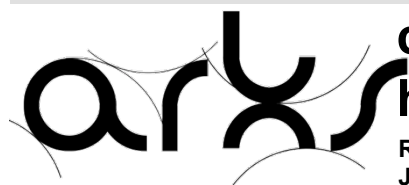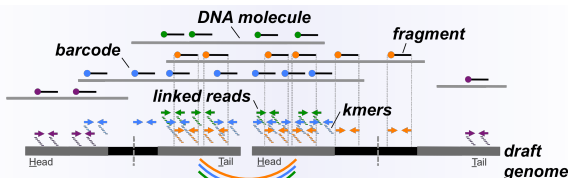**BC Cancer Agency**
Canada's Michael Smith Genome Sciences Centre
www.bcgsc.ca • rwarren@bcgsc.ca

arks — chromosome-scale scaffolding human genomes with linked read kmers

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

DNA molecule / barcode / fragment / linked reads / kmers / Head / Tail / draft genome

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
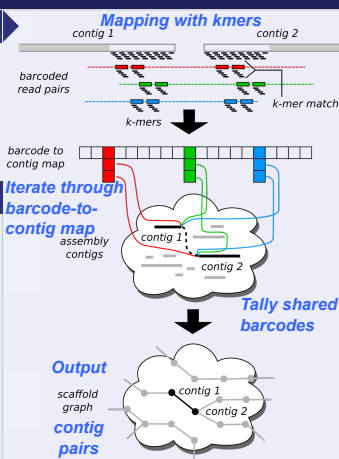
Gap size estimation

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

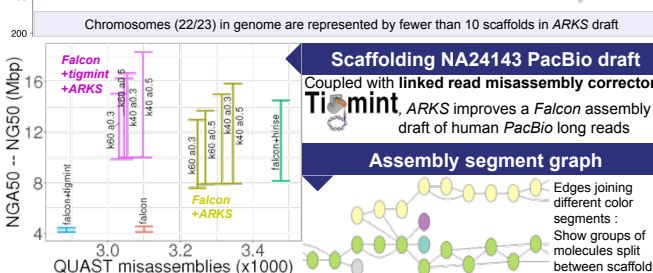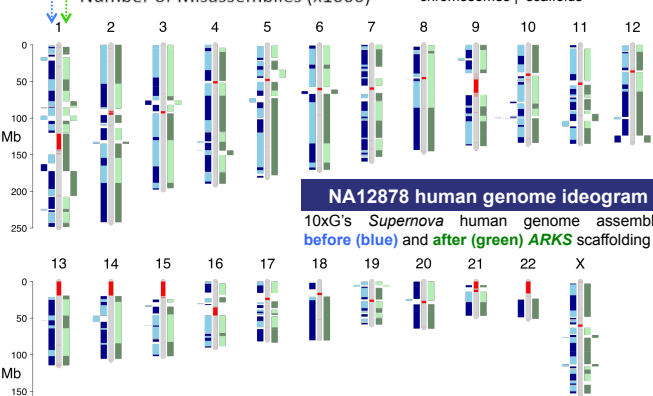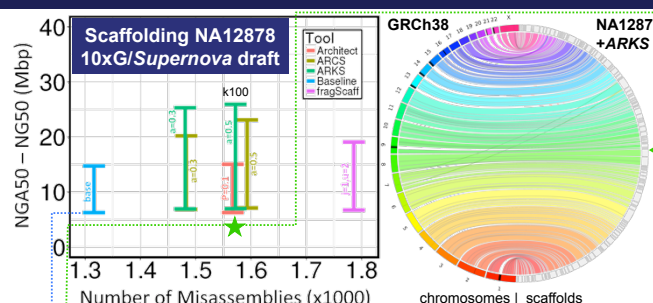$$J(x,y) = \frac{|barcodes(x) \cap barcodes(y)|}{|barcodes(x) \cup barcodes(y)|}$$

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

$$D_{min}(x,y) = Q_{0.01}\{D(x_i,y_i)\| \underset{i_1,i_2,\dots,i_{20}}{argmin} \sum |J(x,y) - J(x_i,y_i)|\}$$
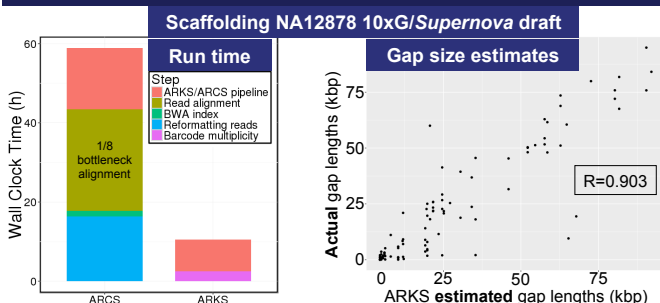
$$D_{max}(x,y) = Q_{0.99}\{D(x_i,y_i)\| \underset{i_1,i_2,\dots,i_{20}}{argmin} \sum |J(x,y) - J(x_i,y_i)|\}$$

Mapping with kmers
contig 1 / contig 2 / barcoded read pairs / k-mers / k-mer match / barcode to contig map

Iterate through barcode-to-contig map
assembly contigs / contig 1 / contig 2

Tally shared barcodes

Output
scaffold graph / **contig pairs** / contig 1 / contig 2

Parameters

ARCS / ARKS — pairing
- -z
- -e
- -c
- -m

Draft sequences

- $-k$ kmer length
- $-j$ min. Jaccard index: read-to-contig map
- $-D$ enables distance estimates
- $-B$ N closest Jaccard indices — retrieve distance

Allow barcoded reads within frequency range
count / reads / barcode

ARKS

scaffolding
- -l
- -a

Create/visit edge with >= l links

Visit dominant edge $\frac{B}{A} \le a$

Performance

Scaffolding NA12878 10xG/Supernova draft

Run time — Wall Clock Time (h). Step: ARKS/ARCS pipeline, Read alignment, BWA index, Reformatting reads, Barcode multiplicity. 1/8 bottleneck alignment. ARCS vs ARKS.

Gap size estimates — Actual gap lengths (kbp) vs ARKS estimated gap lengths (kbp). R=0.903

Results

Scaffolding NA12878 10xG/Supernova draft — NGA50 – NG50 (Mbp) vs Number of Misassemblies (x1000). Tool: Architect, ARCS, ARKS, Baseline, fragScaff. k100.

GRCh38 / NA12878 +ARKS. chromosomes | scaffolds

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

NGA50 – NG50 (Mbp) vs QUAST misassemblies (x1000). Falcon+tigmint+ARKS, Falcon+ARKS, Falcon, falcon-tigmint, falcon-shine.

Scaffolding NA24143 PacBio draft

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

Assembly segment graph

Edges joining different color segments : Show 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