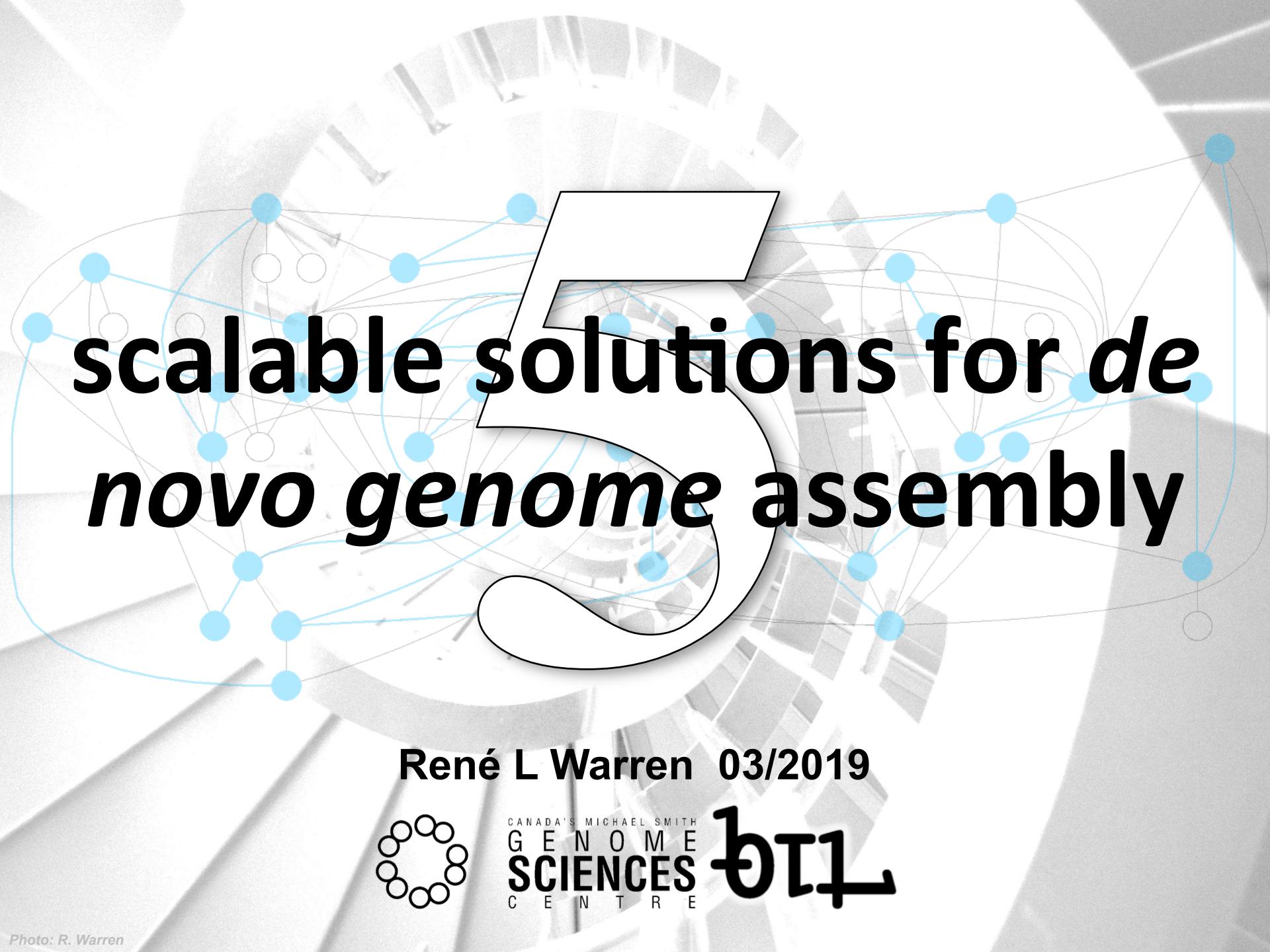


# scalable solutions for *de novo genome assembly*

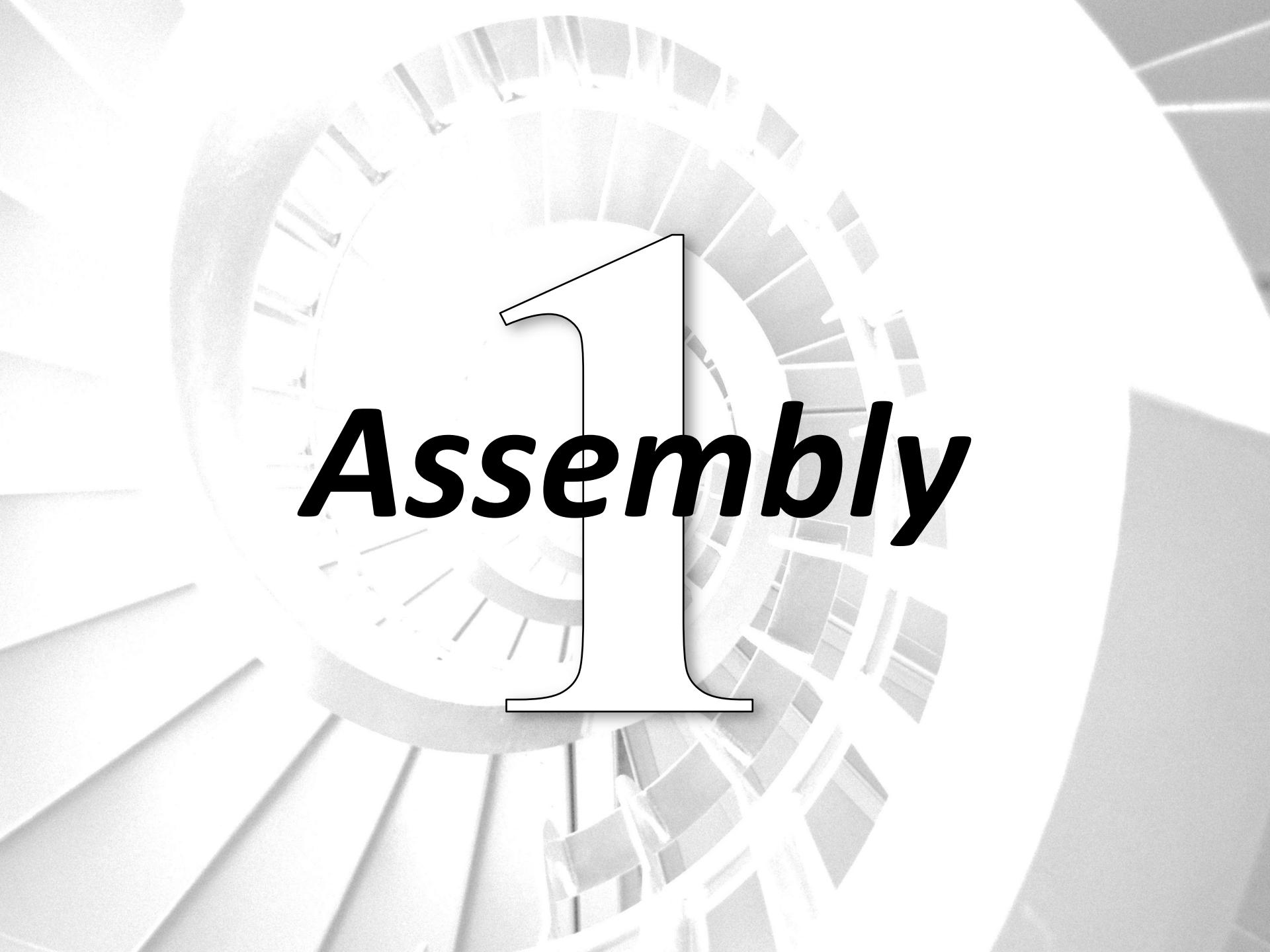


René L Warren 03/2019



CANADA'S MICHAEL SMITH  
GENOME  
SCIENCES  
CENTRE

b1



# Assembly



## 2009 : Parallel DBG assembler

MPI to aggregate memory

Assembled 20Gb spruce genome

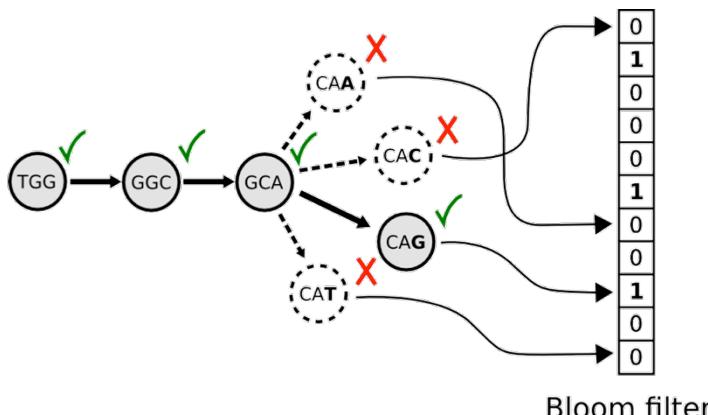
GENOME  
RESEARCH  
Resource

ABySS: A parallel assembler for short read sequence data

Jared T. Simpson,<sup>1</sup> Kim Wong, Shaun D. Jackman, Jacqueline E. Schein, Steven J.M. Jones, and İnanç Birol<sup>2</sup>

## 2017 : Bloom filter representation

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



GENOME  
RESEARCH  
Method

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

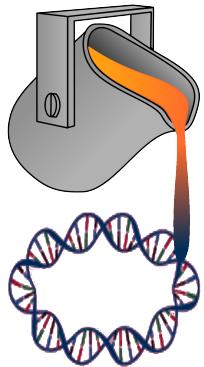
Shaun D. Jackman,<sup>1</sup> Benjamin P. Vandervalk,<sup>1</sup> Hamid Mohamadi, Justin Chu, Sarah Yeo, S. Austin Hammond, Golnaz Jahesh, Hamza Khan, Lauren Coombe, Rene L. Warren, and Inanc Birol

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



**Correction**

# Tigmint



## linked reads misassembly correction

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

BMC Bioinformatics

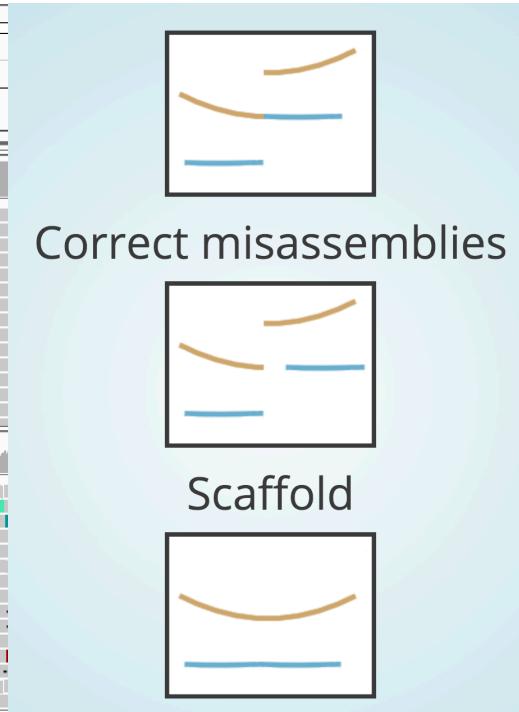
Open Access



SOFTWARE

### Tigmint: correcting assembly errors using linked reads from large molecules

Shaun D. Jackman<sup>1\*</sup> , Lauren Coombe<sup>1</sup>, Justin Chu<sup>1</sup>, Rene L. Warren<sup>1</sup>, Benjamin P. Vandervalk<sup>1</sup>, Sarah Yeo<sup>1</sup>, Zhuyi Xue<sup>1</sup>, Hamid Mohamadi<sup>1</sup>, Joerg Bohlmann<sup>2</sup>, Steven J.M. Jones<sup>1</sup> and Inanc Birul<sup>1</sup>

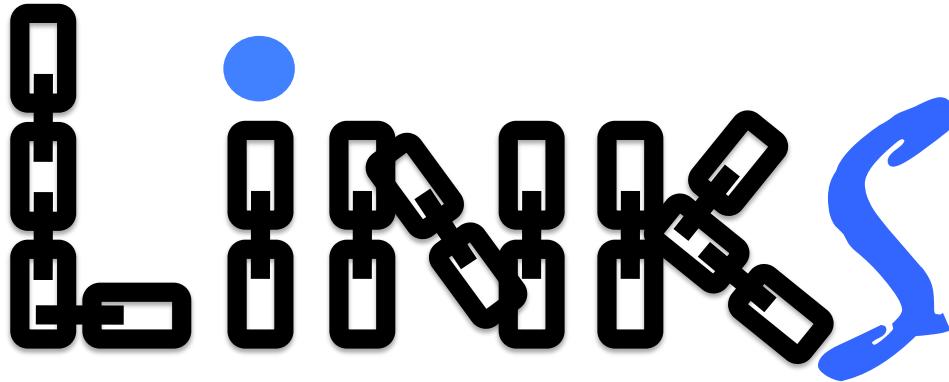


IGV screenshot: Tigmint breakpoint in human genome NA24143

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



# **Scaffolding**



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

(GIGA)<sup>n</sup>  
SCIENCE

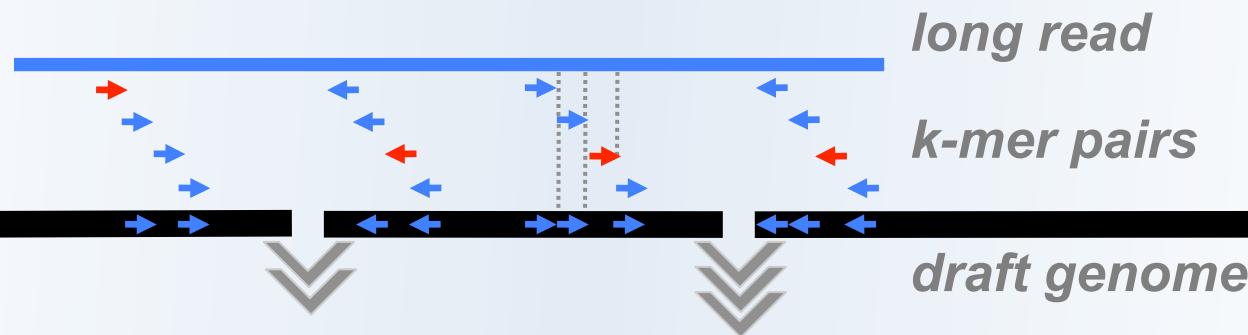
RESEARCH Open Access



## 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



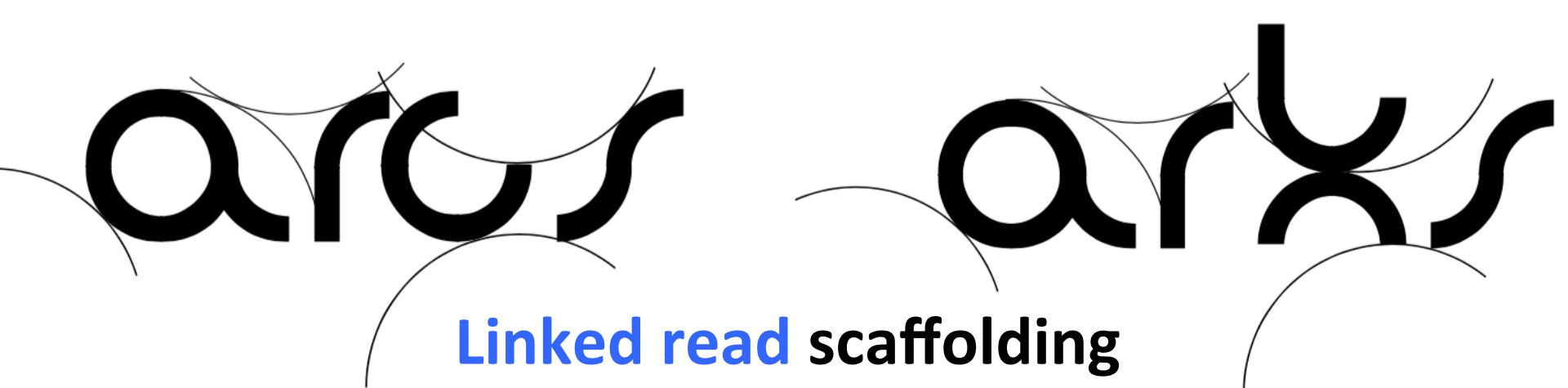
- **Scaffolder** : order & orient sequences
- ***k*-mer based** : no alignments
- **Vast *k*-mer space** : no fragment length limitation
- **Versatile** : long-reads, draft sequences, MPET

# length

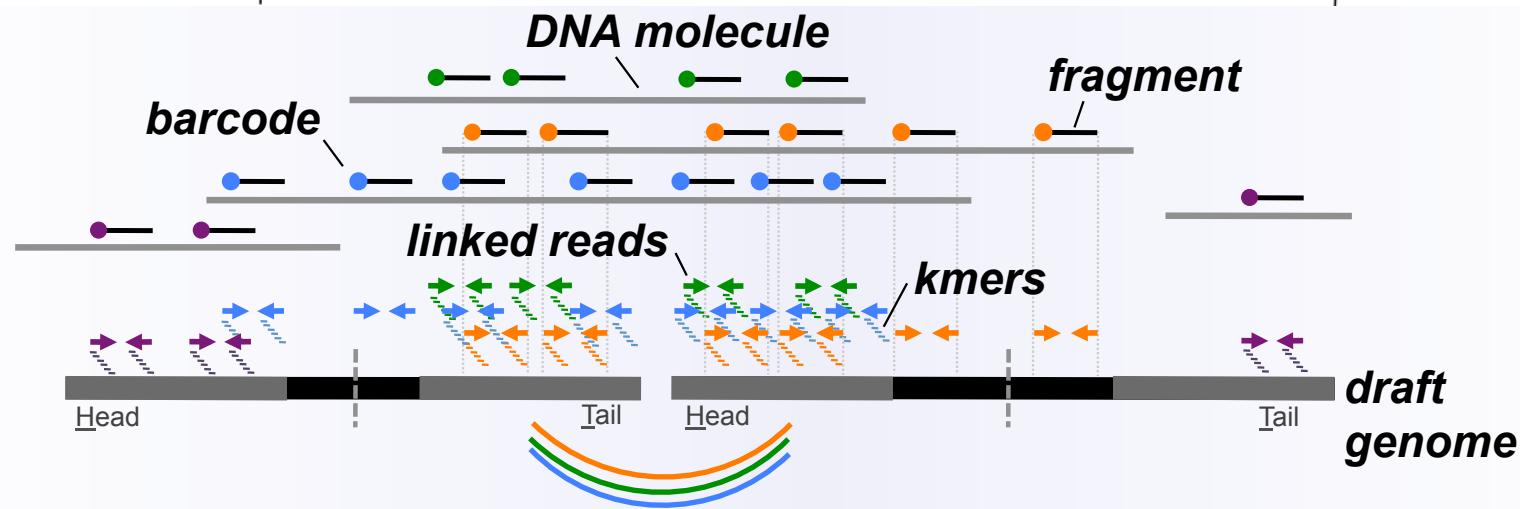
# errors

∅ 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>

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<sup>†</sup>, Jessica Zhang<sup>†</sup>, Benjamin P. Vandervalk, Justin Chu, Shaun D. Jackman, Inanc Birol and René L. Warren<sup>\*</sup>

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

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

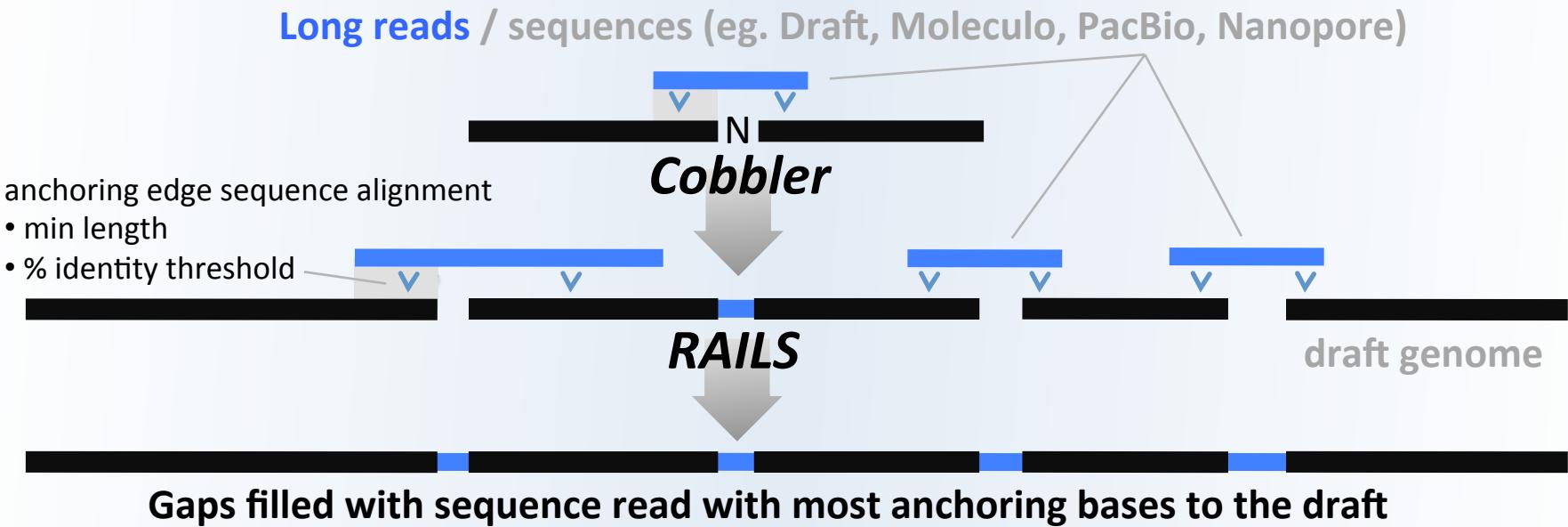
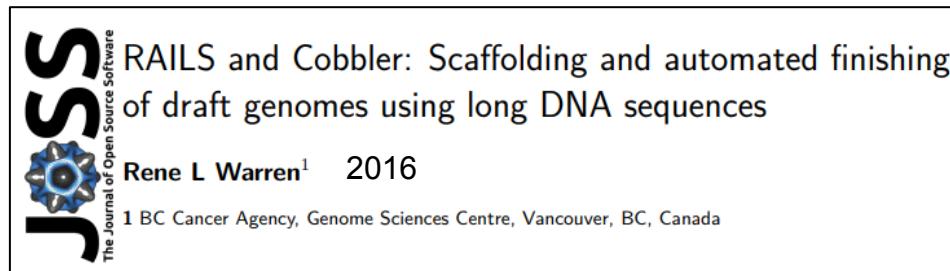


**Gap-filling**



# Scaffolding and gap-filling

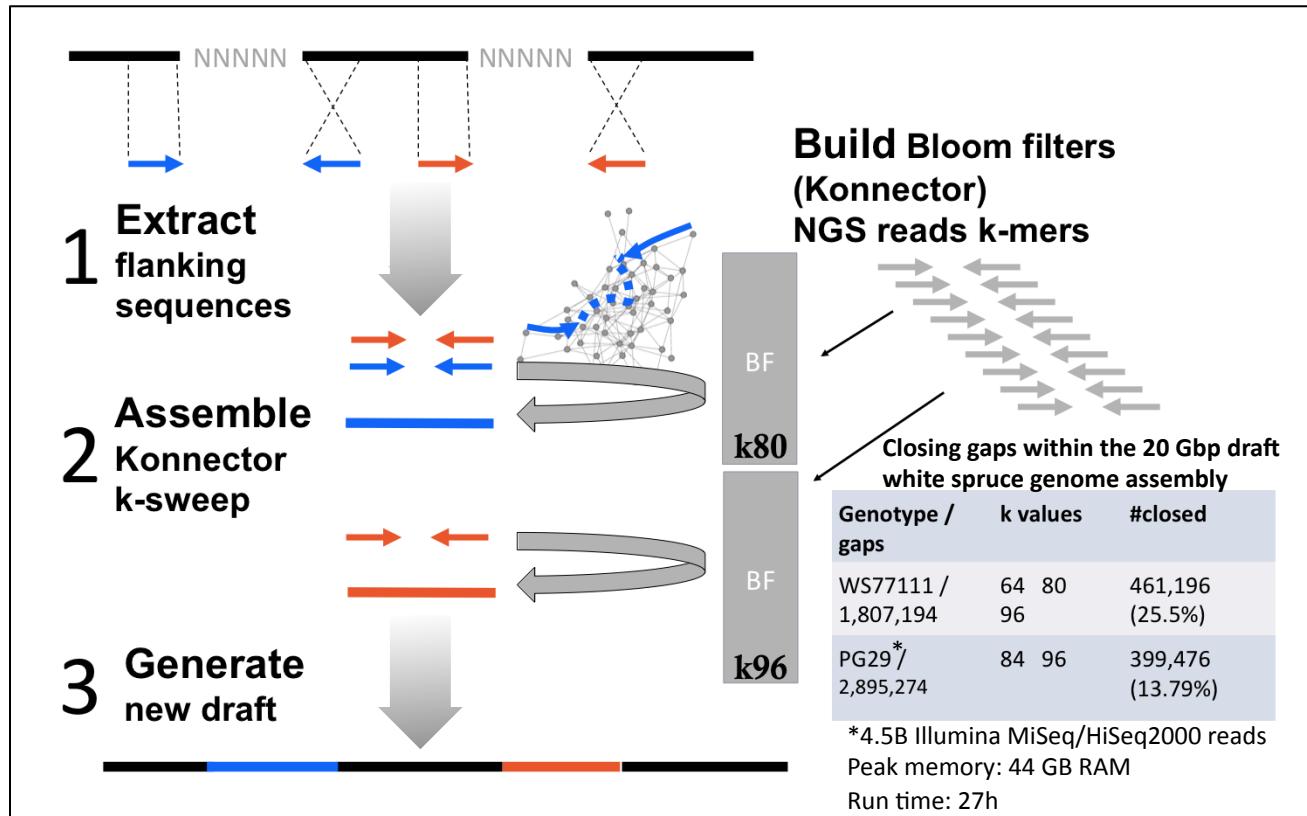
Uses LINKS scaffolding algorithm



# 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

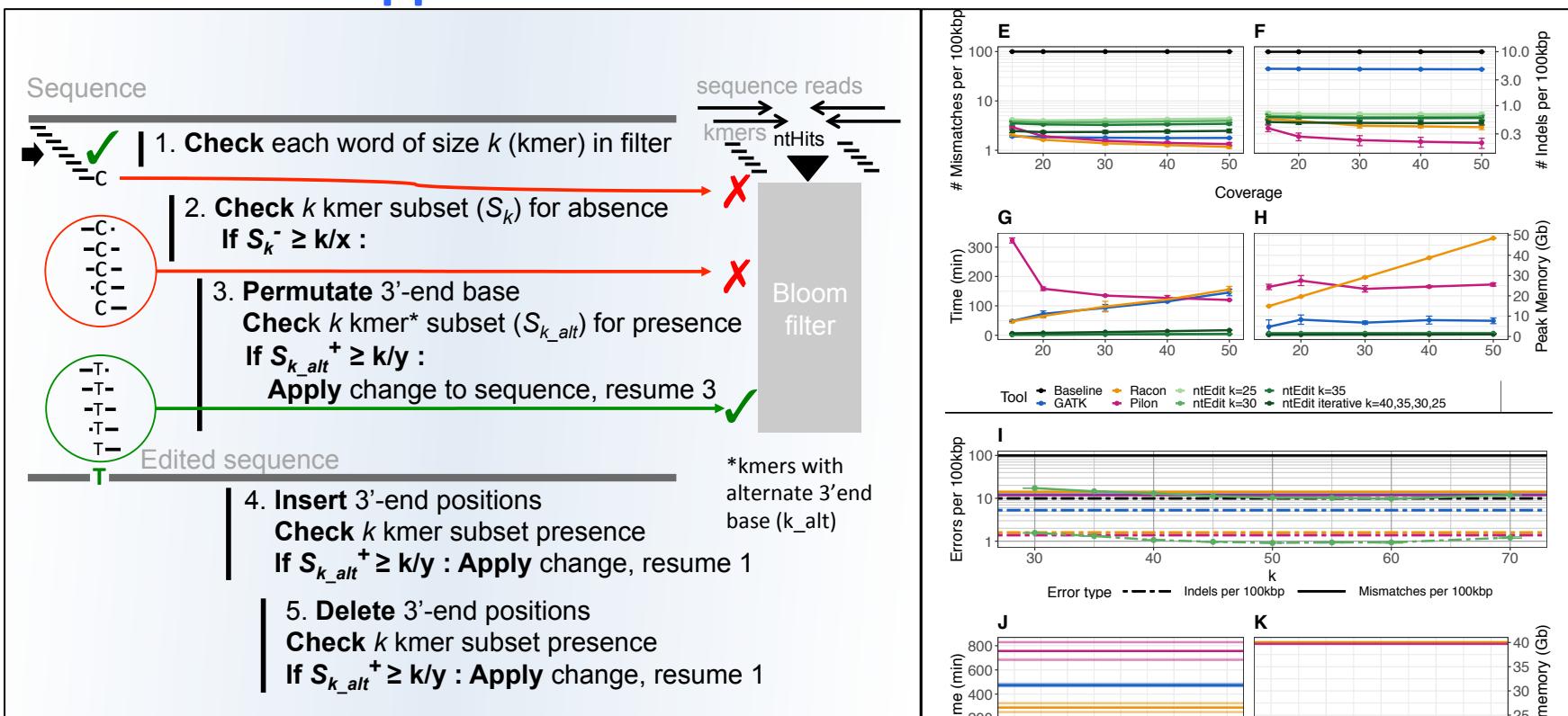




# *Polishing*

# ntEdit

## Fast homozygous error correction / genome “haploidization” Approach Results



ntEdit: scalable genome sequence polishing  
2019



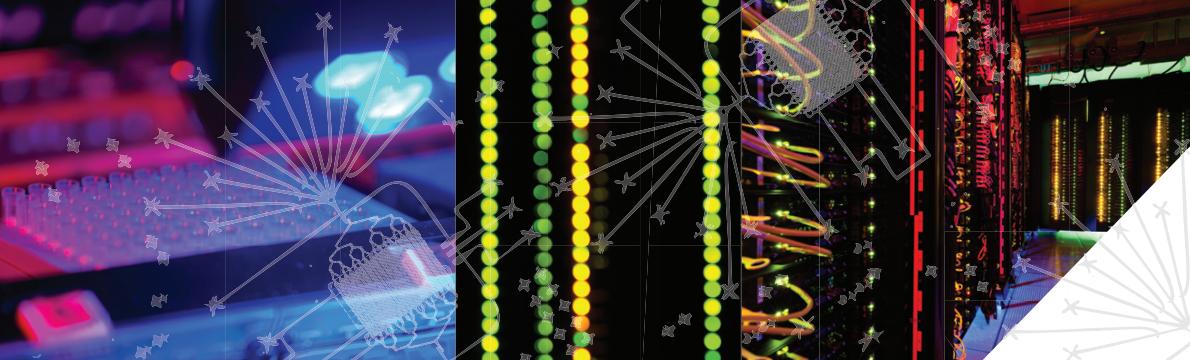
bioRxiv  
THE PREPRINT SERVER FOR BIOLOGY

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>

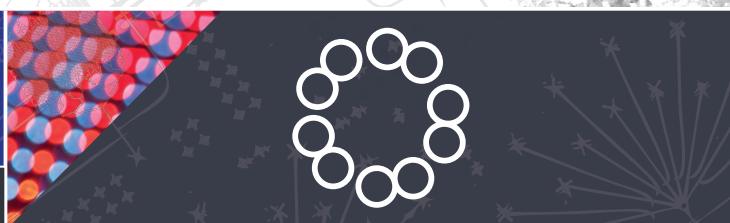
*C. elegans*  
*H. Sapiens chr21*



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



Bioinformatics Technology Lab



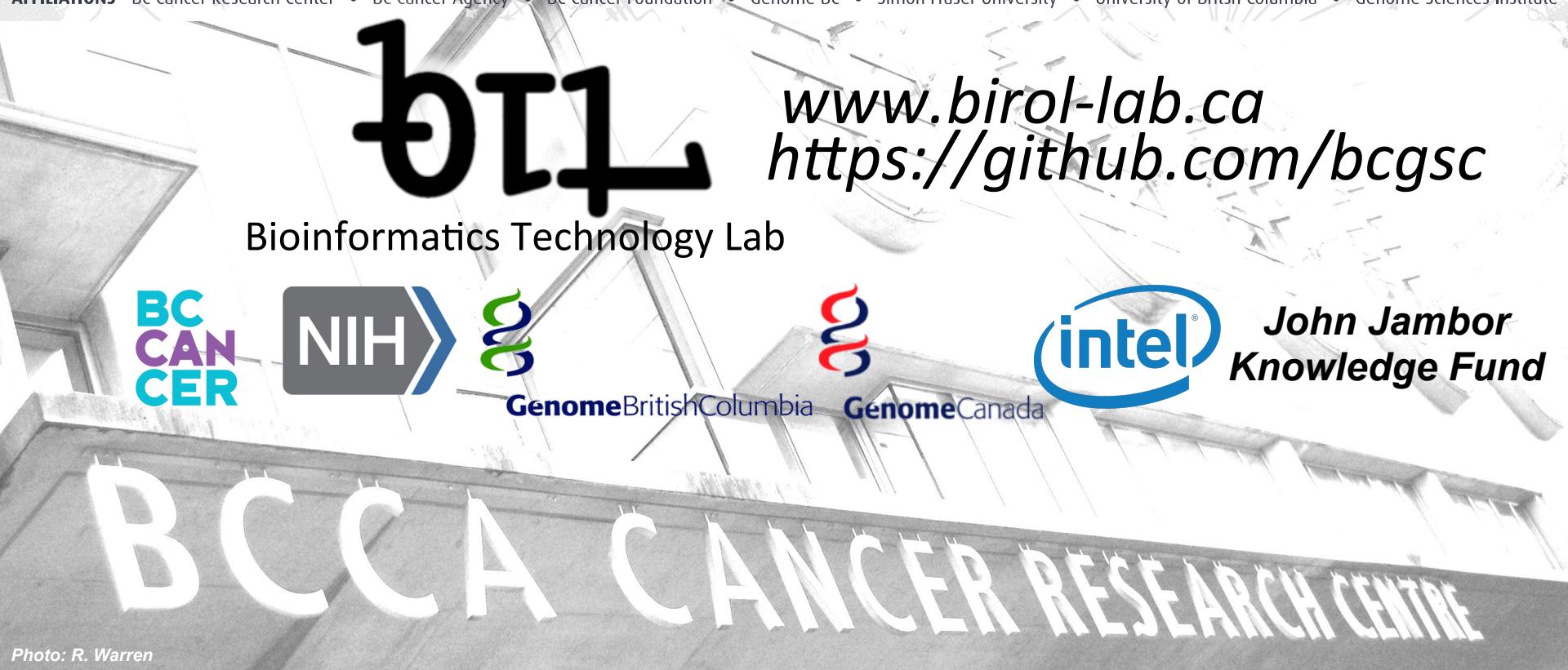
GenomeBritishColumbia



GenomeCanada



John Jambor  
Knowledge Fund



BCCA CANCER RESEARCH CENTRE

