Massively parallel assemblers for massively parallel DNA sequencers

Length: 1 hour

Sébastien Boisvert
Ph.D. student, Laval University
CIHR doctoral scholar

UNIVERSITÉ LAVAL

CIHR IRSC
Canadian Institutes of Health Research
Institute of Genetics
Instituts de recherche en santé du Canada
L'Institut de génétique

Élénie Godzaridis
Strategic Technology Projects
Bentley Systems, Inc.

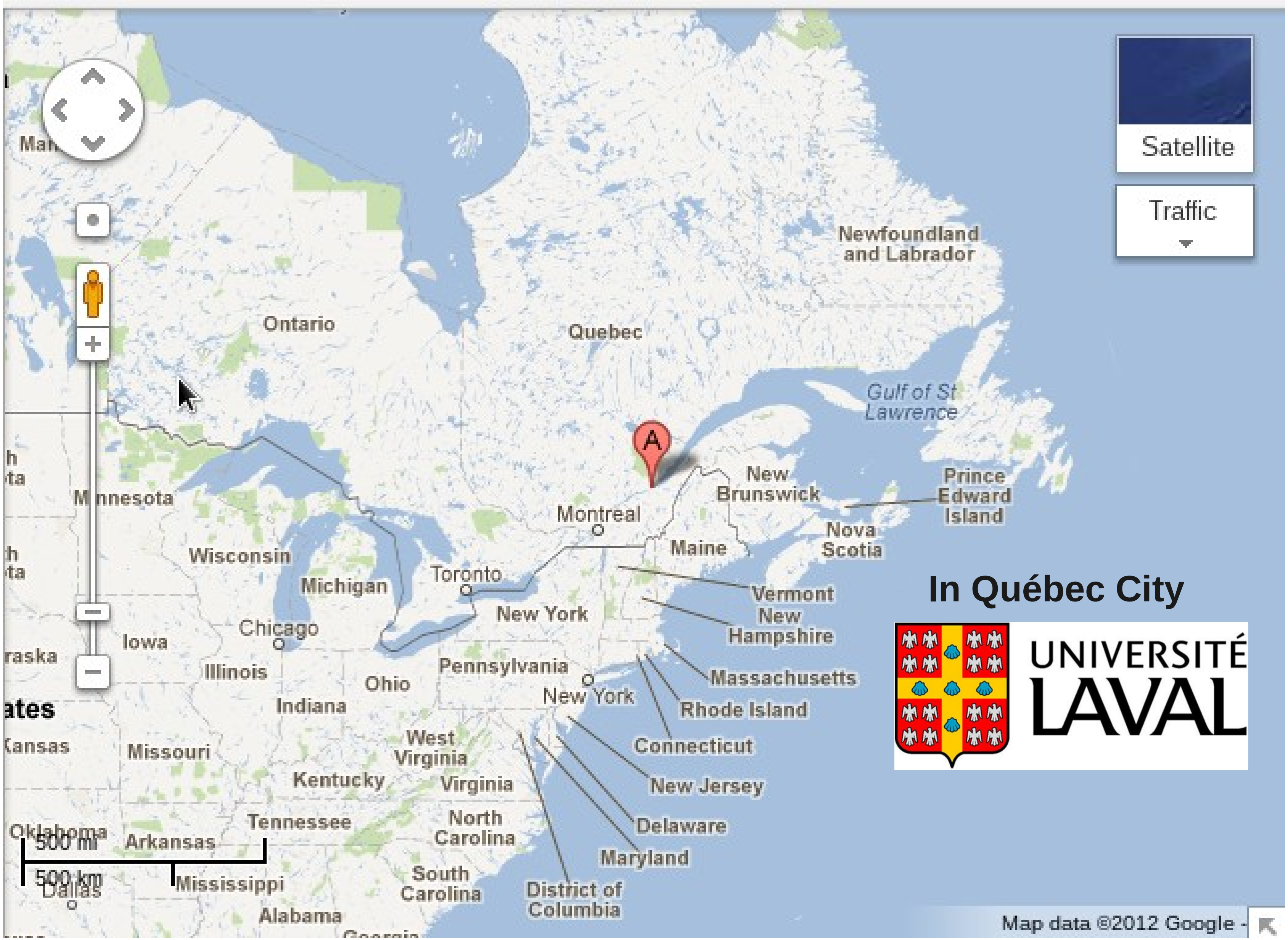
Meta-data

- Invited by Daniel Gruner (SciNet, Compute Canada)
- Start: 2012-11-27 12:00 End: 2012-11-27 14:00
- Location: SciNet offices at 256 McCaul Street, Toronto, 2nd Floor.
- https://support.scinet.utoronto.ca/courses/?q=node/94
- SciNet Seminar by Sébastien Boisvert and Élénie Godzaridis, developers of the parallel genome assembler "Ray".

Introductions

- Who are we ?
- **Sébastien:** message passing, software development, biological systems, repeats in genomes, usability, scalability, correctness, open innovation, Linux
- **Élénie:** software engineering, blueprints, designs, books, biochemistry, life, rendering engines, geometry, web technologies, cloud, complex systems

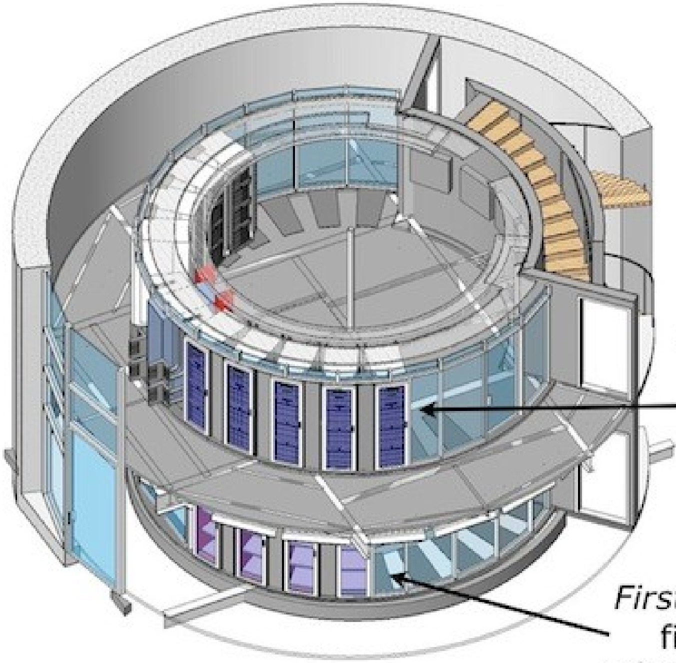
Where is Laval University ?



In Québec City

UNIVERSITÉ LAVAL

Super computing at Laval University



Racks aligned in a circle around a central hot core; outside ring is a cold air plenum

Second floor contains all compute racks + core networking switches

First floor contains file system & infrastructure nodes

colosse
#314 top500 06/2012
7616 Intel Xeon X5560 cores
Mellanox Technologies MT26428
332 kW

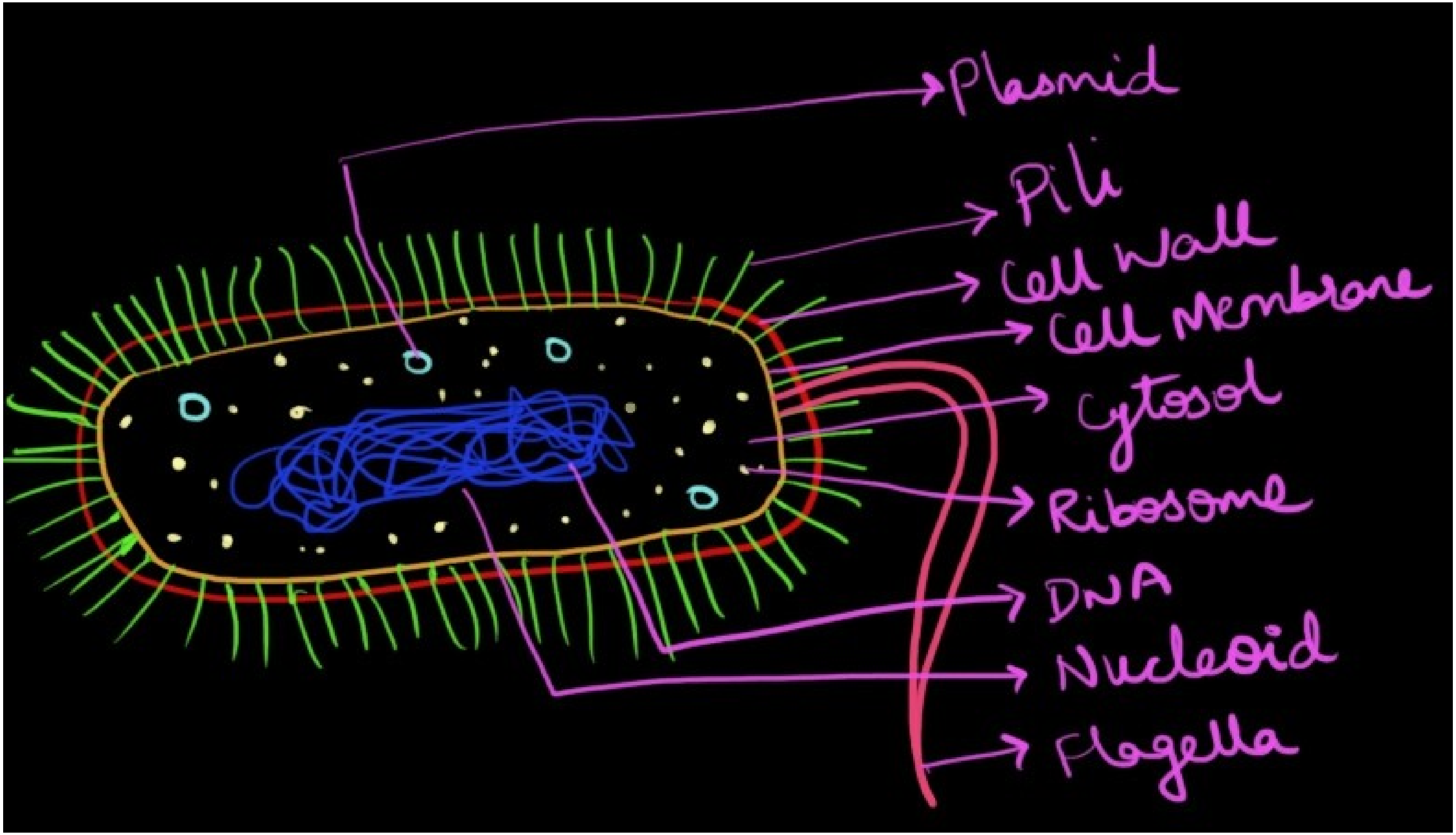
2010 InfoWorld GREEN 15

Contents

- Current-generation DNA sequencers
- Survey of assemblers
- Why parallel is important
- Ray, Ray Meta, Ray Communities
- Workflows with Ray
- Test on Amazon EC2
- Ray Cloud Browser (HTML5 de Bruijn graph explorer)

- Current-generation DNA sequencers

Why bother with DNA ?



Current-generation DNA sequencers

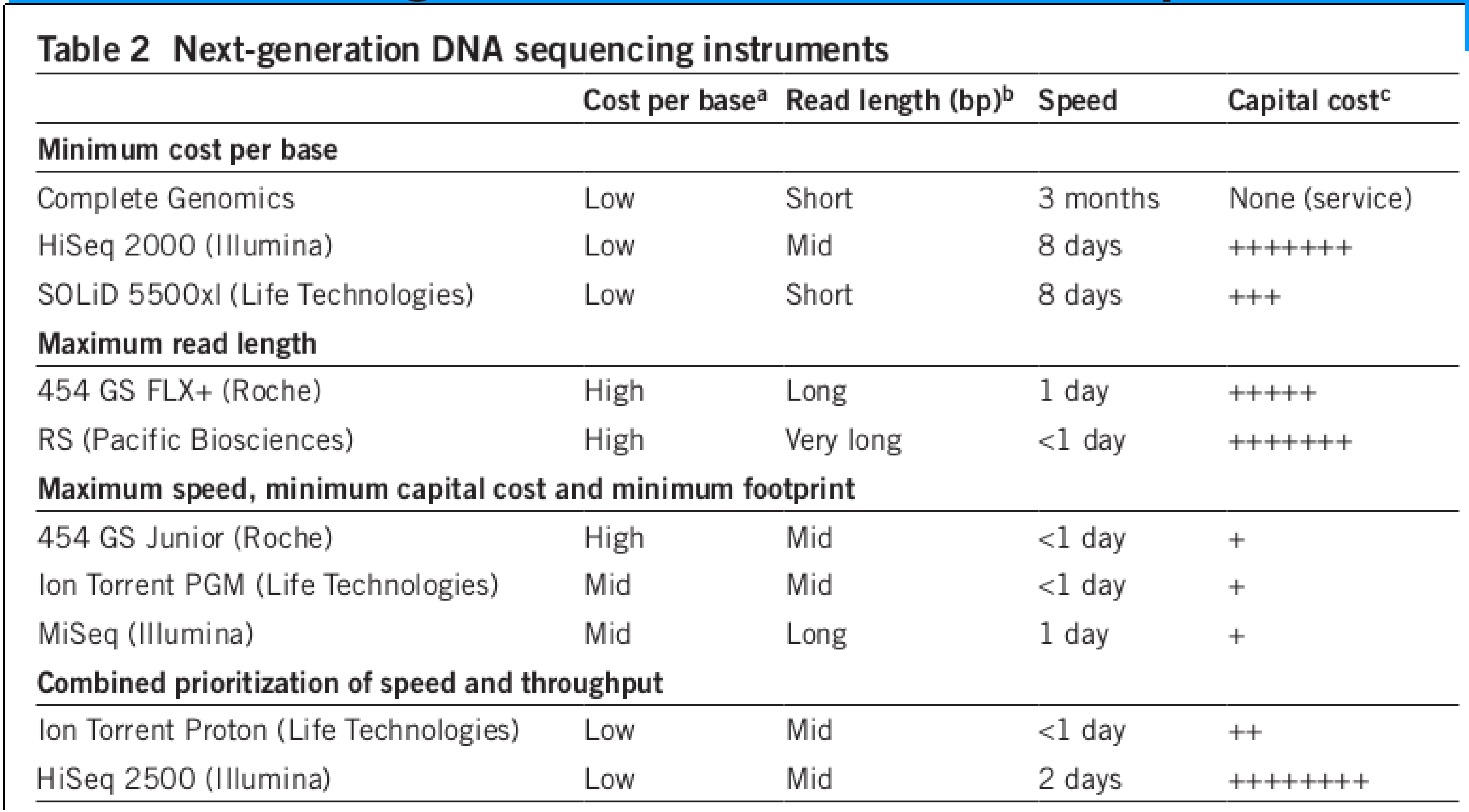


Table 2 Next-generation DNA sequencing instruments

	Cost per base[a]	Read length (bp)[b]	Speed	Capital cost[c]
Minimum cost per base				
Complete Genomics	Low	Short	3 months	None (service)
HiSeq 2000 (Illumina)	Low	Mid	8 days	+++++++
SOLiD 5500xl (Life Technologies)	Low	Short	8 days	+++
Maximum read length				
454 GS FLX+ (Roche)	High	Long	1 day	+++++
RS (Pacific Biosciences)	High	Very long	<1 day	+++++++
Maximum speed, minimum capital cost and minimum footprint				
454 GS Junior (Roche)	High	Mid	<1 day	+
Ion Torrent PGM (Life Technologies)	Mid	Mid	<1 day	+
MiSeq (Illumina)	Mid	Long	1 day	+
Combined prioritization of speed and throughput				
Ion Torrent Proton (Life Technologies)	Low	Mid	<1 day	++
HiSeq 2500 (Illumina)	Low	Mid	2 days	++++++++

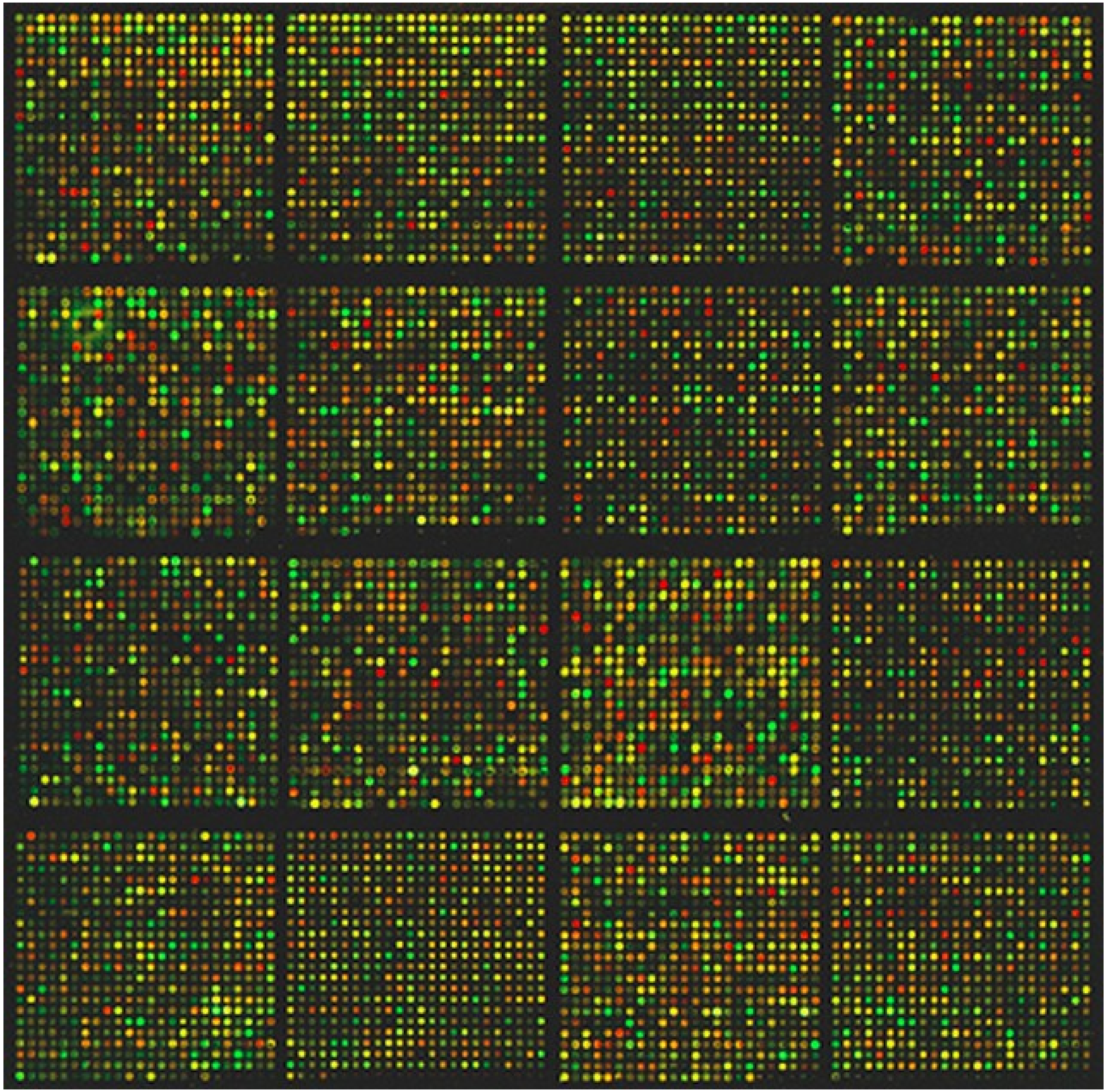
Jay Shendure & Erez Lieberman Aiden
Nature Biotechnology 30, 1084–1094 (2012)

Illumina HiSeq 2000

Read Length	Single Flow Cell Run Time	Dual Flow Cell Run Time	Output*
1 × 36 bp	~ 1.5 days	~ 2 days	105 Gb
2 × 50 bp	~ 4.5 days	~ 5.5 days	270-300 Gb
2 × 100 bp	~ 8.5 days	~ 11 days	540-600 Gb

Run Type	Reads Passing FIlter
Single Read	Up to 3 billion
Paired-End Read	Up to 6 billion

Arrays of bio objects



AV-0101-5194 Dr. Jason Kang, NCI (Lance Miller)

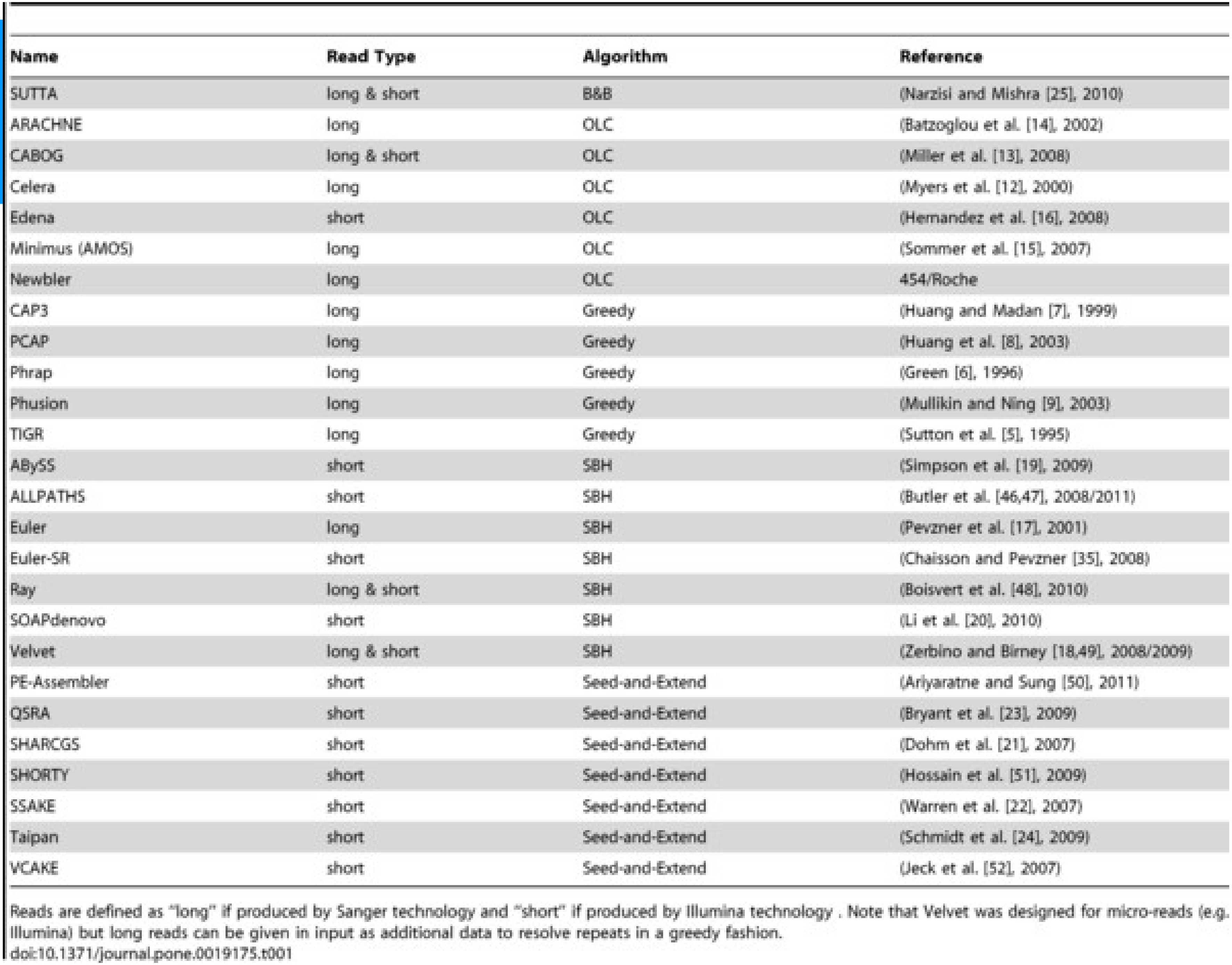
- Survey of assemblers

de novo genome assembly

What is the desire of biologists regarding NGS** analysis

- Features of a biologist-friendly tool:
  - **Correctness** of results
  - **Usability** (**fun to use** versus painful to get started)
  - **Scalability**
    - Can I use more computing power if I have more data ?
    - And does the software scale well ?
  - **Versatility** – Can I reuse the same tool for various related tasks ?
  - **Open**: improve / redistribute the product ?

**NGS = Next Generation Sequencing

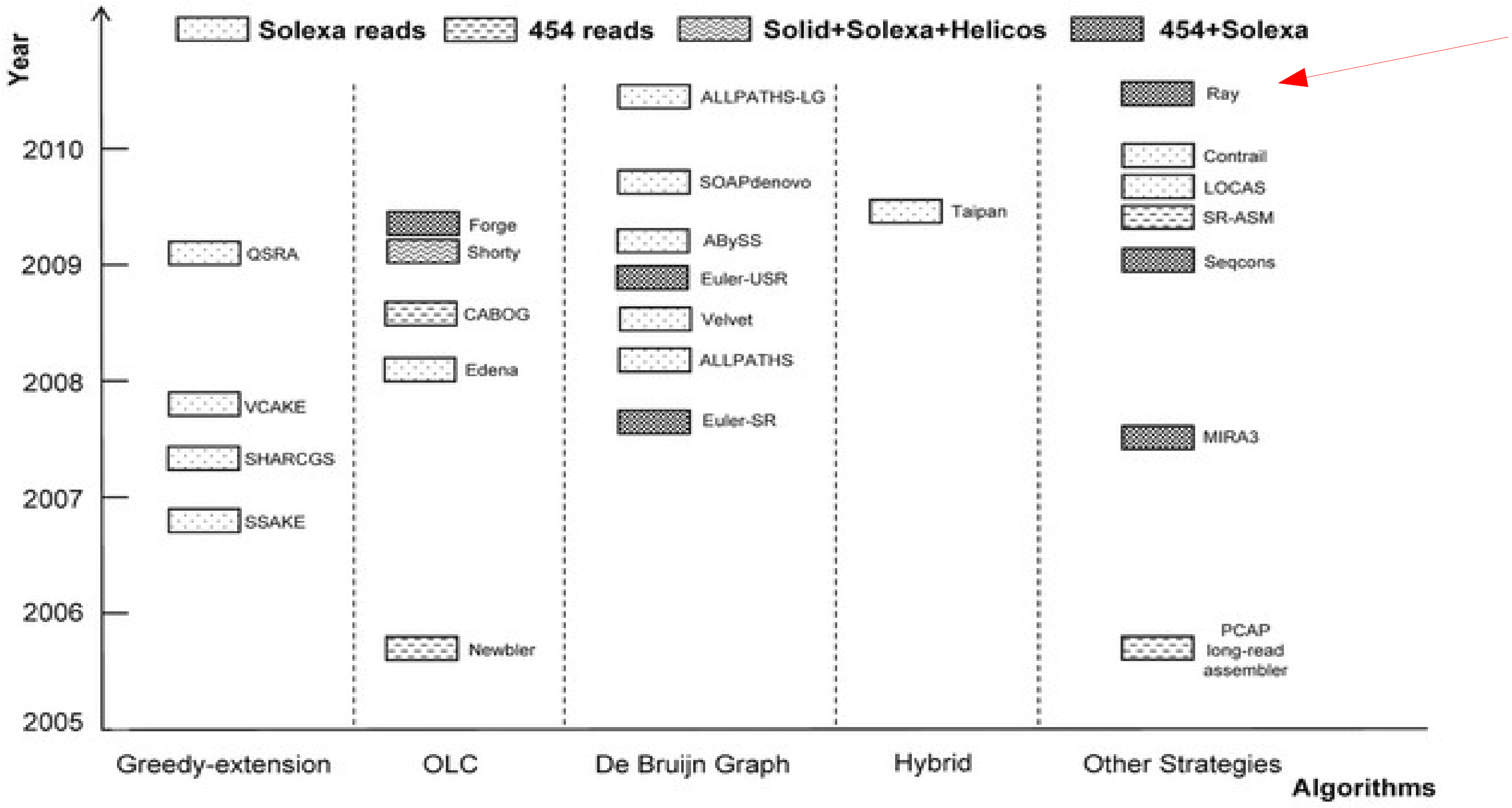


Name	Read Type	Algorithm	Reference
SUTTA	long & short	B&B	(Narzisi and Mishra [25], 2010)
ARACHNE	long	OLC	(Batzoglou et al. [14], 2002)
CABOG	long & short	OLC	(Miller et al. [13], 2008)
Celera	long	OLC	(Myers et al. [12], 2000)
Edena	short	OLC	(Hernandez et al. [16], 2008)
Minimus (AMOS)	long	OLC	(Sommer et al. [15], 2007)
Newbler	long	OLC	454/Roche
CAP3	long	Greedy	(Huang and Madan [7], 1999)
PCAP	long	Greedy	(Huang et al. [8], 2003)
Phrap	long	Greedy	(Green [6], 1996)
Phusion	long	Greedy	(Mullikin and Ning [9], 2003)
TIGR	long	Greedy	(Sutton et al. [5], 1995)
ABySS	short	SBH	(Simpson et al. [19], 2009)
ALLPATHS	short	SBH	(Butler et al. [46,47], 2008/2011)
Euler	long	SBH	(Pevzner et al. [17], 2001)
Euler-SR	short	SBH	(Chaisson and Pevzner [35], 2008)
Ray	long & short	SBH	(Boisvert et al. [48], 2010)
SOAPdenovo	short	SBH	(Li et al. [20], 2010)
Velvet	long & short	SBH	(Zerbino and Birney [18,49], 2008/2009)
PE-Assembler	short	Seed-and-Extend	(Ariyaratne and Sung [50], 2011)
QSRA	short	Seed-and-Extend	(Bryant et al. [23], 2009)
SHARCGS	short	Seed-and-Extend	(Dohm et al. [21], 2007)
SHORTY	short	Seed-and-Extend	(Hossain et al. [51], 2009)
SSAKE	short	Seed-and-Extend	(Warren et al. [22], 2007)
Taipan	short	Seed-and-Extend	(Schmidt et al. [24], 2009)
VCAKE	short	Seed-and-Extend	(Jeck et al. [52], 2007)

Reads are defined as "long" if produced by Sanger technology and "short" if produced by Illumina technology . Note that Velvet was designed for micro-reads (e.g. Illumina) but long reads can be given in input as additional data to resolve repeats in a greedy fashion.

doi:10.1371/journal.pone.0019175.t001

Narzisi G, Mishra B (2011) PLoS ONE 6(4) e19175



Zhang W, Chen J, Yang Y, Tang Y, Shang J, et al. (2011) PLoS ONE 6(3): e17915

Quality of results

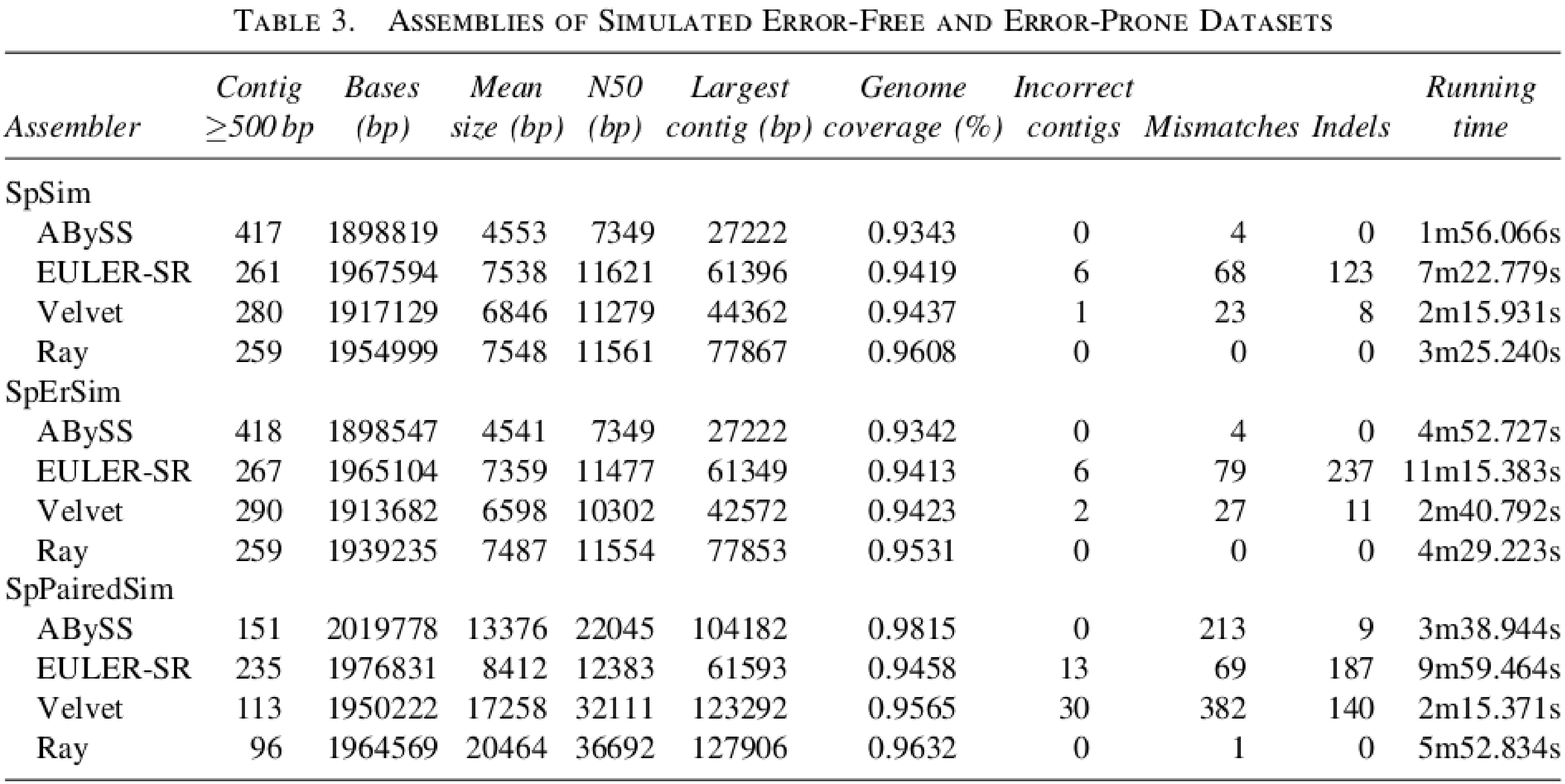


Table 3. Assemblies of Simulated Error-Free and Error-Prone Datasets

Assembler	*Contig ≥500 bp*	*Bases (bp)*	*Mean size (bp)*	*N50 (bp)*	*Largest contig (bp)*	*Genome coverage (%)*	*Incorrect contigs*	*Mismatches*	*Indels*	*Running time*
SpSim										
ABySS	417	1898819	4553	7349	27222	0.9343	0	4	0	1m56.066s
EULER-SR	261	1967594	7538	11621	61396	0.9419	6	68	123	7m22.779s
Velvet	280	1917129	6846	11279	44362	0.9437	1	23	8	2m15.931s
Ray	259	1954999	7548	11561	77867	0.9608	0	0	0	3m25.240s
SpErSim										
ABySS	418	1898547	4541	7349	27222	0.9342	0	4	0	4m52.727s
EULER-SR	267	1965104	7359	11477	61349	0.9413	6	79	237	11m15.383s
Velvet	290	1913682	6598	10302	42572	0.9423	2	27	11	2m40.792s
Ray	259	1939235	7487	11554	77853	0.9531	0	0	0	4m29.223s
SpPairedSim										
ABySS	151	2019778	13376	22045	104182	0.9815	0	213	9	3m38.944s
EULER-SR	235	1976831	8412	12383	61593	0.9458	13	69	187	9m59.464s
Velvet	113	1950222	17258	32111	123292	0.9565	30	382	140	2m15.371s
Ray	96	1964569	20464	36692	127906	0.9632	0	1	0	5m52.834s

Sébastien Boisvert, François Laviolette, and Jacques Corbeil.

Journal of Computational Biology. November 2010, 17(11): 1519-1533.

Quality of results

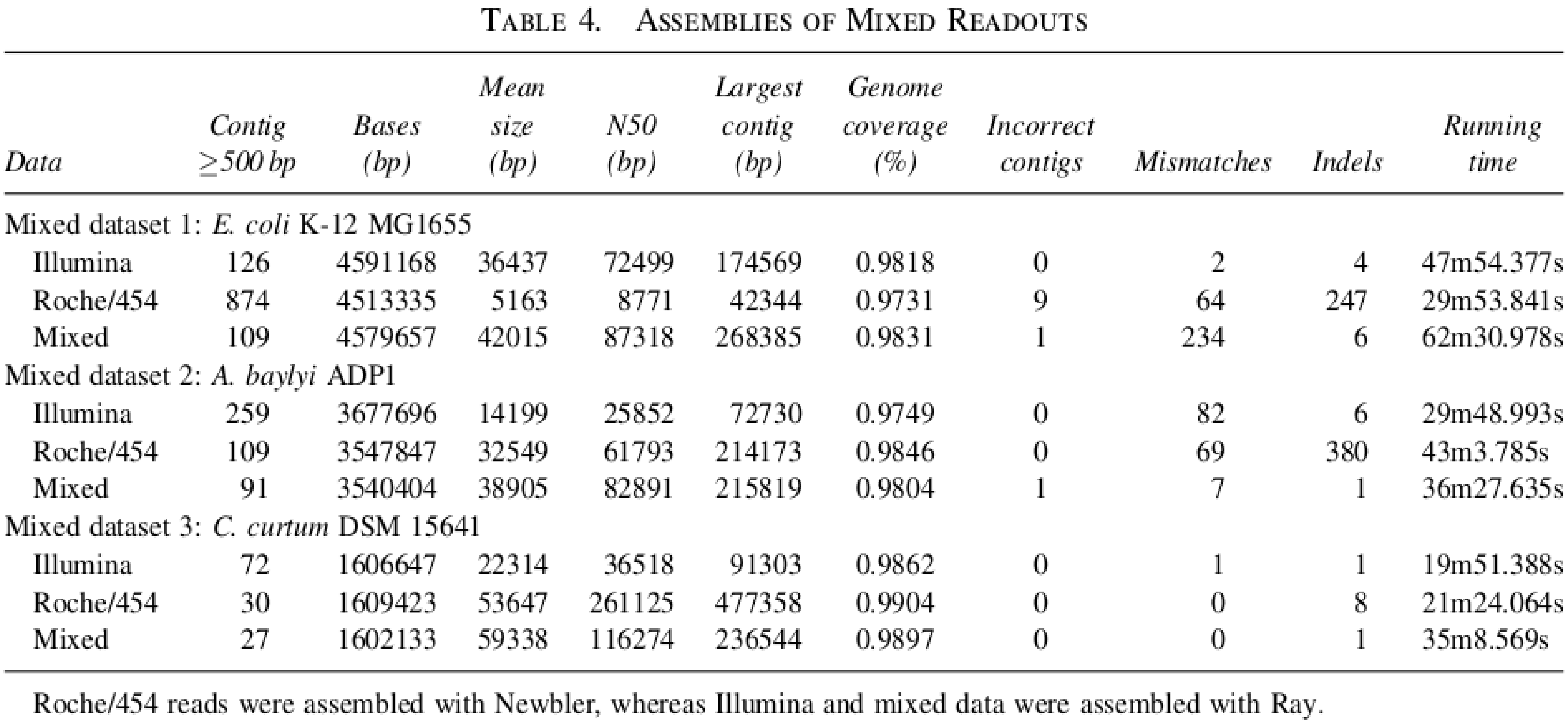


TABLE 4. ASSEMBLIES OF MIXED READOUTS

Data	Contig ≥500 bp	Bases (bp)	Mean size (bp)	N50 (bp)	Largest contig (bp)	Genome coverage (%)	Incorrect contigs	Mismatches	Indels	Running time
Mixed dataset 1: *E. coli* K-12 MG1655										
Illumina	126	4591168	36437	72499	174569	0.9818	0	2	4	47m54.377s
Roche/454	874	4513335	5163	8771	42344	0.9731	9	64	247	29m53.841s
Mixed	109	4579657	42015	87318	268385	0.9831	1	234	6	62m30.978s
Mixed dataset 2: *A. baylyi* ADP1										
Illumina	259	3677696	14199	25852	72730	0.9749	0	82	6	29m48.993s
Roche/454	109	3547847	32549	61793	214173	0.9846	0	69	380	43m3.785s
Mixed	91	3540404	38905	82891	215819	0.9804	1	7	1	36m27.635s
Mixed dataset 3: *C. curtum* DSM 15641										
Illumina	72	1606647	22314	36518	91303	0.9862	0	1	1	19m51.388s
Roche/454	30	1609423	53647	261125	477358	0.9904	0	0	8	21m24.064s
Mixed	27	1602133	59338	116274	236544	0.9897	0	0	1	35m8.569s

Roche/454 reads were assembled with Newbler, whereas Illumina and mixed data were assembled with Ray.

Sébastien Boisvert, François Laviolette, and Jacques Corbeil.

Journal of Computational Biology. November 2010, 17(11): 1519-1533.

Ray in 2012

- Our main claim is scalability
- For correctness: ALLPATHS
- For memory usage: sga

Ray in 2012 and beyond

- Ray Meta for metagenomics
- Metagenome assemblers: Genovo, Meta-IDBA, MetaVelvet, Ray Meta
- Boisvert et al. 2012 Genome Biology (accepted)

Some results with Ray Meta

- All these results are on Colosse
- Round-trip in-application point-to-point latency > 100 microseconds for 512-process jobs
- 3 000 000 000 reads from a 1000-bacterium metagenome, 15 hours on 1024 cores
- 400 000 000 reads from 100-bacterium metagenome, 14 hours, 128 cores
- Includes also k-mer based profiling (genome abundance, taxonomy, gene ontology)

Steps for 1000-genome

- Network testing: 3 minutes, 55 seconds
- Counting sequences to assemble: 2 minutes, 12 seconds
- Sequence loading: 24 minutes, 32 seconds
- K-mer counting: 32 minutes, 50 seconds
- Coverage distribution analysis: 3 seconds
- Graph construction: 1 hours, 21 minutes, 35 seconds
- Null edge purging: 28 minutes, 3 seconds
- Selection of optimal read markers: 44 minutes, 11 seconds
- Detection of assembly seeds: 46 minutes, 58 seconds
- Estimation of outer distances for paired reads: 23 minutes, 36 seconds
- Bidirectional extension of seeds: 3 hours, 25 minutes, 50 seconds
- Merging of redundant paths: 4 hours, 27 minutes, 55 seconds
- Generation of contigs: 5 minutes, 48 seconds
- Scaffolding of contigs: 2 hours, 4 minutes, 7 seconds
- Counting sequences to search: 19 seconds
- Graph coloring: 18 minutes, 18 seconds
- Counting contig biological abundances: 3 minutes, 44 seconds
- Counting sequence biological abundances: 31 minutes, 50 seconds
- Loading taxons: 22 seconds
- Loading tree: 14 seconds
- Processing gene ontologies: 6 seconds
- Computing neighbourhoods: 0 seconds
- Total: 15 hours, 46 minutes, 41 seconds

- **Why parallel is important**

Parallel sequencers, computers, & software tools

- DNA sequencers are parallel with distributed clusters on a array (Illumina) or on beads (454)
- Computers are parallel and distributed -- think IBM Blue Gene/Q, Cray XE6, IBM iDataPlex, or Beowulf clusters
- Next-generation gap between sequencing and processing hardware and analysis software

John D McPherson
Nature Methods 6, S2 - S5 (2009)

Processors are parallel too !

- AMD Opteron 6200 has 16 cores, 16 threads
- Intel Xeon E5-2690 has 8 cores & 16 threads
- IBM PowerPC A2 has 16 cores, 64 threads

Parallel compute tracks

- Ray uses all available tracks on computing infrastructure
- Ray's parallelism matches the parallelism of super computers and DNA sequencers

- Ray, Ray Meta, Ray Communities

Why care about Ray?

Does quality control

De novo bacterial genome assembly

De novo metagenome assembly

Plant genomes*

Open source git repository GNU GPLv3

Mammal genomes*

Runs on netbooks or super computers

Well engineered

Portable C++ 1998 MPI

Supports paired reads

Runs on 1 or more processes

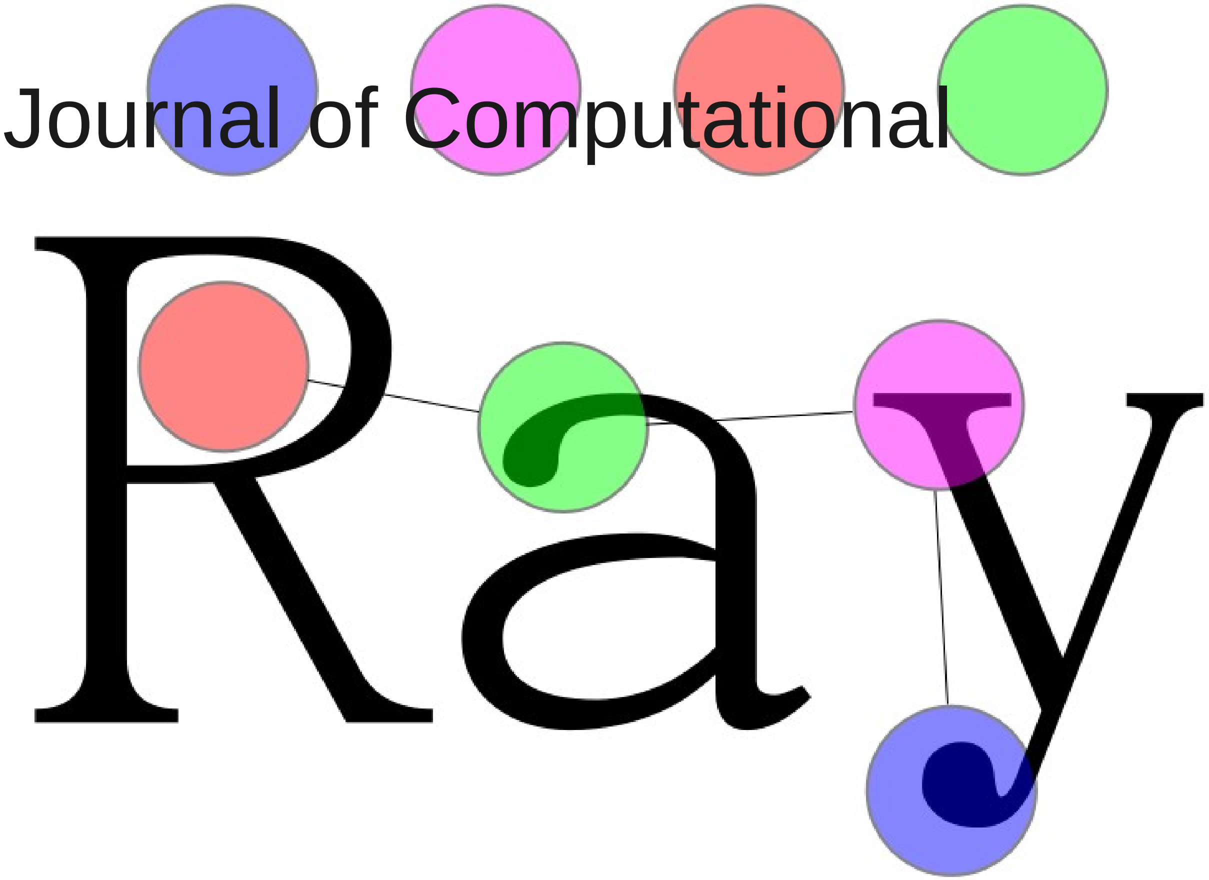
1 single executable
Easy to install
Easy to run

Supports compressed gz and bz2 files

*Results may vary

Learn more about Ray

- Boisvert et al. Genome Biology 2012 (accepted)
- Boisvert et al. Journal of Computational Biology 2010



- One-stop resource: http://DeNovoAssembler.SF.NET

- + Mailing list

- **Workflows with Ray**

- Easy to install
- Easy to use
- 1 program called Ray

Ray *de novo* assembly of single genomes

Ray$_{\textbf{Méta}}$ *de novo* assembly of metagenomes

Ray$_{\textbf{Communities}}$ microbe abundance + taxonomic profiling

Ray$_{\textbf{Ontologies}}$ gene ontology profiling

- Test on Amazon EC2

- Cost Effectiveness Analysis (CEA) of running Ray on Amazon EC2

https://github.com/sebhtml/Ray-in-Amazon-EC2-CLOUD

- Sample: SRA001125 (E. coli)
- URL: http://trace.ddbj.nig.ac.jp/DRASearch/submission?acc=SRA001125
- DNA reads: 34911784 (2 * 17455892)
- Read length (nt): 36
- Technology: Illumina Genome Analyzer

- Why use Ray?
-
- 1. It gives correct (excellent) results.
- 2. It's 0 $.
- 3. It's free software (freedom).
- 4. It runs on all the cores you give it.
- 5. It scales.
- 6. It's "cloud-ready".

- API name: m1.large
- 2 Rays
- Running time: 05:28:46
- Pricing: 0.260 $ / h
- Cost: 1.560 $

- API name: m3.xlarge
- 4 Rays
- Running time: 02:31:34
- Pricing: 0.580 $ / h
- Cost: 1.730 $

- API name: cc2.8xlarge
- 32 Rays
- Running time: 00:54:06
- Pricing: 2.400 / h
- Cost: 2.400 $

- Conclusions:
- 1. You get your results faster if you pay more.
- 2. For cc2.8xlarge, 33% (00:19:40) of the time was loading sequences from EBS.
- That's a lot !
- 3. The scalability on this problem is not that good because the
- problem size is not very large.
- 4. Amazon EC2 is really affordable for de novo assemblies of bacterial genomes.

- **Ray Cloud Browser (HTML5 de Bruijn graph explorer)**

https://github.com/sebhtml/Ray-Cloud-Browser

Conclusion

- Compute Canada is Infrastructure as a Service, free for academics!
- Automation is everything
  - DNA sequencing is automated
  - Compute infrastructure is automated
  - Ray is automated genome assembly in parallel/distributed infrastructure

Acknowledgements / Invitation

- Daniel Gruner (invitation and arrangements)
- Ramses van Zon (reviewed slides)

Acknowledgements / Funding

- SB: Doctoral award, **Canadian Institutes for Health Research**, September 2010 – August 2013
- Jacques Corbeil: **Canada Research Chair in Medical Genomics**
- Discovery Grants Program (Individual, Team and Subatomic Physics Project) from the **Natural Sciences and Engineering Research Council of Canada** (grant 262067 to François Laviolette)

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada
Institute of Genetics
L'Institut de génétique

NSERC
CRSNG

Acknowledgements / Product team

- Sébastien Boisvert (designer, developer, release technician, community manager)
- Élénie Godzaridis (parallel designs, works in the industry)
- Prof. François Laviolette (graph specialist)
- Prof. Jacques Corbeil (genomician)
- Maxime Boisvert (design tricks, consultant in the industry)
- Dr. Frédéric Raymond (end user / stakeholder)
- Pier-Luc Plante (intern)

Acknowledgements / CPU time

- 2011: 50 core-years on Colosse
- 2012: 250 core-years on Colosse
- Compute Canada (Colosse, Mammouth Parallèle II, Guillimin)
- Calcul Québec, CLUMEQ, RQCHP
- Canadian Foundation for innovation for the 32-core 128-GB SMP machine
- Collaboration with Cray Inc. for the Cray XE6 (with Carlos Sosa)

Questions