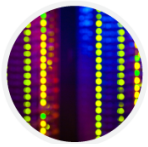
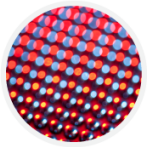


CANADA'S MICHAEL SMITH GENOME SCIENCES CENTRE

LINKS

Long Interval Nucleotide K-mer caffolder

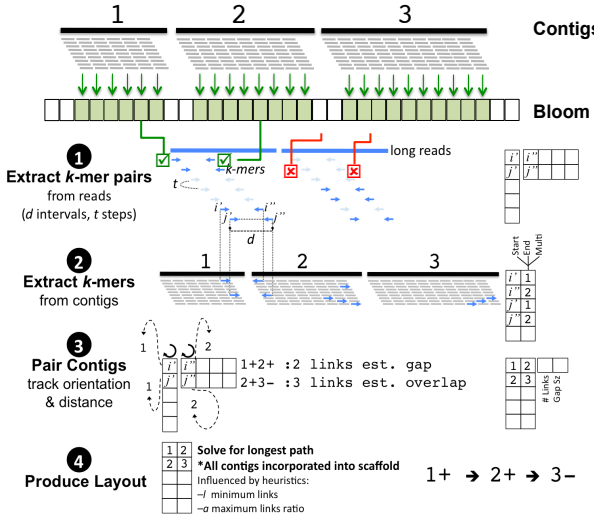
Scalable Algorithms for Long Read Assembly and Scaffolding

René L. Warren,
Chen Yang, Benjamin P. Vandervalk,
Bahar Behsaz, Albert Lagman Steven JM Jones,
Inanç Birol

Abstract

Routine reconstruction of complex genomes from experimental data is not a solved problem, owing to long repeats that are not resolvable by short reads. Established and emerging long read technologies hold the potential to address present limitations, but their current high errors typically require base correction and/or additional pre-processing before use. Here we present LINKS, a method that exploits the sequence properties of long reads for scaffolding high-quality genome drafts.

Algorithm

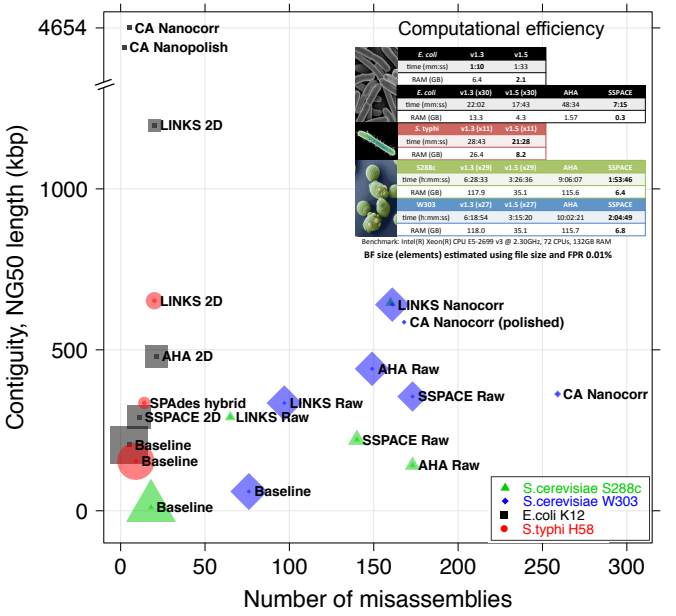


Error Model

Mismatch: $P_m \sim a_m$ Poisson(l_m) + (1-a_m) Geometric(p)
Insertion: $P_i \sim \alpha_i$ Weibull(l_i, κ_i) + (1-α_i) Geometric(p_i)
Deletion: $P_d \sim \alpha_d$ Weibull(l_d, κ_d) + (1-α_d) Geometric(p_d)

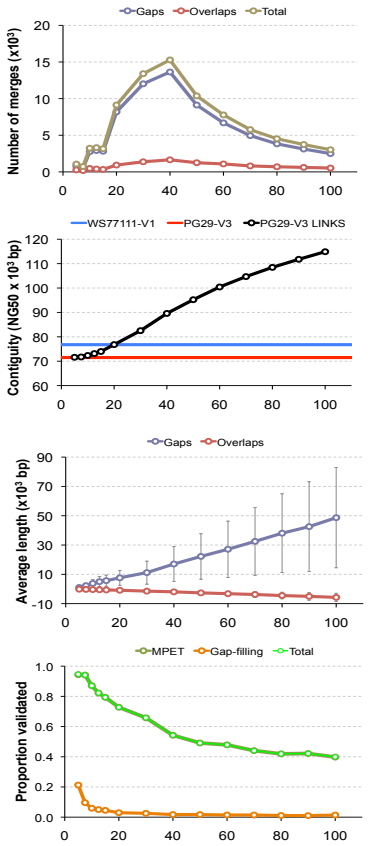
	Mismatch			Insertion				Deletion			
	a_m	l_m	p_m	a_i	l_i	k_i	p_i	a_d	l_d	k_d	p_d
E. coli R7	0.248	0.480	0.711	0.850	1.004	0.968	0.418	0.870	0.986	1.026	0.403
E. coli R7.3	0.138	0.441	0.476	0.900	1.045	1.473	0.272	0.959	1.059	1.682	0.249
S. cerevisiae R7	0.177	0.499	0.479	0.961	1.024	1.613	0.194	0.891	1.066	1.814	0.207

Performance



Point size is proportional to number of resulting scaffolds, normalized from baseline assembly

Scalability



QUAST analysis of a baseline *E. coli* assembly and re-scaffolded assemblies using Oxford Nanopore 2D (R7 chem.) or raw (R7.3 chem.) reads

Stats based on sequences ≥ 500 bp	A. ABySS Contigs	B. ABySS Scaffolds	C. LINKS k15d4000	D. LINKS k15d4000	E. SSPACE -LR g200	F. LINKS 30x k15d500-16kbp	G. LINKS 30x k15d500-16kbp	H. LINKS 30x k15d500-16kbp
Input Data	Illumina MiSeq	A. +MiSeq	A. +ONT Full 2D R7	B.+ONT Full 2D R7	B. +ONT Full 2D R7	B. +ONT Full 2D R7	B. +ONT All 2D R7	B. +raw ONT R7.3
Time (h:mm:ss)	-	-	0:01:35	0:01:32	0:01:09	0:17:43	1:50:25	3:04:11
Memory (GB)	-	-	2.1	2.1	0.7	4.3	27.2	46.7
Total sequences	67	61	49	48	43	27	16	26
Largest (bp)	358,719	406,793	633,204	633,147	628,411	1,057,556	1,286,148	1,286,419
NG50 length (bp)	179,720	206,356	270,992	293,925	226,696	633,147	1,197,321	645,796
Misassemblies	5	5	5	5	8	11	20	9
Genes + parts	4,442 + 63	4,443 + 62	4,443 + 62	4,443 + 62	4,448 + 57	4,443 + 62	4,440 + 62	4,443 + 62
Max.alignment (bp)	358,223	405,659	486,572	486,527	405,659	760,934	760,934	759,131
NGA50 (bp)	177,531	179,569	228,879	228,879	226,324	299,206	299,206	486,527
NA50 (bp)	146,850	177,531	226,324	226,324	215,056	293,772	294,667	344,280

Iterative scaffolding of the 20 Gbp *Picea glauca* (white spruce) assembly using draft assemblies of two genotypes

Genotype 1 (GCA_000411955.3) 4.2M scaffolds
Genotype 2 (PRJNA242552) 4.3M scaffolds

- Found 84,529 total merges
- Validated final LINKS assembly with MPET reads and a gap-filling tool (Sealer, Paulino *et al.*, in review)

References

Birol, K. *et al.* (2014) Assembling Large Genomes with Single-Molecule Sequencing and Locality Sensitive Hashing. *bioRxiv* doi: http://dx.doi.org/10.1101/008003
Boetzer, M. and Pirovano, W. (2014) SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics*. **15**, 211.
Clarke, J. *et al.* (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.*, **4**, 265–70.
Goodwin, S. et al. (2015) Oxford Nanopore Sequencing and de novo Assembly of a Eukaryotic Genome. doi: http://dx.doi.org/10.1101/013490.
Gurevich, A. *et al.* (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, **29**,1072-5
Koren, S. and Phillippy, A. M. (2014) One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Curr.Opin.Microbiol.*, **23**C,110-120.
Quick, J. *et al.* (2014) A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer. *GIgascience*, **3**, 22.
Simpson, J.T. *et al.* (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19**,1117-23.
Warren, R.L. *et al.* (2007) Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*, **23**, 500-1.

Funding

NIH BC CANCER FOUNDATION partners in discovery Genome BritishColumbia GenomeCanada intel

Genome Sciences Centre • British Columbia Cancer Agency • 100 – W 7th Ave Vancouver BC V5Z 4S6 Canada • tel. 604-707-5900 • fax. 604-876-3561 • www.bcgsc.ca