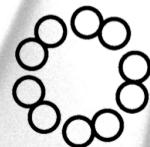


A toolkit for *de novo* assembly

LINKS scaffold graph / *E. coli* K12

René L Warren, 2018



CANADA'S MICHAEL SMITH
GENOME
SCIENCES
CENTRE

b1

Tig mint

arcs
arks

LINKS

Konnection

&

Sealer



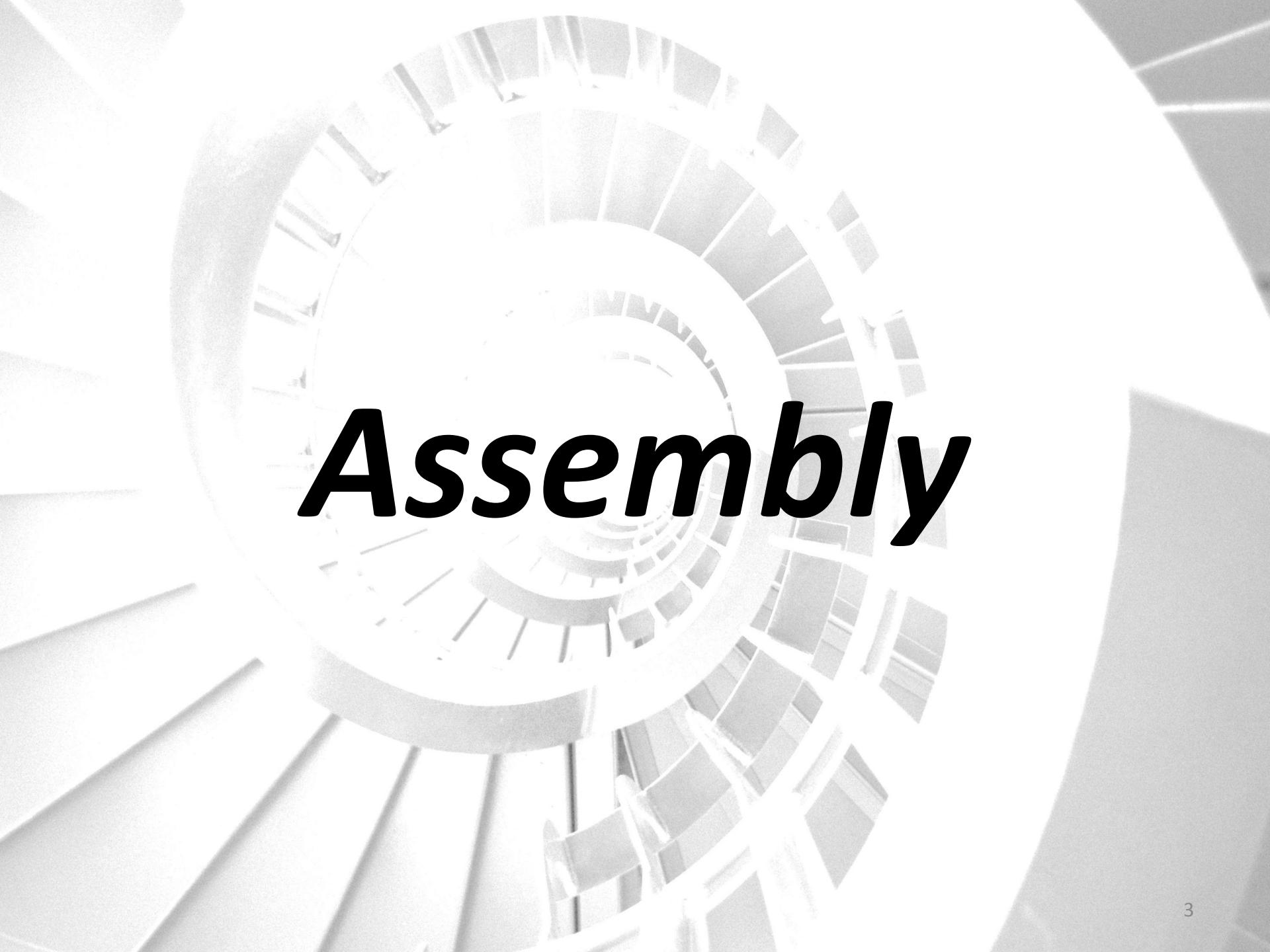
BBT



KOLLECTOR

MATCHVIEW





Assembly



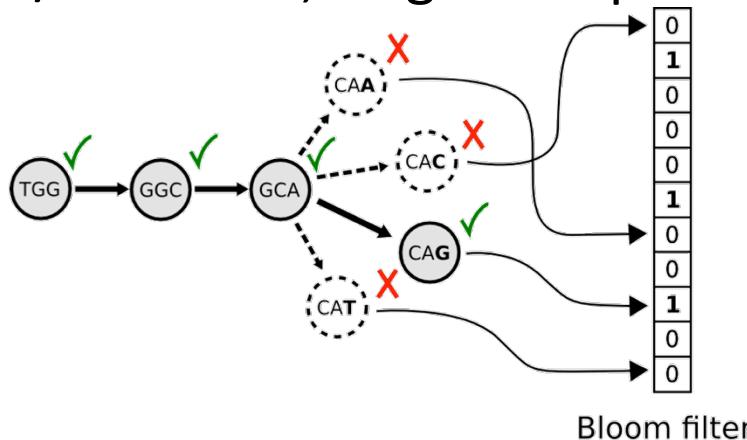
2009 : Parallel DBG assembler

MPI to aggregate memory

Assembled 20Gb spruce genome

2017 : Bloom filter representation

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



GENOME
RESEARCH
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 Inanc Birol²

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

Inanc Birol, Anthony Raymond, Shaun D. Jackman, Stephen Pleasance, Robin Coope ...

Bioinformatics, Volume 29, Issue 12, 15 June 2013, Pages 1492–1497,

<https://doi.org/10.1093/bioinformatics/btt178>

PLOS

DIDA: Distributed Indexing Dispatched Alignment

Hamid Mohamadi, Benjamin P Vandervalk, Anthony Raymond, Shaun D Jackman, Justin Chu, Clay P Breshears, Inanc Birol

Published: April 29, 2015 • <https://doi.org/10.1371/journal.pone.0126409>

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

KOLLECTOR

Targeted *de novo* assembly of gene loci

Using a progressive Bloom filter

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,

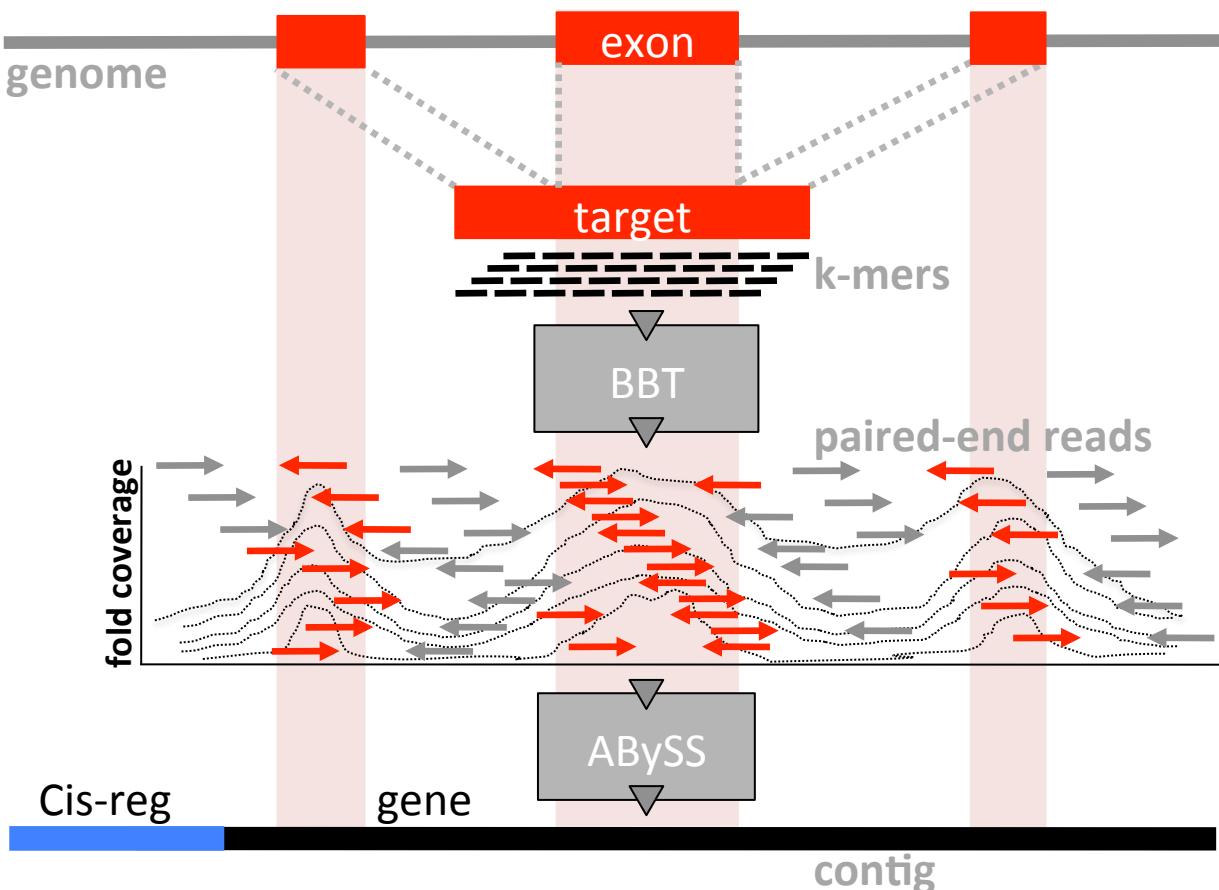


Article | OPEN | Published: 10 November 2017

The North American bullfrog draft genome provides insight into hormonal regulation of long noncoding RNA

S. Austin Hammond, René L. Warren, Benjamin P. Vandervalk, Erdi Kucuk, Hamza Khan, Ewan A. Gibb, Pawan Pandoh, Heather Kirk, Yongjun Zhao, Martin Jones, Andrew J. Mungall, Robin Coope, Stephen Pleasance, Richard A. Moore, Robert A. Holt, Jessica M. Round, Sara Ohora, Branden V. Walle, Nik Veldhoen, Caren C. Helbing & Inanc Birol 

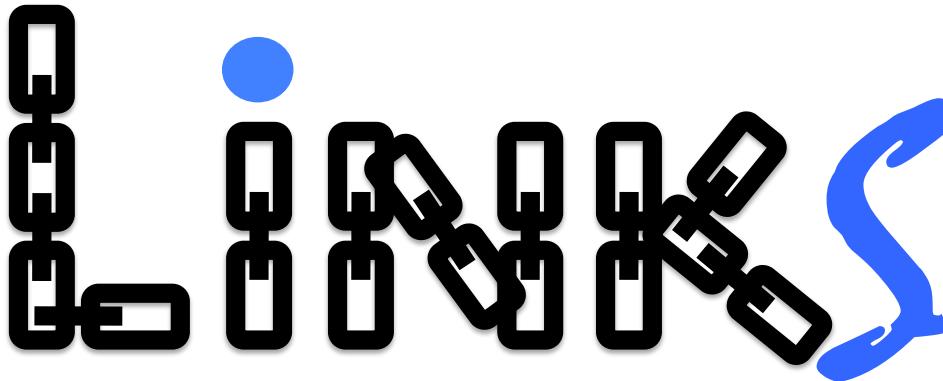
792 K bullfrog transcripts



5 iterations

624 K (~80%) genes
reconstructed

Scaffolding



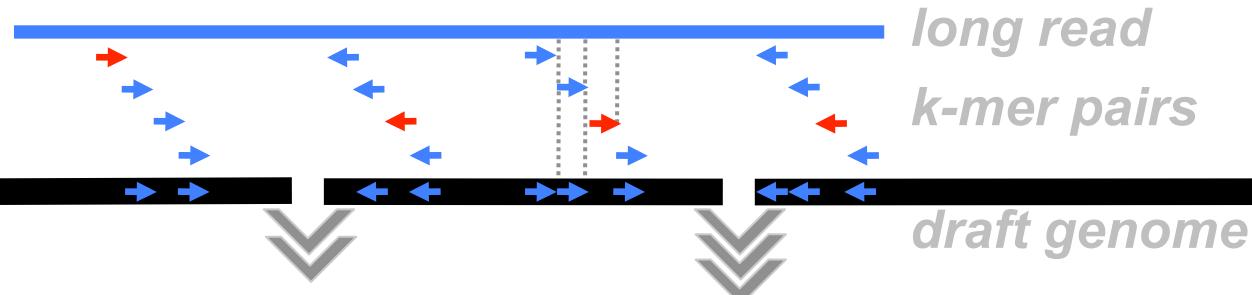
RESEARCH

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 Birol

Long read kmer scaffolding

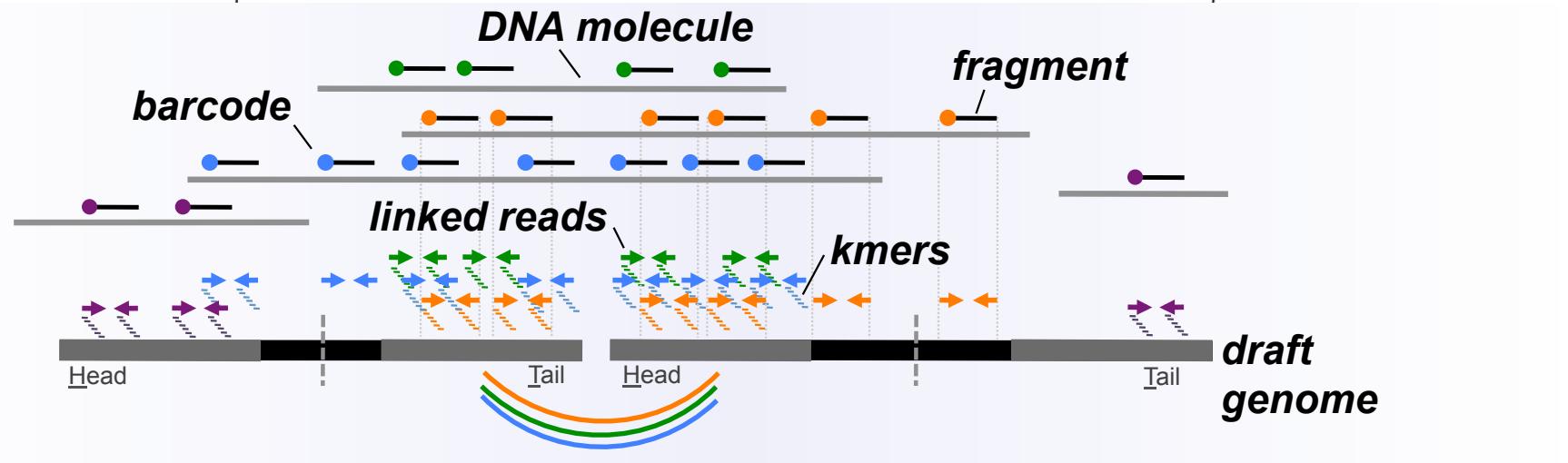
- **Scaffolder** : order & orient sequences
- ***k-mer* based** : no alignments
- **Vast *k-mer* space** : no fragment length limitation
- **Versatile** : long-reads, draft sequences, MPET
- **# length**
- **# errors**
- **🚫 base correction**



arcs

arks

Linked read scaffolding



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

BMC Bioinformatics

SOFTWARE

Open Access

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>

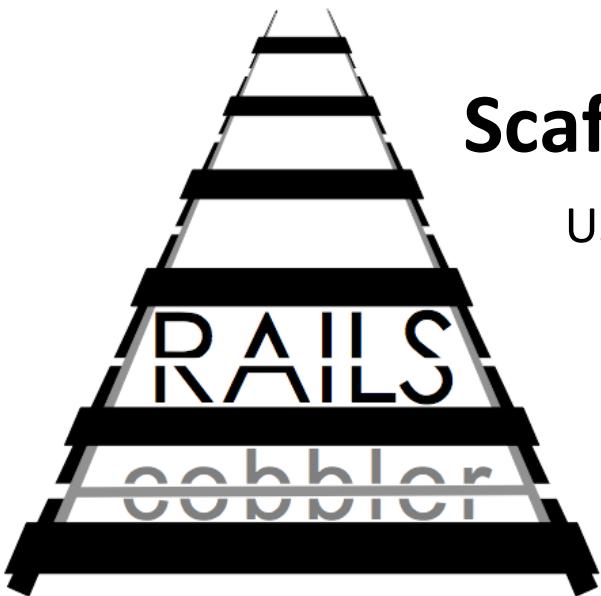
ARKS: chromosome-scale scaffolding of human genome drafts with linked read kmers



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



Gap-filling



Scaffolding and gap-filling

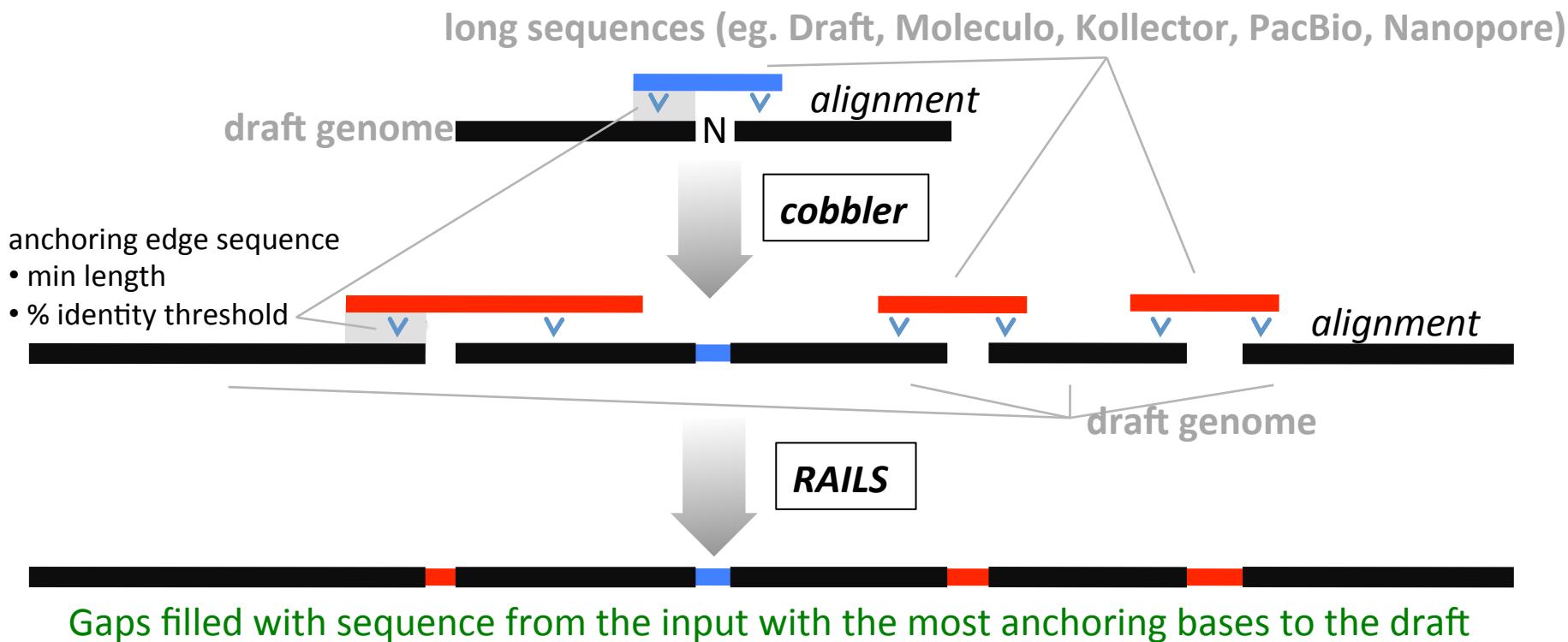
Uses LINKS scaffolding algorithm



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

Rene L Warren¹ 2016

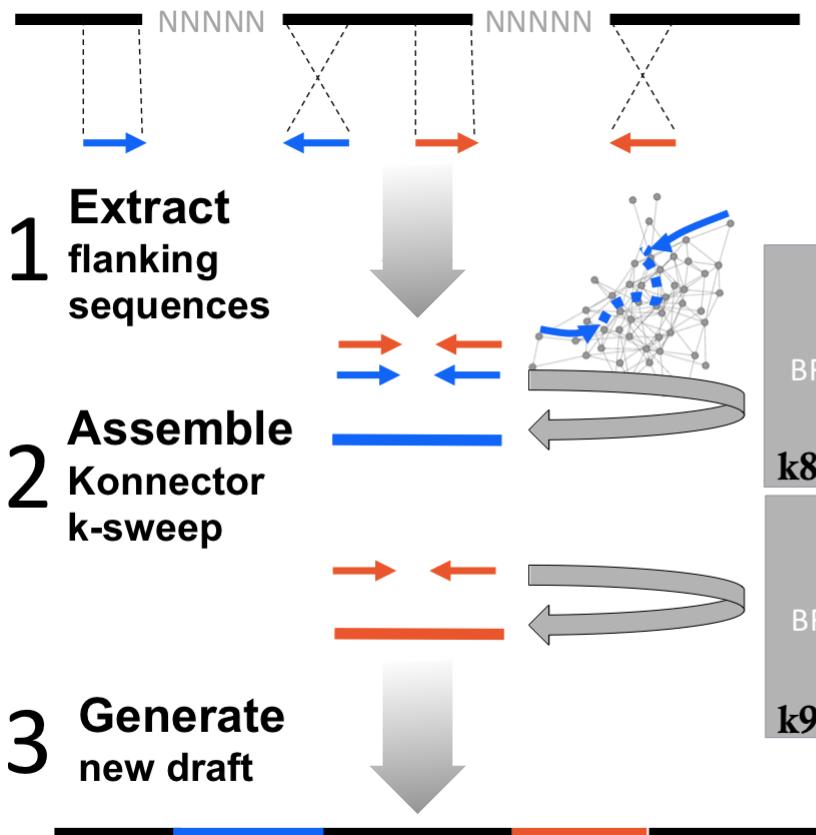
¹ BC Cancer Agency, Genome Sciences Centre, Vancouver, BC, Canada



Sealer

Automated genome finishing

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



RESEARCH

Open Access

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



Application of Konnecter

Build Bloom filters
(Konnecter)
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

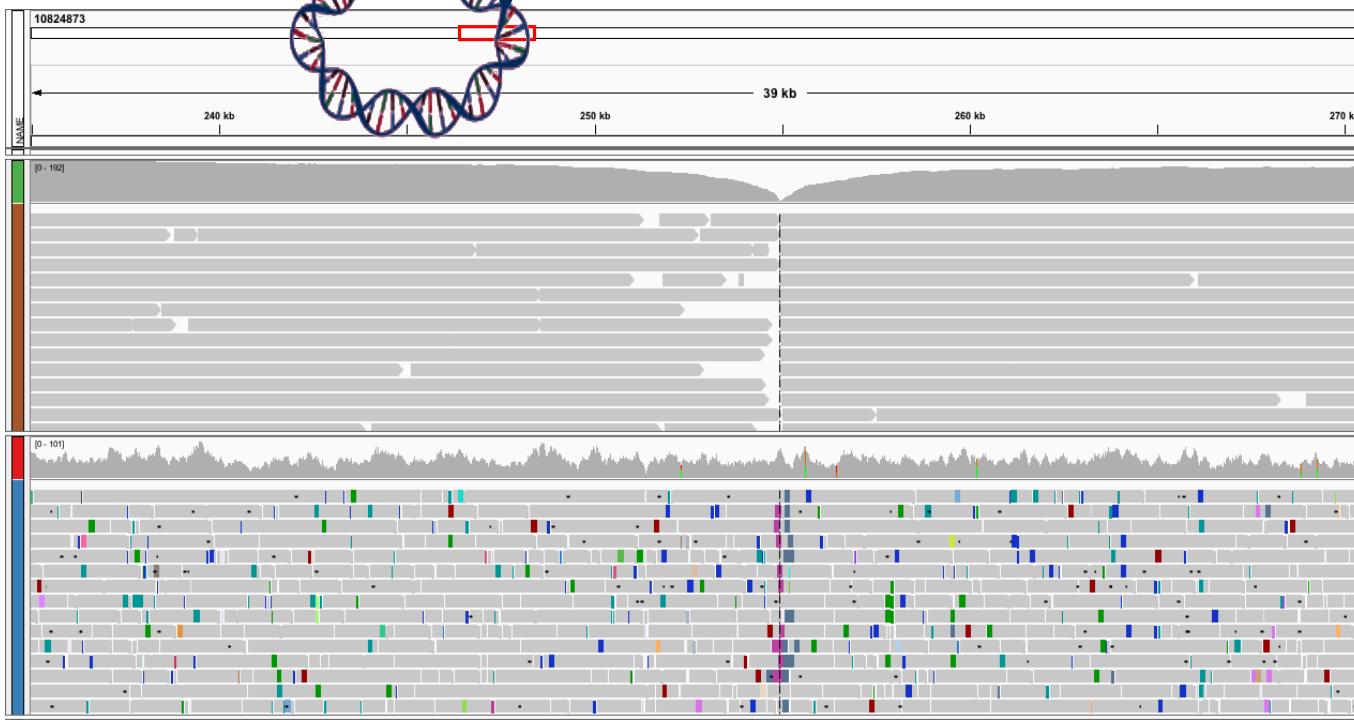
Run time: 27h



Analysis

Tigmint

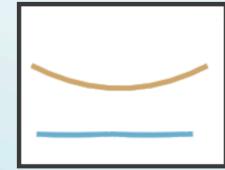
misassembly
correction with
linked reads



Correct misassemblies



Scaffold



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

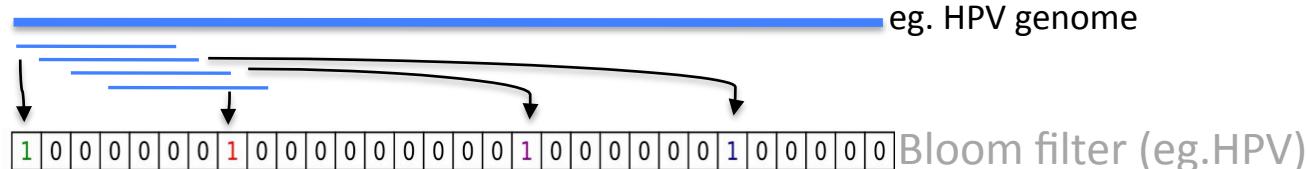
Sequence filtering

contaminant screening

pathogen discovery

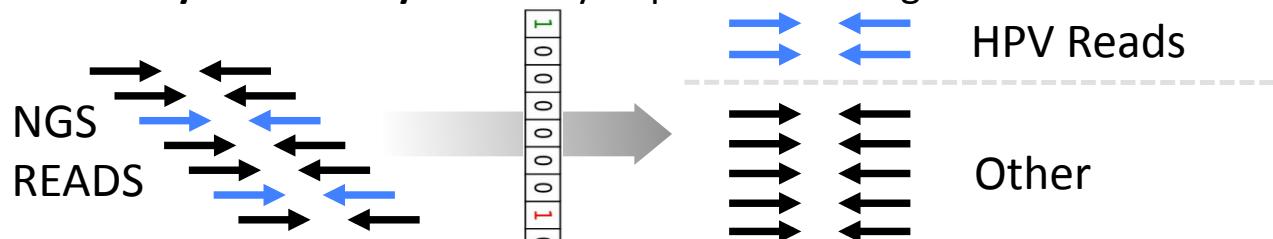
BioBloom-Maker

- 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** MiBF



Categorizer

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

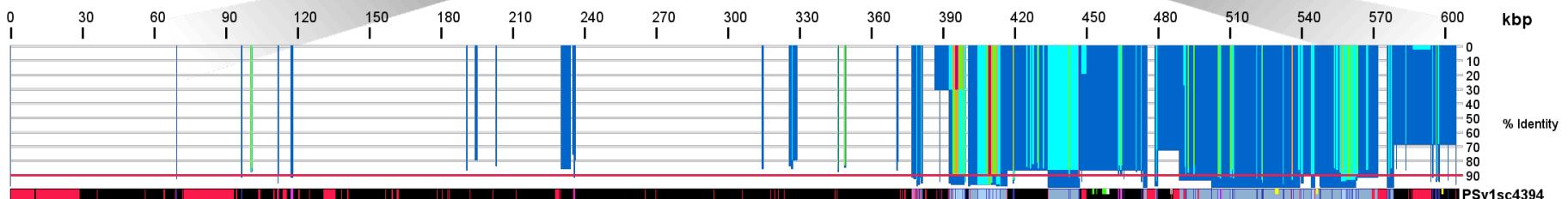


BioBloom tools: fast, accurate and memory-efficient host species sequence screening using bloom filters

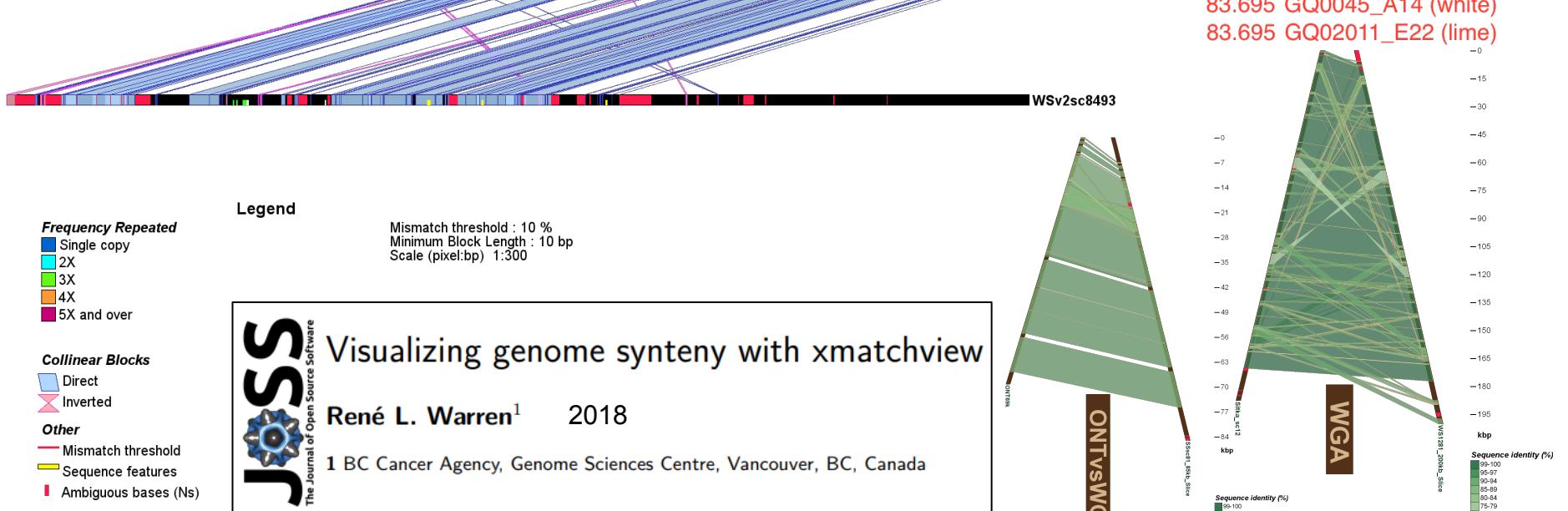
Justin Chu , Sara Sadeghi, Anthony Raymond, Shaun D. Jackman, Ka Ming Nip, Richard Mar, Hamid Mohamadi, Yaron S. Butterfield, A. Gordon Robertson, Inanç Birol
Author Notes

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)



Additional **b1**T Technologies

Visualization	ABySS-explorer	: Visualizing assembly graphs
QC	gNAVIGATOR	: Assembly completeness (cDNA) and QC (genetic map)
De novo assembly	ABYSS-LR	: Linked read <i>de novo</i> assembler
	TAILR	: Targeted, hybrid <i>de novo</i> assembler informed by linked read
	ConnectTig	: Nanopore-short read hybrid <i>de novo</i> assembler
Analysis	PAVfinder	: Structural variant finder (genomes/transcriptomes)
	AMPlify	: Antimicrobial peptide assessment with deep learning
Comparative	ABYSS-Bloom	: Comparative genomics with kmer Bloom filters
Reads	NanoSim	: Nanopore read simulator, models on experimental data
	DIDA	: Distributed Indexing & alignment on a compute farm
RNA	Chop-Stitch	: <i>De novo</i> exon annotation, splice graph construction
	Trans-ABYSS	: Transcriptome Assembler with short reads
	RNA-Bloom	: Resource-efficient transcriptome assembler
	KLEAT	: Analysis of Alternative Poly-Adenylation
	TransNanoSim	: Transcript nanopore read simulator
Data structure	miBF/BBT2	: Multi-Index Bloom Filters
Algorithms	ntHash	: Fast nucleotide sequence hashing
	ntCard	: Sequencing data cardinality estimator
	ntHit	: Sequencing data repeat detection

<https://github.com/bcgsc>

} ntpack

bt1 Funded Projects Landscape

	Granting Agency	Subject	Years	
*	Genome BC	Characterization and analysis of amphibian host defence peptides as potential antibiotics	10/2017 – 03/2019	AMPlify, AMPassayDB
*	NIH	De Novo Assembly Tools: Research with Unbiased Engines-Renewal (DNA-TRUER)	09/2017 – 07/2021	<i>de novo</i> assembly Tools, incl. ABYSS2, TAILR, ConnectTig, etc
*	NIH	De Novo Assembly Tools: Research with Unbiased Engines (DNA-TRUE)	03/2014 – 01/2017	
*	Genome BC	Automation of the Clinical Bioinformatics Pipeline at the Centre for Clinical Genomics	04/2017 – 03/2020	TAP
	Genome Canada	New bioinformatics for new sequencing technologies: Genome characterization and variation detection using long reads	10/2016 – 09/2018	TransNanosim, RNA-Bloom
*	Genome Canada	Spruce-Up: Advanced spruce genomics for productive and resilient forests	10/2016 – 09/2020	ABYSS-LR, gNavigator, XMVC
*	CIHR	Development of an automated end-to-end next generation sequencing assay to detect all classes of genetic variant in a single diagnostic test	04/2016 – 03/2019	PAVfinder
	NIH	Pan-cancer survey of candidate non-coding RNA transcripts on the cloud using a targeted <i>de novo</i> assembly approach	01/2016 – 08/2016	TASR-KLEAT
	Genome Canada	Methods and Technology Development at the Sequencing Platform at the BC Cancer Agency Genome Sciences Centre	10/2015 – 09/2017	ntPack, Konnector RAILS
	BC Cancer Foundation	Next Generation Bioinformatics for Clinical Genomics: using <i>de novo</i> assembly in personalized medicine	06/2015 – 06/2016	RNA-Bloom
	Genome BC	Advanced Assembly Assessment and Annotation	10/2014 – 03/2016	ABYSS-Bloom, Kollector
	Genome Canada	SMarTForests: Spruce Marker Technologies for Sustainable Forestry	07/2011 – 09/2015	DIDA, LINKS

*active

Translational Aspect



Robert Warren, 1979



René Warren, 2007



Noémie Warren, 2017

Article | OPEN | Published: 10 November 2017

The North American bullfrog draft genome provides insight into hormonal regulation of long noncoding RNA

S. Austin Hammond, René L. Warren, Benjamin P. Vandervalk, Erdi Kucuk, Hamza Khan, Ewan A. Gibb, Pawan Pandoh, Heather Kirk, Yongjun Zhao, Martin Jones, Andrew J. Mungall, Robin Coope, Stephen Pleasance, Richard A. Moore, Robert A. Holt, Jessica M. Round, Sara Ohora, Branden V. Walle, Nik Veldhoen, Caren C. Helbing & Inanc Birol

genes



Article

The Genome of the Beluga Whale (*Delphinapterus leucas*)

Steven J. M. Jones ^{1,2,3,*}, Gregory A. Taylor ¹, Simon Chan ¹, René L. Warren ¹, S. Austin Hammond ¹, Steven Bilobram ¹, Gideon Mordecai ^{4,5}, Curtis A. Suttle ^{4,5,6,7}, Kristina M. Miller ⁸, Angela Schulze ⁸, Amy M. Chan ^{4,5}, Samantha J. Jones ^{1,3}, Kane Tse ¹, Irene Li ¹, Dorothy Cheung ¹, Karen L. Mungall ¹, Caleb Choo ¹, Adrian Ally ¹, Noreen Dhalla ¹, Angela K. Y. Tam ¹, Armelle Troussard ¹, Heather Kirk ¹, Pawan Pandoh ¹, Daniel Paulino ¹, Robin J. N. Coope ¹, Andrew J. Mungall ¹, Richard Moore ¹, Yongjun Zhao ¹, Inanc Birol ^{1,3}, Yussanne Ma ¹, Marco Marra ^{1,3} and Martin Haulena ⁹

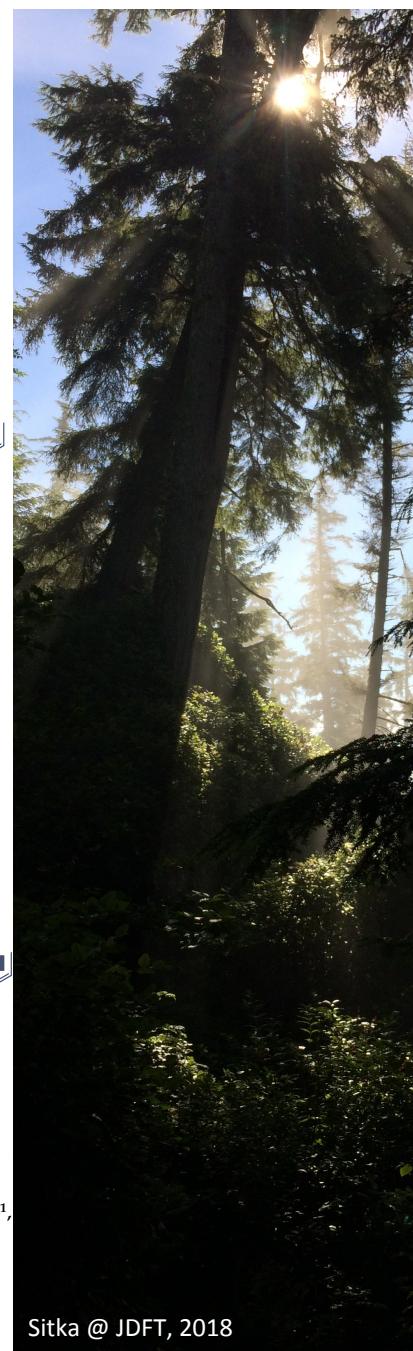
genes



Article

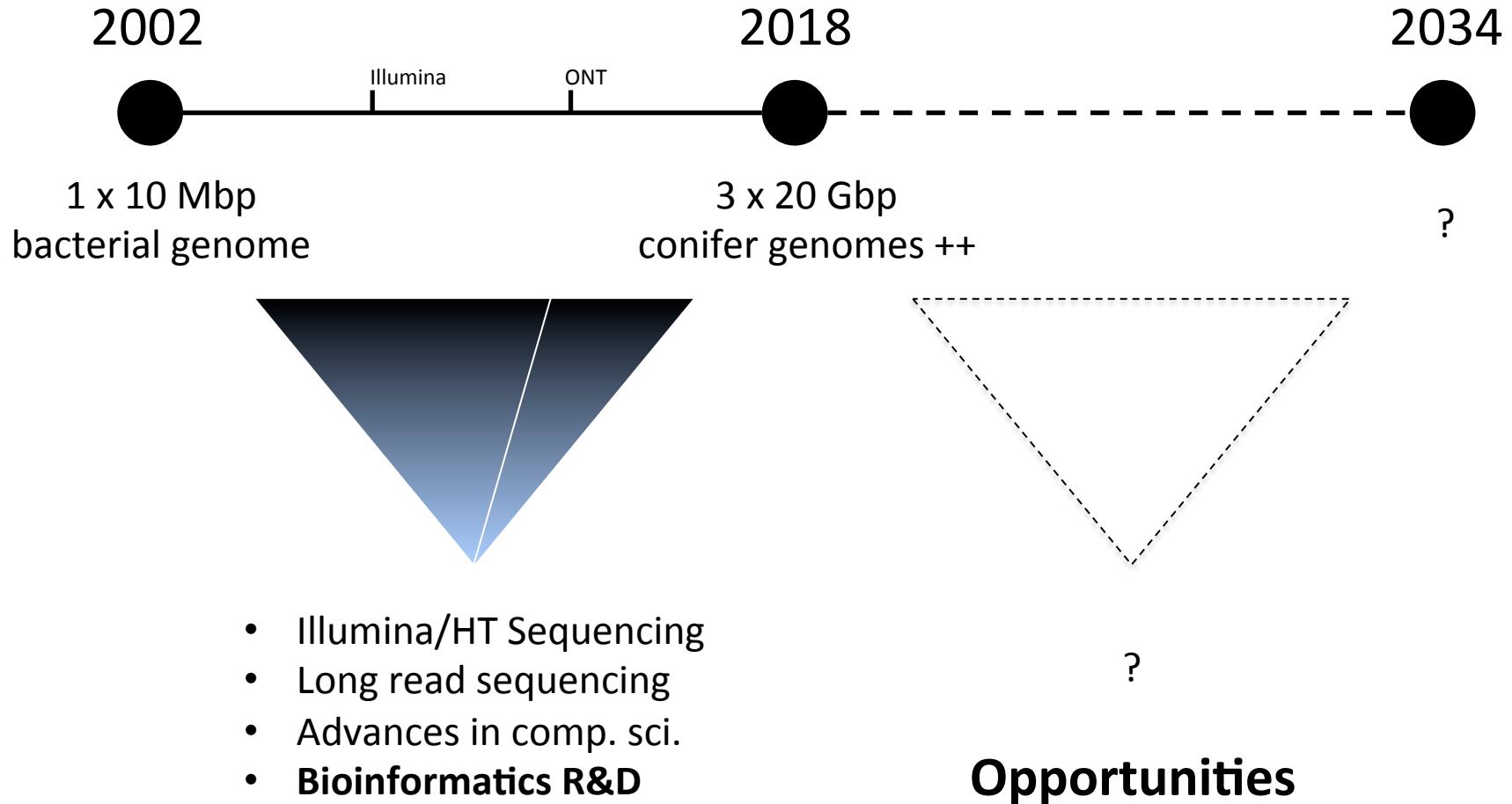
The Genome of the Northern Sea Otter (*Enhydra lutris kenyoni*)

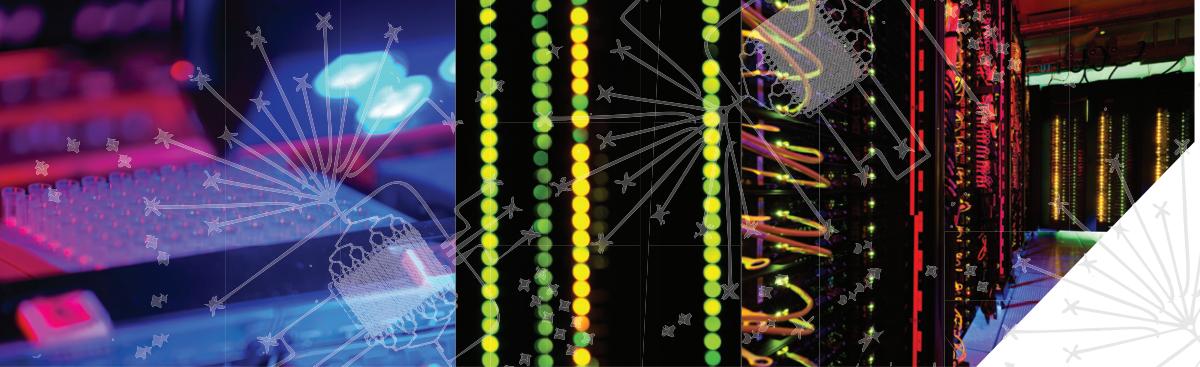
Samantha J. Jones ^{1,2}, Martin Haulena ³, Gregory A. Taylor ¹, Simon Chan ¹, Steven Bilobram ¹, René L. Warren ¹, S. Austin Hammond ¹, Karen L. Mungall ¹, Caleb Choo ¹, Heather Kirk ¹, Pawan Pandoh ¹, Adrian Ally ¹, Noreen Dhalla ¹, Angela K. Y. Tam ¹, Armelle Troussard ¹, Daniel Paulino ¹, Robin J. N. Coope ¹, Andrew J. Mungall ¹, Richard Moore ¹, Yongjun Zhao ¹, Inanc Birol ^{1,2}, Yussanne Ma ¹, Marco Marra ^{1,2} and Steven J. M. Jones ^{1,2,4,*}



Sitka @ JDFT, 2018

Perspective





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 British Columbia • Genome Sciences Institute

Ben Vandervalk | Zhuyi Xue | Austin Hammond | Lauren Coombe

Darcy Sutherland | Sauparna Palchowdhury | Jessica Zhang | Diana Lin | Eric Chen

Justin Chu | Shaun Jackman | Chen Yang | Kristina Gagalova | Ka Ming Nip

Golnar Sheikhshab | Saber Hafezqorani | Chenkai Li | Yee Fay Lim | Figali Taho

Amirhossein Afshinard | Readman Chiu | Hamid Mohamadi

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

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



GenomeBritishColumbia



GenomeCanada



SpruceUp

John Jambor Knowledge Fund

BCCA CANCER RESEARCH CENTRE



github.com/bcgsc

birollab.ca
b11

