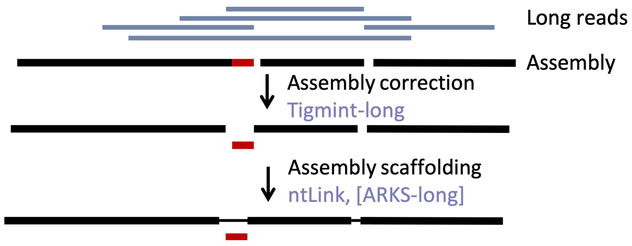
longStitch: High-quality genome assembly correction and scaffolding using long reads

Lauren Coombe, Janet X Li, Theodora Lo, Johnathan Wong, Vladimir Nikolic, René L Warren and Inanc Birol

lcoombe@bcgsc.ca
www.birollab.ca

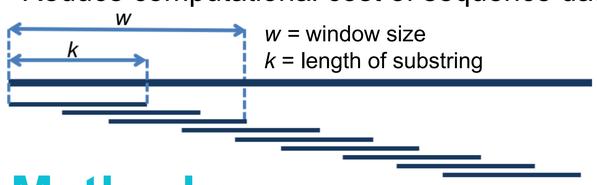
LongStitch



- **Long read** *de novo* assembly correction/scaffolding (3 steps)
- **ntLink**: newly developed long-read scaffolder
- **Tigmint-long**: correction
- **ARKS-long**: scaffolding

Minimizer sketches

Reduce computational cost of sequence data storage and manipulation[1]



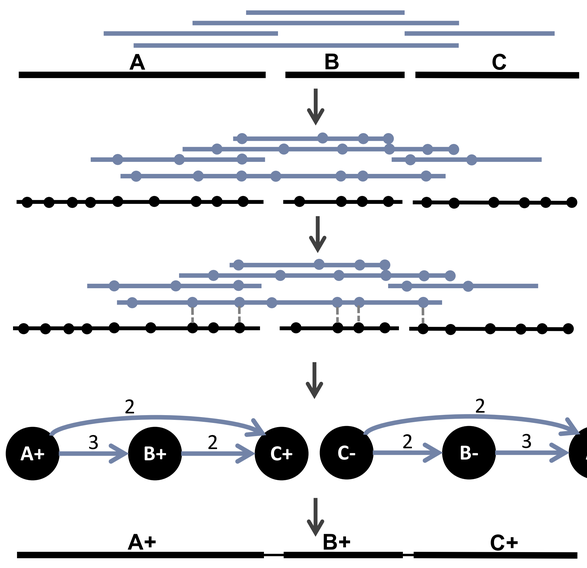
w = window size
k = length of substring

For each window of *w* adjacent *k*-mers:
- Compute hash values of each *k*-mer
- Window's minimizer = smallest hash value

Generates ordered list of minimizers per sequence

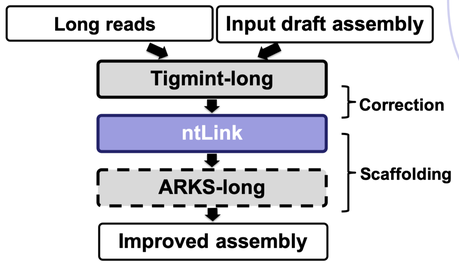
Methods

ntLink algorithm

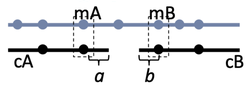


Circles: minimizers
Dotted lines: identical minimizers

Pair	# support reads
A+ → B+	3
B+ → C+	2
A+ → C+	2



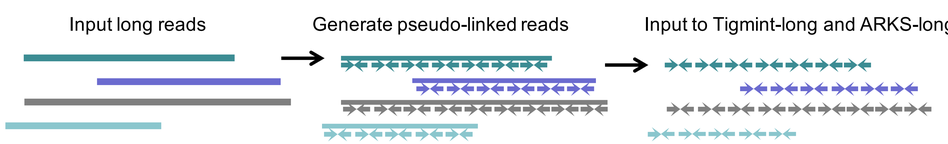
ntLink gap size estimation

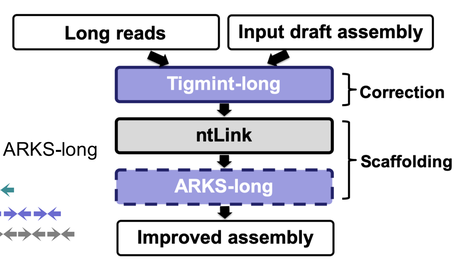


$$gap\ size = |pos(mA, long\ read)| - |pos\ (mB, long\ read)| - a - b$$

Adapting Tigmint and arks for long reads

Originally developed for linked reads[2,3,4]





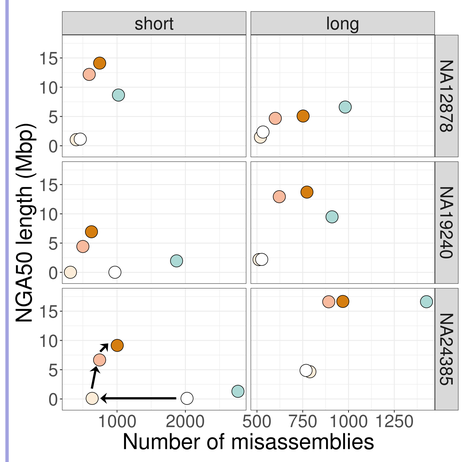
Results

Running LongStitch (human data)

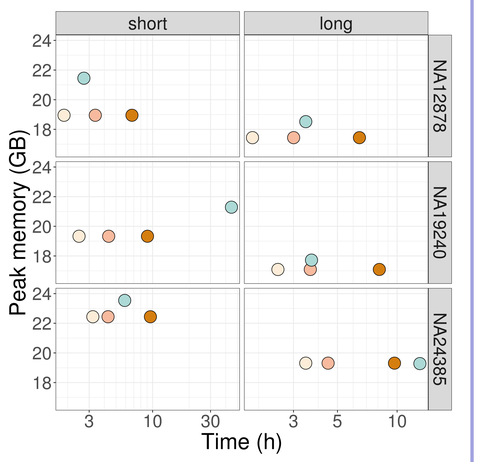
Short: short-read ABySS[5] assembly
Long: long-read Shasta[6] assembly
Comparing LongStitch to state-of-the-art long read scaffolder LRScaf[7]

Individual	NA12878	NA19240	NA24385
Nanopore read coverage	39X	49X	51X

Contiguity and Correctness

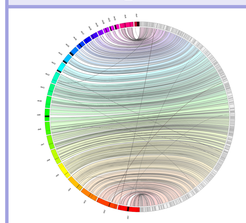
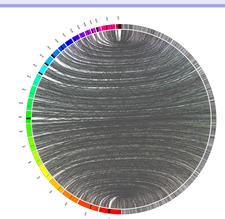
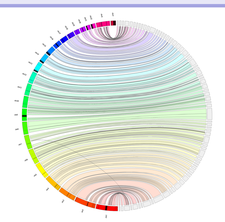
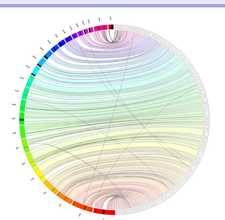


Benchmarking



Visualizing large-scale misassemblies with Jupiter[8] plots – NA24385

Large misassemblies evident as interrupting ribbons

short		long	
LongStitch	LRScaf	LongStitch	LRScaf
			

Conclusions

- LongStitch: scalable assembly correction and scaffolding
- Leverages the rich information in long reads
- Generates high-quality genome assemblies

- Paper describing LongStitch: *Coombe, L. et al. (2021) bioRxiv, 2021.06.17.448848.*

References

1. Roberts,M. et al. (2004) *Bioinformatics*, 20, 3363–3369.
2. Jackman,S.D. et al. (2018) *BMC Bioinformatics*, 19, 1–10.
3. Yeo,S. et al. (2018) *Bioinformatics*, 34, 725–731.
4. Coombe,L. et al. (2018) *BMC Bioinformatics*, 19, 1–10.
5. Jackman,S.D. et al. (2017) *Genome Res.*, 27, 768–777.
6. Shafin,K. et al. (2020) *Nat. Biotechnol.*, 38, 1044–1053.
7. Qin,M. et al. (2019) *BMC Genomics*, 20, 1–12.
8. Chu,J. (2018) *Zenodo*. doi:10.5281/ZENODO.1241235.

Software Availability

https://github.com/bcgsc/longstitch
https://github.com/bcgsc/ntlink
conda install -c bioconda longstitch

Funding

John Jambor Education Fund
Genome British Columbia
Genome Canada
National Institutes of Health