Towards haplotype-resolved assemblies with Canu

Sergey Koren
Staff Scientist, Genome Informatics Section, NHGRI

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

@sergekoren

TrioBinning: Trio-based assembly

How I stopped worrying and learned to love the F1

Sergey Koren
Staff Scientist, Genome Informatics Section, NHGRI

National Human Genome Research Institute
Advancing human health through genomics research

@sergekoren

Variant Terminology

- Megabubbles

  - Variants output separately
  - Phased but short
  - Homozygous regions are single-copy
  - Falcon associated “haplotigs” report only one half of bubble

- Pseudohaplotypes

  - Random path through variants
  - Not phased but long
  - Falcon primary contigs are an example

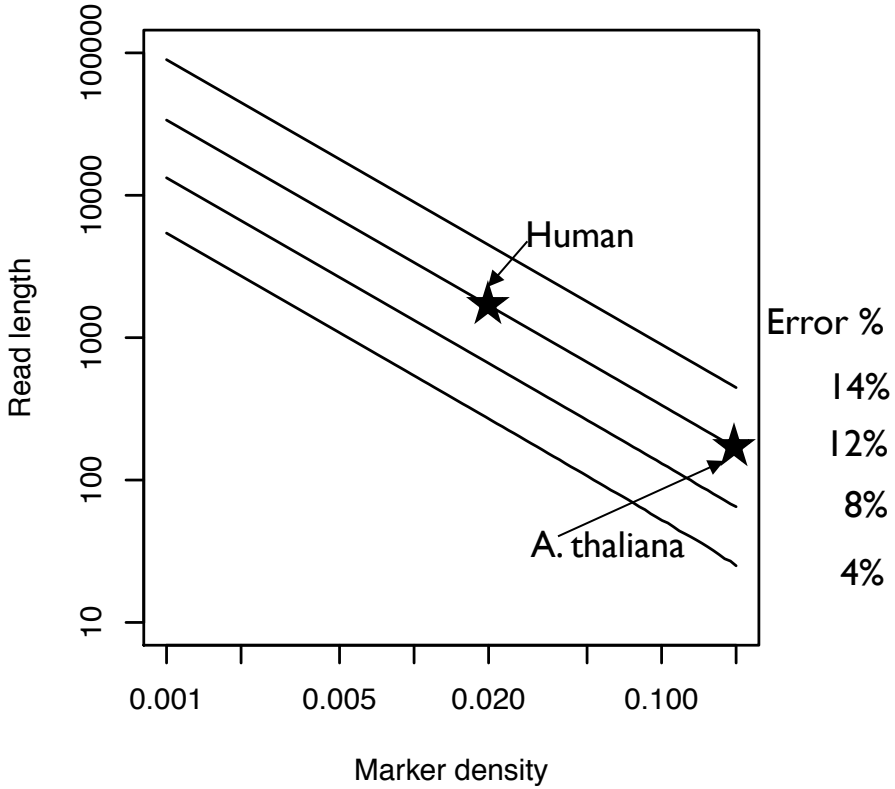
- Haplotigs

  - Consistent path through each haplotype
  - Homozygous regions represented twice
  - Each set of haplotigs is a complete representation of a single haplotype

- https://support.10xgenomics.com/de-novo-assembly/software/pipelines/latest/output/generating

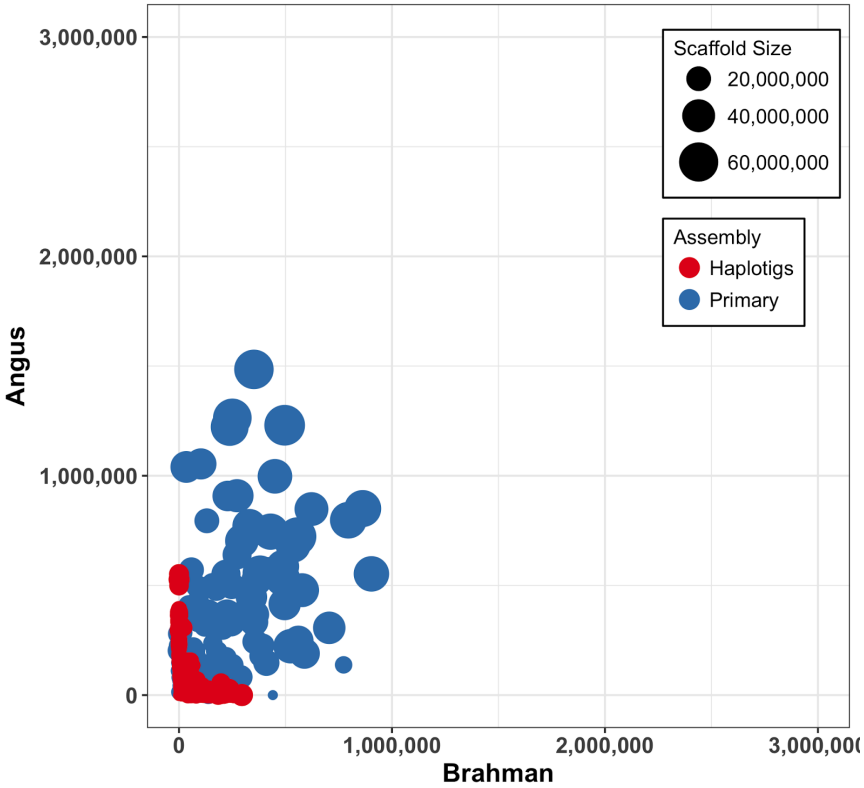
Classification with sequencing error

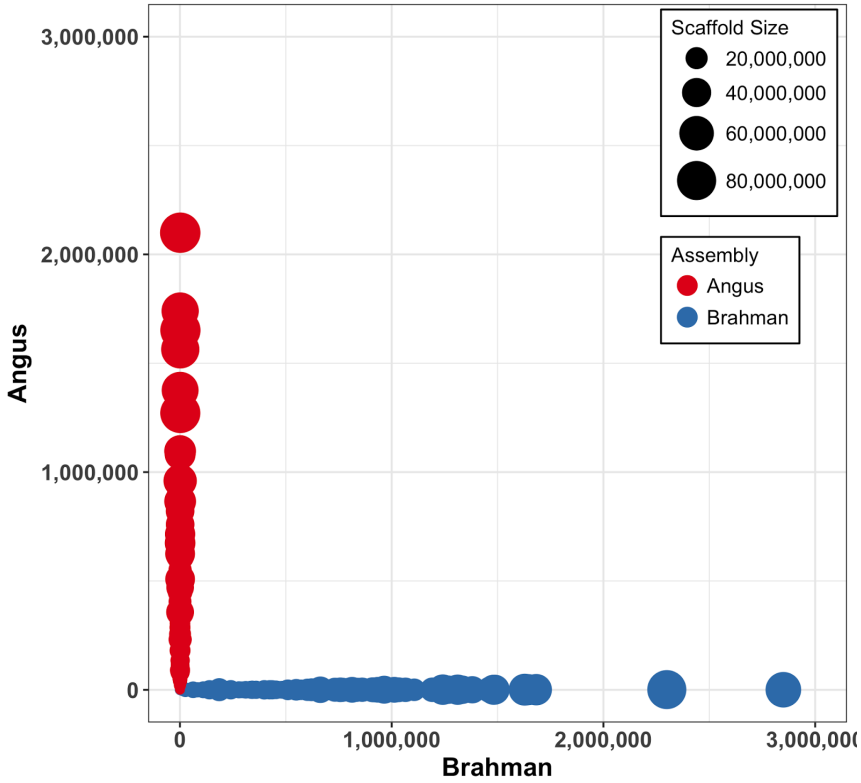
- K-mers sensitive to SVs and SNPs
  - Each SNP == k k-mers
- Expect
  - 90% confidence reads ≥ 5 kbp have at least one k-mer
- Observe
  - 87.4% of all bases
    - avg read length 12 kbp
  - 90% of all bases >= 5kbp



- Pick minimum k-mer given genome size to avoid random collision to maximize survival

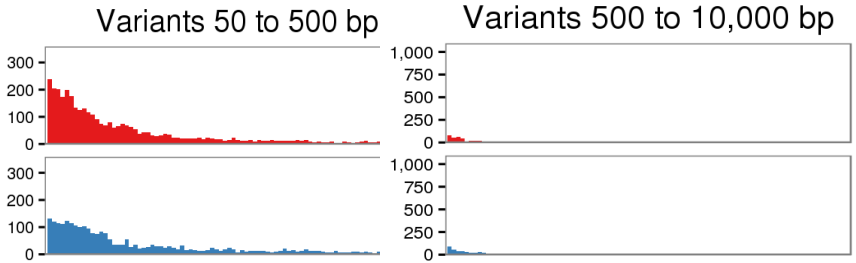
B. taurus Falcon-unzip vs TrioBinning



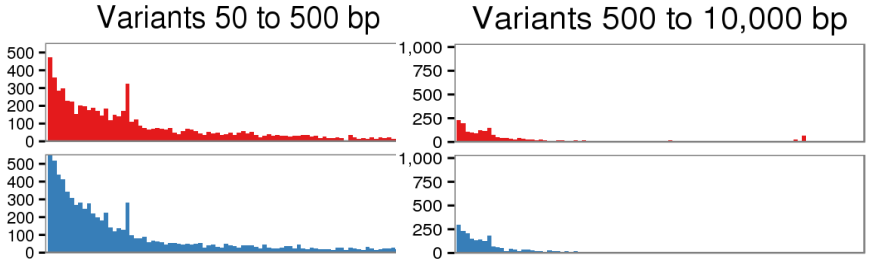


- #10MbHaplotigClub

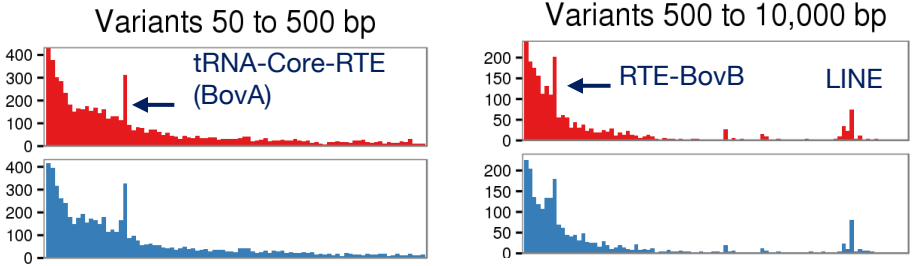
What do you miss with a poor reference?



- UMD3 vs Nelore (*B. indicus*)
  - No variants >200 bp

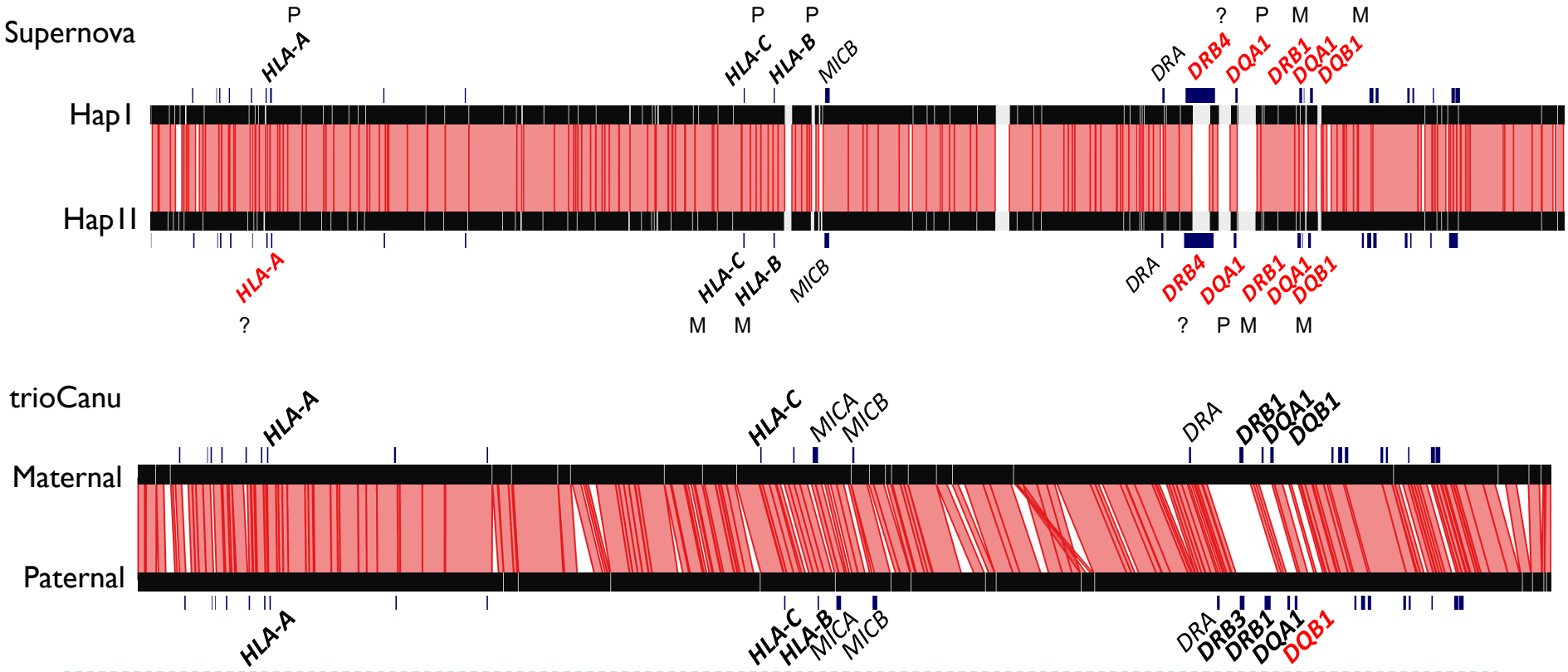


- UMD3 vs Brahman (maternal)
  - No variants > 1kbp



- Father (*B. taurus*) vs Mother (*B. indicus*)
  - Complete profile

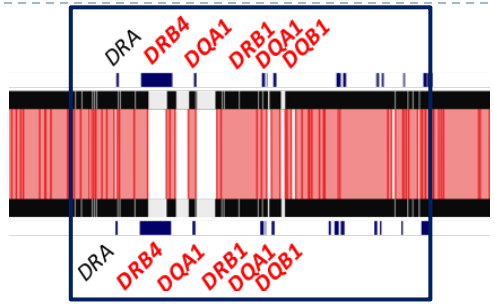
MHC Comparison

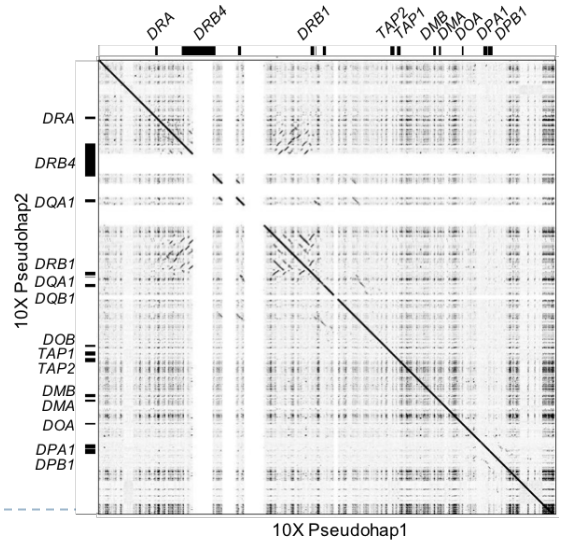


- 10X average edit distance: 45.25 bp, TrioBinning average edit distance: 0.1 bp

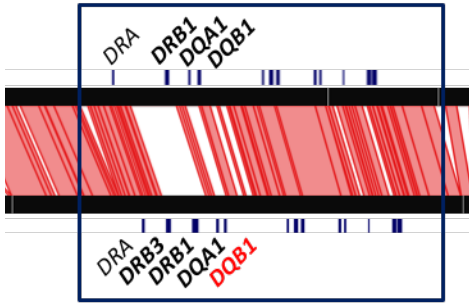
Class II

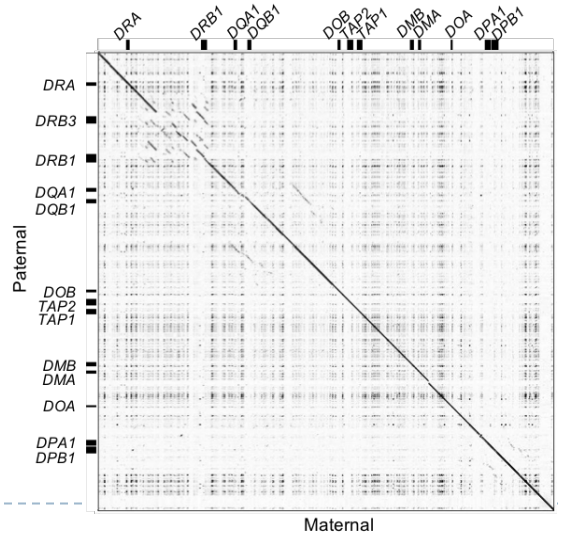
Supernova





trioCanu





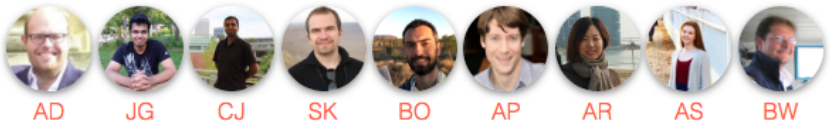
A new strategy to generate references?

- No inbreeding is ever perfect
  - Time consuming
  - Wrong strategy
- Select most **outbred** individual along with parents to improve haplotype resolution
  - Get two full haplotypes phased across full genome
  - Greater continuity than assembling without trio information with sufficient coverage
  - Minimal additional cost of two Illumina libraries
  - Can also work with population data
  - Limited in regions of parent and child homozygosity (e.g. 0/1 genotype in all)
    - Trio approach cannot resolve unless spanned by reads
      - Select more outbred individual
      - Sequence with longer reads
- Sequence/assembler agnostic
  - Polish/gap-fill as before using haplotype-assigned sequences
- Combine with Hi-C to get haplotype resolved chromosomes

Acknowledgements

genomeinformatics.github.io

- Adam Phillippy
- Brian Walenz
- Alexander Dilthey
- **Arang Rhie**
- Brian Ondov



canu.readthedocs.io

- Adam Phillippy
- Brian Walenz
- Konstantin Berlin
- Jason Miller

- Cow F1 collaborators
  - **Tim Smith**
  - John Williams
  - Sarah Kingan