

Genome Assembly Toolkit

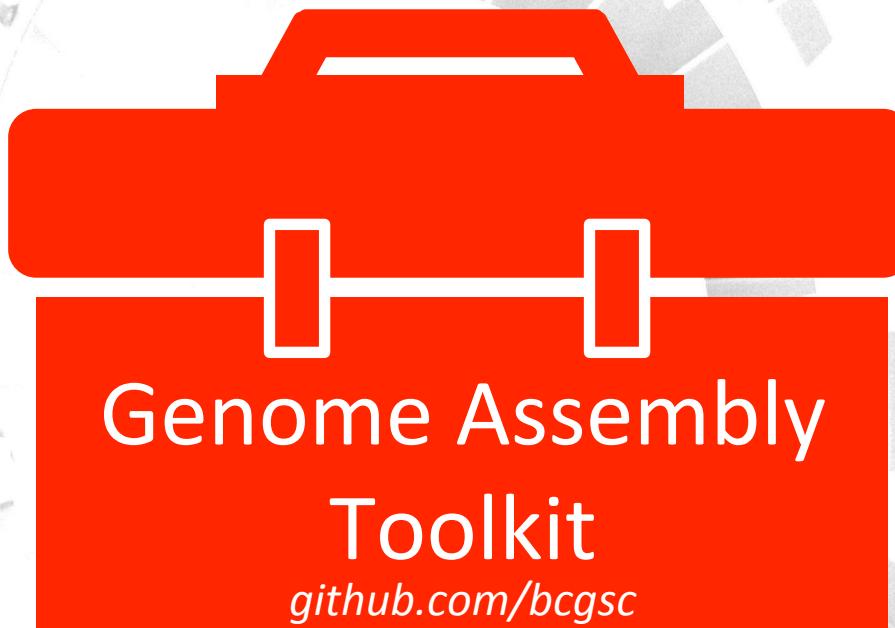


René L Warren
Bioinformatics Technology Lab
2018

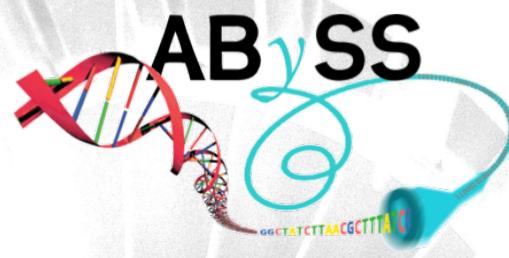
LINKS scaffold graph / *E. coli* K12



CANADA'S MICHAEL SMITH
**GENOME
SCIENCES
CENTRE**
WWW.BCGSC.CA



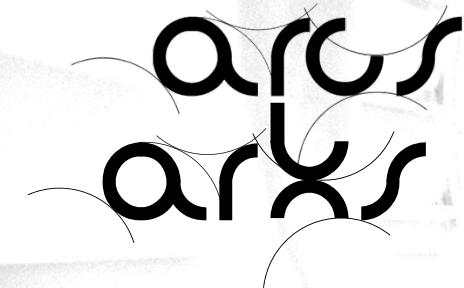
Tigmint



BBT



KOLLECTOR



LINKS

►Konnector

&

Sealer

MATCHVIEW



De novo Assembly



de novo genome assembly with short reads

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

Journal Article

Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data

[İnanç Birol, Anthony Raymond, Shaun D. Jackman, Stephen Pleasance, Robin Cope ...](#)

Bioinformatics, Volume 29, Issue 12, 15 June 2013, Pages 1492–1497,
<https://doi.org/10.1093/bioinformatics/btt178>

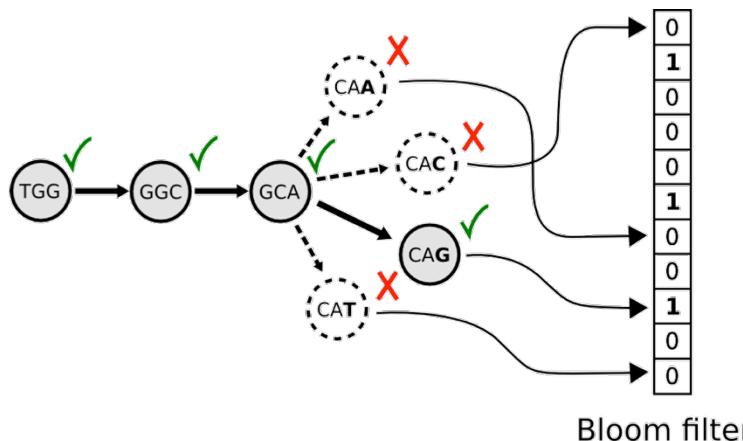
2009: Parallel DBG assembler

MPI to aggregate memory

Assembled 20Gb spruce genome

2017: Bloom filter representation

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



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 İnanç Birol

KOLLECTOR

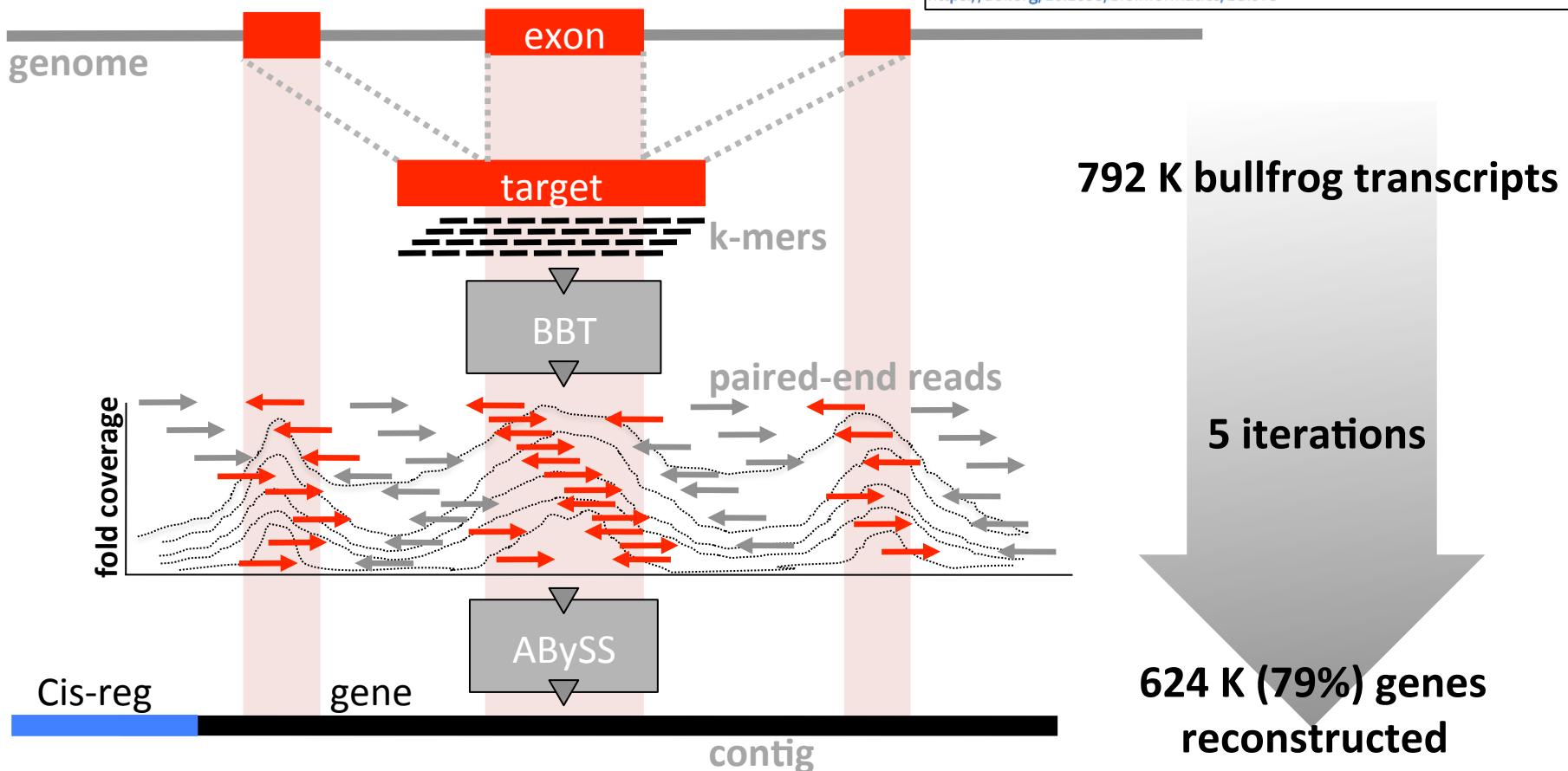
Targeted *de novo* assembly of gene loci

Using a progressive Bloom filter

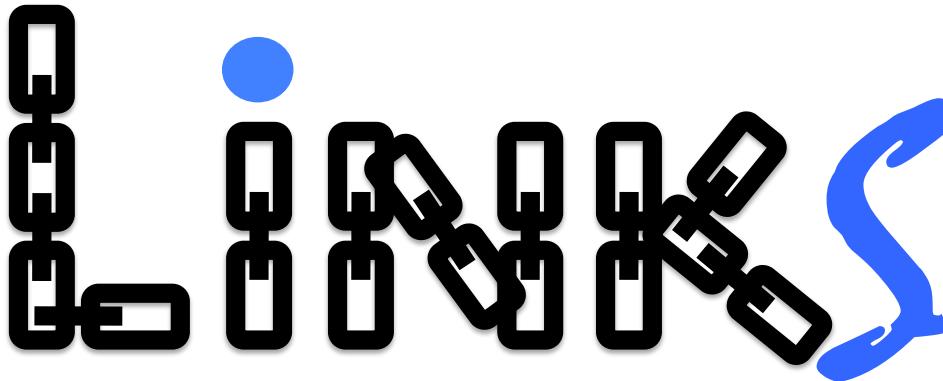
Journal Article

Kollector: transcript-informed, targeted de novo assembly of gene loci [@](#)
Erdi Kucuk, Justin Chu, Benjamin P. Vandervalk, S. Austin Hammond, René L. Warren ...

Bioinformatics, Volume 33, Issue 12, 15 June 2017, Pages 1782–1788,
<https://doi.org/10.1093/bioinformatics/btx078>



Scaffolding



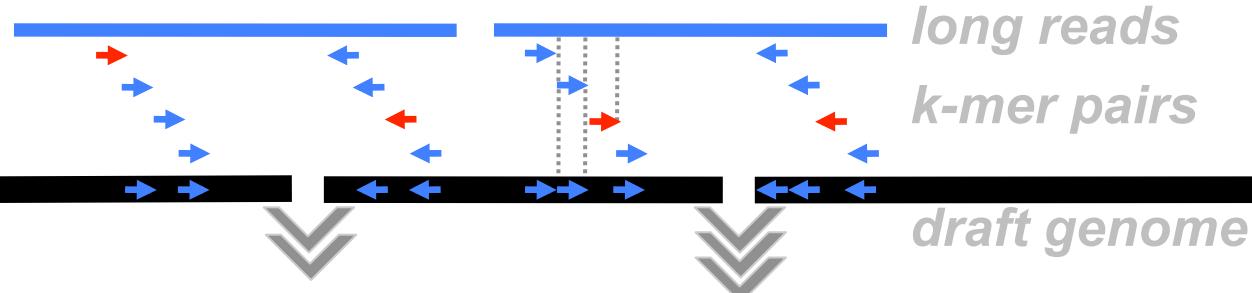
RESEARCH

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 Biröldü

Long read kmer scaffolding

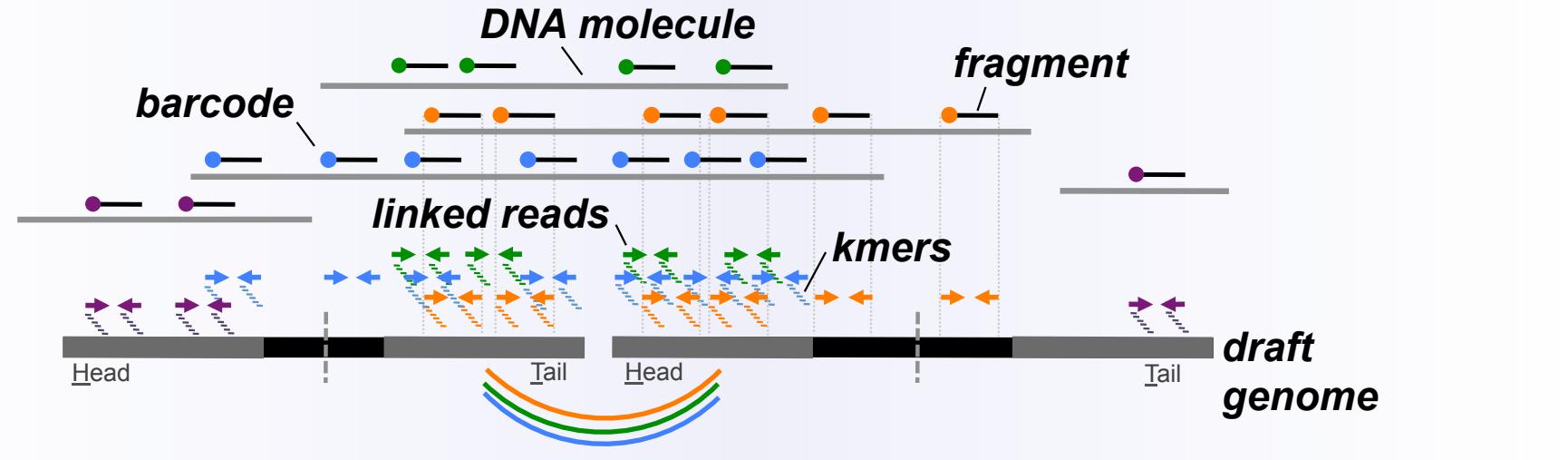
- **Scaffolder** : order & orient sequences
- ***k-mer* based** : no alignments
- **Vast *k-mer* space** : no fragment length limitation
- **Versatile** : long-reads, draft sequences, MPET
- **# length**
- **# errors**
- **🚫 base correction**



arcs

arks

Linked read scaffolding



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

BMC Bioinformatics

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>

SOFTWARE

Open Access

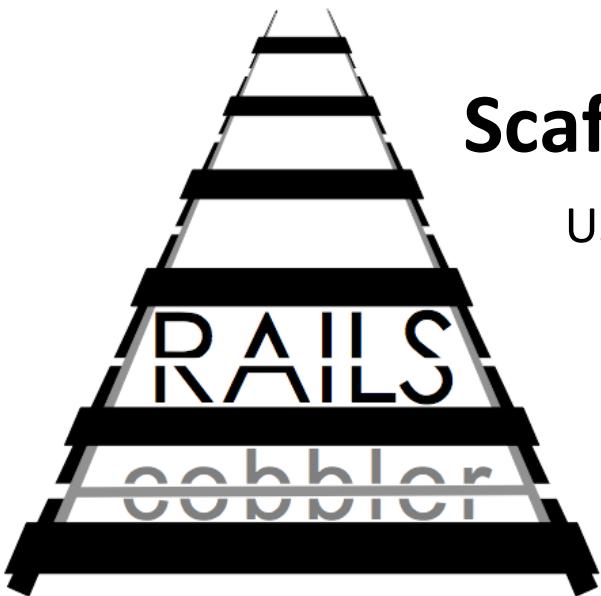
ARKS: chromosome-scale scaffolding of human genome drafts with linked read kmers



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



Gap-filling



Scaffolding and gap-filling

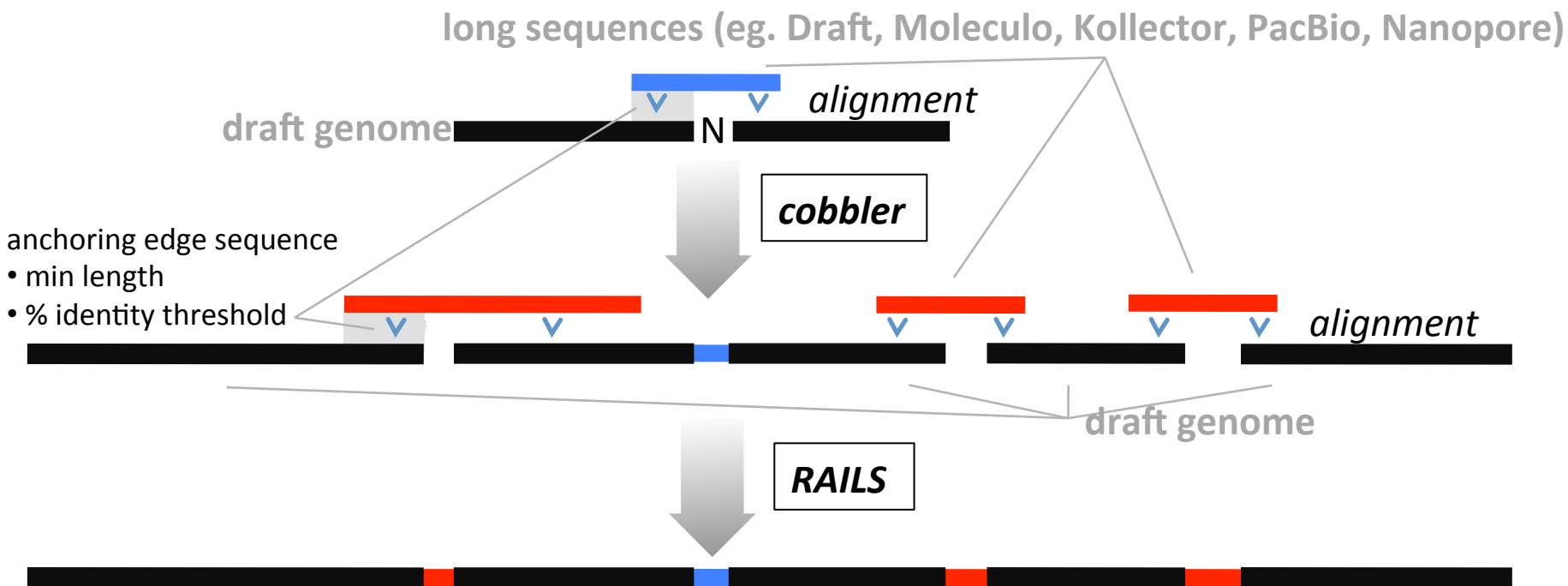
Uses LINKS scaffolding algorithm



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

Rene L Warren¹ 2016

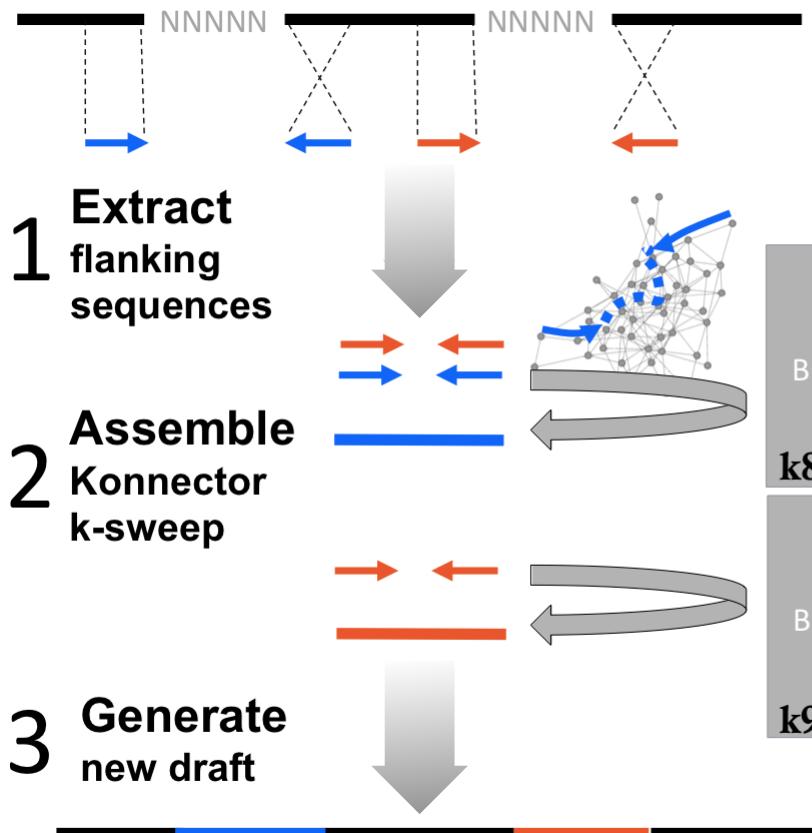
¹ BC Cancer Agency, Genome Sciences Centre, Vancouver, BC, Canada



Sealer

Automated genome finishing

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



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



SOFTWARE

Open Access

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



Application of Konnector

Build Bloom filters
(Konnector)
NGS reads k-mers

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

Misassembly Correction & Assessment

Tigmint

Linked-Read
Misassembly
Correction



IGV screenshot: a Tigmint breakpoint in human genome NA24143

Pre-print: Jackman et al., “Tigmint: Correcting Assembly Errors Using Linked Reads From Large Molecules”, bioRxiv, <https://doi.org/10.1101/304253>.

Bio Bloom Tools

Sequence classification with Bloom filters

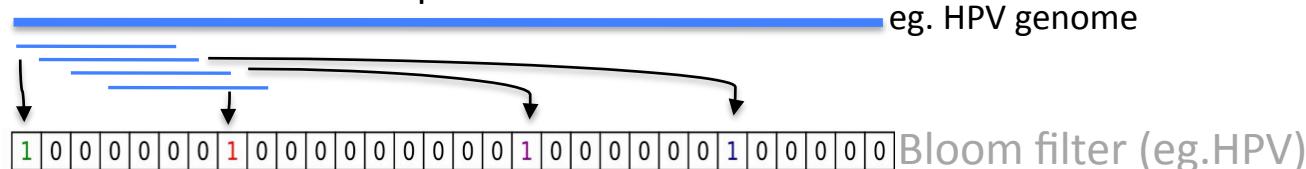
Sequence filtering

contaminant screening

pathogen discovery

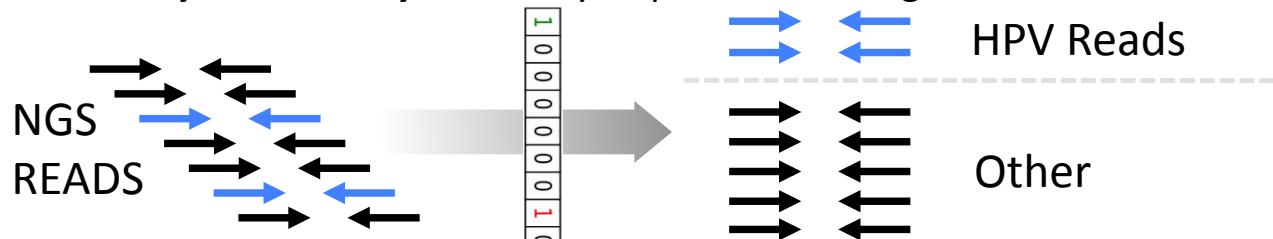
BioBloom-Maker

- Build filters : **Re-usable** loadable binary file, human readable text file from input sequences
- Customizable : **Flexible** adjust k score threshold FPR #hash functions
- Multi-filter : **Concurrent** BloomMap



Categorizer

- Bins sequences : **Analysis summary** - hits tally to particular categories

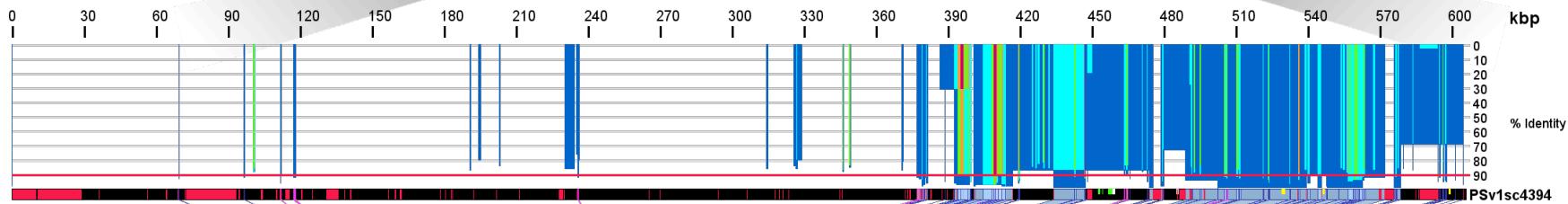


BioBloom tools: fast, accurate and memory-efficient host species sequence screening using bloom filters

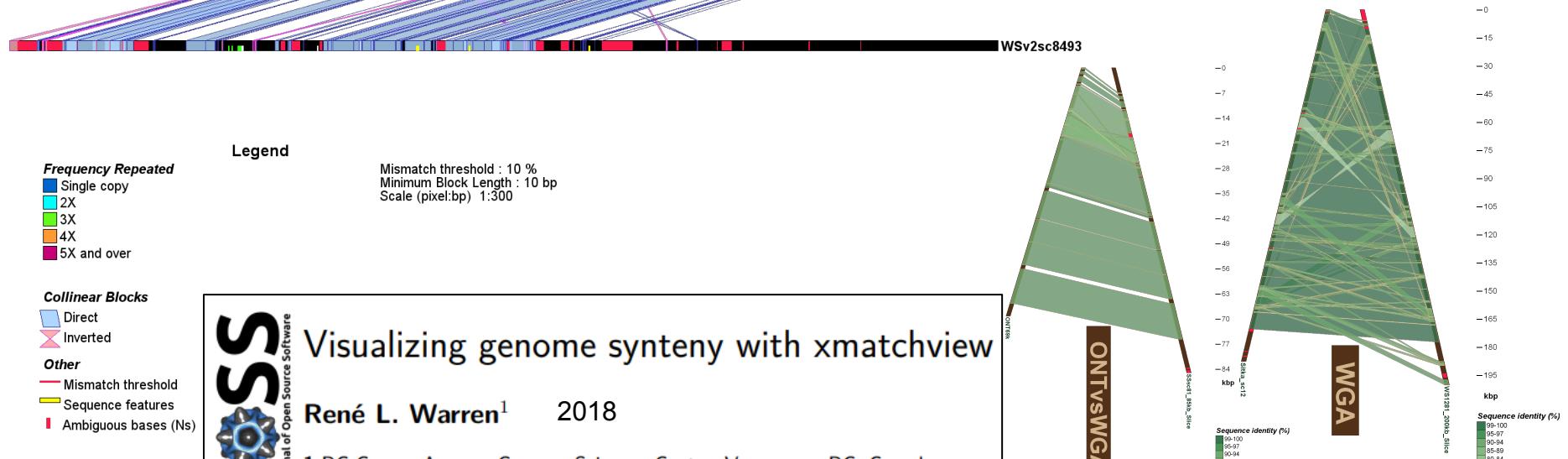
Justin Chu , Sara Sadeghi, Anthony Raymond, Shaun D. Jackman, Ka Ming Nip, Richard Mar, Hamid Mohamadi, Yaron S. Butterfield, A. Gordon Robertson, İnanç Birol
Author Notes

Bioinformatics, Volume 30, Issue 23, 1 December 2014, Pages 3402–3404,

MATCHVIEW



LG11 (←)
 83.694 GQ03115_N13 (yellow)
 83.695 GQ0045_A14 (white)
 83.695 GQ02011_E22 (lime)



b1T1 Projects

Visualization ABYSS-explorer : Visualizing assembly graphs <https://github.com/bcgsc>

QC gNAVIGATOR : Assembly completeness (cDNA) and QC (genetic map)

De novo assembly ABYSS-LR : Linked-read *de novo* assembler
TAILR : Targeted, hybrid *de novo* assembler informed by LR
ONTig : Nanopore, hybrid *de novo* assembler

Analysis PAVfinder : Structural variant finder (genome/transcriptomes)

Comparative ABYSS-Bloom : Comparative genomics with kmer Bloom filters

Reads NanoSim : Nanopore read simulator, models on ONT data
DIDA : Distributed Indexing & alignment on a compute farm

RNA Chop-Stitch : Exon annotation, splice graph construction

Trans-ABYSS : Transcriptome Assembler with short reads

RNA-Bloom : Resource-efficient transcriptome assembler

KLEAT : Analysis of APA events using transcriptomes

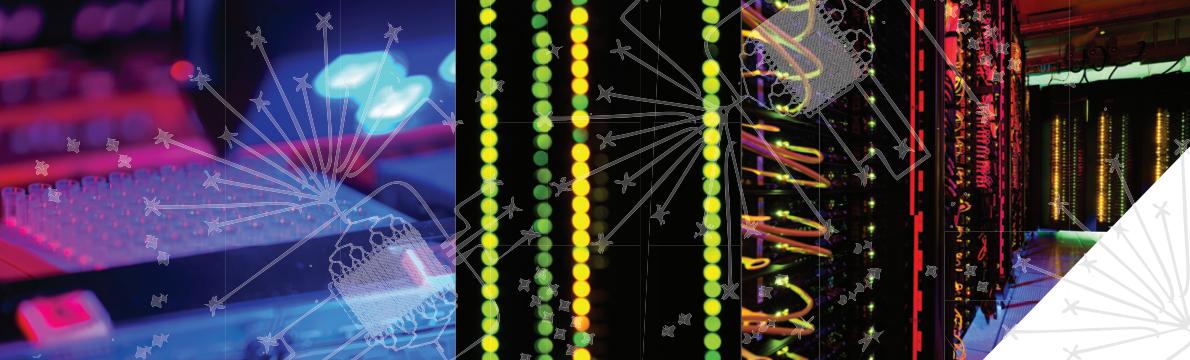
TransNanoSim : Nanopore transcriptome simulator

Data structure miBF : Multi-Index Bloom Filters

Algorithms ntHash : Fast nucleotide sequence hashing

ntCard : kmer cardinality estimations

ntHit : kmer repeat detection

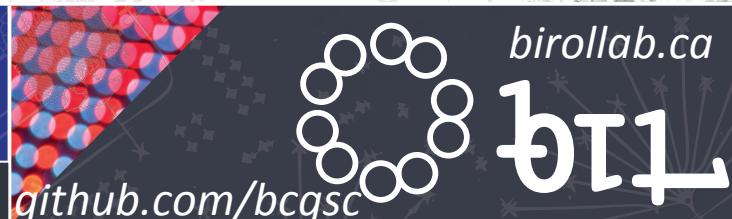


CANADA'S MICHAEL SMITH

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 10 MINUTES • HIGH-PERFORMANCE COMPUTING



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

**Lauren Coombe | Austin Hammond | Ben Vandervalk | Zhuyi Xue | Readman Chiu
Darcy Sutherland | Jessica Zhang | Jeffrey Tse | Justin Chu | Shaun Jackman
Kristina Gagalova | Chen Yang | Saber Hafezqorani | Golnar Sheikhshab
Ka Ming Nip | Yee Fay Lim | Chenkai Li | Hamid Mohamadi**

Rene Warren (rwarren@bcgsc.ca) | Sinead Aherne | Inanc Birol

Tony Raymond | Daniel Paulino | Sarah Yeo | Erdi Kucuk | Hamza Khan

**BC
CAN
CER**

NIH

GenomeBritishColumbia

GenomeCanada

intel

Spruce Up

John Jambor Knowledge Fund

BCCA CANCER RESEARCH CENTRE

