

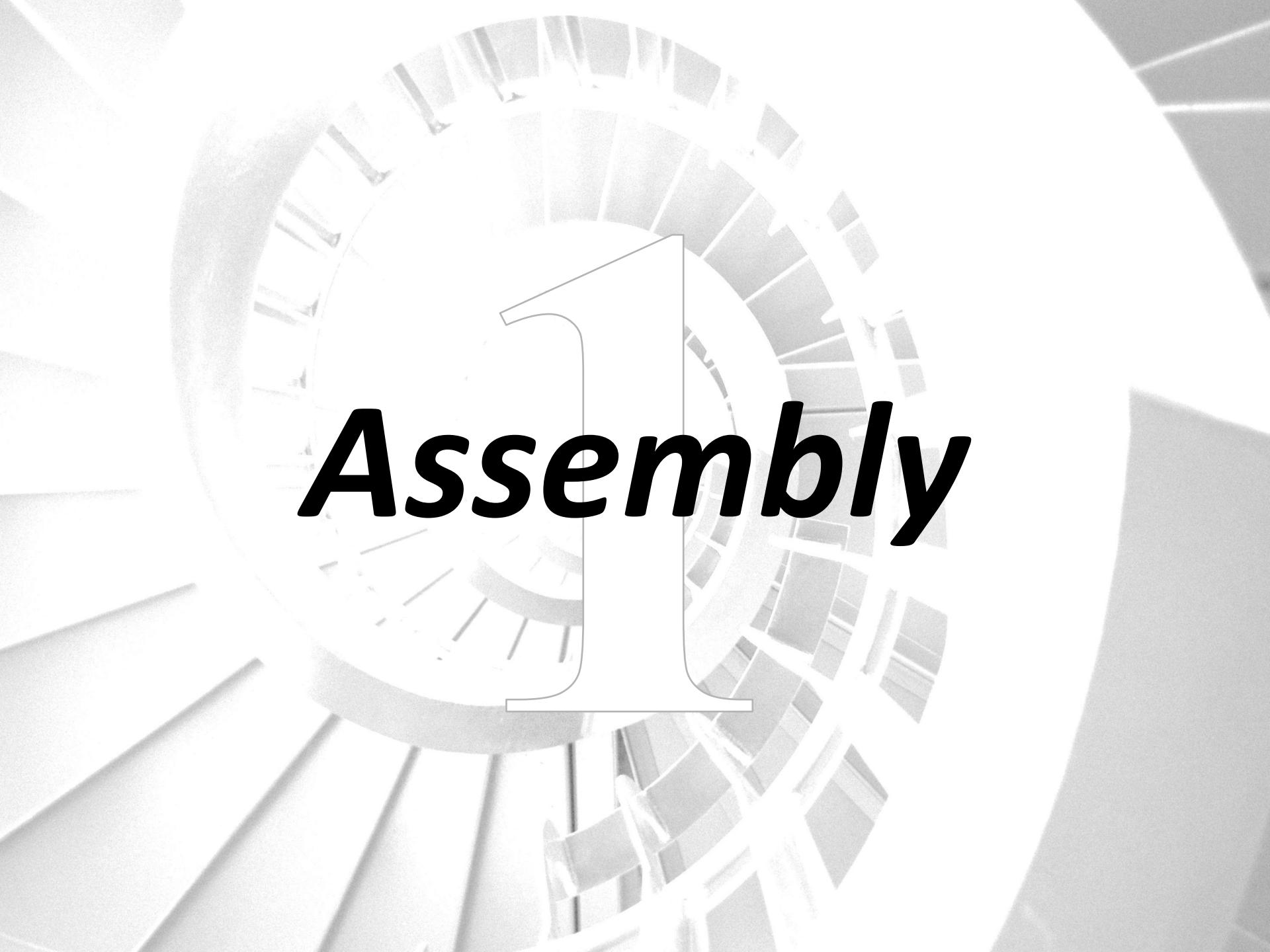
scalable solutions for *de novo* genome assembly

René L Warren 2019



CANADA'S MICHAEL SMITH
GENOME
SCIENCES
CENTRE





Assembly



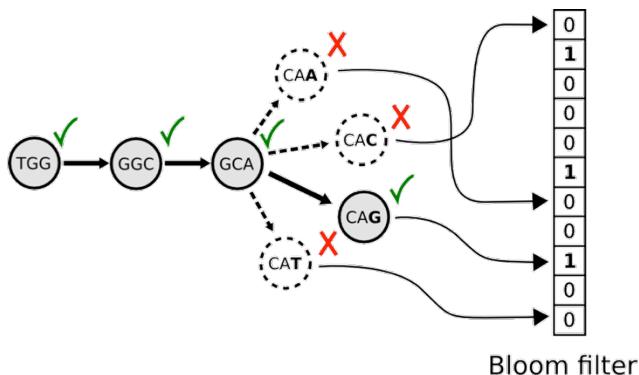
Short reads *de novo* genome assembly

2009 Parallel DBG assembler

- MPI to aggregate memory
- Assembled 20 Gb spruce genome

2017 Bloom filter representation

- 1/10th RAM
- Large genomes, single computer



GENOME
RESEARCH

Resource

ABYSS: A parallel assembler for short read sequence data

Jared T. Simpson,¹ Kim Wong, Shaun D. Jackman, Jacqueline E. Schein, Steven J.M. Jones, and İnanç Birol²

GENOME
RESEARCH

Method

ABYSS 2.0: resource-efficient assembly of large genomes using a Bloom filter

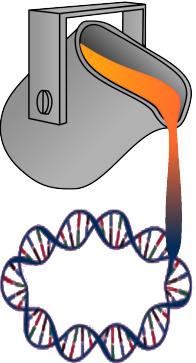
Shaun D. Jackman,¹ Benjamin P. Vandervalk,¹ Hamid Mohamadi, Justin Chu, Sarah Yeo, S. Austin Hammond, Golnaz Jahesh, Hamza Khan, Lauren Coombe, Rene L. Warren, and Inanc Birol

<https://github.com/bcgsc/abyss>



Correction

Tigmint



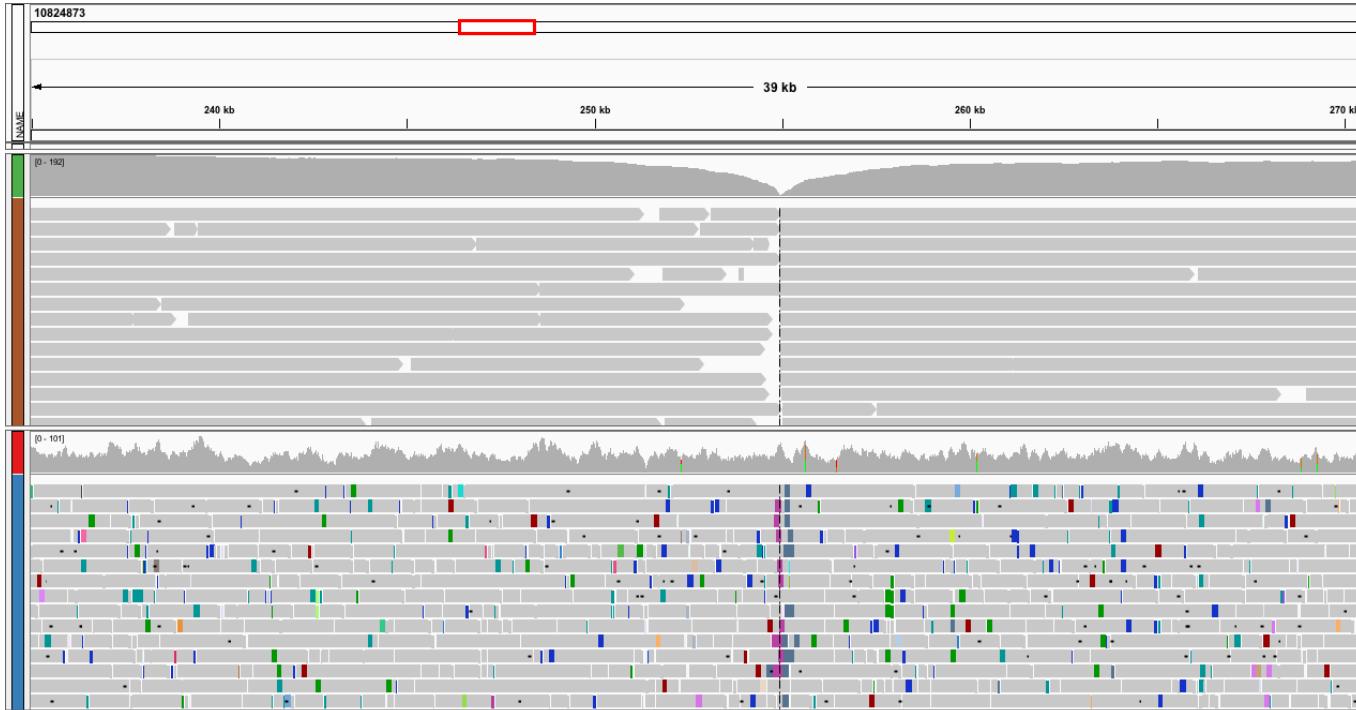
Linked reads misassembly correction



SOFTWARE

Tigmint: correcting assembly errors using linked reads from large molecules

Shaun D. Jackman^{1*} , Lauren Coombe¹, Justin Chu¹, Rene L. Warren¹, Benjamin P. Vandervalk¹, Sarah Yeo¹, Zhuyi Xue¹, Hamid Mohamadi¹, Joerg Bohlmann², Steven J.M. Jones¹ and Inanc Birol¹

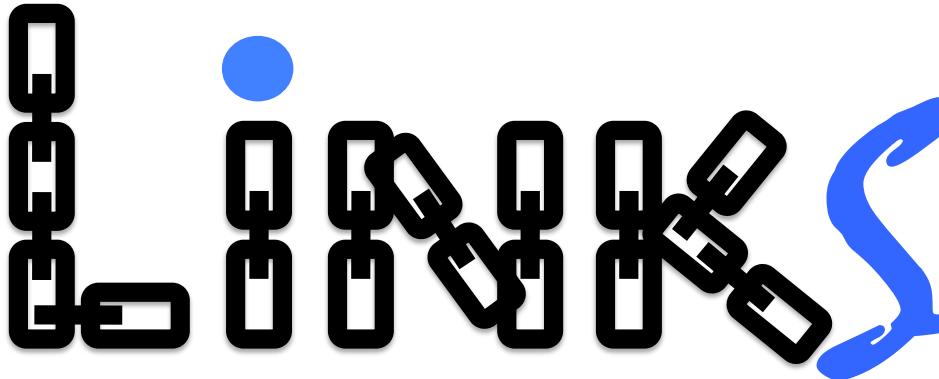


IGV screenshot: Tigmint breakpoint in human genome NA24143

<https://github.com/bcgsc/tigmint>



Scaffolding



Warren et al. *GigaScience* (2015) 4:35
DOI 10.1186/s13742-015-0076-3

(GIGA)ⁿ
SCIENCE

Open Access

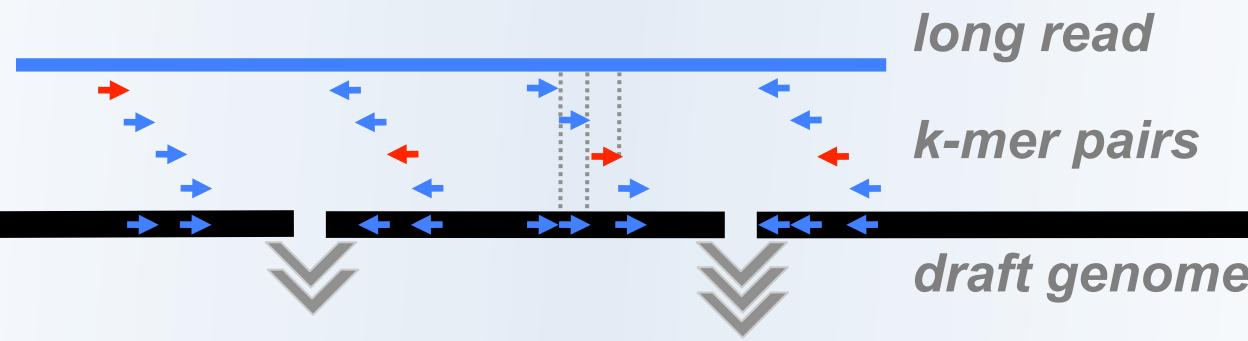


RESEARCH

LINKS: Scalable, alignment-free scaffolding
of draft genomes with long reads

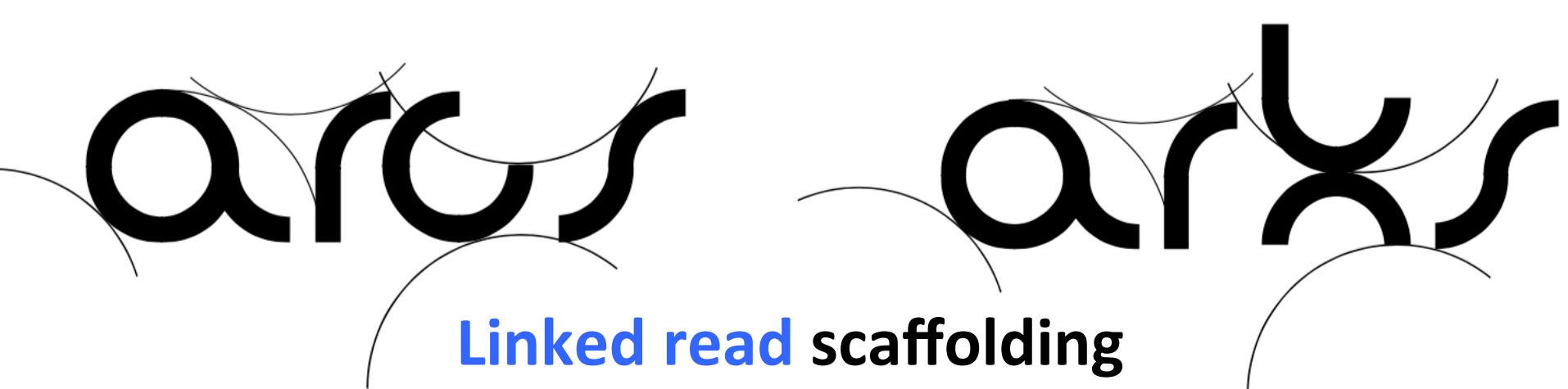
René L. Warren*, Chen Yang, Benjamin P. Vandervalk, Bahar Behsaz, Albert Lagman, Steven J. M. Jones
and Inanç Birol

Long read kmer scaffolding

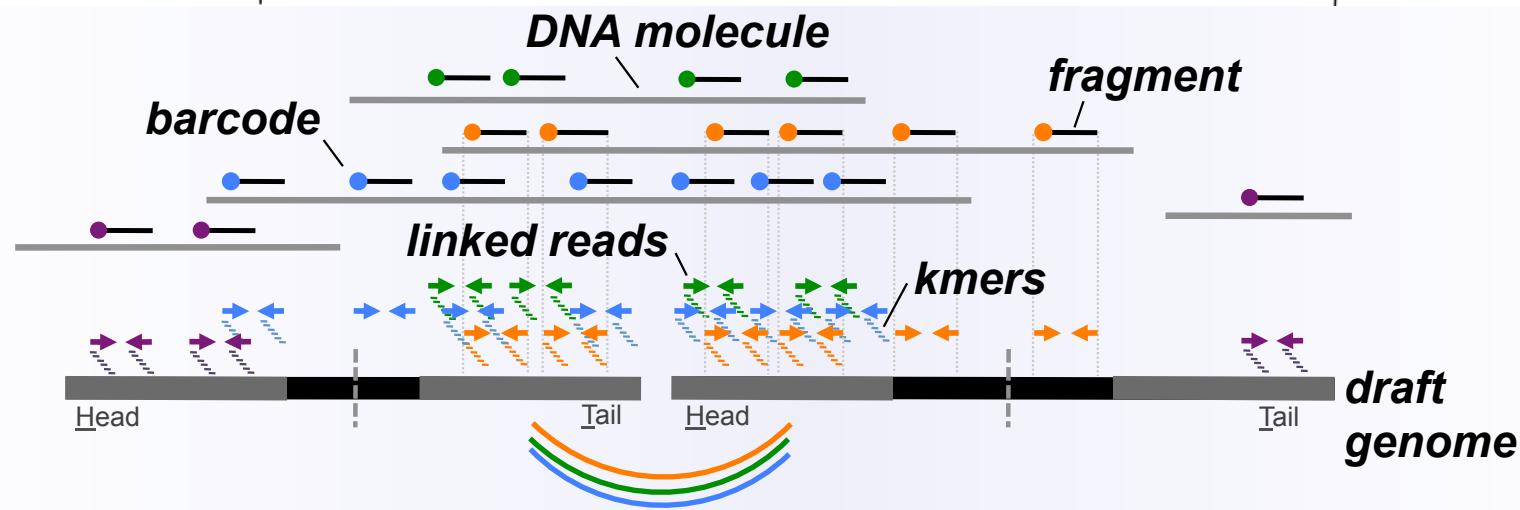


- **Scaffolder** order & orient sequences
- ***k-mer* based** no alignments, error tolerant = no error corrections
- **Vast *k-mer* space** no fragment length limitations
- **Versatile** long-reads, draft sequences, MPET

<https://github.com/bcgsc/links>



Linked read scaffolding



ARCS: scaffolding genome drafts with linked reads



Sarah Yeo, Lauren Coombe, René L Warren ✉, Justin Chu, Inanç Birol Author Notes

Bioinformatics, Volume 34, Issue 5, 1 March 2018, Pages 725–731,

<https://doi.org/10.1093/bioinformatics/btx675>

Coombe et al. BMC Bioinformatics (2018) 19:234
<https://doi.org/10.1186/s12859-018-2243-x>

BMC Bioinformatics

SOFTWARE

Open Access

ARKS: chromosome-scale scaffolding of human genome drafts with linked read kmers



Lauren Coombe[†], Jessica Zhang[†], Benjamin P. Vandervalk, Justin Chu, Shaun D. Jackman, Inanc Birol and René L. Warren^{*}

<https://github.com/bcgsc/arcs>

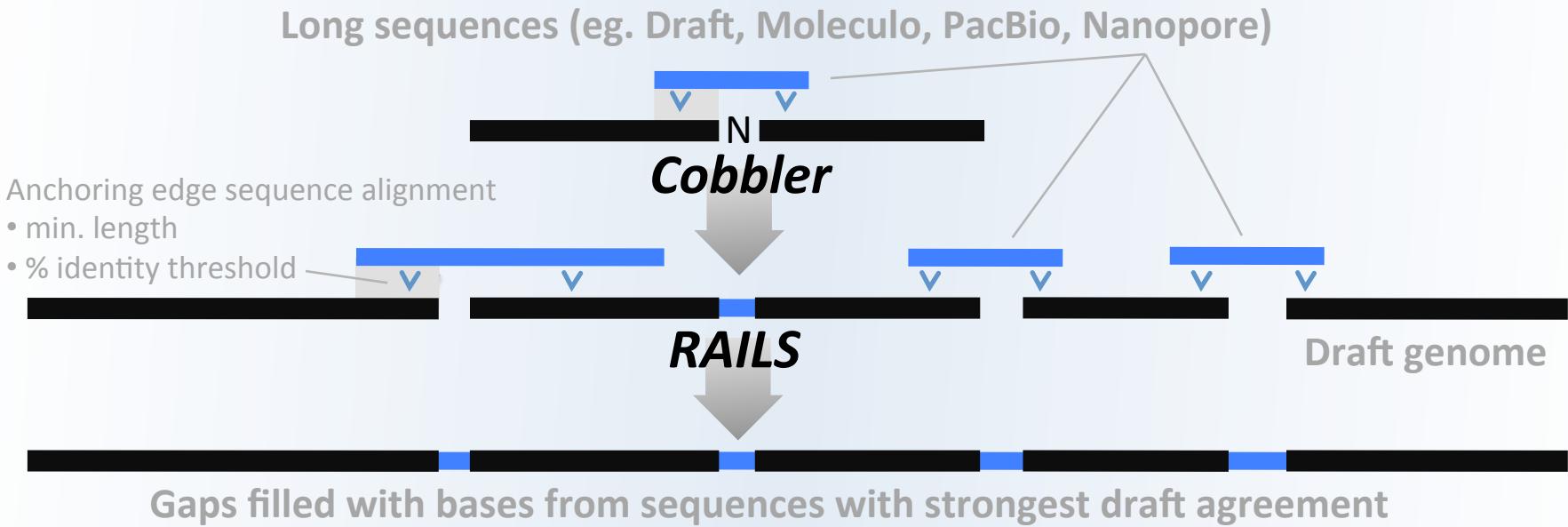
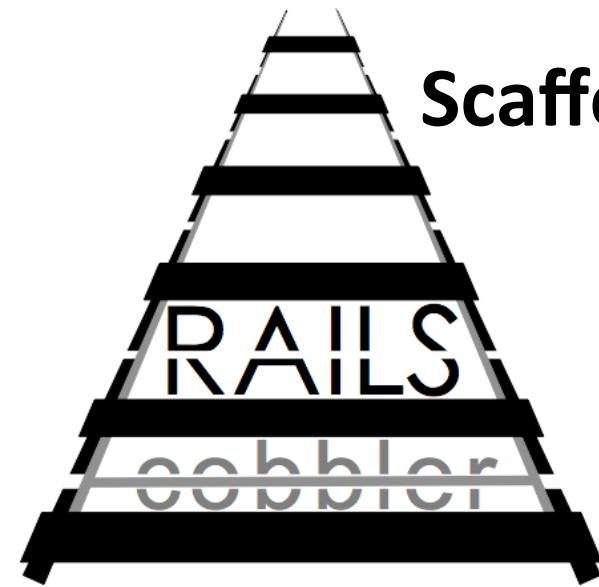
<https://github.com/bcgsc/arks>



Gap-filling

Scaffolding and gap-filling with long reads

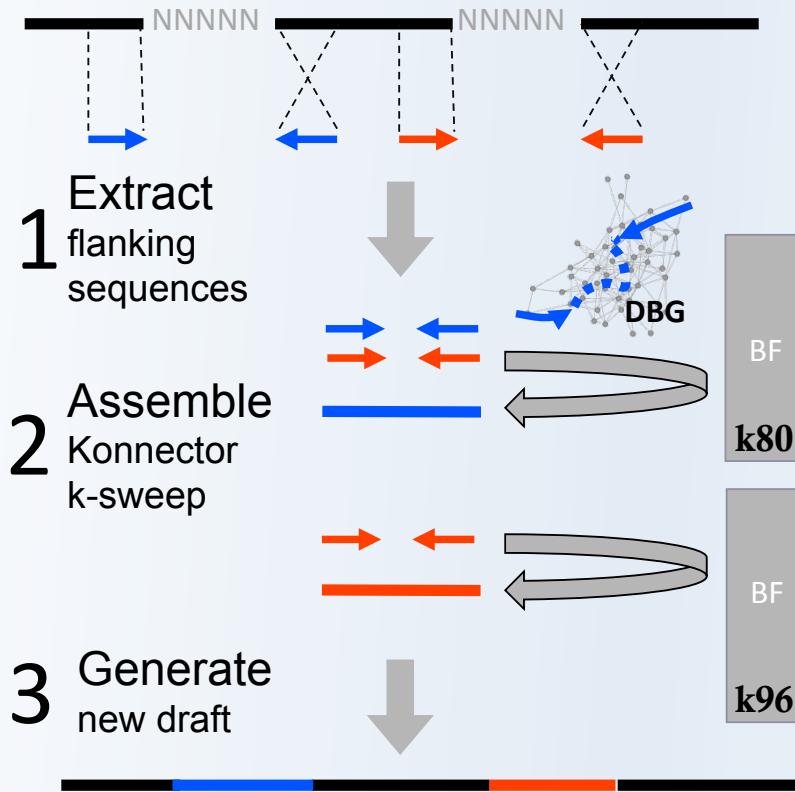
Uses LINKS scaffolding algorithm



<https://github.com/bcgsc/rails>

Sealer

Gap filling with short reads



Vandervalk et al. BMC Medical Genomics 2015, 8(Suppl 3):S1
http://www.biomedcentral.com/1755-8794/8/S3/S1



RESEARCH

Open Access

Konnecter v2.0: pseudo-long reads from paired-end sequencing data

Paulino et al. BMC Bioinformatics (2015) 16:230
DOI 10.1186/s12859-015-0663-4



SOFTWARE

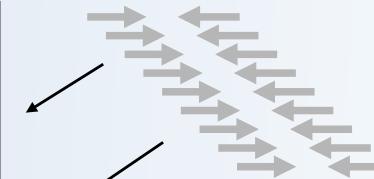
Open Access

Sealer: a scalable gap-closing application for finishing draft genomes



Build Bloom filters (BF) with Konnector

NGS reads k-mers



BF
k80

BF
k96

Genotype/#Gaps	k	#Closed
WS77111/ 1.8M	64 80	461,196
	96	(25.5%)
PG29/ 2.9M	84 96	399,476
		(13.79%)

4.5B Illumina reads Peak RAM: 44GB Time: 27h

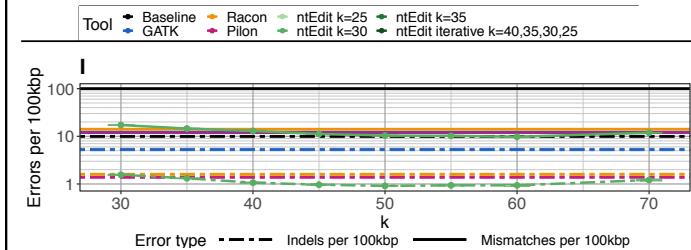
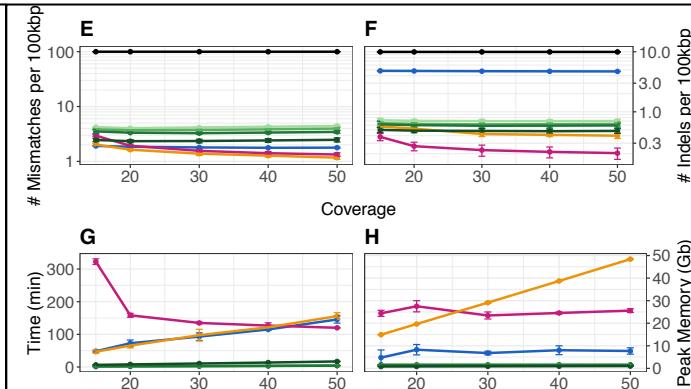
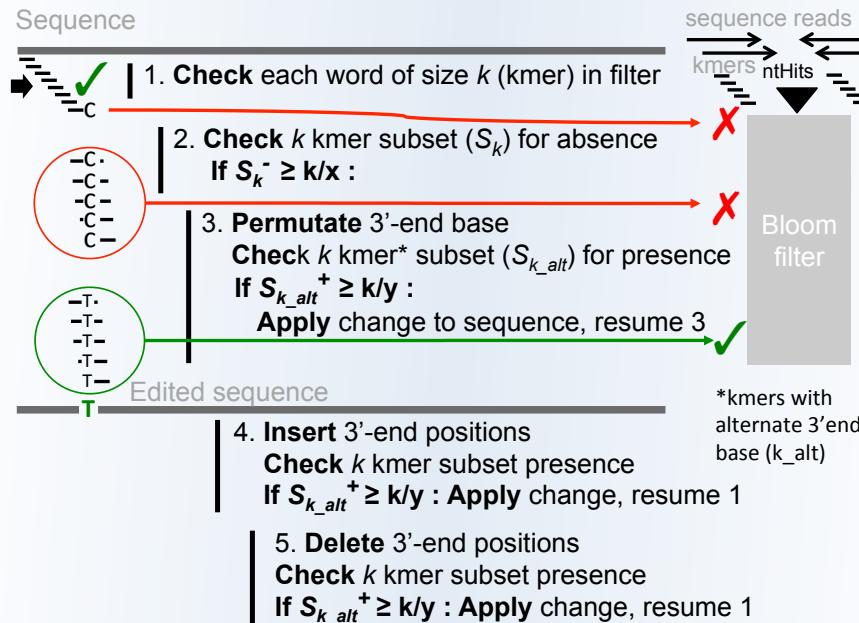
<https://github.com/bcgsc/abyss/tree/master/Sealer>

The background features a complex geometric pattern of concentric circles and radiating lines in shades of gray. The circles are composed of numerous small, rectangular segments that create a textured, gear-like appearance. Radiating from the center are several thick, light-gray lines that intersect the circles at various angles.

Polishing

WTD

Fast genome polishing with short reads



C. elegans

H. Sapiens chr21

- Human and spruce genomes in 4 and 25 minutes
- Fix frameshift errors in nanopore/pacbio assemblies

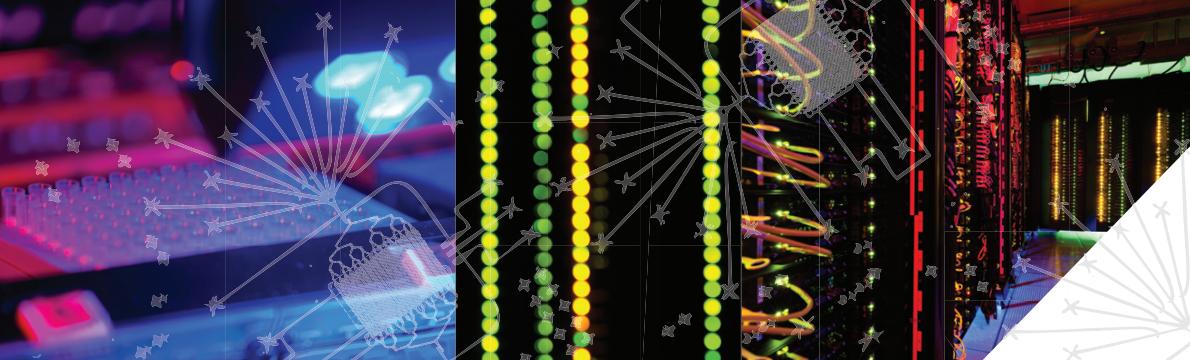
ntEdit: scalable genome sequence polishing

2019

René L Warren, Lauren Coombe, Hamid Mohamadi, Jessica Zhang, Barry Jaquish, Nathalie Isabel, Steven JM Jones, Jean Bousquet, Joerg Bohlmann, Inanç Birol
doi: <https://doi.org/10.1101/565374>



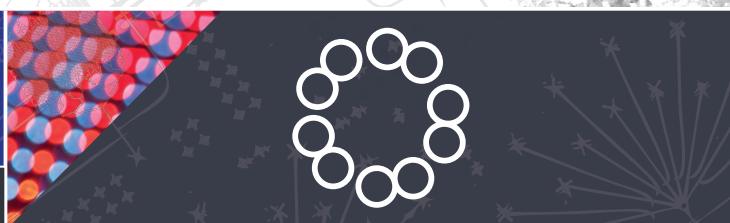
bioRxiv
THE PREPRINT SERVER FOR BIOLOGY



CANADA'S MICHAEL SMITH
GENOME
SCIENCES
CENTRE

A leading international centre for genomics and bioinformatics research committed to advancing knowledge of cancer-related diseases, improving human health through disease prevention, diagnosis and therapeutic approaches, and realizing social and economic benefits of genomics research.

>2 PETABASES SEQUENCED • A HUMAN GENOME EVERY 15 MINUTES • HIGH-PERFORMANCE COMPUTING



Bioinformatics Technology Lab



Genome British Columbia



Genome Canada



BCCA CANCER RESEARCH CENTRE

