Scalable solutions for *de novo genome* assembly

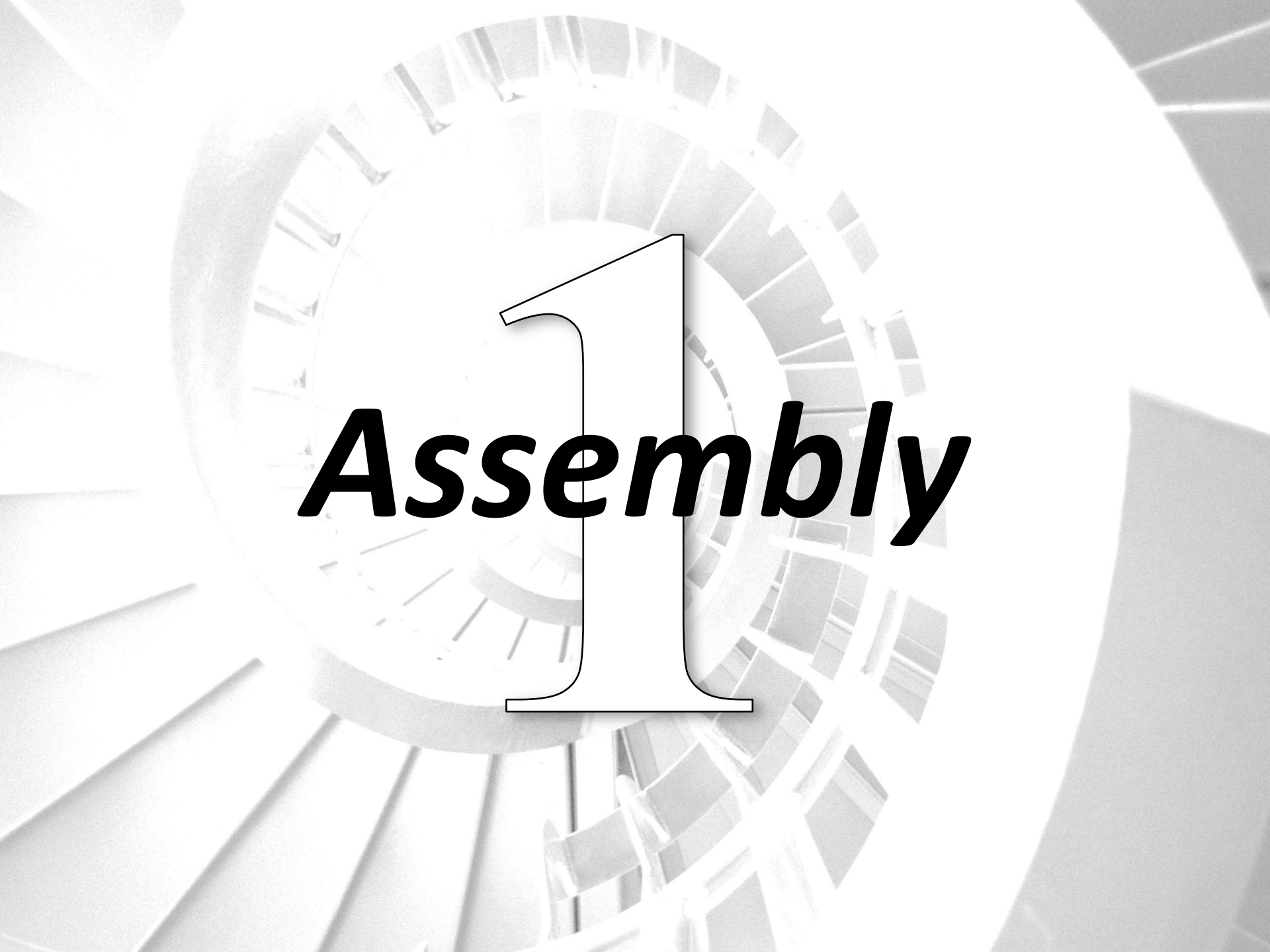
– in 5 easy steps

René L Warren 03/2019

CANADA'S MICHAEL SMITH GENOME SCIENCES CENTRE

bcgsc

Photo: R. Warren


1
Assembly

ABySS

de novo genome assembly with short reads

2009 : **Parallel DBG assembler**

MPI to aggregate memory

Assembled 20Gb spruce genome

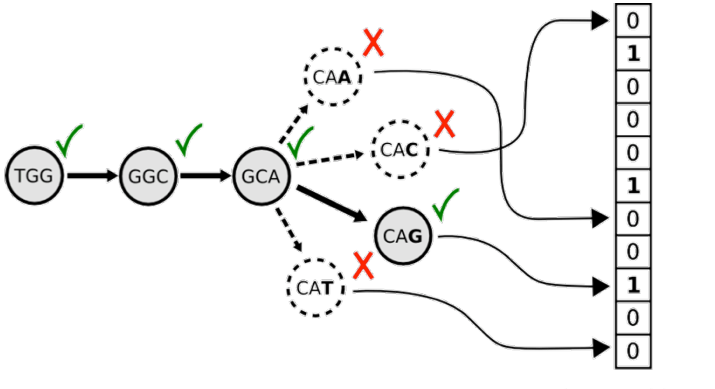
GENOME RESEARCH
Resource

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 Birol[2]

2017 : **Bloom filter representation**

1/10th RAM, single computer, scalable to 20 Gbp spruce



Bloom filter

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 2

Tigmint

Jackman *et al. BMC Bioinformatics* (2018) 19:393
https://doi.org/10.1186/s12859-018-2425-6

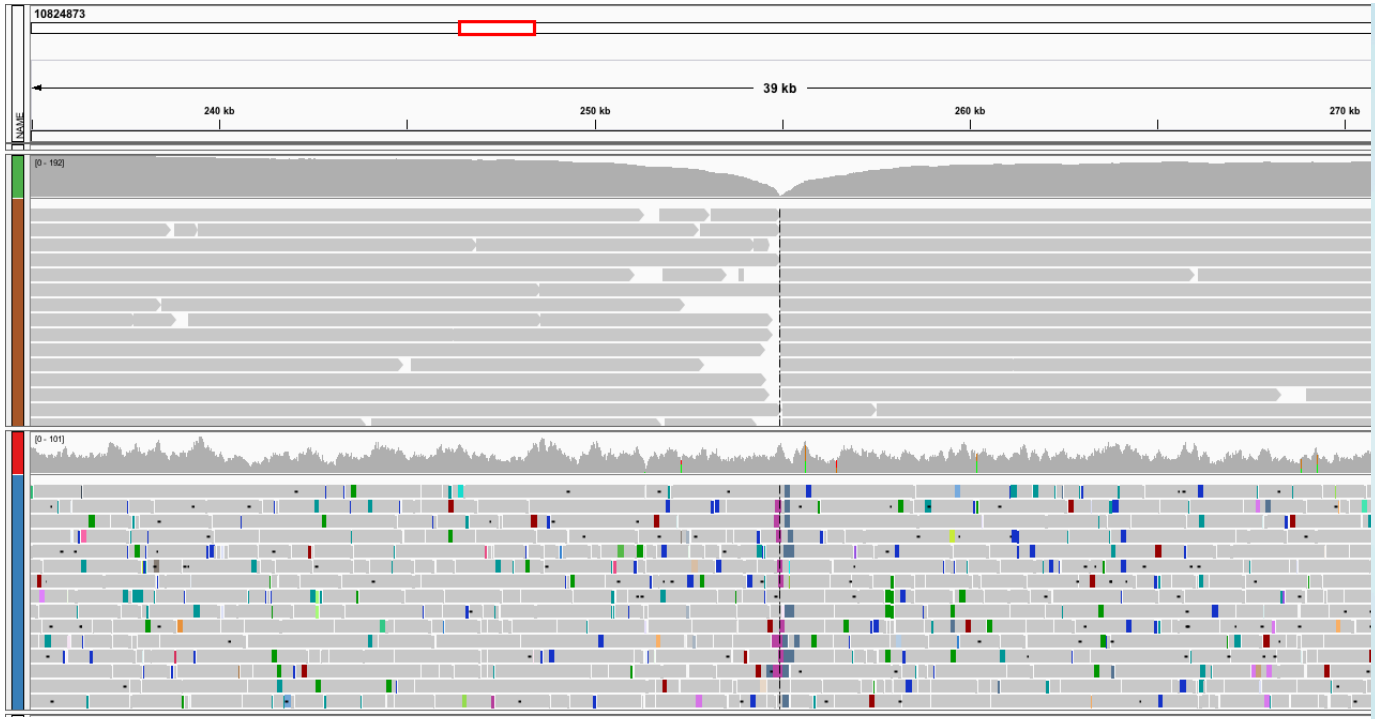
BMC Bioinformatics

SOFTWARE Open Access

Tigmint: correcting assembly errors using linked reads from large molecules

Shaun D. Jackman[1*], Lauren Coombe[1], Justin Chu[1], Rene L. Warren[1], Benjamin P. Vandervalk[1], Sarah Yeo[1], Zhuyi Xue[1], Hamid Mohamadi[1], Joerg Bohlmann[2], Steven J.M. Jones[1] and Inanc Birol[1]

linked reads misassembly correction



IGV screenshot: Tigmint breakpoint in human genome NA24143

https://github.com/bcgsc/tigmint


3
Scaffolding

LINKS

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

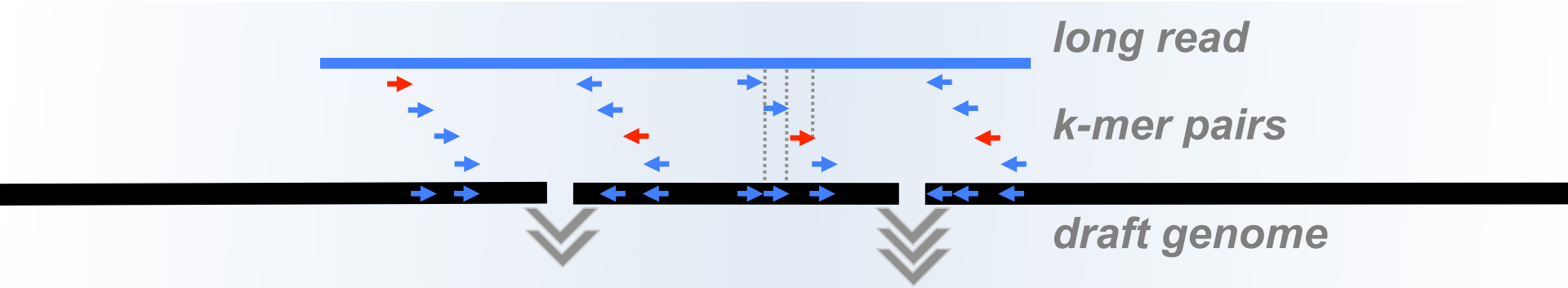
(GIGA)n SCIENCE

RESEARCH Open Access

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

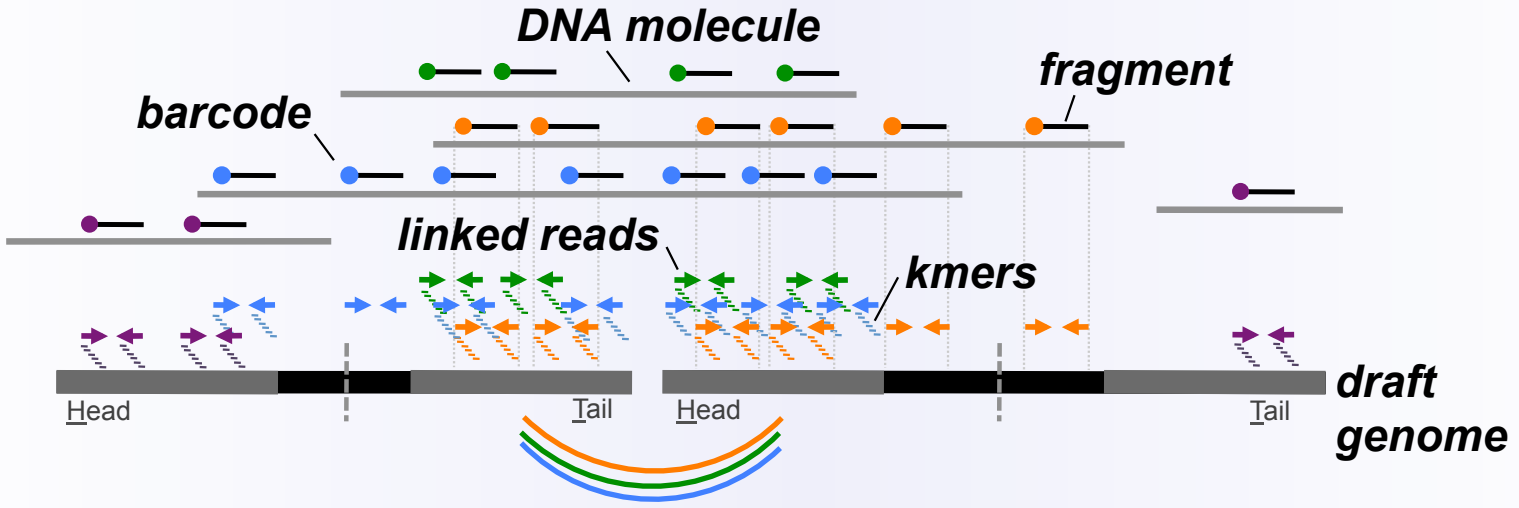


- **Scaffolder** : order & orient sequences
- ***k-mer* based** : no alignments
- **Vast k-mer space** : no fragment length limitation
- **Versatile** : long-reads, draft sequences, MPET

≠ length **≠ errors** **⃠ base correction**

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

https://github.com/bcgsc/arcs

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/arks

4 Gap-filling

Scaffolding and gap-filling

Uses LINKS scaffolding algorithm

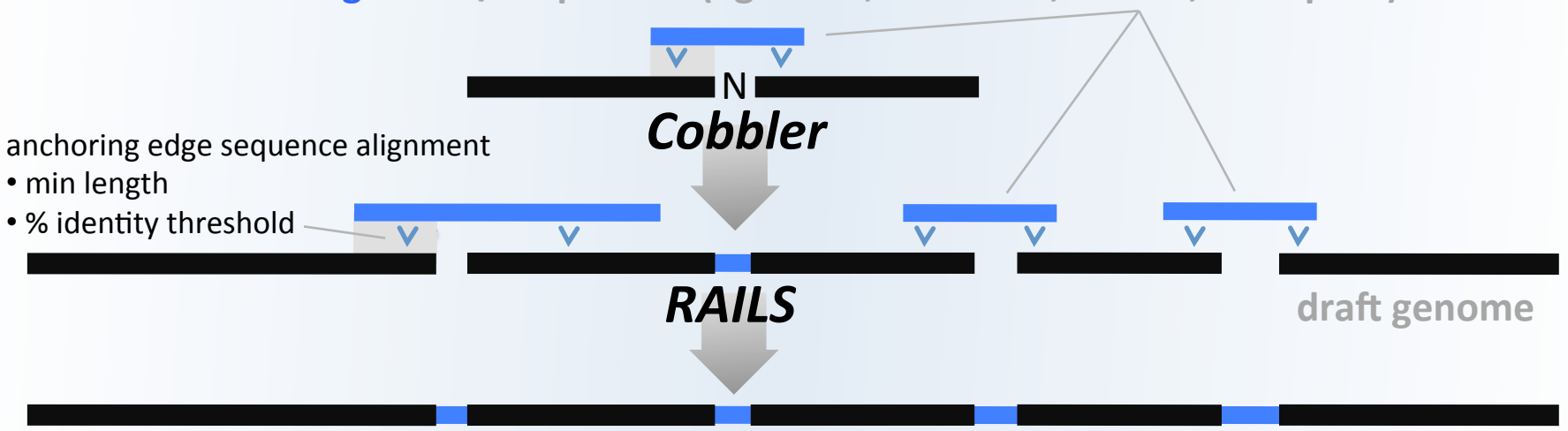
JSS The Journal of Open Source Software

RAILS and Cobbler: Scaffolding and automated finishing of draft genomes using long DNA sequences

Rene L Warren[1] 2016

1 BC Cancer Agency, Genome Sciences Centre, Vancouver, BC, Canada

Long reads / sequences (eg. Draft, Moleculo, PacBio, Nanopore)



Gaps filled with sequence read with most anchoring bases to the draft

https://github.com/bcgsc/rails

Sealer

Automated genome finishing

- Gap-filler (resolve *Ns*)
- Implements Bloom filter de Bruijn graph

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

BMC Medical Genomics

RESEARCH Open Access

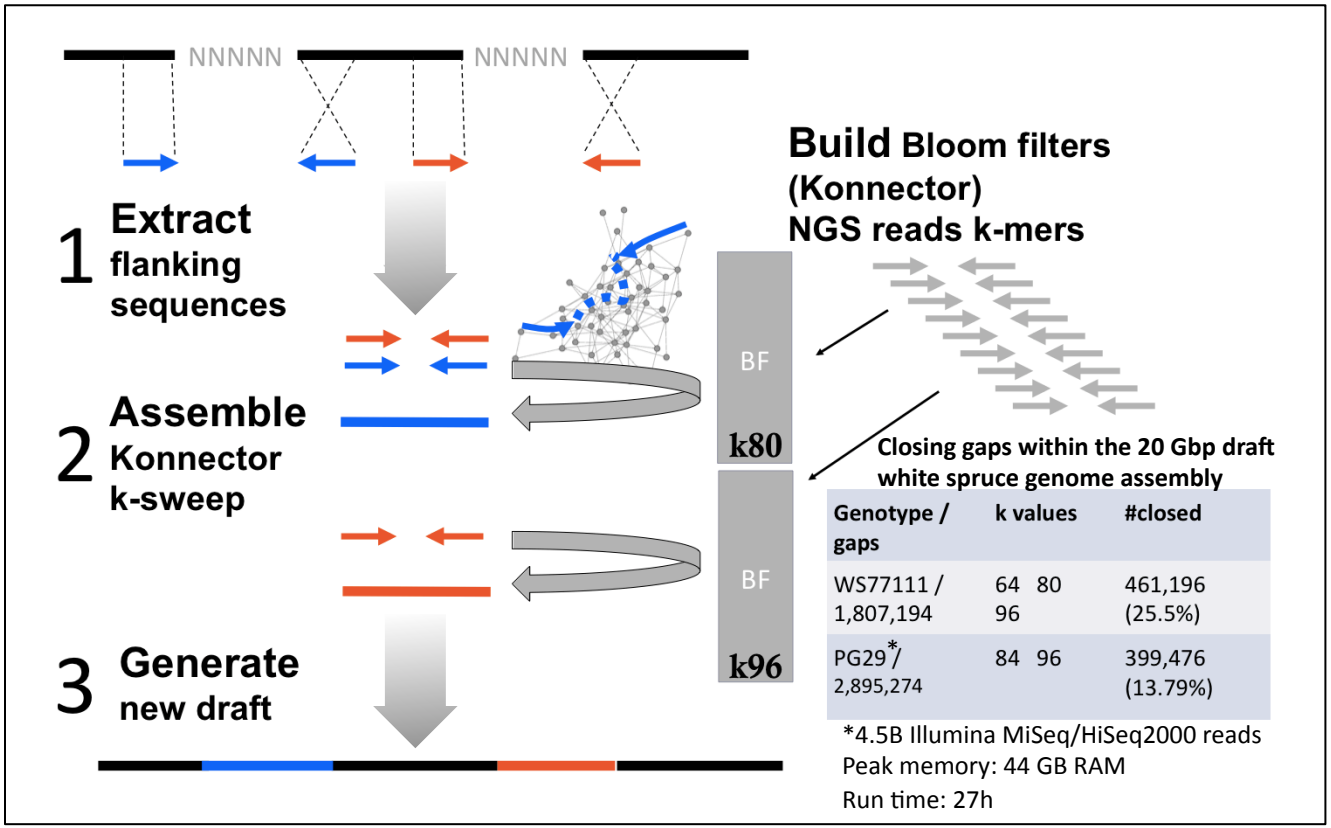
Konnector v2.0: pseudo-long reads from paired-end sequencing data

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

BMC Bioinformatics

SOFTWARE Open Access

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



Closing gaps within the 20 Gbp draft white spruce genome assembly

Genotype / gaps	k values	#closed
WS77111 / 1,807,194	64 80 96	461,196 (25.5%)
PG29* / 2,895,274	84 96	399,476 (13.79%)

*4.5B Illumina MiSeq/HiSeq2000 reads
Peak memory: 44 GB RAM
Run time: 27h

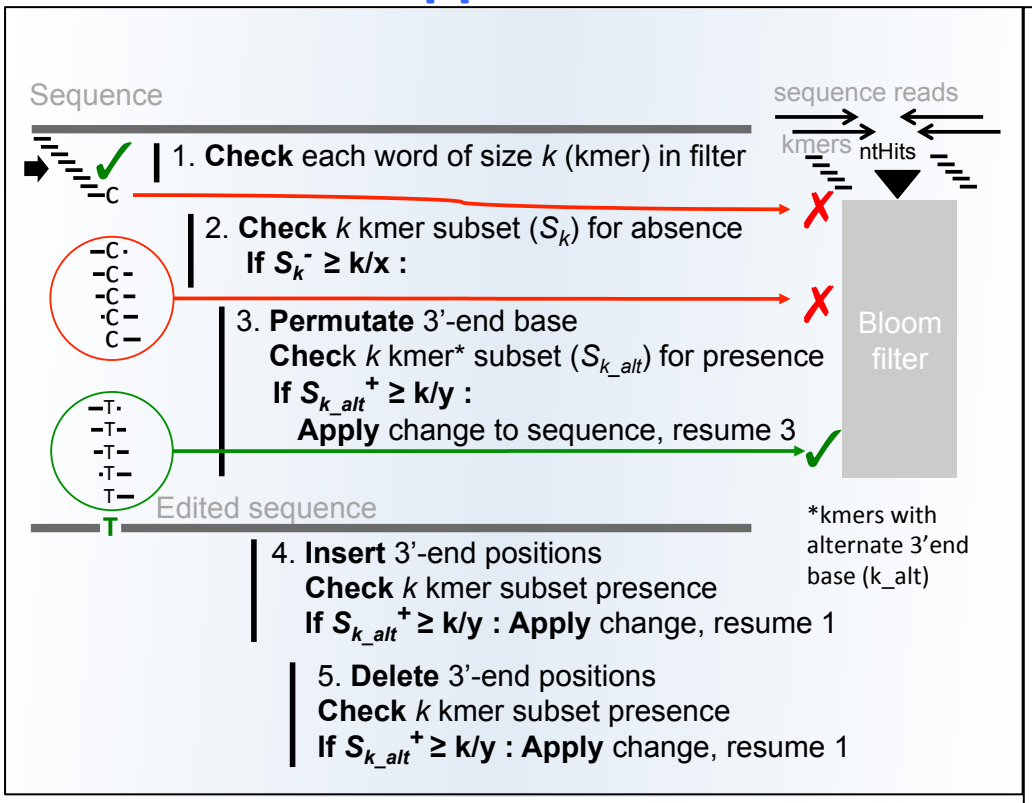
https://github.com/bcgsc/abyss/tree/master/Sealer


5
Polishing

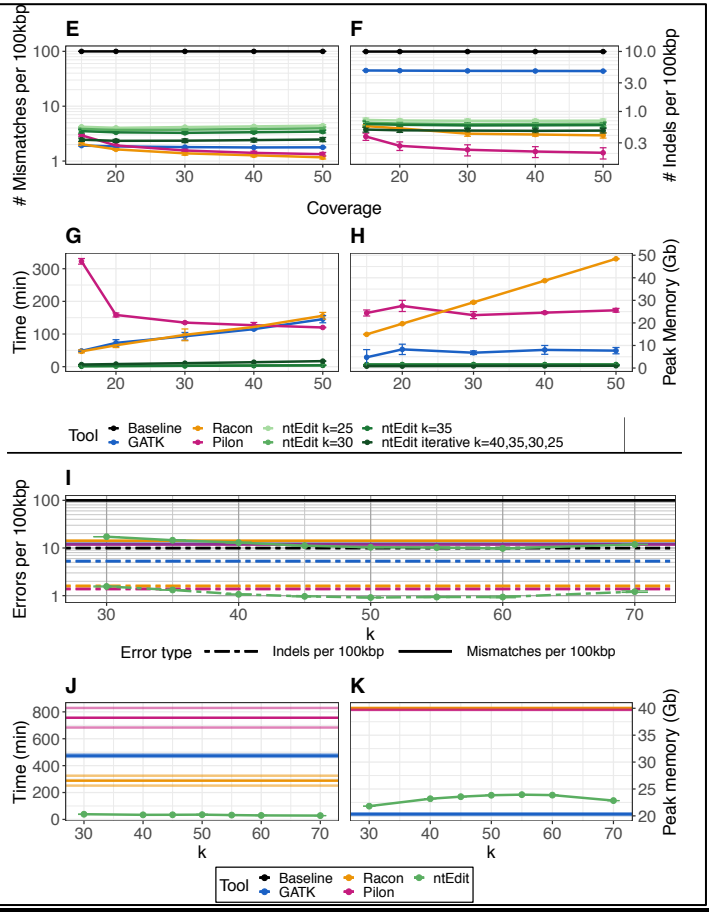
NtEdit

Fast homozygous error correction / genome "haploidization"

Approach



Results



C. elegans

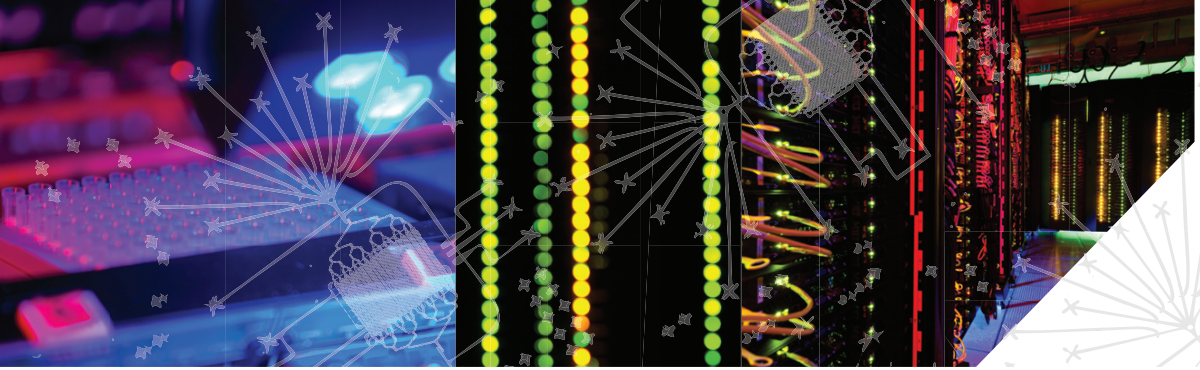
H. Sapiens chr21

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

https://github.com/bcgsc/ntedit



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

AFFILIATIONS BC Cancer Research Center • BC Cancer Agency • BC Cancer Foundation • Genome BC • Simon Fraser University • University of Britsh Columbia • Genome Sciences Institute

Bioinformatics Technology Lab

www.birol-lab.ca
https://github.com/bcgsc

BC CANCER

NIH

GenomeBritishColumbia

GenomeCanada

intel

John Jambor Knowledge Fund

Photo: R. Warren