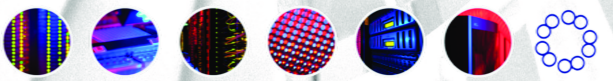
Genome Assembly Toolkit

LINKS scaffold graph / *E. coli* K12

René L Warren

ISMB, July 2016



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

Photo: RL Warren

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

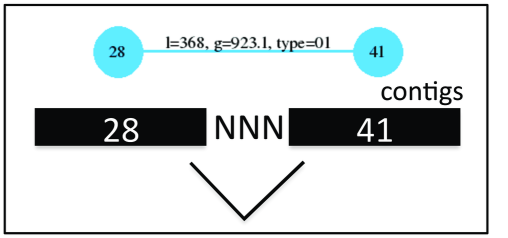
GIGA SCIENCE

2015

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



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

BMC Medical Genomics

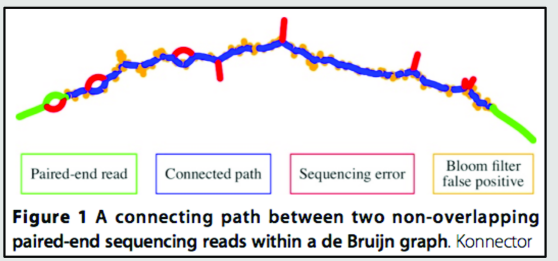


Figure 1 A connecting path between two non-overlapping paired-end sequencing reads within a de Bruijn graph. Konnector

RESEARCH Open Access

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

Benjamin P Vandervalk, Chen Yang, Zhuyi Xue, Karthika Raghavan, Justin Chu, Hamid Mohamadi, Shaun D Jackman, Readman Chiu, René L Warren, Inanç Birol*

From IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2014)
Belfast, UK. 2-5 November 2014

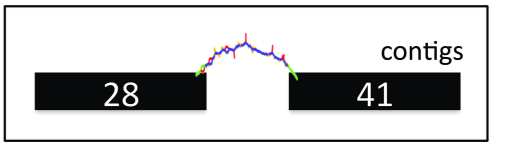
Paulino *et al. BMC Bioinformatics* (2015) 16:230
DOI 10.1186/s12859-015-0663-4

BMC Bioinformatics

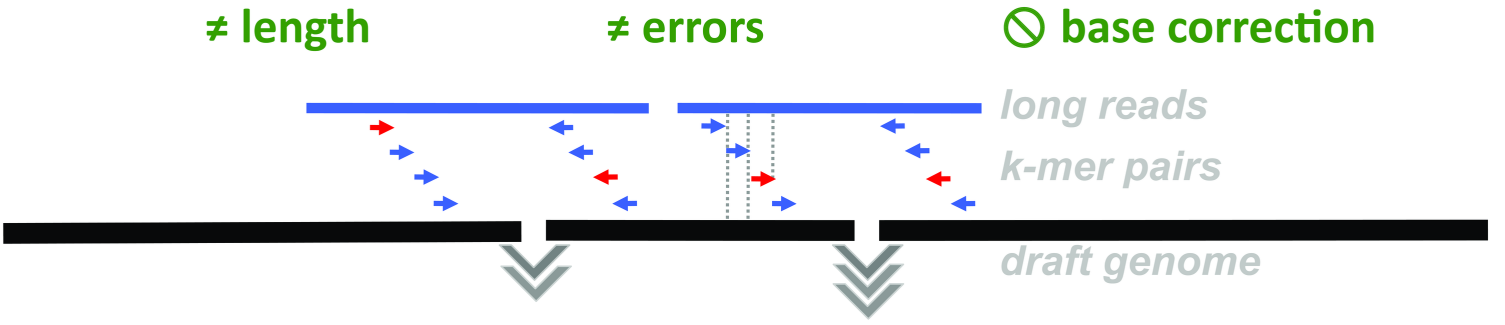
SOFTWARE Open Access

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

Daniel Paulino[1†], René L. Warren[1†], Benjamin P. Vandervalk[1], Anthony Raymond[1], Shaun D. Jackman[1] and Inanç Birol[1,2*]



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

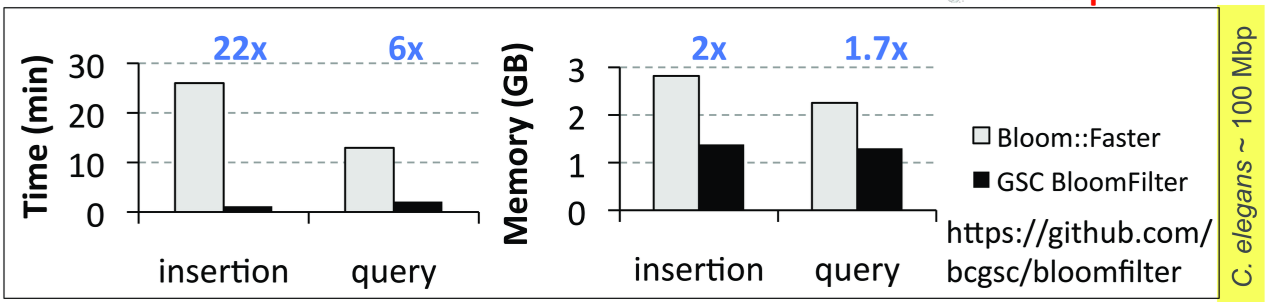


New Features

v1.6 : Custom Bloom filter

- *ntHash* recursive *k-mer* hashing : fast

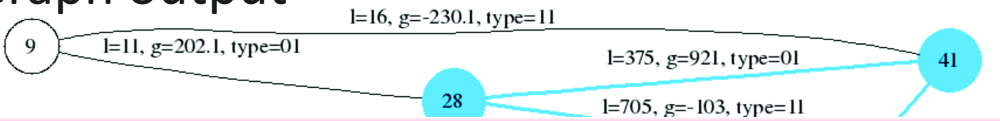
Mohamadi *et al.*
ISMB poster



- Build large genomes filters

Species	Genome (Gb)		Time	RAM (GB)
• human	3	:	40m	26
• bullfrog	6	:	2h	37
• spruce	20	:	5h	178

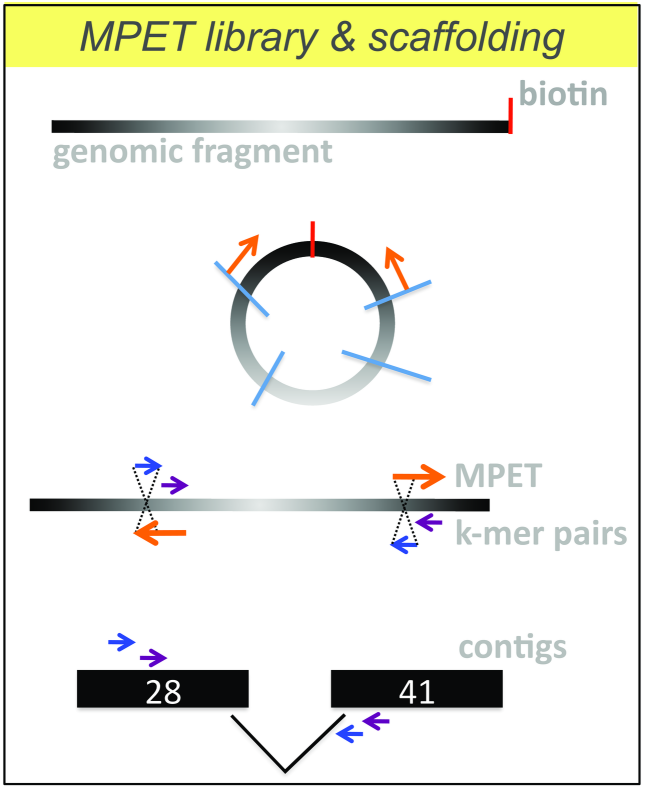
- Graph output



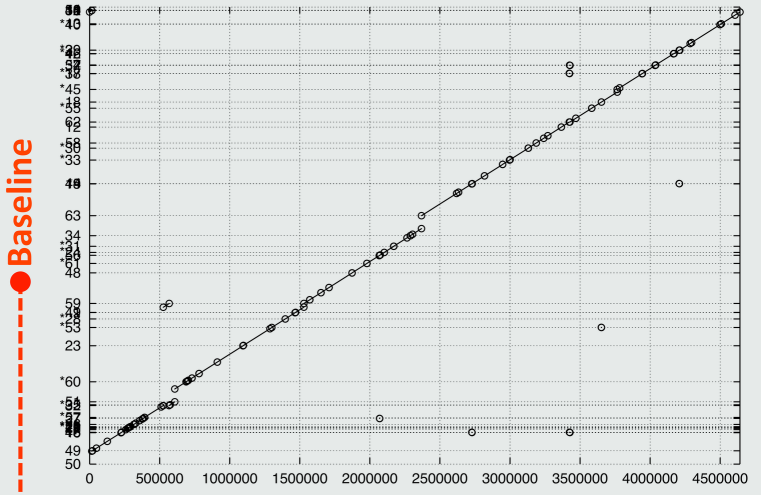
New Features

v1.7 : MPET Scaffolding

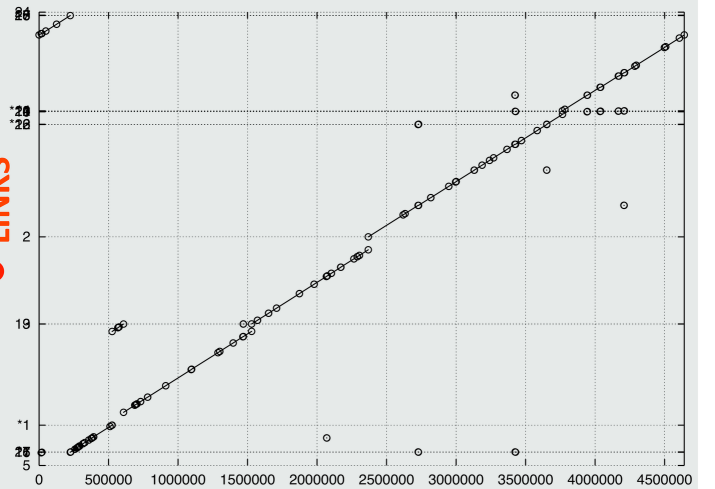
- Long-range pairing
- Massive depth, linkage support
- [Still] no alignments
- Scaffolding checkpoints
- Support for zip/gzip



MPET Scaffolding



E. coli K12 MG1655



E. coli K12 MG1655

E. coli K12	n:500	NGA50
Baseline contig	61	180 kbp
LINKS v1.7	27	300 kbp

Trimmed 4kb MPET: 260-fold physical coverage

H. sapiens	NA19238 NGA50 *	GIAB HG004 NGA50 ^
Baseline contig	56 kbp	48 kbp
LINKS v1.7	473 kbp	1.5 Mbp

10h @ 400 GB RAM

*Trimmed 2kb MPET: 150-fold physical coverage
^Trimmed 6kb MPET: 175-fold physical coverage

New Features

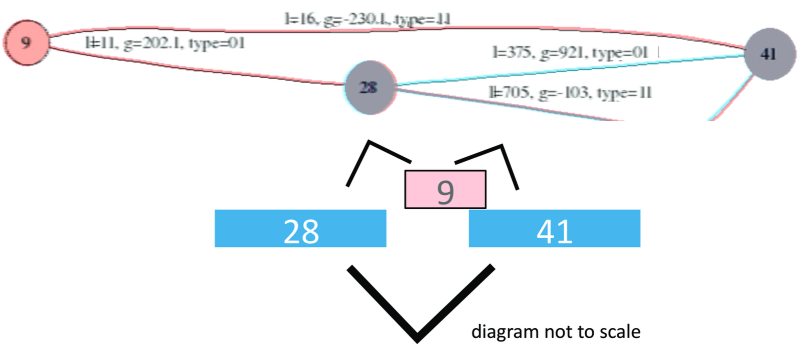
v1.8 : Native k-mer length *intervals*

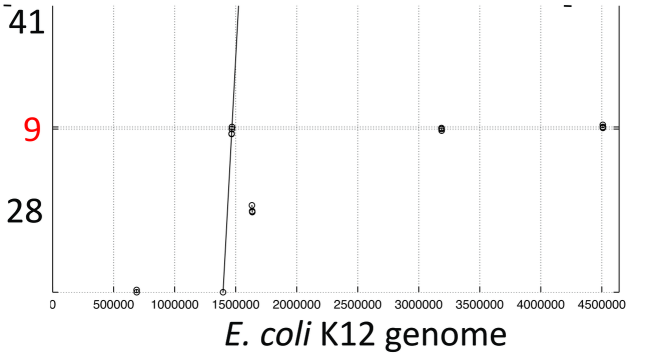
- Single-run, Δ*i*

E. coli K12 MinION full2D ONT	paper (v1.5) 30 iterations	v1.8 1 iteration 15 intervals*	Fold change
NG50 length (bp)	633,147	**1,152,663**	1.8
Memory (GB)	**4.3**	**4.3**	-
Time (mm:ss)	17:21	**02:54**	5.5

*1-10:1kb, 12-20:2kb

- Prioritizes proximal contigs





E. coli K12 genome

LINKS

BBT

KOLLECTOR

Konnector
&
Sealer

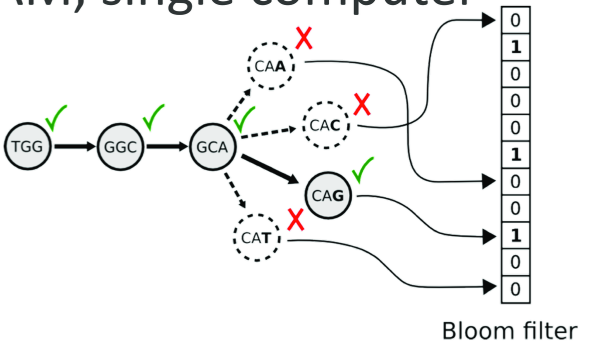
RAILS

ABySS

***de novo* assembly of large genomes**

Human Genome First:

- Then: parallel, short read *k-mer* assembler
  - MPI to aggregate memory (Simpson *et al*, 2009)
- Now: de Bruijn graph Bloom filter representation
  - 1/10th RAM, single computer



Vandervalk *et al.*
ISMB poster
HitSeq talk/poster

BBT

Bio Bloom Tools

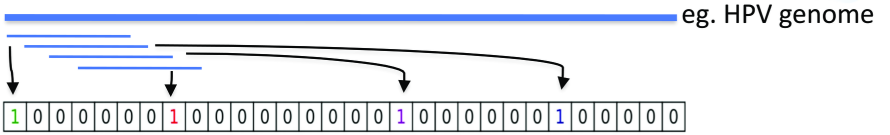
Sequence classification with Bloom filters

general purpose filtering *contaminant screening* *pathogen discovery*

BioBloom-

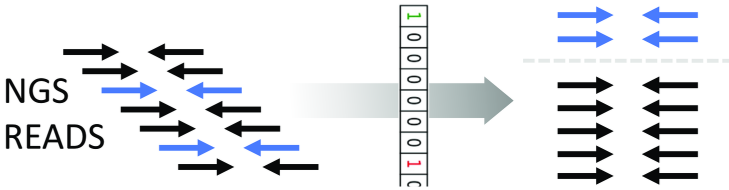
Maker

- Build filters : **Re-usable**
- Customizable : **Flexible**
- Multi-filter : **Concurrent**



Categorizer

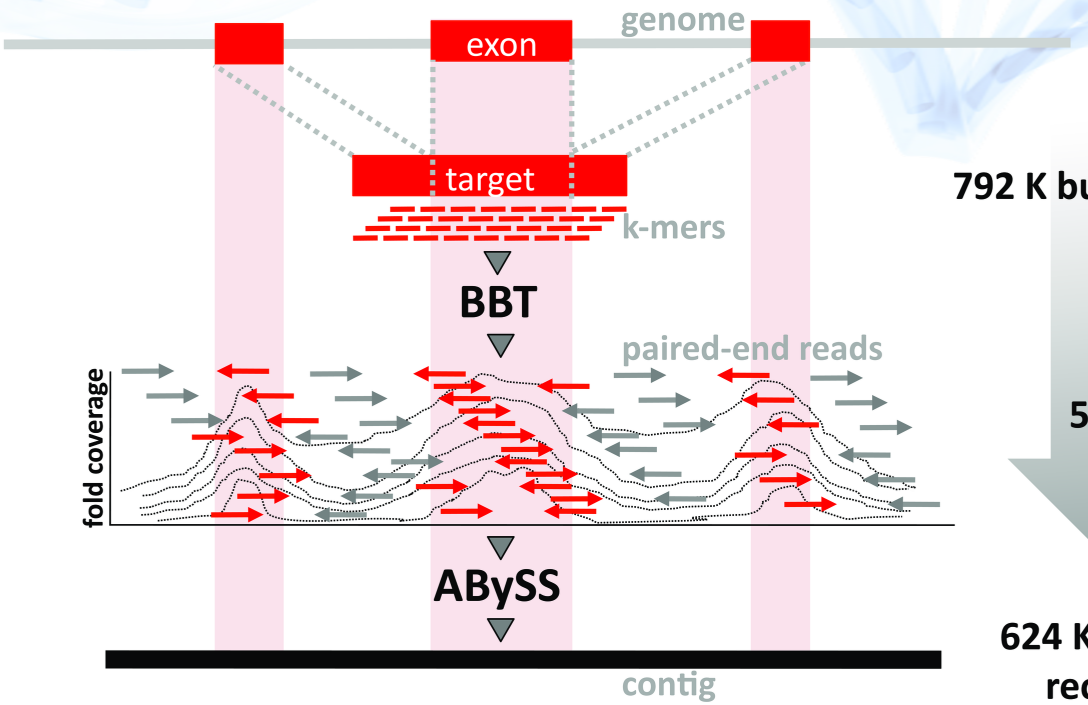
- Bins sequences : **Analysis summary**



Chu *et al.* 2014
HitSeq
poster/talk

KOLLECTOR

Targeted *de novo* assembly of gene loci from RNA



792 K bullfrog transcripts

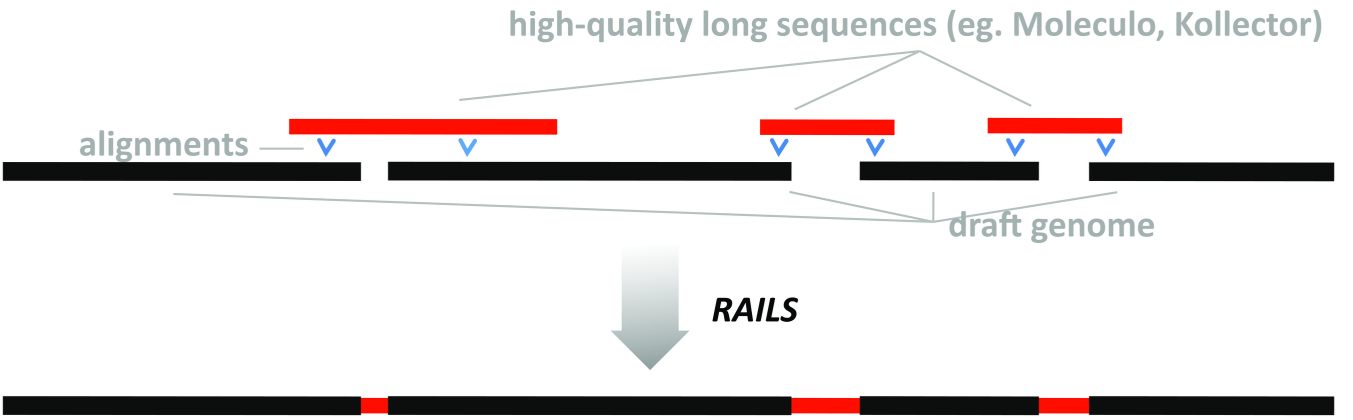
5 iterations

624 K (~80%) genes reconstructed

RAILS

Radial Assembly Improvement by Long Sequence Scaffolding

- Scaffolding **and** gap-filling
- Uses LINKS algorithm

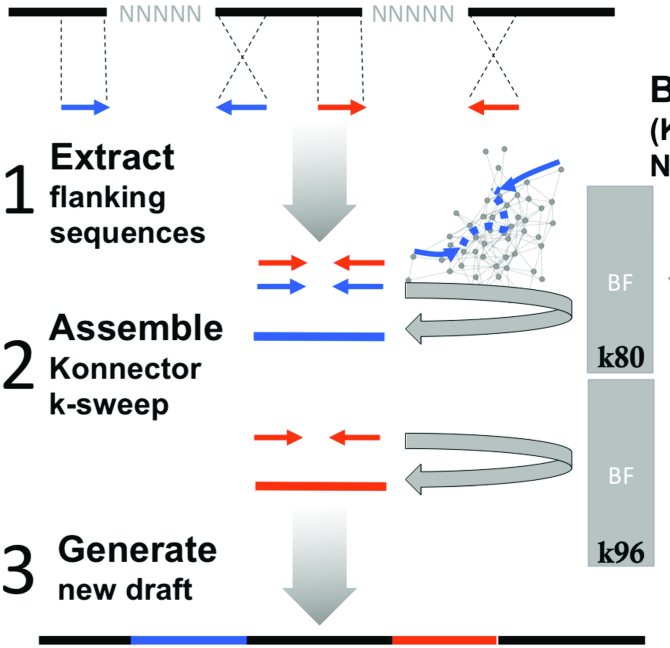


Sealer

Automated genome finishing

- Gap-filler (resolve *Ns*) Application of ***Konnector***
- Bloom filter de Bruijn graph (Scalable)

Vandervalk *et al*, 2015
Paulino *et al*, 2015



Build Bloom filters (Konnector) NGS reads k-mers

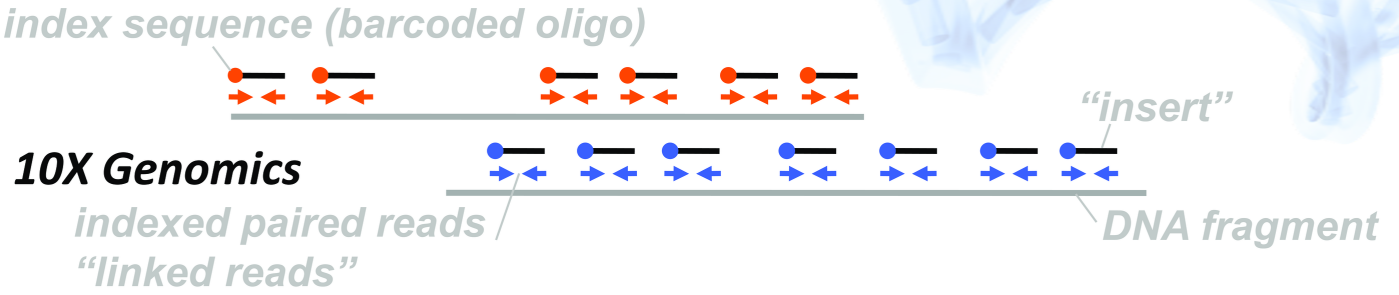
Spruce

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

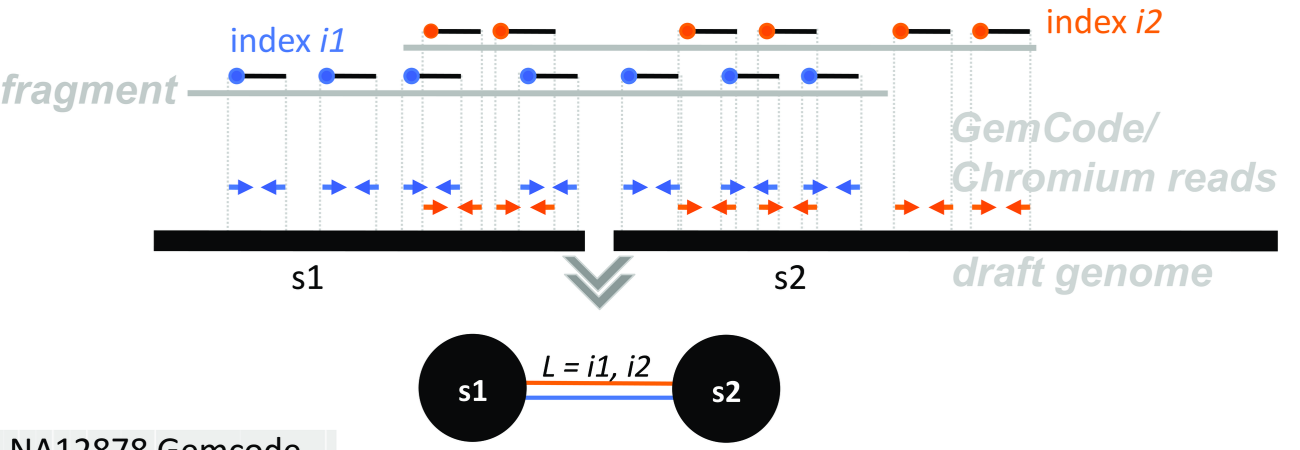
Assembly Roundup by *Chromium* Scaffolding

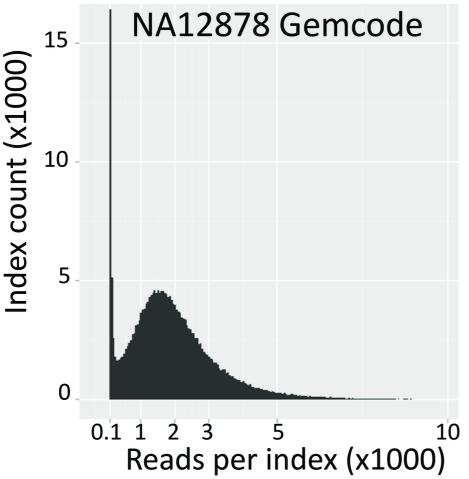
gemCode



- **Capture** long-range info
- **Identify** co-located scaffolds in genome draft
- **Improve** *de novo* assembly

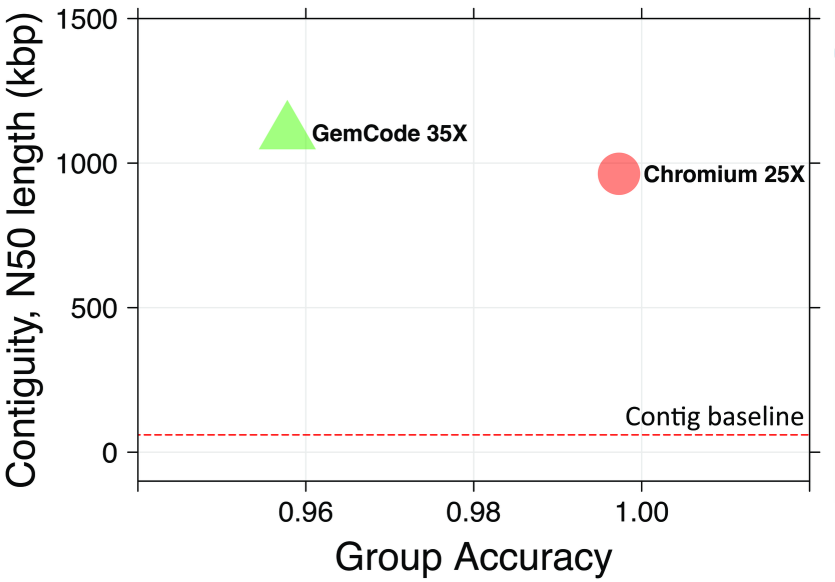
Method

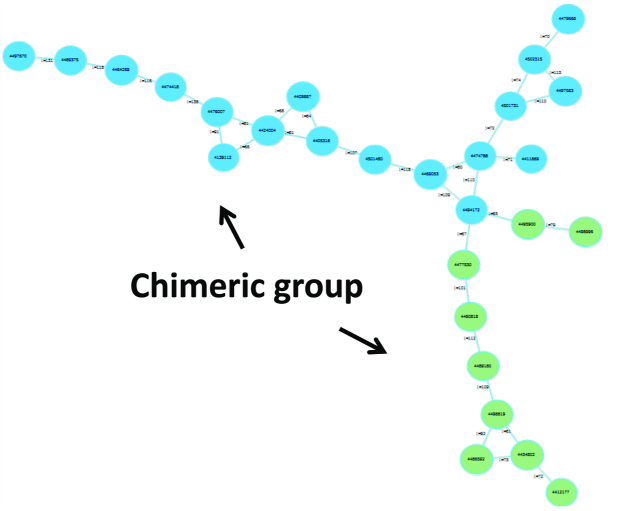




- **Pair** sequences, **Build** graph
  - **Allow** indexed reads within frequency range
  - **Link** 2 sequences >= reads map per index
  - **Create** edge with >= *L* links
- Connected components = groups

Grouping, human NA12878





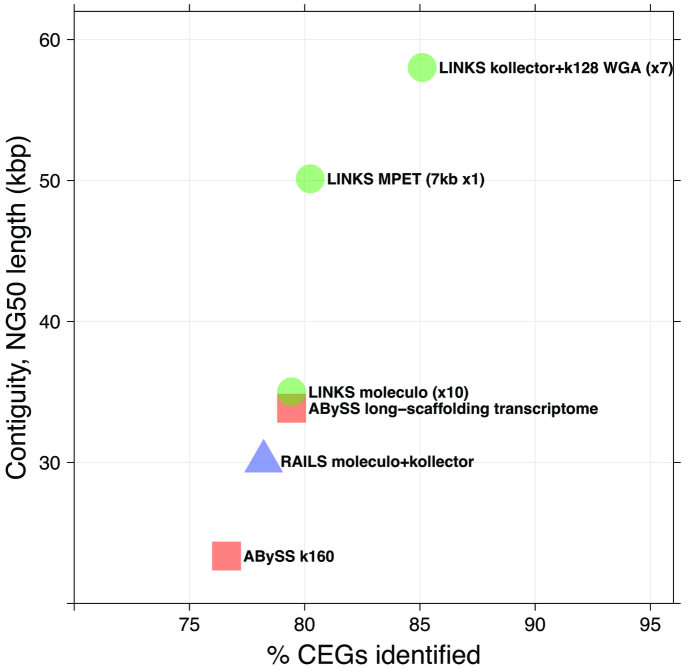
- **Accuracy of groups assessed with BBT**

Application to large genomes

image source: https://en.wikipedia.org/wiki/American_bullfrog

Scaffolding the bullfrog genome

Assemblies	ABySS k160	RAILS	Long-Scaffolding	LINKS
#Scaffold merges	NA	36,137	NA	177,506
NG50 length (bp)	23,361	30,085	33,847	58,021
Core Eukaryotic Genes	77%	78%	79%	85%



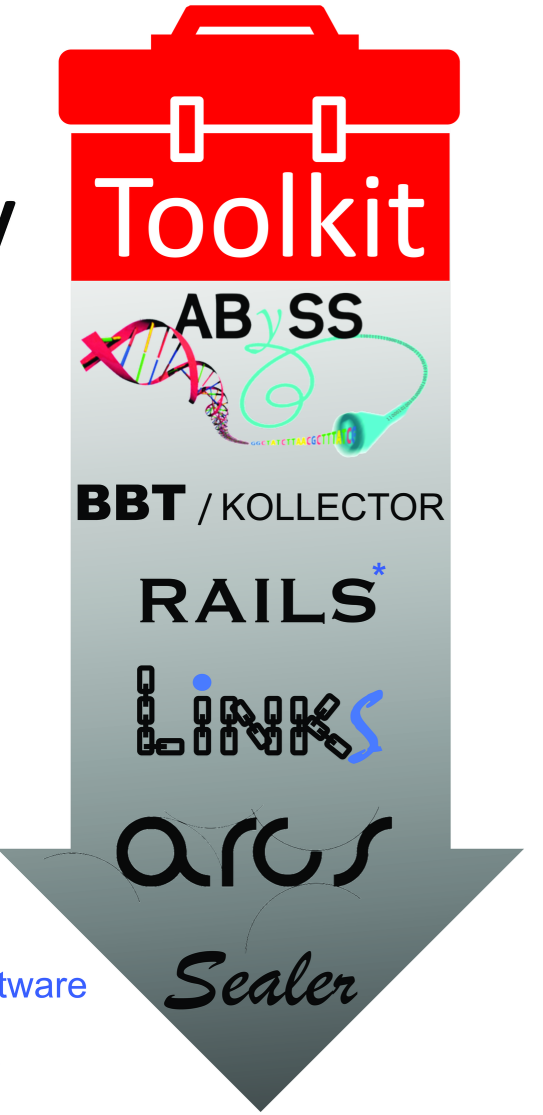
Data

- *Moleculo (TruSeq) – 0.4X coverage*
- *Bullfrog Transcripts*
- *Kollector sequences*
- *MPET (613.5M reads, 61Gb)*
- *Draft assembly (k128)*

- 61K gaps closed (13.9%) with *Sealer*
- Bullfrog draft submitted to genbank

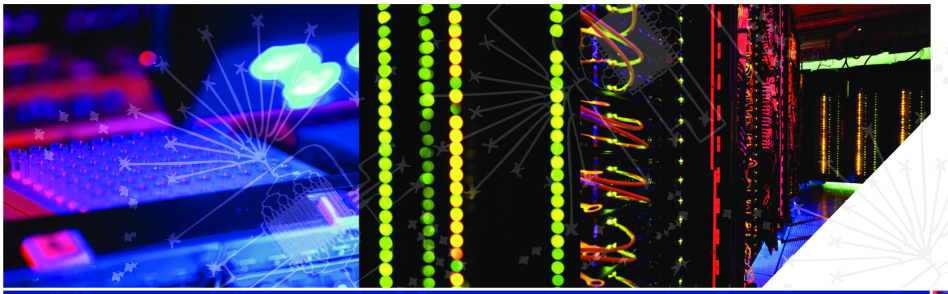
ϟwork with Caren Helbing, U. Victoria

Genome Assembly



http://www.bcgsc.ca/platform/bioinfo/software

*ftp://ftp.bcgsc.ca/supplementary/RAILS



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.

54 TERABASES SEQUENCED • A HUMAN GENOME EVERY 17 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

Thank You!

Austin Hammond | Ben Vandervalk | Chen Yang | Daniel Paulino | Erdi Kucuk
Golnaz Jahesh | Hamid Mohamadi | Inanc Birol | Justin Chu | Lauren Coombe
René Warren | Sarah Yeo | Shaun Jackman | Tony Raymond

NIH

BC CANCER FOUNDATION
partners in discovery

GenomeBritishColumbia

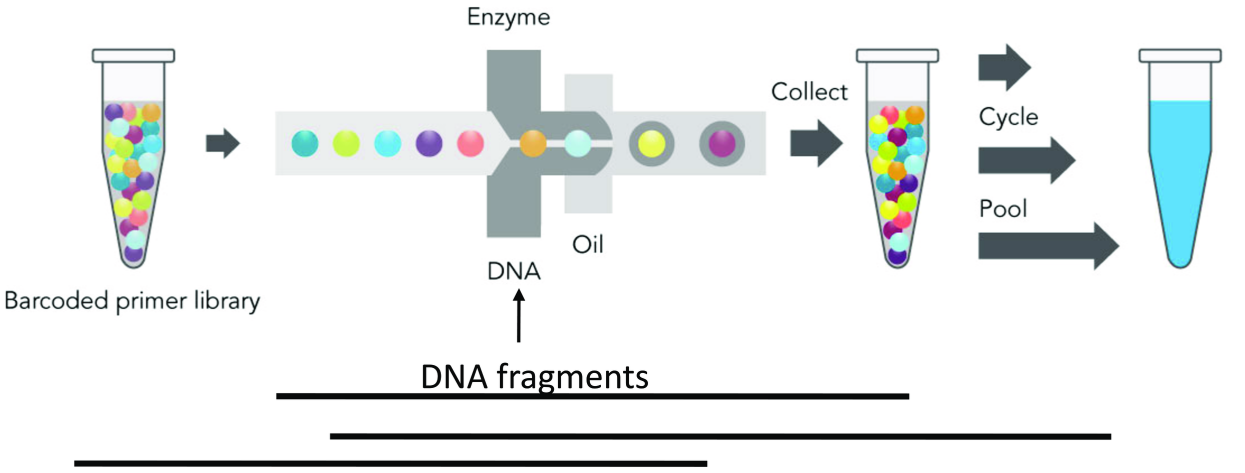
GenomeCanada

intel

John Jambor Knowledge Fund

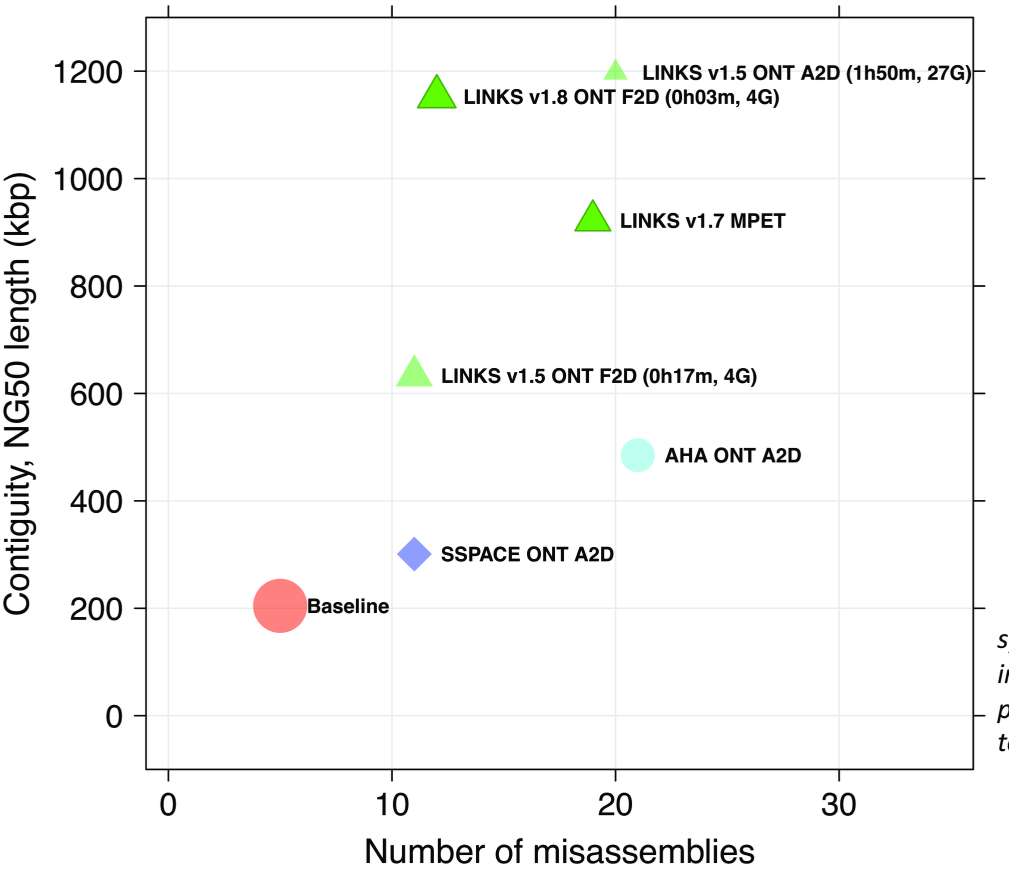
Photo: RL Warren

Overview of 10X Genomics' GemCode Platform



Picture: http://2.bp.blogspot.com/-9bC-Fg0qMEQ/VPR0yO2y5hI/AAAAAAAABjw/B6M3OLAvqLE/s1600/Screenshot%2B2015-03-02%2B14.36.11.png

Summary, *E. coli*



symbol size inversely proportional to #scaffolds