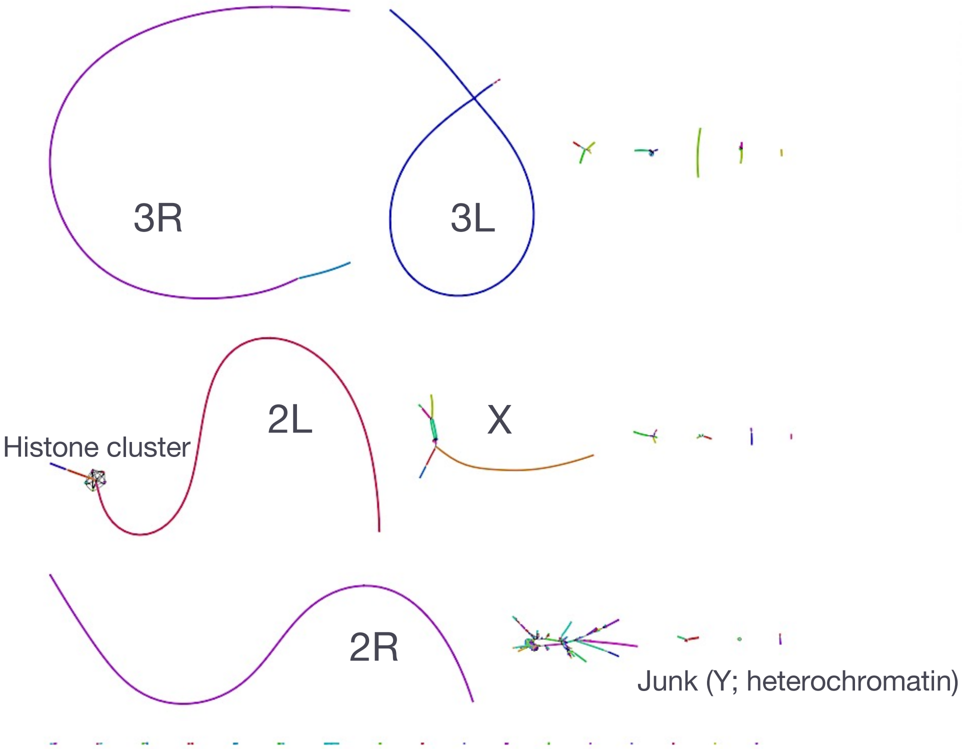
Long reads to the rescue?

Can you Canu?

- Long read data is noisy
  - Base errors
  - Chimeric reads
  - *Solution:* read clustering, correction, and trimming
- Overlaps are long, and graph is big
  - All-pairs alignment is slow
  - Full graph is a giant tangle (due to repeats)
  - *Solution:* MinHash “best” overlap graph
- *D. melanogaster* results
  - Celera Assembler v8: **630,000** CPU hours, 15 Mbp NG50
  - Canu v1: **500** CPU hours, 21 Mbp NG50

Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation.
Koren et al. *Genome Research* (2017)

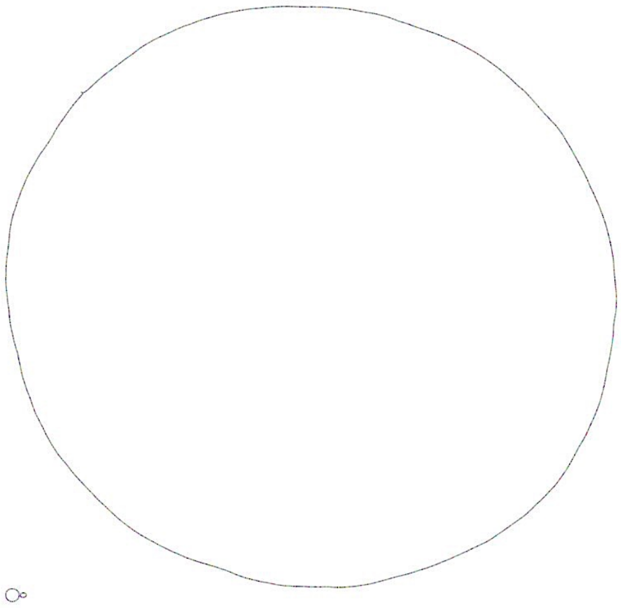
Complete *D. melanogaster* assembly



- **Assembling large genomes with single-molecule sequencing and locality-sensitive hashing.** Berlin et al. *Nature Biotechnology* (2015)

Can long reads solve assembly?

- 2012: Bacteria (10^6 bp)
- 2014: Yeast (10^7 bp)
- 2014: Drosophila (10^8 bp)
- ????: Human (10^9 bp)

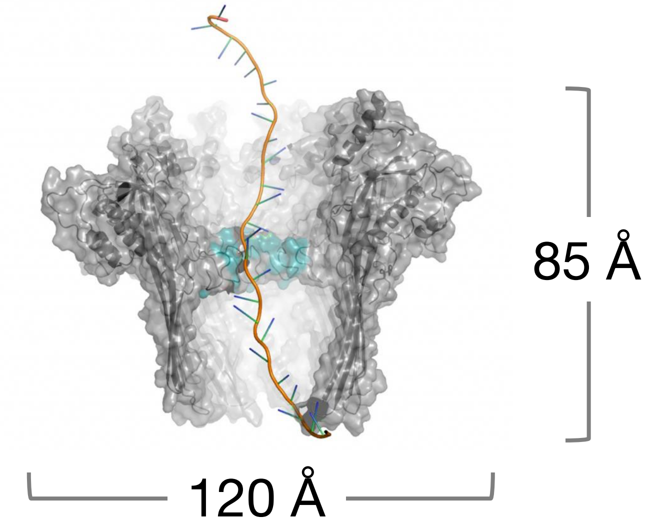


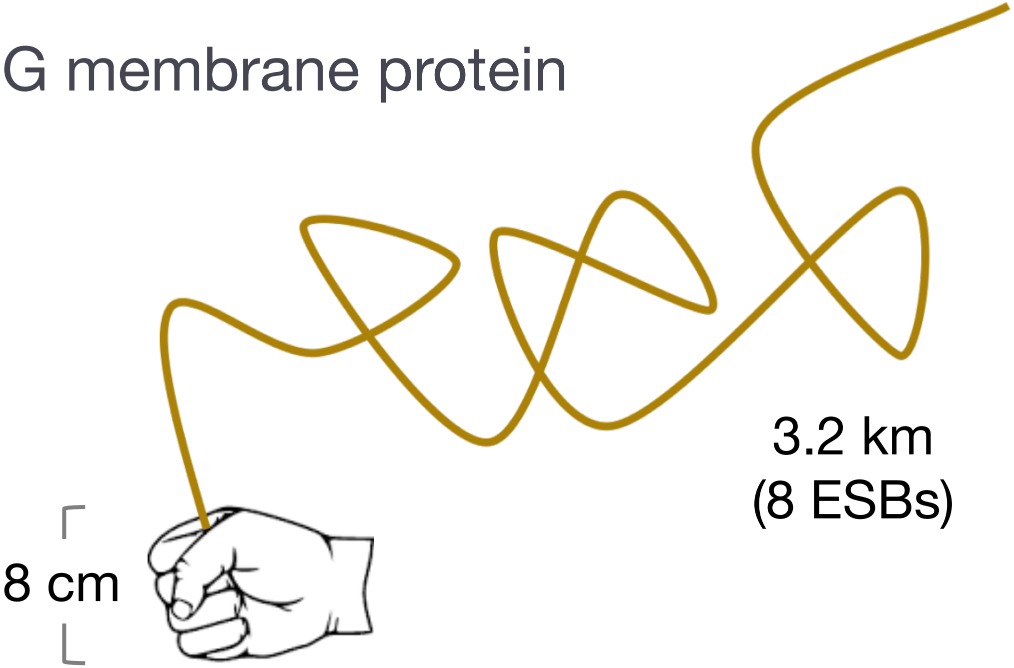
- **New advances in sequence assembly.** Phillippy. *Genome Research* (2017)

Ultra-long reads

Nanopore dimensions

- ONT R9 pore
  - Engineered *E. coli* CsgG membrane protein



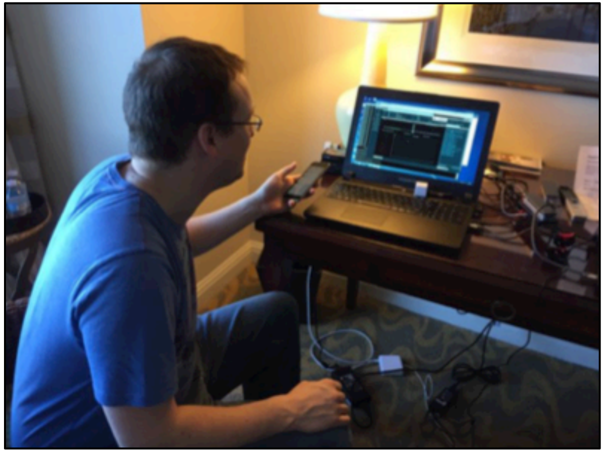


- *Assuming 3.4 Å per bp, 1 Mbp = 3,400,000 Å = 40,000x height of the pore

Nanopore sequencing of human genomes

- GM12878 Utah/Ceph
  - 35x MinION R9.4
  - 11 kb N50 read len
  - 3 Mbp N50 contig len

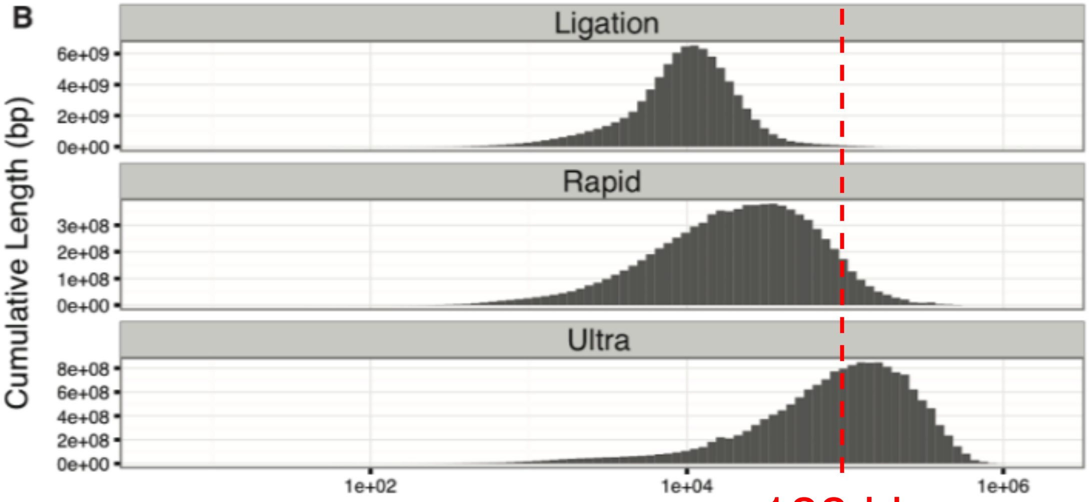
- Clive Brown, ONT
  - 60x MinION R9.4
  - 19 kb N50 read len
  - 30 Mbp N50 contig len



- **Nanopore sequencing and assembly of a human genome with ultra-long reads.**
  Jain et al. *Nature Biotechnology* (2018)

Ultra-long reads

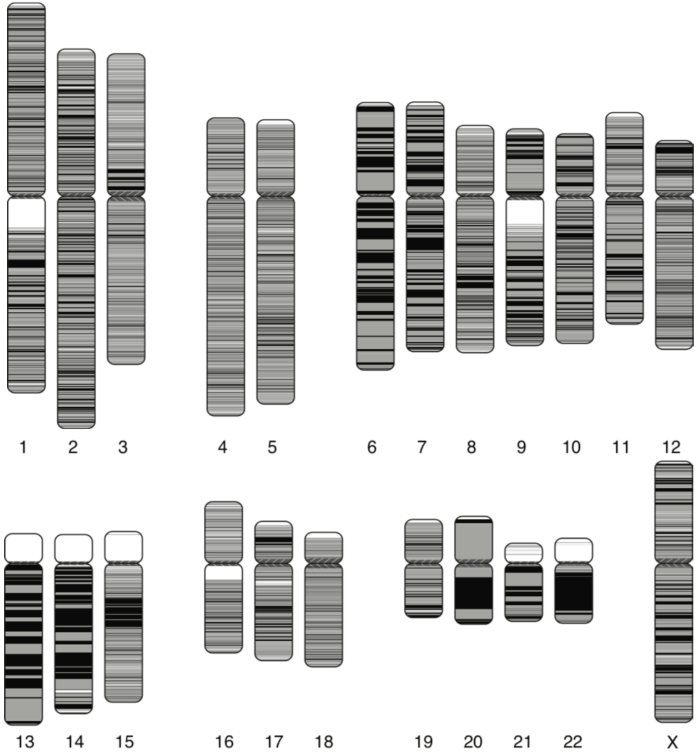
- 100 kb read N50, max close to 1 Mb!
  - Sambrook and Russel phenol-chloroform prep
  - Minimal pipetting, high input to rapid (transposase) kit



100 kb

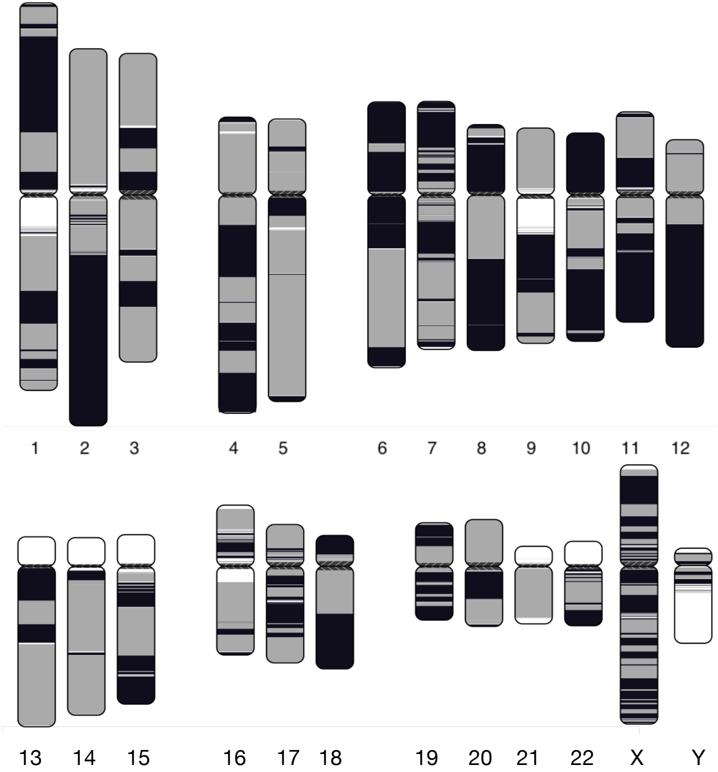
- http://lab.loman.net/2017/03/09/ultrareads-for-nanopore/ (Josh Quick & Nick Loman, U. Birmingham)

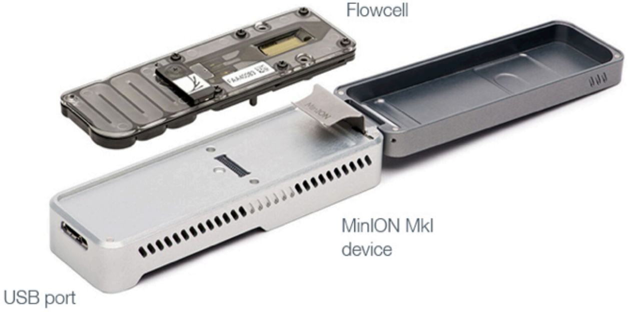
Human genome, 2001



- ref28 / hg10 : N50 0.5 Mbp

Cliveome, 2017





- Cliveome 60x : NG50 29.5 Mbp

Not so fast…

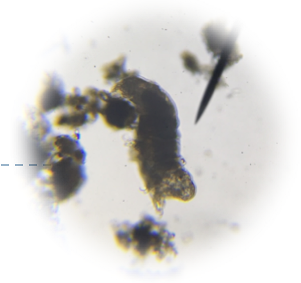
Clive Brown is not an insect

The perils

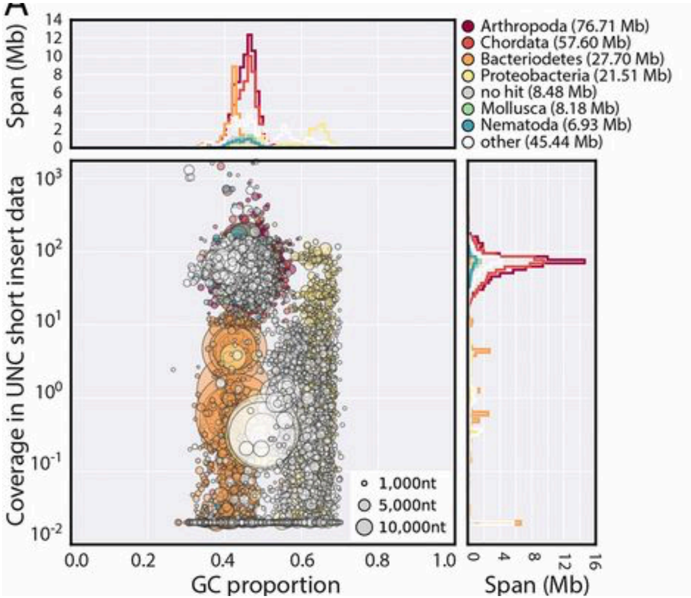
- Tiny bugs
  - Can’t sequence a single individual
  - Contamination risk
- Repeats
  - Every genome is different
- Diversity
  - A pot of bugs is a metagenome

Contamination

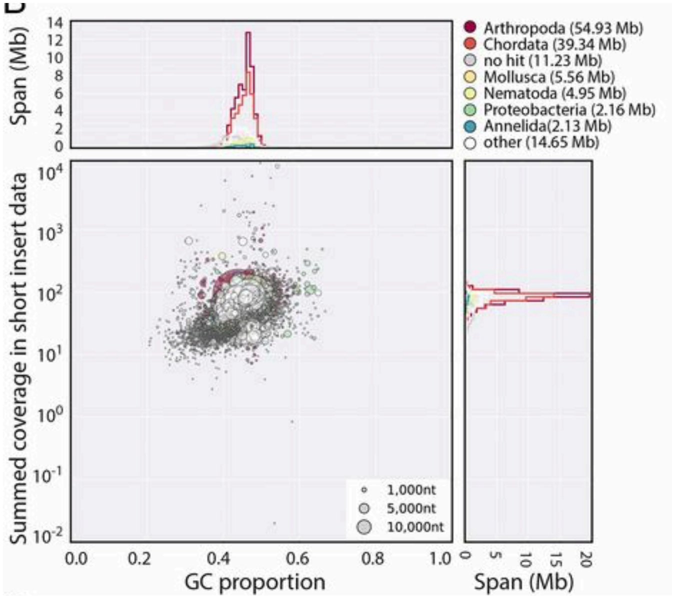
- “Tardigate”



UNC



Edinburgh



- **No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*.** Koutsovoulos et al. *PNAS* (2016)

Repeats

- Mealworm beetle
  - Brenda Oppert, USDA
  - Why isn’t Canu finishing?

- Runaway satellite
  - 60% of genome is a 142 nt repeat
  - Required adjusting Canu parameters for repeat weighting/screening

- **Distribution and sequence homogeneity of an abundant satellite DNA in the beetle, *Tenebrio molitor.***
Davis and Wyatt, *Nucleic Acids Research* (1989)

Diversity

- Heterozygous diploids
  - Some bugs hard to inbreed
  - Large populations, large diversity

(c) Alex Wild

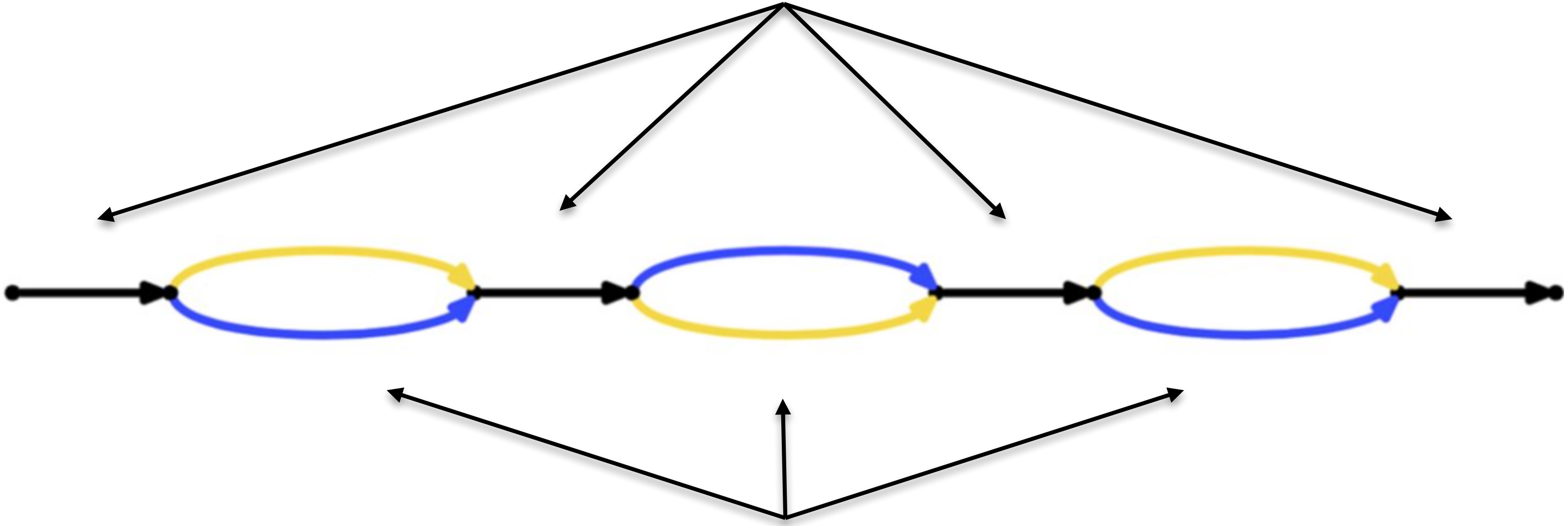
- Grind up and sequence a pot of bugs
  - 100+ mosquitos
  - ≥2 alleles at each locus?
  - Polymorphic inversions & integrations?

- **Improved *Aedes aegypti* mosquito reference genome assembly enables biological discovery and vector control.** Matthews et al. *bioRxiv* (2017)

Dealing with heterozygosity

Diploid assembly graph

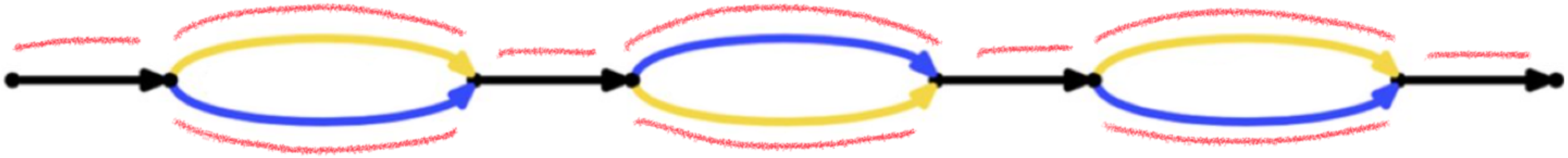
Homozygous alleles (collapsed)



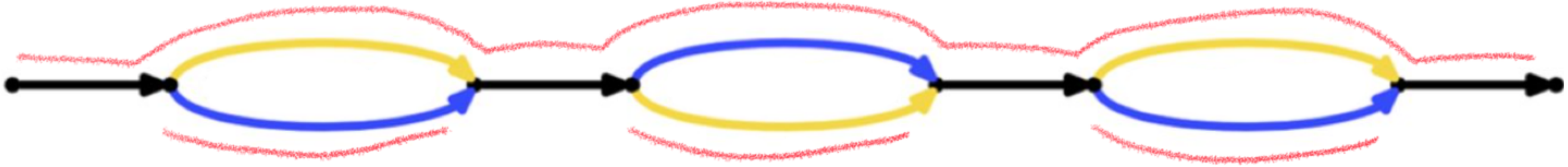
Heterozygous alleles (bubbles)

- Weisenfeld 2017

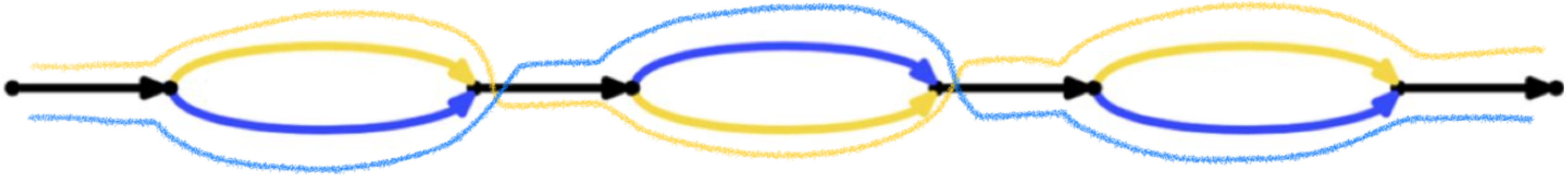
Haplotigs



Pseudo-haplotype + alts

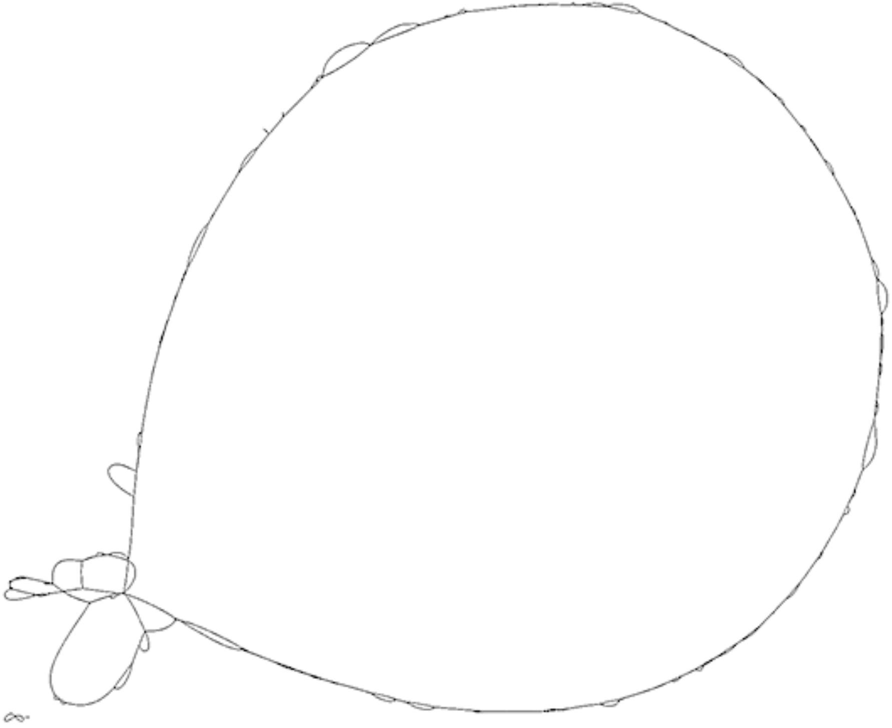


Complete haplotypes



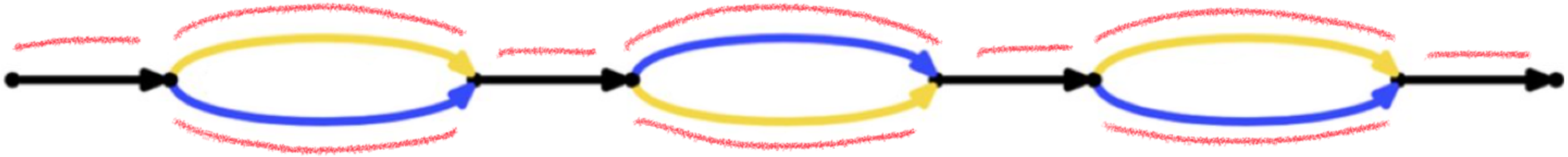
Reality not so simple

- Two *E. coli* strains
- Imagine now…
  - *N* alleles mixed at different abundances
  - Plus, long high-copy repeat families



Aedes aegypti example

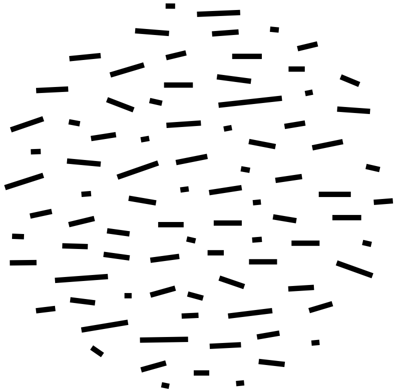
- Genome size ~1.3 Gbp
- Assembly size
  - FALCON-Unzip primary: 1.7 Gbp
  - FALCON-Unzip primary + alts: 2.0 Gbp
  - Canu: 2.8 Gbp
- "Deduplicated" with Hi-C and contig alignments



- **Improved *Aedes aegypti* mosquito reference genome assembly enables biological discovery and vector control.** Matthews et al. *bioRxiv* (2017)

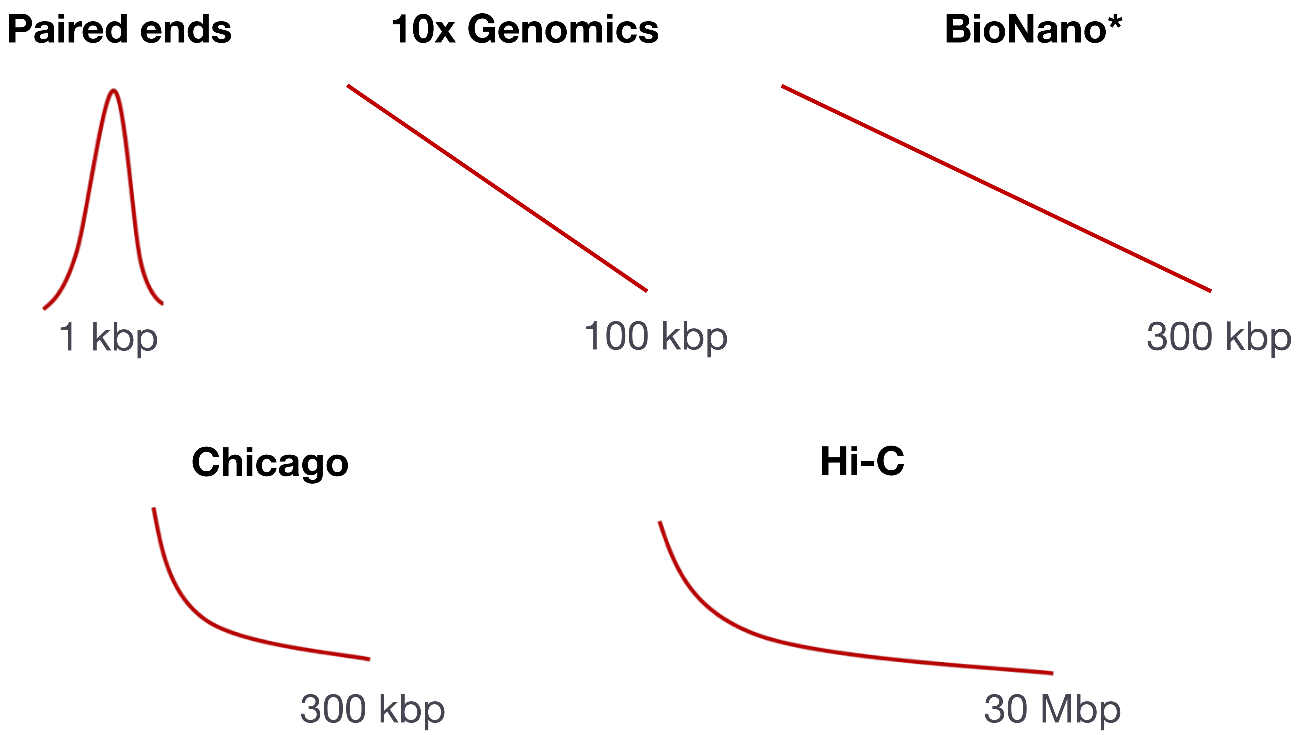
De novo reference genomes

Contigs ≠ Chromosomes



≠

Scaffolding options



Paired ends

1 kbp

10x Genomics

100 kbp

BioNano*

300 kbp

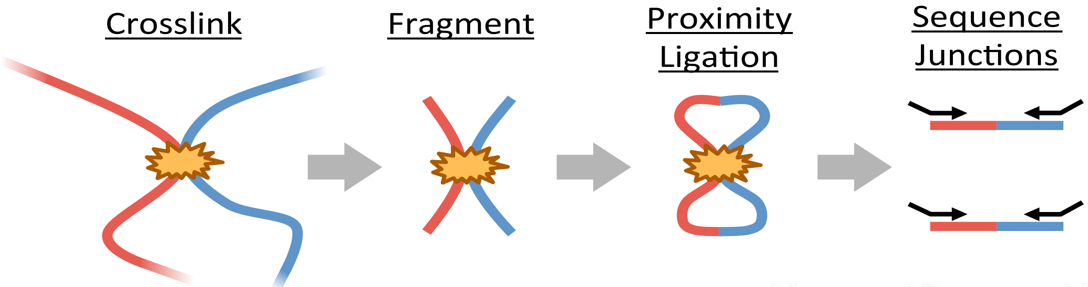
Chicago

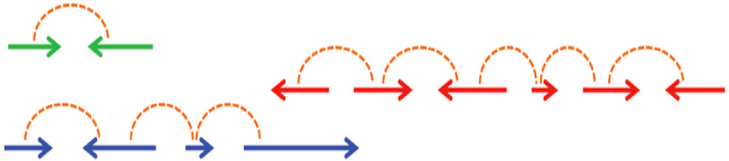
300 kbp

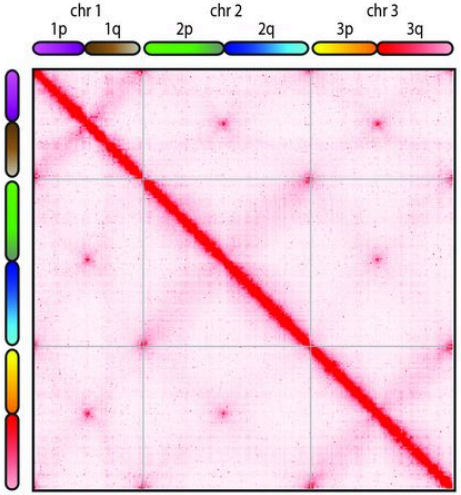
Hi-C

30 Mbp

Hi-C chromatin conformation capture







Rabl configuration

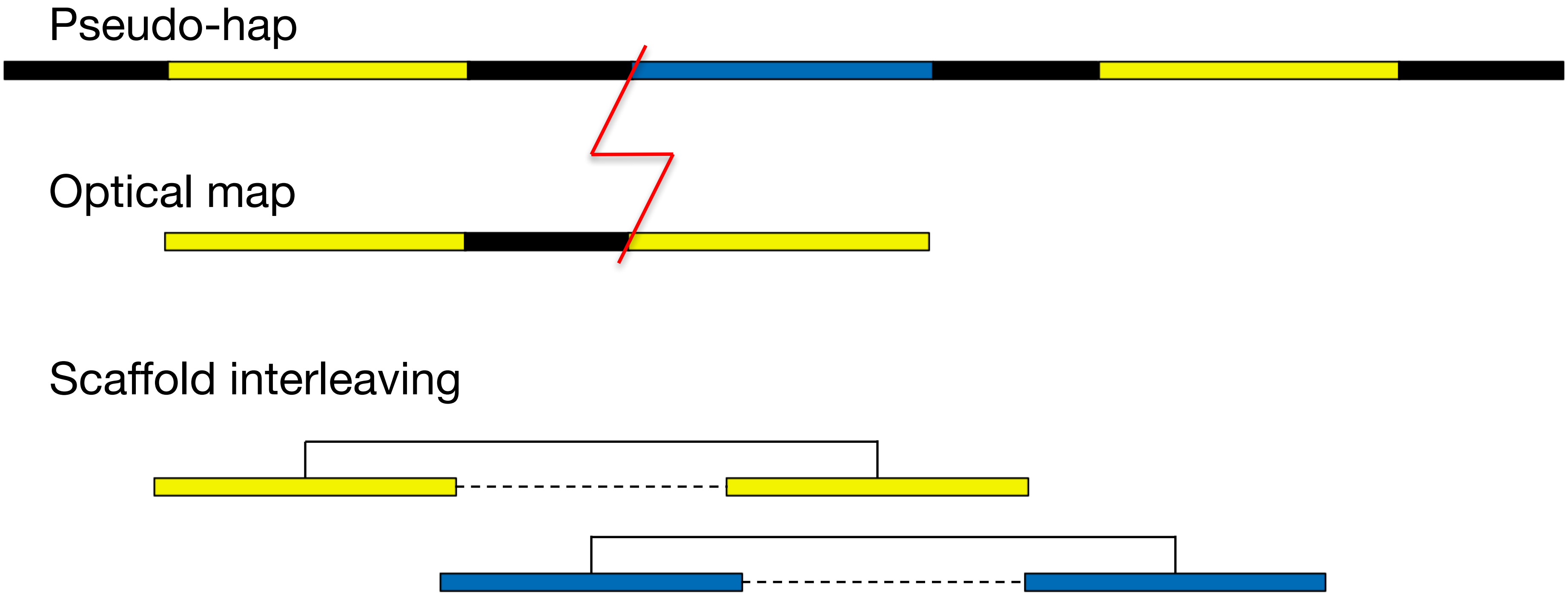
- Fig credit: Phase Genomics (top/left), Dudchenko et al. *Science* (2017) (bottom right)

VGP ordinal sequencing recipe

- Observations
  - PacBio : contigs
  - 10XG : scaffolds, phasing, and polishing
  - BioNano : scaffolds and validation
  - Hi-C: chromosome-scale scaffolds and phasing
- What’s essential for reference genomes?
  - Start with long reads, add others as needed
  - Thorough validation
  - DO NOT ignore haplotype variation… (Korlach & Jarvis 2017)

- **Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome.** Bickhart et al. *Nature Genetics* (2017)

Scaffolding pseudo haplotypes is not fun

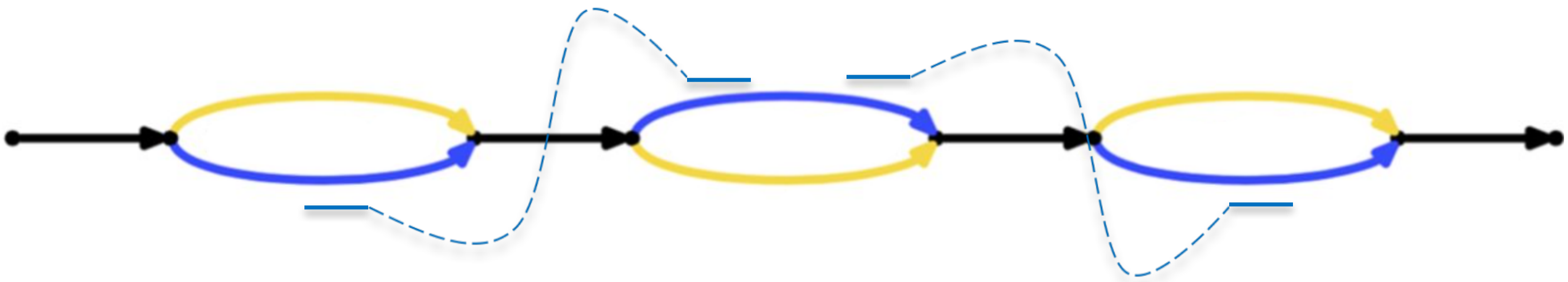


Pseudo-hap

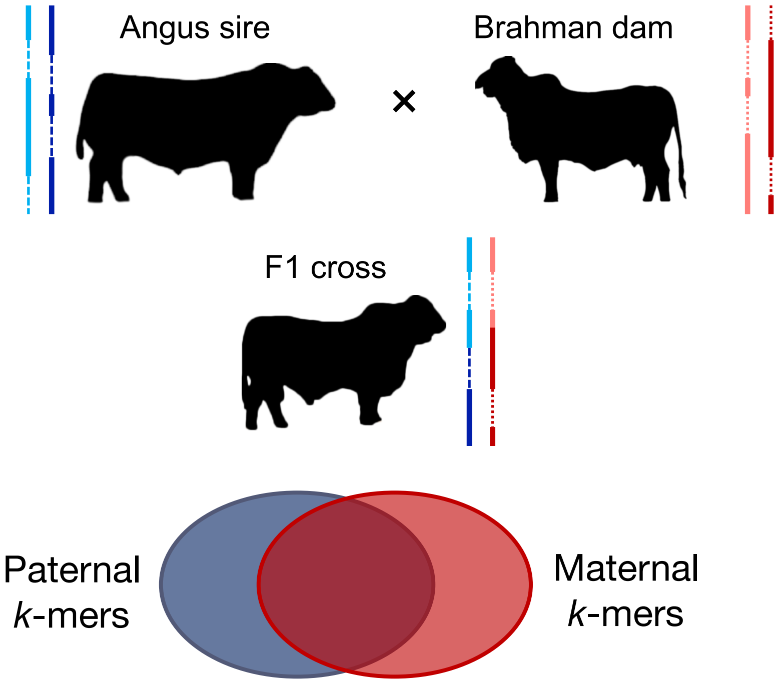
Optical map

Scaffold interleaving

Hard solution: scaffold the graph



Easy solution: trio binning



Angus sire

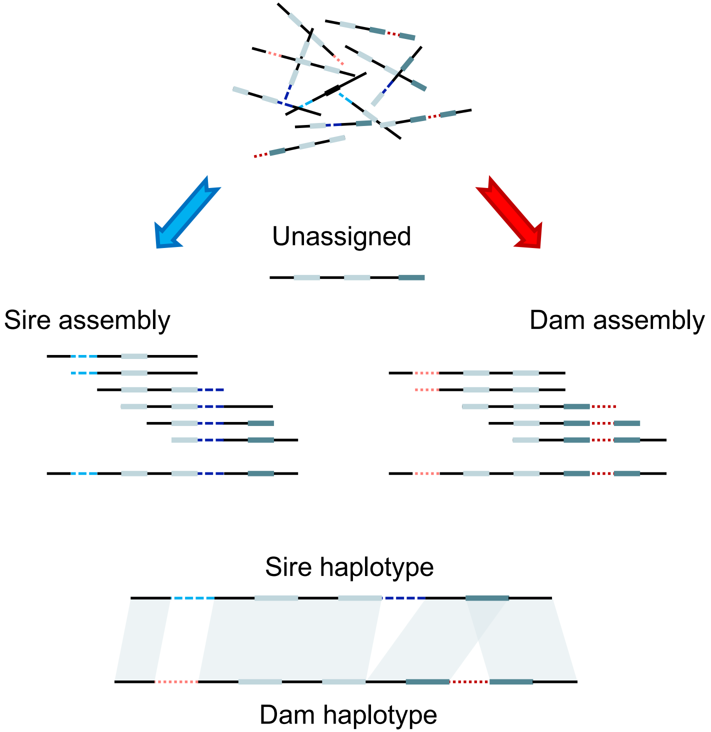
×

Brahman dam

F1 cross

Paternal *k*-mers

Maternal *k*-mers



Unassigned

Sire assembly

Dam assembly

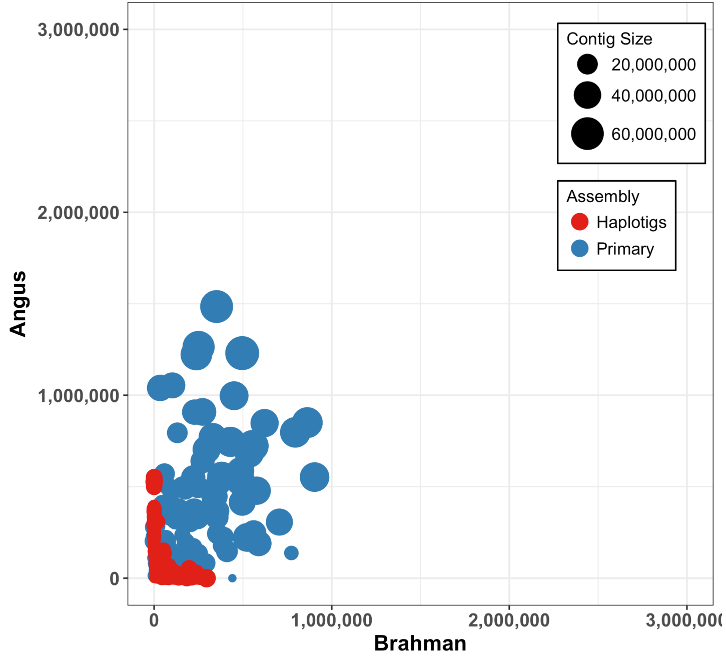
Sire haplotype

Dam haplotype

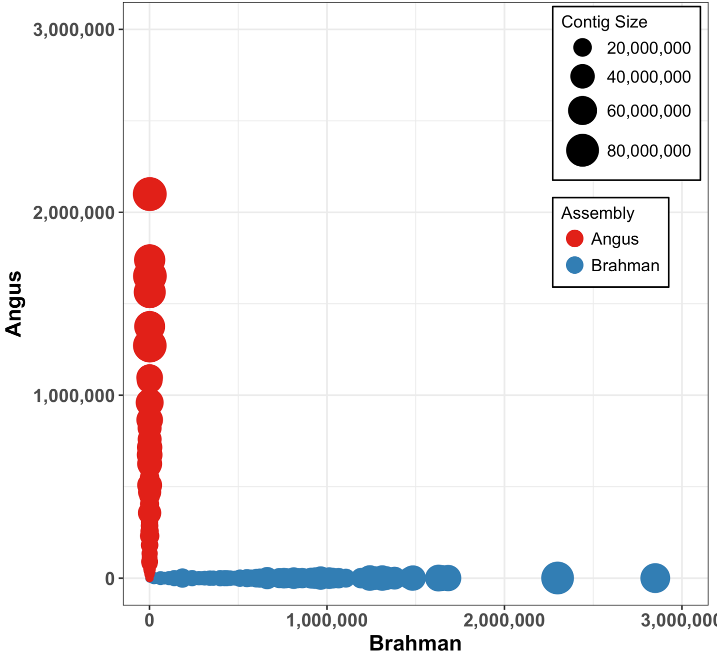
- Koren, Rhie, et al. (in preparation)

Pseudo vs complete haplotypes

- FALCON-Unzip

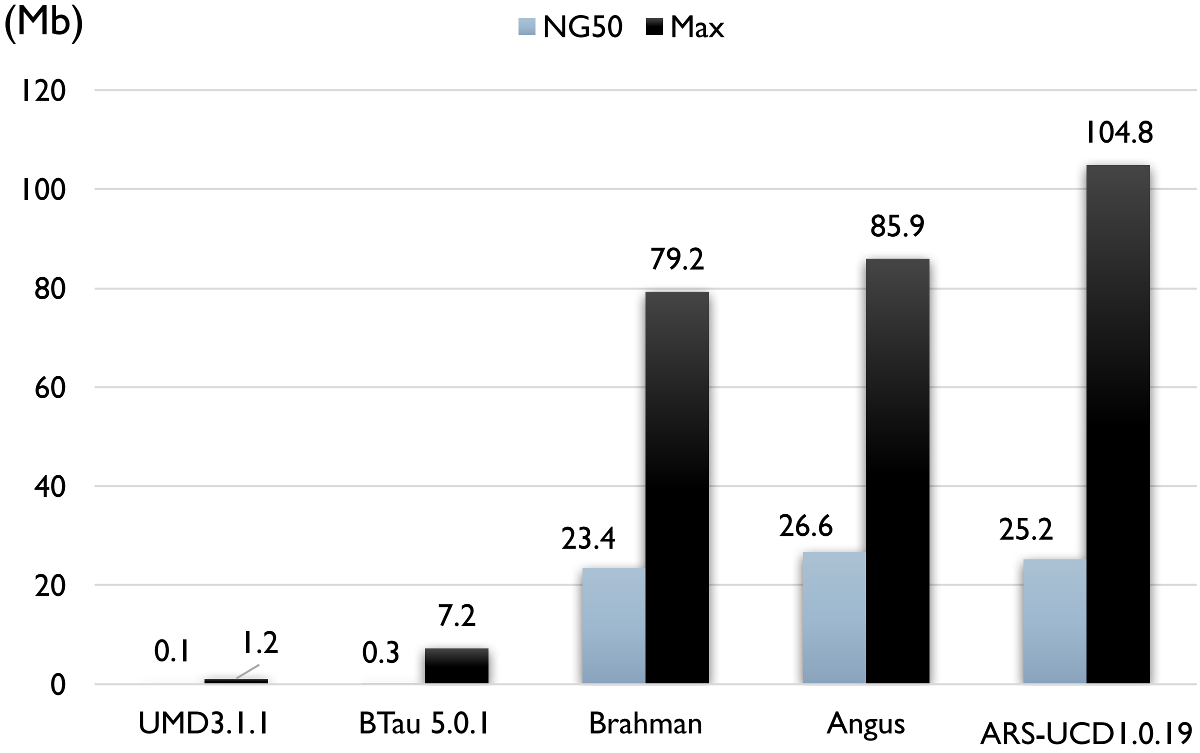


- TrioCanu



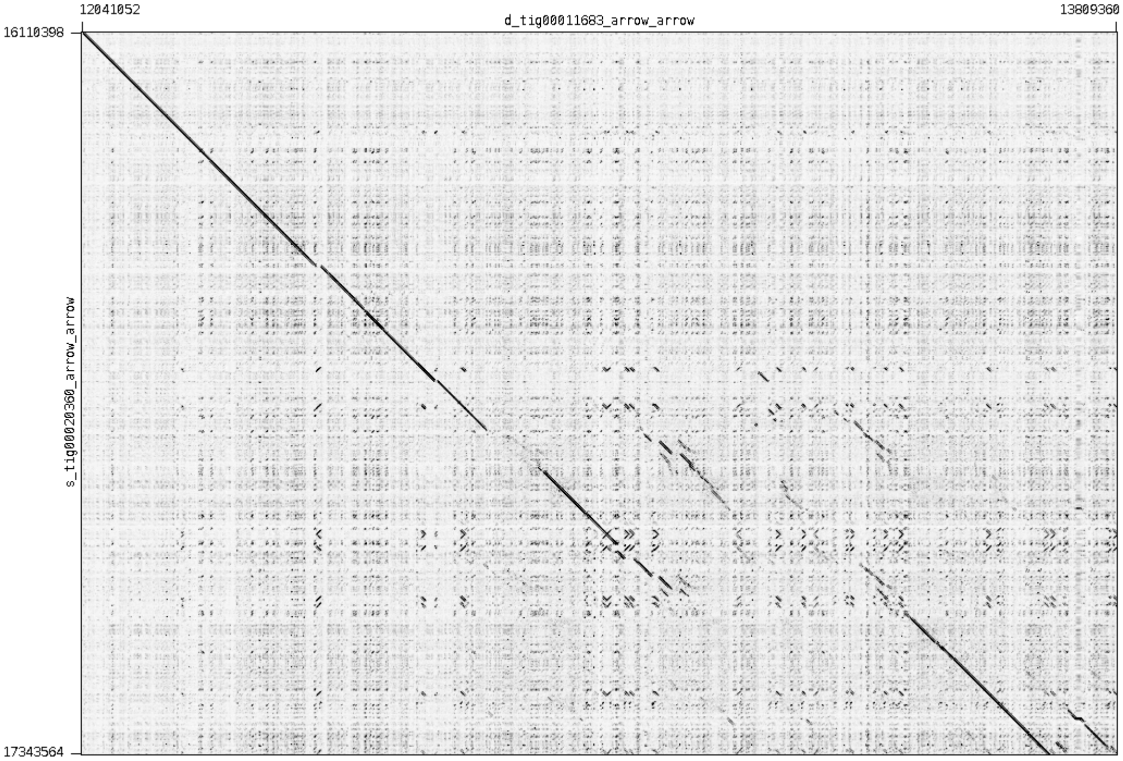
- Koren, Rhie, et al. (in preparation)

Excellent continuity of both haplotypes



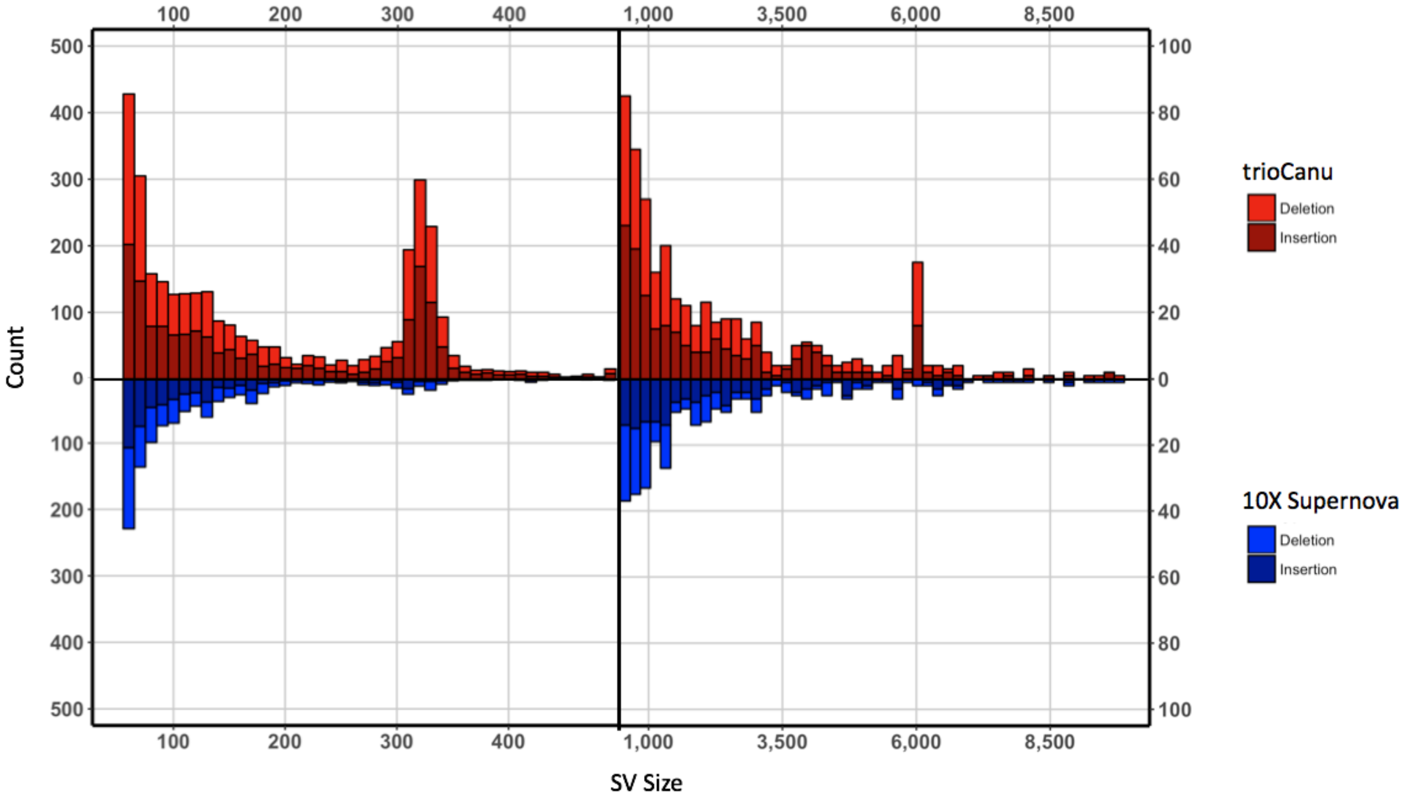
- Koren, Rhie, et al. (in preparation)

Complex haplotype variation



- Y-axis: Angus paternal haplotype, X-axis: Brahman maternal haplotype (MHC class II)

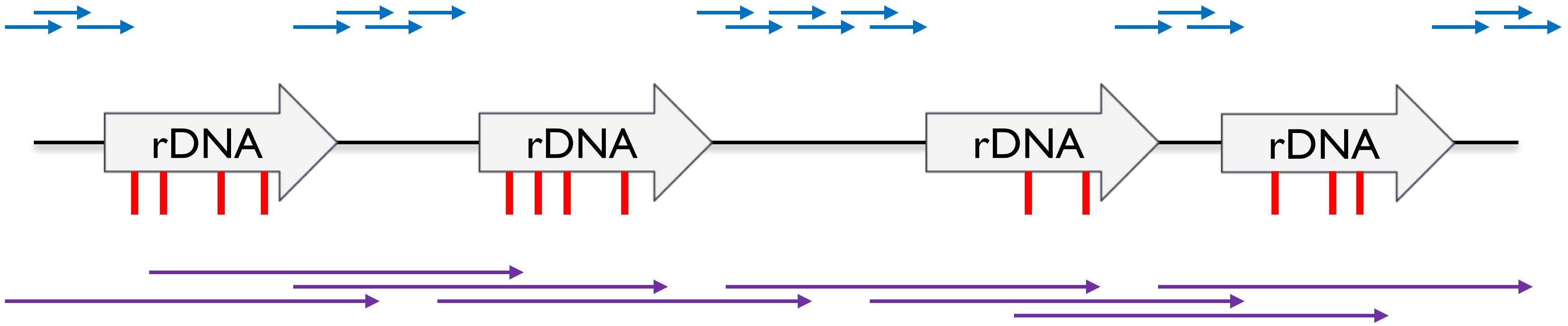
Short reads miss large variation



- Corrected phase block NG50: TrioCanu: 12.92 Mbp, 10x: 4.26 Mbp

Long read polishing is essential

- Cannot map short reads to repeats and errors
  - Therefore, cannot polish/assemble repeats with short reads
  - Long read assemblies more accurate in repeats
  - Beware of haplotype variation



- In some regions, short-read polishing can actually harm the assembly

All assemblies are wrong, some are useful

Tools

- Long-read assembly
  - FALCON-Unzip, **Canu**, Flye, wtdbg
- Scaffolding
  - **Salsa**, 3D-DNA, HiRise*, Scaff10x, ARCS, BioNano
- Polishing
  - Quiver/Arrow, Nanopolish*, FreeBayes, Pilon, PBJelly*
- QC & Validation
  - BioNano, BUSCO, **Mash**, BlobTools, Juicebox
  - GenomeScope, KAT, Assemblytics, IGV

- Tools in bold from the Phillippy lab

Summary

- *Haploid* assembly is solved by long reads
  - But most sequencing samples are not haploid
- Reads will get longer and cheaper
  - Nanopore promising, but behind in consensus quality
- Remaining assembly challenges
  - **Complete haplotype recovery**
  - Diploids, polyploids, and populations
  - Heterochromatin and large duplications
  - New representations and tools

Acknowledgements

genomeinformatics.github.io

- Sergey Koren
- Brian Walenz
- Alexander Dilthey
- Arang Rhie
- Jay Ghurye

AD

JG

CJ

SK

BO

AP

AR

AS

BW