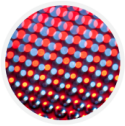







CANADA'S MICHAEL SMITH GENOME SCIENCES CENTRE
WWW.BCGSC.CA

High-Performance Computing for Assembly and Analysis of Big Genomics Data

René Warren, Benjamin Vandervalk, Anthony Raymond, Shaun Jackman, Hamid Mohamadi, Daniel Paulino, Justin Chu, Ewan Gibb and İnanç Birol
British Columbia Cancer Agency, Genome Sciences Centre, Vancouver, BC V5Z 4S6

Abstract

DNA Sequencing technology is developing at an unprecedented pace, surpassing the rate of advances in computer hardware development. Limited compute resources for storing, processing and analyzing omics data have spurred the improvements of file compression formats, low memory footprint data structures, algorithms that use communication protocols for parallel programming, and astute approaches for handling large data on commodity hardware. Our research team oversees the development of such bioinformatics technologies. Past accomplishments include: enabling the first assembly with millions of very short sequence reads (Warren et al. 2006), assembly of the human genome from short reads with the first parallel assembler (Simpson et al. 2009) and last year, assembly of the then largest genome, that of the 20 Gbp white spruce (Birol et al. 2013). We discuss key enabling algorithms, specifically introducing data structures, processes, compression schemes within **ABySS** (Simpson et al. 2009), **BBT** (Chu et al. 2014), **DIDA** (Mohamadi et al. submitted), **Konnector** (Vandervalk et al. 2014) and **TASR** (Warren et al. 2011) that are tailored to the needs of today's big sequence data reality.

Funding

BC Cancer Foundation — Supporting research & care at BC Cancer Agency
GenomeCanada GenomeBritishColumbia
Research reported in this publication was also partly supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number R01HG007182. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

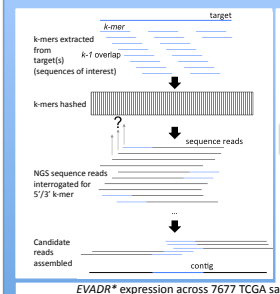
References

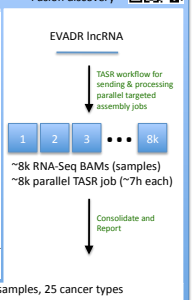
-Birol I, Raymond A, Jackman SD, Pleasance S, Coope R, Taylor GA, Yuen MM, Keeling CI, Brand D, Vandervalk BP et al. 2013. Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data. Bioinformatics. 29:1492
-Bloom BH. 1970. Space/time trade-offs in hash coding with allowable errors. Commun ACM. 13:422
-Chu J, Sadeghi S, Raymond A, Jackman SD, Nip KM, Mar R, Mohamadi H, Butterfield YS, Robertson AG, Birol I. 2014. BioBloom tools: fast, accurate and memory-efficient host species sequence screening using bloom filters. Bioinformatics. 30:3402
-Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19:1117
-Vandervalk B, Jackman SD, Raymond A, Mohamadi H, Yang C, Attali D A, Chu J, Warren RL, Birol I. 2014. Konnector: connecting paired-end reads using a Bloom filter de Bruijn graph. in Bioinformatics and Biomedicine (BIBM 2014), Belfast UK, 2014.
-Warren RL, Holt RA: Targeted assembly of short sequence reads. PLoS ONE 2011, 6:e19816.
-Warren RL, Sutton GG, Jones SJ, Holt RA. 2007. Assembling millions of short DNA sequences using SSAKE. Bioinformatics. 23:500

TASR

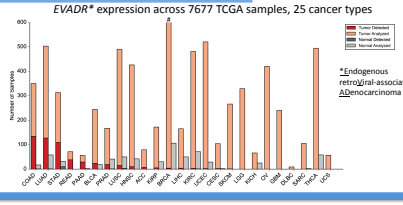
Targeted Assembly of Sequence Reads

Applications
•SNV profiling
•Transcript profiling
•Fusion discovery





*EVADR** expression across 7677 TCGA samples, 25 cancer types



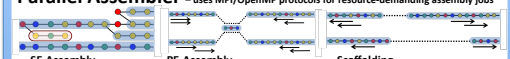
*Endogenous retroViral-associated ADenocarcinoma RNA

ABySS

Assembly By Short Sequences

Applications
•De novo assembly

Parallel Assembler – uses MPI/OpenMP protocols for resource-demanding assembly jobs

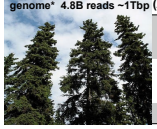


SE-Assembly
k-mer extension on a de Bruijn graph

PE-Assembly
search for unambiguous contig merging along paths

Scaffolding
search for unambiguous linkage across distant contigs

New features: Strand specificity, more efficient input handling (20% speed-up), long-range scaffolding, FM-index read mapping



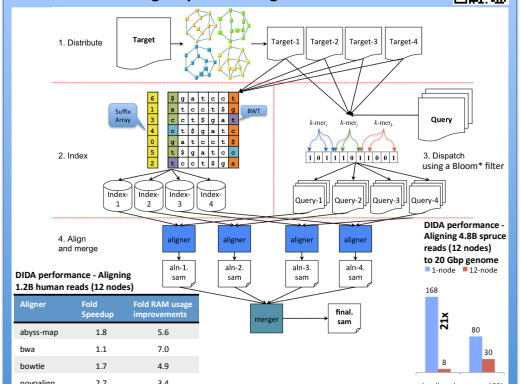
*Assembly resources required

(ABySS/DIDA)	k=116	unitig	contig	scaffold
Peak #CPU cores		184	3,168	144
#Comp. nodes		15	264	12
Compute time (hh:mm)		13:32	51:18	46:54
			~2 days	~2 days
Peak memory (GB)		3,128	360	360

DIDA

Distributed Indexing Dispatched Alignment

Applications
•Read alignments



DIDA performance - Aligning 1.2B human reads (12 nodes)

Aligner	Fold Speedup	Fold RAM usage improvements
abyss-map	1.8	5.6
bwa	1.1	7.0
bowtie	1.7	4.9
novoalign	2.7	3.4

DIDA performance - Aligning 4.8B spruce reads (12 nodes) to 20 Gbp genome

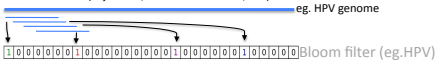
BBT

Bio Bloom Tools

Applications
•Sequence filtering
•Contaminant screening
•Pathogen discovery

Set of tools to generate [Bloom] filters for –and categorize- sequences

BioBloomMaker (BBM)
-Builds filters - Generates loadable binary file and human readable text file from any fasta file
-Highly customizable filters (adjust FPR, hash functions used, etc.)

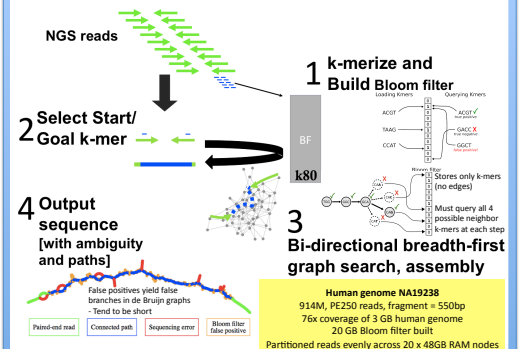


BioBloomCategorizer (BBC)
-Categorizes sequences – Outputs an analysis summary of hits to a particular organism and bins reads into separate fasta files

Konnector

Bloom Filter DBG Assembler

Applications
•De novo assembly
•SNV profiling
•Gap filling (below)

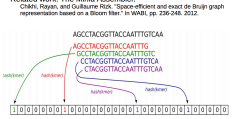


Human genome NA19238
914M, PE250 reads, fragment = 550bp
76x coverage of 3 GB human genome
20 GB Bloom filter built
Partitioned reads evenly across 20 x 48GB RAM nodes
Wall clock time: 4.5 days / Over 50% connected

Bloom Filter de Bruijn Graph

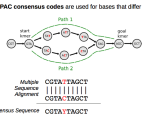
- Assemblers typically use hash tables to store kmers
- A **Bloom filter** is a more memory efficient alternative
  - Like a hash table, but uses one bit per bucket
- Related work: The *Minia* Assembler.

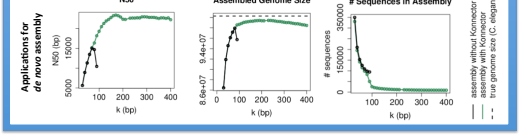
Chikhi, Rayan, and Guillaume Rizk. "Space-efficient and exact de Bruijn graph representation based on a Bloom filter." In WABI, pp. 236-248. 2012.



Reconciling Alternate Paths

- Konnector exhaustively searches for all connecting paths
- If multiple paths are found, a multiple sequence alignment is performed
  - **IUPAC consensus codes** are used for bases that differ

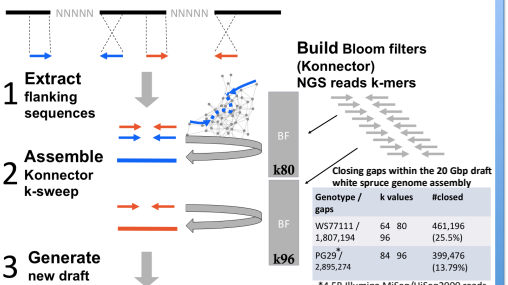




Sealer

Scalable Gap Filler

Applications
•Gap filling



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

Genome Sciences Centre • British Columbia Cancer Agency • 100 – W 7th Ave Vancouver BC V5Z 4S6 Canada • tel. 604-707-5900 • fax. 604-876-3561 • www.bcgsc.ca