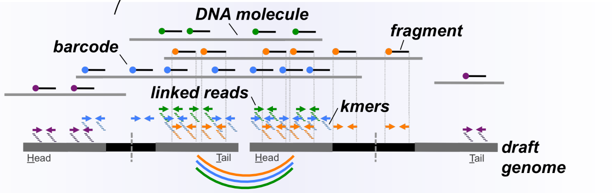
BC Cancer Agency
Canada's Michael Smith Genome Sciences Centre
www.bcgsc.ca • rwarren@bcgsc.ca

arks chromosome-scale scaffolding human genomes with linked read kmers

René Warren • Lauren Coombe • Jessica Zhang • Ben Vandervalk
Justin Chu • Shaun Jackman • Jeffrey Tse • Inanç Birol



Builds on our arcs scaffolder [1]

- Alignment-free linked read scaffolder
- Order and orient genomic contigs
  - Goal : 1 scaffold / chromosome
  - Recover complete genes
  - Estimate gap size
- Uses 10x Genomics (10xG) Chromium
- Similar tools
  - *fragScaff* [2,3] HiC
  - *Architect* [4] Moleculo read cloud
  - *Supernova* [5] Chromium

KMER MAPPING ★ STREAMLINED ★ FASTER ★ GAP SIZE ESTIMATES ★ IMPROVES 10xG DRAFTS

Approach

Linked read mapping

Requires a contig end to match a minimum fraction of the read kmers (parameter *-j*, 0.55, default) :

$$score_j(contig, read) = \frac{|kmers(contig) \cap kmers(read)|}{|kmers(read)|}$$

Higher specificity : both reads/pair must map same target & kmers with multiple memberships discarded

Gap size estimation

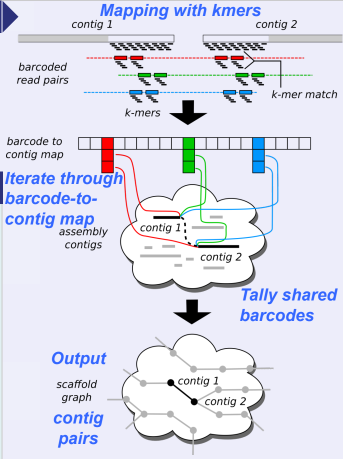
- Train on distances (*D*) between contig head and tail
- Record *D* and Jaccard index (*J*) for shared barcodes

$$J(x, y) = \frac{|barcodes(x) \cap barcodes(y)|}{|barcodes(x) \cup barcodes(y)|}$$

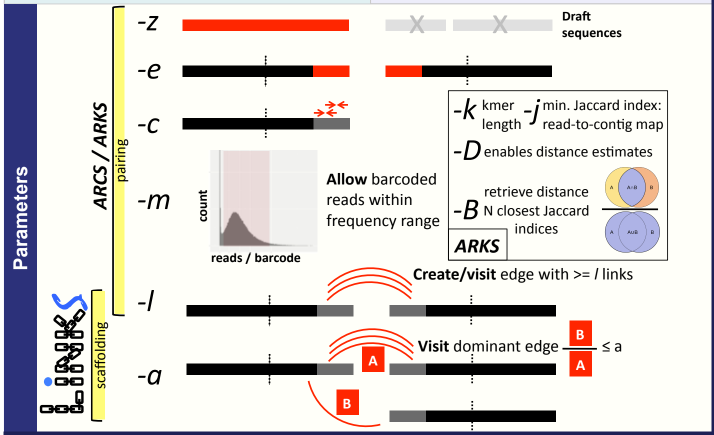
- Retrieve intra-contig *D* samples with the N closest *J*
- Output 1st / 99th centile of *D* as upper / lower bounds

$$D_{min}(x, y) = Q_{0.01}\{D(x_i, y_i) \| \underset{i_1, i_2, \ldots, i_{20}}{\operatorname{argmin}} \sum |J(x, y) - J(x_i, y_i)|\}$$

$$D_{max}(x, y) = Q_{0.99}\{D(x_i, y_i) \| \underset{i_1, i_2, \ldots, i_{20}}{\operatorname{argmin}} \sum |J(x, y) - J(x_i, y_i)|\}$$



Parameters



Allow barcoded reads within frequency range

-k kmer length *-j* min. Jaccard index: read-to-contig map

-D enables distance estimates

-B retrieve distance N closest Jaccard indices

ARKS

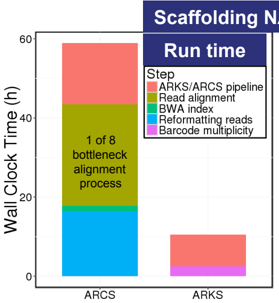
Create/visit edge with >= *l* links

Visit dominant edge B/A ≤ a

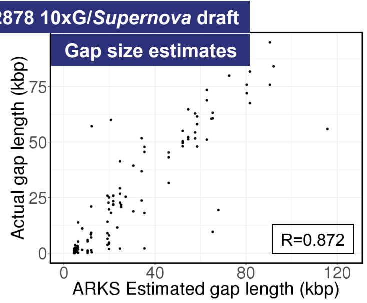
Performance

Scaffolding NA12878 10xG/*Supernova* draft

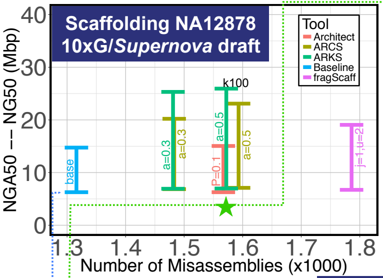
Run time

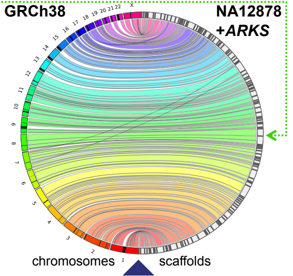


Gap size estimates

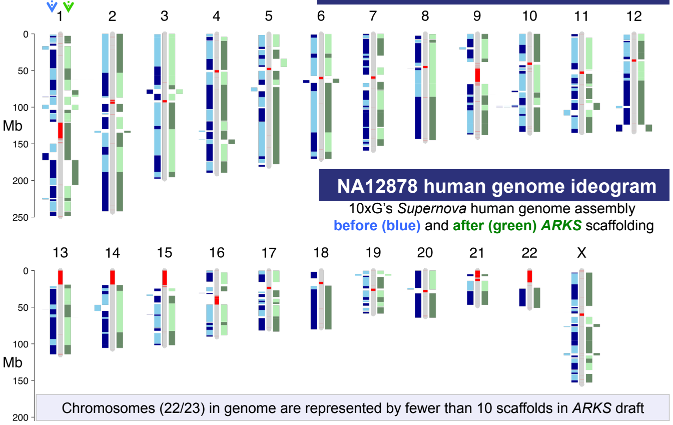


Results





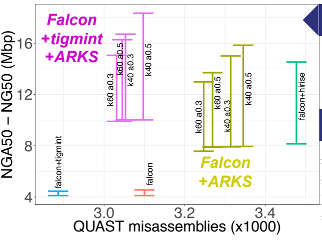
Comparison to human reference



NA12878 human genome ideogram

10xG's *Supernova* human genome assembly **before (blue)** and **after (green)** *ARKS* scaffolding

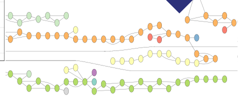
Chromosomes (22/23) in genome are represented by fewer than 10 scaffolds in *ARKS* draft



Scaffolding NA24143 PacBio draft

Coupled with **linked read misassembly corrector** Tigmint, *ARKS* improves a *Falcon* assembly of human *PacBio* long reads

Assembly segment graph



Edges joining different color segments : groups of molecules split between scaffolds

Acknowledgements

References

1. Yeo, *et al.* 2017. *Bioinformatics* 34, 725
2. Adey, *et al.* 2014. *Genome Res.* 24, 2041
3. Mostovoy, *et al.* 2016. *Nat. Methods* 13, 587
4. Kuleshov, *et al.* 2016. *Bioinformatics* 32, i216
5. Weisenfeld, *et al.* 2017. *Genome Res.* 27, 757

Software

https://github.com/bcgsc/
arks
arcs
links
tigmint

NIH
Funding
GenomeBritishColumbia GenomeCanada
BC CANCER FOUNDATION