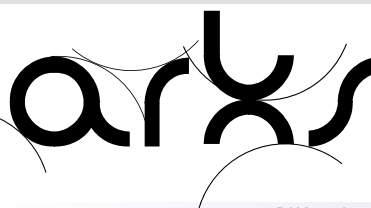




BC Cancer Agency

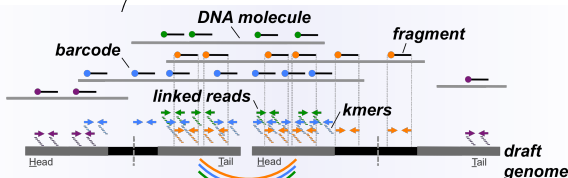
Canada's Michael Smith Genome Sciences Centre

www.bcgsc.ca • rwarren@bcgsc.ca



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



Builds on our *ARKS* scaffolder ¹

- Alignment-free linked read scaffolder
- Order and orient genomic contigs
- Uses 10x Genomics (10xG) Chromium
- Similar tools

- Goal : 1 scaffold / chromosome
- Recover complete genes
- Estimate gap size
- *fragScaff* ^{2,3} HiC
- *Architect* ⁴ Molecule 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 γ , 0.55, default) :

$$\text{score}_j(\text{contig}, \text{read}) = \frac{|\text{kmers}(\text{contig}) \cap \text{kmers}(\text{read})|}{|\text{kmers}(\text{read})|}$$

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

Gap size estimation

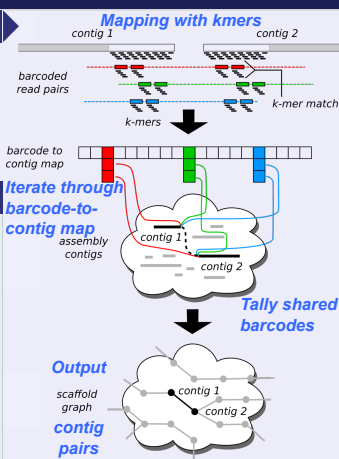
- Train on distances (D) between contig head and tail
- Record D and Jaccard index (J) for shared barcodes

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

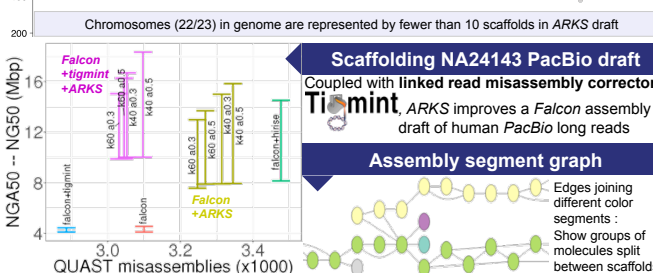
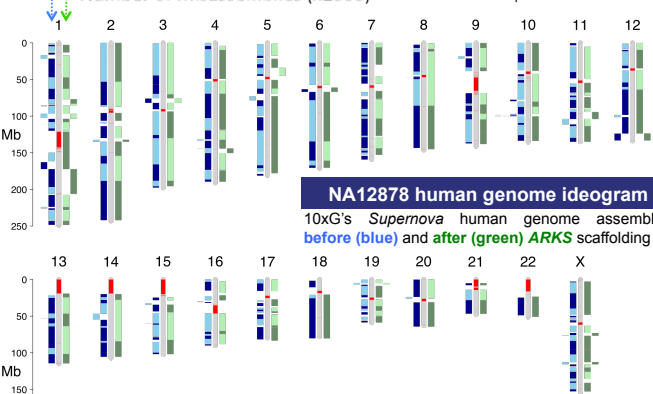
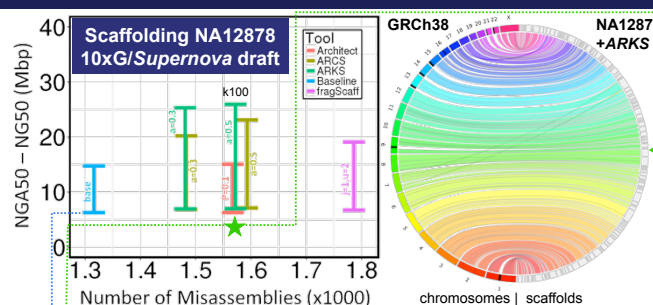
- Retrieve intra-contig D samples with the N closest J
- Output 1st / 99th centile of D as upper / lower bounds

$$D_{\min}(x, y) = Q_{0.01}\{D(x_i, y_i) \mid \arg\min_{i_1, i_2, \dots, i_{20}} \sum |J(x_i, y_i) - J(x_{i_1}, y_{i_1})|\}$$

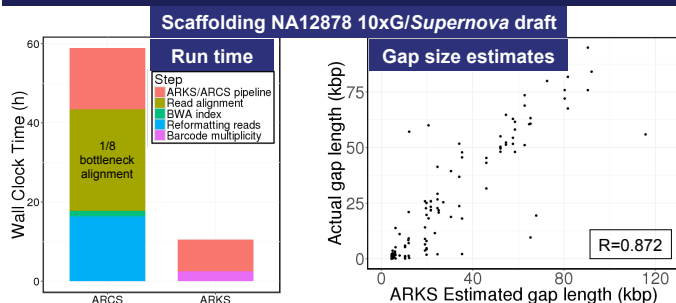
$$D_{\max}(x, y) = Q_{0.99}\{D(x_i, y_i) \mid \arg\min_{i_1, i_2, \dots, i_{20}} \sum |J(x_i, y_i) - J(x_{i_1}, y_{i_1})|\}$$



Results



Performance



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

