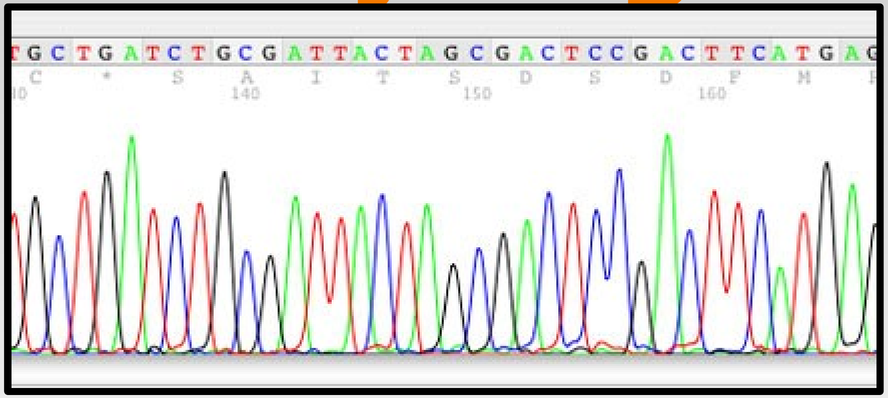
Robert Cedergren Bioinformatics Colloquium 2009,
room S1-139, Jean-Coutu Bldg.

On the fifth of November, 2009, 11h00 - 11h30

OpenAssembler: assembly of reads from a mix of high-throughput sequencing technologies

Sébastien Boisvert François Laviolette Jacques Corbeil

Sequencing and analyzing DNA

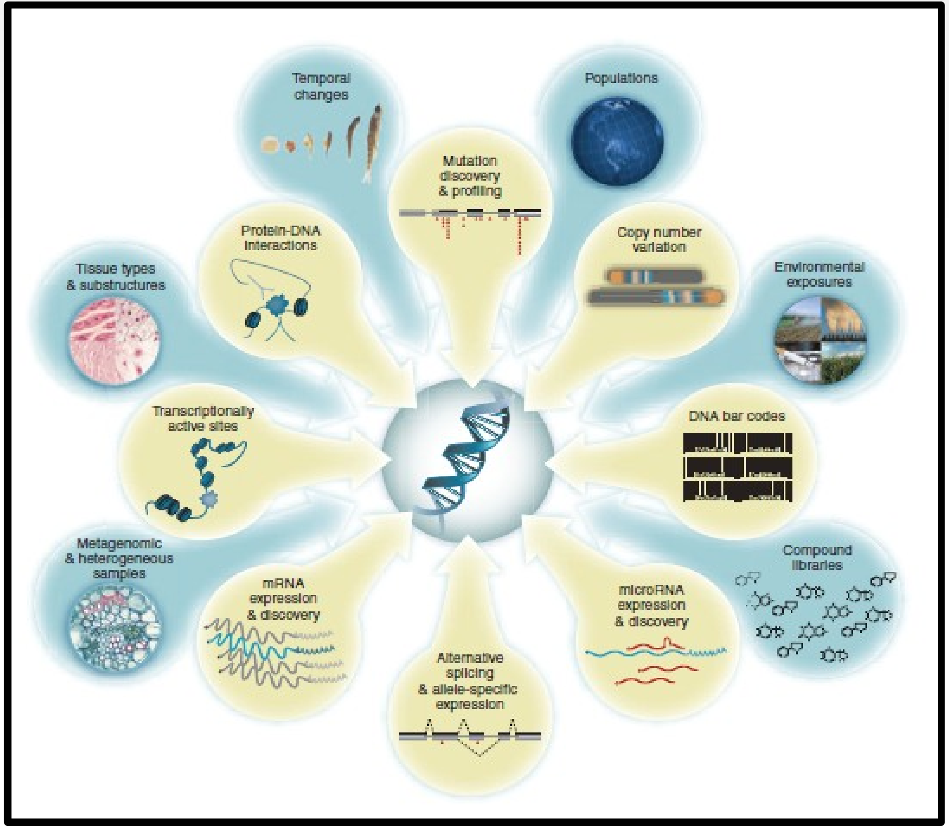


- Sequencing reads DNA
- Determine the primary structure of DNA
- Algorithms can help us!
- Hutchinson (1969) had foreseen the power of graph theory in sequence analysis
- Graph theory is everywhere

Evaluation of polymer sequence fragment data using graph theory.
Hutchinson G.
Bull Math Biophys. 1969 Sep;31(3):541-62.

Why do we decode life?

- Explain and treat genetic diseases (dystonia, huntington disease, Alzheimer's disease,...)
- Rapid detection of pathogenic agents (flu, H1N1, *C. difficile*, *S. pneumoniae*,...)
- Study evolution
- Study speciation
- Bridge the proteome and genome
- Study gene splicing
- Study **genome variation**


Temporal changes
Populations
Mutation discovery & profiling
Protein-DNA interactions
Copy number variation
Tissue types & substructures
Environmental exposures
Transcriptionally active sites
DNA bar codes
Metagenomic & heterogeneous samples
mRNA expression & discovery
Alternative splicing & allele-specific expression
microRNA expression & discovery
Compound libraries

What would you do if you could sequence everything?
Kahvejian A, Quackenbush J, Thompson JF.
Nat Biotechnol. 2008 Oct;26(10):1125-33.

Limits of sequencing

- Uneven genome coverage
- Reproducible errors (example: Roche/454's homopolymer-located errors)
- Contaminations
- Read length shorter than genome length

Technology	Read length (in bases)
Sanger	800
Roche/454	400
Illumina	50

The new paradigm of flow cell sequencing.
Holt RA, Jones SJ.
Genome Res. 2008 Jun;18(6):839-46.

Genome assembly

- DNA assemblers piece together reads to build larger contiguous sequences
- NP-Hard (according to Pop 2009)
- Genome finishing is lengthy
- Minimizing assembly errors is relevant (to avoid the laborious finishing step)

Genome assembly reborn: recent computational challenges.
Pop M.
Brief Bioinform. 2009 Jul;10(4):354-66.

Hybrid assemblies

More than one technology...

A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes.
Goldberg SM et al.
Proc Natl Acad Sci U S A. 2006 Jul 25;103(30):11240-5.

High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies.
Aury JM et al.
BMC Genomics. 2008 Dec 16;9:603.

De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data.
Diguistini S et al.
Genome Biol. 2009 Sep 11;10(9):R94.

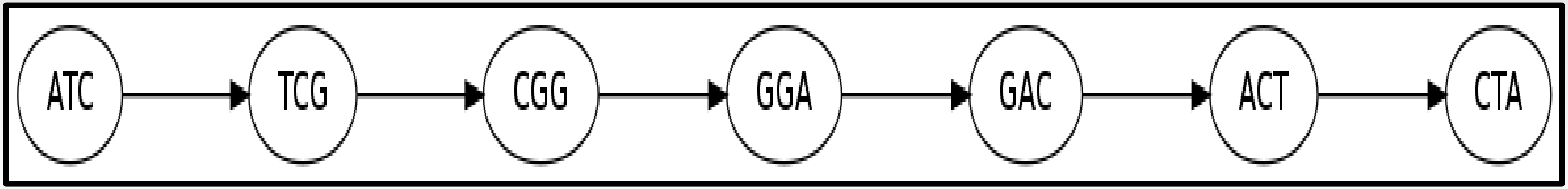
Drawbacks

- These approaches use several tools
- Reads obtained by different technologies are assembled separately
- Each assembler is tailored to a particular technology
- They consider reads from different technologies as being fundamentally different.
- All reads should be born equal!
- Graphs make that possible

de Bruijn and his graphs

Nucleotide space: ATCGGACTA

Graph space (with k=3):



- de Bruijn property: k-1 overlap between adjacent vertices
- Reads naturally induce a de Bruijn graph (with a fixed k)
- An assembly is a set of walks

http://en.wikipedia.org/wiki/De_Bruijn_graph

Assembly with Eulerian paths

- Uses a de Bruijn graph
- Equivalent transformations
- Polynomial
- Very sensitive to errors

An Eulerian path approach to DNA fragment assembly.
Pevzner PA, Tang H, Waterman MS.
Proc Natl Acad Sci U S A. 2001 Aug 14;98(17):9748-53.

De novo fragment assembly with short mate-paired reads: Does the read length matter?
Chaisson MJ, Brinza D, Pevzner PA.
Genome Res. 2009 Feb;19(2):336-46.

Velvet

- Tailored for Illumina
- Similar to EULER-SR
- Error correction
- Very fast

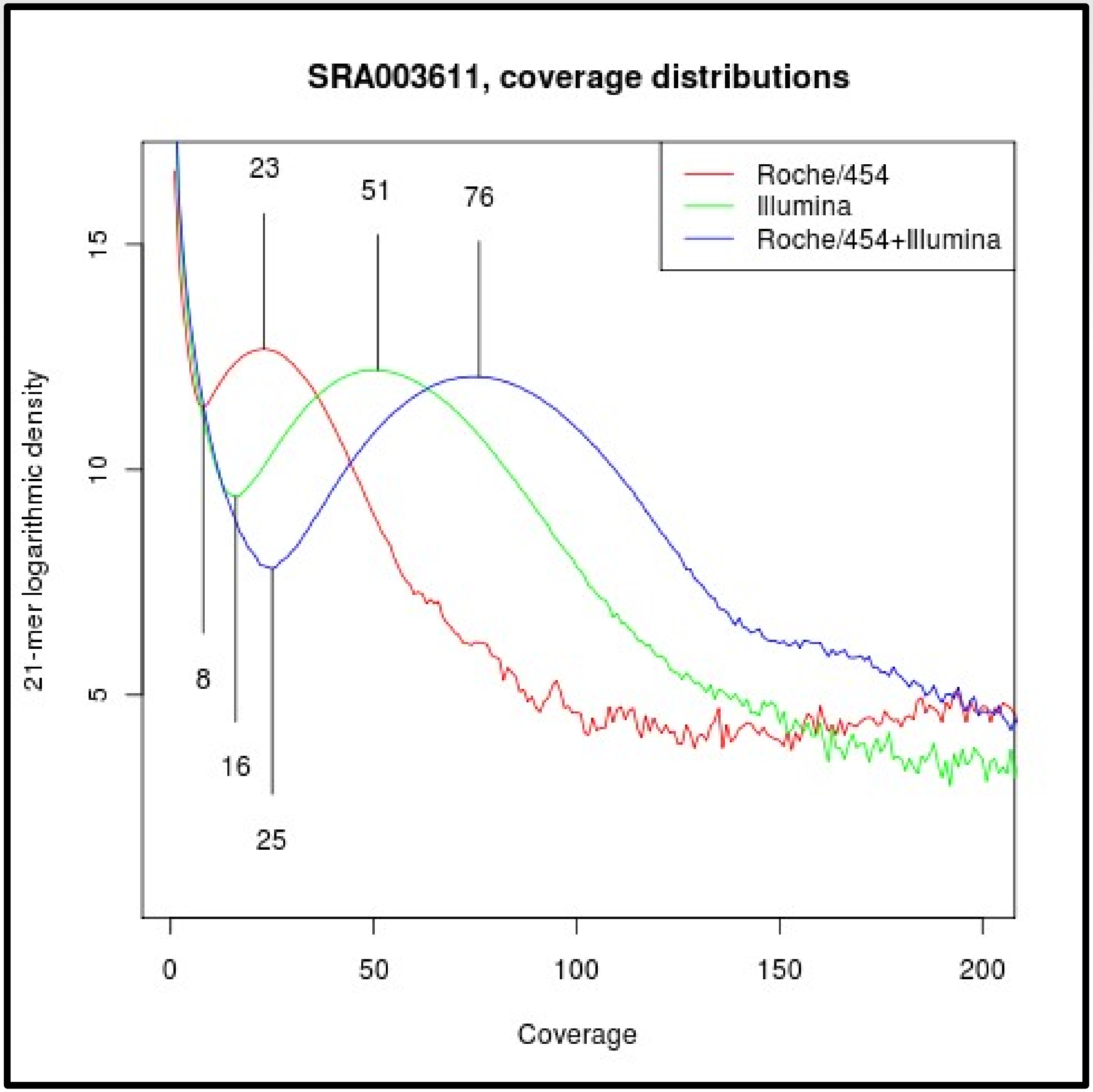
Velvet: algorithms for de novo short read assembly using de Bruijn graphs.
Zerbino DR, Birney E.
Genome Res. 2008 May;18(5):821-9.

OpenAssembler

- No eulerian paths
- No equivalent transformations
- Greedy (owing to the NP-hard nature of the problem)
- All reads have the same rights.

Coverage

- Each vertex of the graph has its depth of coverage – its number of occurences in reads



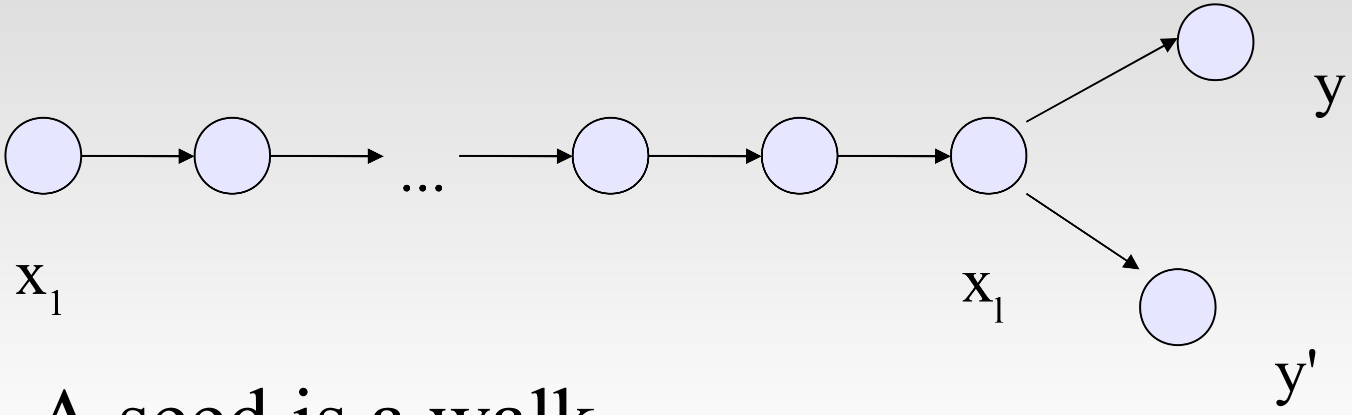
Mixing 454 and Illumina Improves the distribution.

Minimum and peak coverages are important.

Priming the assembly

- Seed coverage: average between minimum and peak coverages
- Seeds: maximal walks with only vertices of in-degree 1 and out-degree 1, and with a depth of coverage a least ”seed coverage”

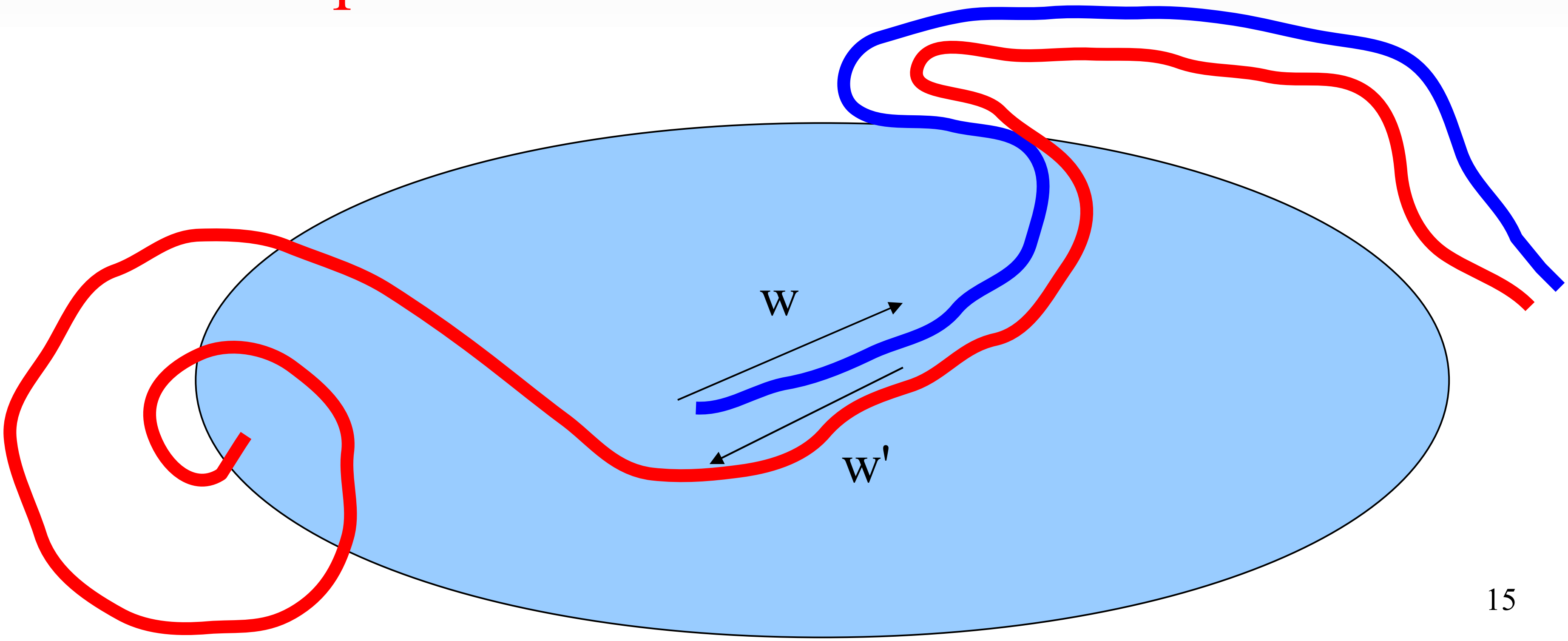
When a seed becomes a grown-up contig



- A seed is a walk.
- Given a walk $<x_1,x_2,\ldots,x_l>$, and two arcs $<x_l,y>$ and $<x_l,y'>$, our algorithm decides which vertex (y or y') is the next to visit
- If the choice is deemed as 'too risky', the extension is stopped.

Bilateral growth

- Each walk w is associated to its reverse-complement walk w'
- Extend w (call the result **w***), and then extend the reverse-complement of w*



OpenAssembler at a glance

- Load reads
- Build the de Bruijn graph (k=21)
- Compute the seeds
- Extend each seed in both directions
- Skip any previously encountered seed
- Write the assembly

- Implemented in c++

The assembler championship

- Two sets of competitions: simulated and real
- Five contenders
- Stringent metrics

Metrics

- Number of contiguous sequences
- Number of bases
- Mean contig length
- Largest contig length
- Genome coverage
- Number of incorrect (chimeric) contigs
- Number of mismatches
- Number of insertions and deletions

Contenders

- The “parallel” AbySS
- The “Eulerian” EULER-SR
- The “commercial” 454 Newbler
- The “greedy” OpenAssembler
- The “fast” Velvet

Living in a virtual world – simulated datasets

- Simulation offers great control – we know the reference sequence.
- SpSim: S. pneumoniae, 50-nt reads, 50 X
- SpErSim: S. pneumoniae, 50-nt reads, 50 X, 1% random mismatch
- SpPairedSim: S. pneumoniae, 50-nt reads, 50 X, paired (fragment length=200)
- EcoliSim: E. coli, 400-nt reads, 50 X

Simulated reads

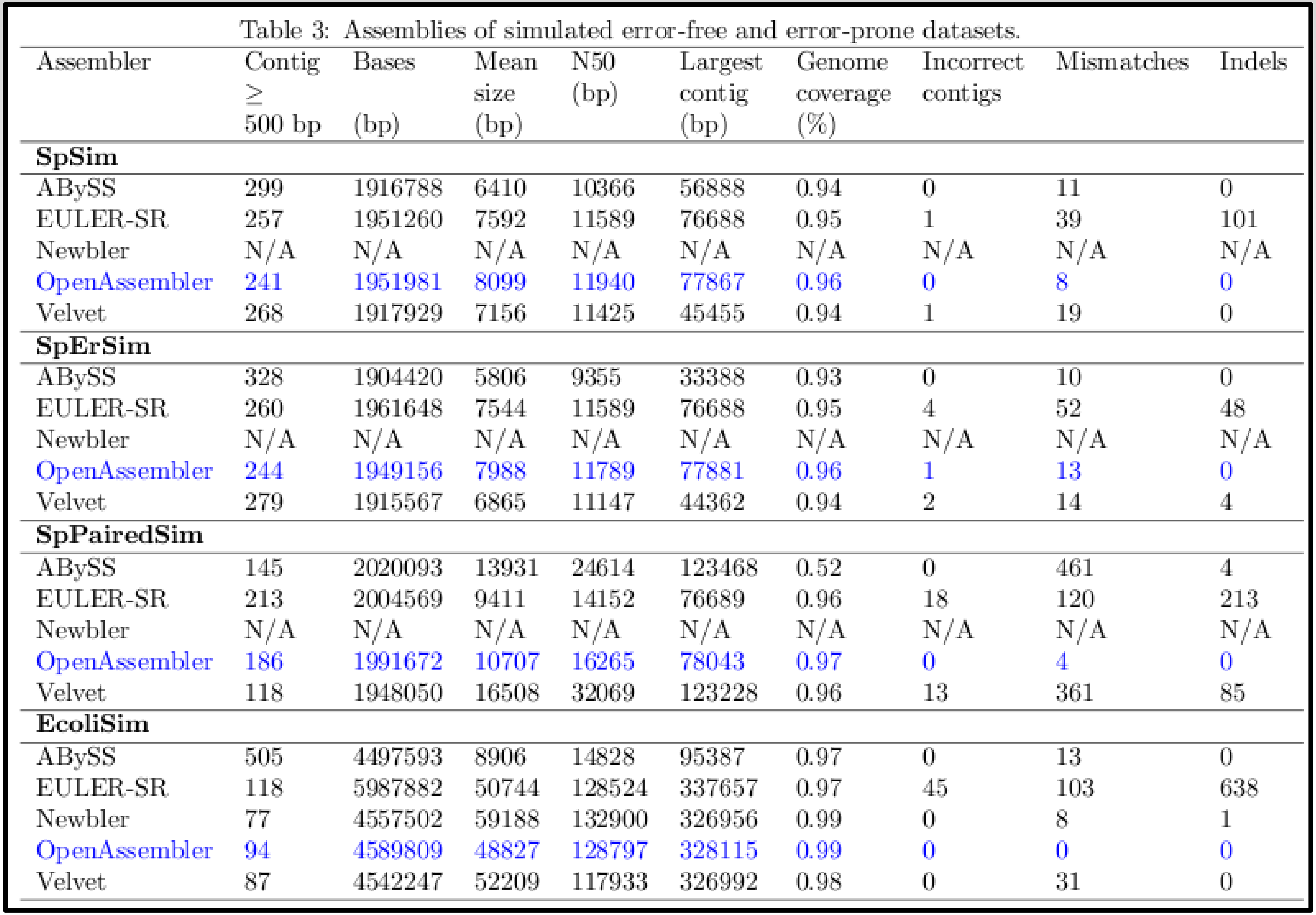
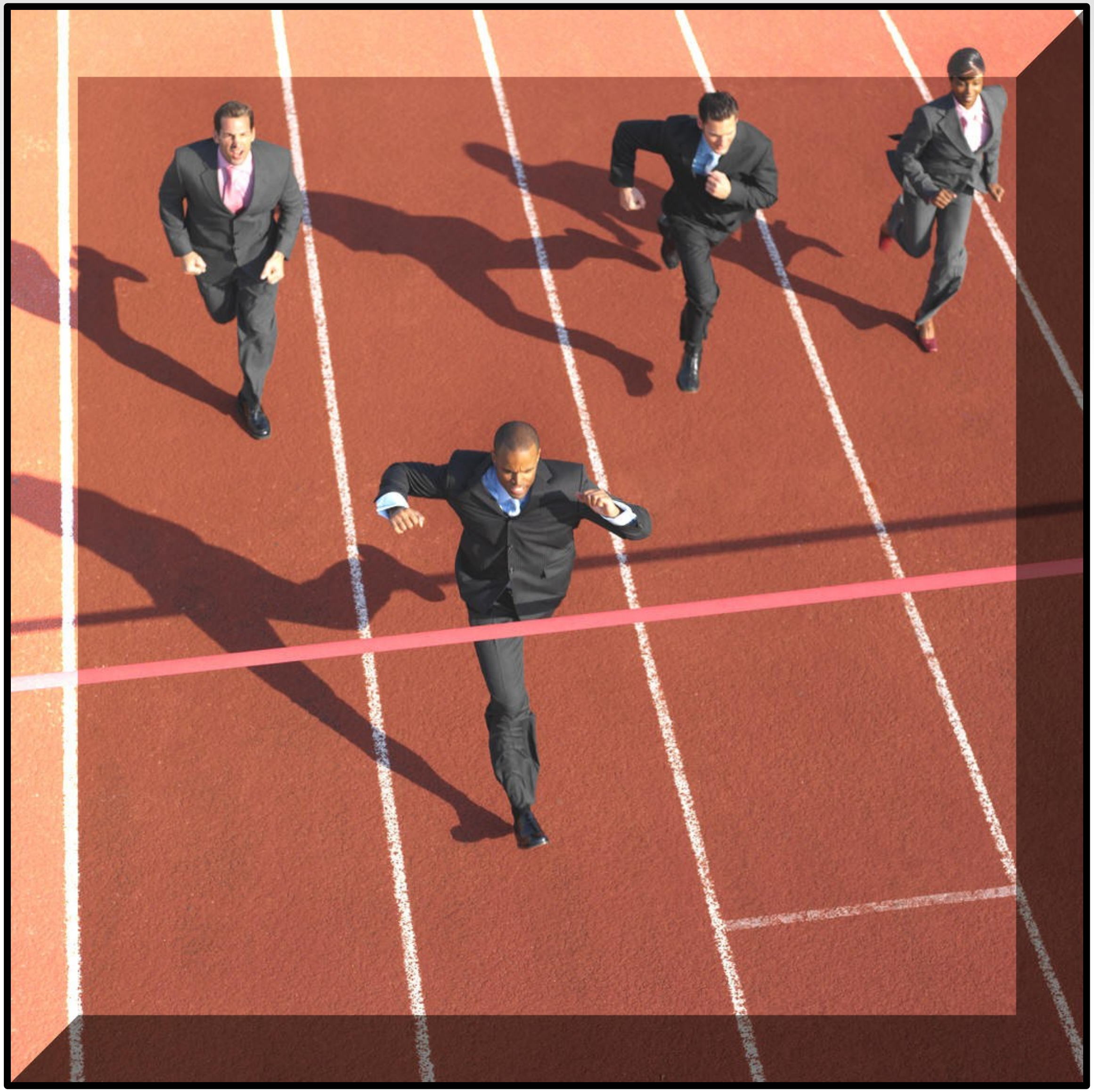


Table 3: Assemblies of simulated error-free and error-prone datasets.

Assembler	Contig ≥ 500 bp	Bases (bp)	Mean size (bp)	N50 (bp)	Largest contig (bp)	Genome coverage (%)	Incorrect contigs	Mismatches	Indels
SpSim									
ABySS	299	1916788	6410	10366	56888	0.94	0	11	0
EULER-SR	257	1951260	7592	11589	76688	0.95	1	39	101
Newbler	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
OpenAssembler	241	1951981	8099	11940	77867	0.96	0	8	0
Velvet	268	1917929	7156	11425	45455	0.94	1	19	0
SpErSim									
ABySS	328	1904420	5806	9355	33388	0.93	0	10	0
EULER-SR	260	1961648	7544	11589	76688	0.95	4	52	48
Newbler	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
OpenAssembler	244	1949156	7988	11789	77881	0.96	1	13	0
Velvet	279	1915567	6865	11147	44362	0.94	2	14	4
SpPairedSim									
ABySS	145	2020093	13931	24614	123468	0.52	0	461	4
EULER-SR	213	2004569	9411	14152	76689	0.96	18	120	213
Newbler	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
OpenAssembler	186	1991672	10707	16265	78043	0.97	0	4	0
Velvet	118	1948050	16508	32069	123228	0.96	13	361	85
EcoliSim									
ABySS	505	4497593	8906	14828	95387	0.97	0	13	0
EULER-SR	118	5987882	50744	128524	337657	0.97	45	103	638
Newbler	77	4557502	59188	132900	326956	0.99	0	8	1
OpenAssembler	94	4589809	48827	128797	328115	0.99	0	0	0
Velvet	87	4542247	52209	117933	326992	0.98	0	31	0

Competition results

- OpenAssembler wins



Facing reality – real datasets

- Simulated reads are useless for real-life applications
- EcoliIllumina: Illumina paired reads, lots of coverage
- A. baylyi ADP1 data: Ab454, AbIllumina, and AbMix
- **Is the mix worth it?**

Real data

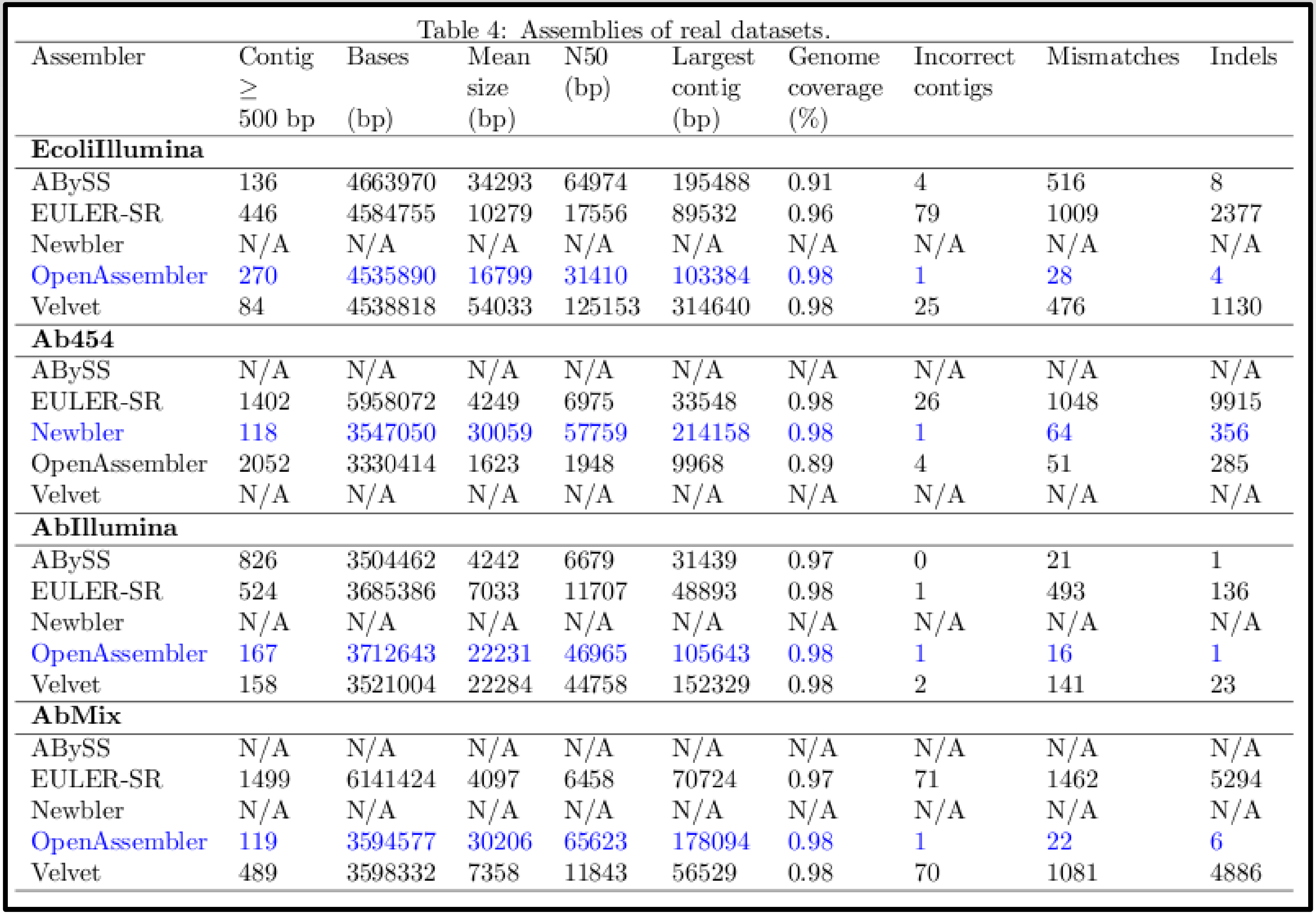


Table 4: Assemblies of real datasets.

Assembler	Contig	Bases	Mean	N50	Largest	Genome	Incorrect	Mismatches	Indels
	≥		size	(bp)	contig	coverage	contigs		
	500 bp	(bp)	(bp)		(bp)	(%)			
EcoliIllumina									
ABySS	136	4663970	34293	64974	195488	0.91	4	516	8
EULER-SR	446	4584755	10279	17556	89532	0.96	79	1009	2377
Newbler	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
OpenAssembler	270	4535890	16799	31410	103384	0.98	1	28	4
Velvet	84	4538818	54033	125153	314640	0.98	25	476	1130
Ab454									
ABySS	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
EULER-SR	1402	5958072	4249	6975	33548	0.98	26	1048	9915
Newbler	118	3547050	30059	57759	214158	0.98	1	64	356
OpenAssembler	2052	3330414	1623	1948	9968	0.89	4	51	285
Velvet	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
AbIllumina									
ABySS	826	3504462	4242	6679	31439	0.97	0	21	1
EULER-SR	524	3685386	7033	11707	48893	0.98	1	493	136
Newbler	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
OpenAssembler	167	3712643	22231	46965	105643	0.98	1	16	1
Velvet	158	3521004	22284	44758	152329	0.98	2	141	23
AbMix									
ABySS	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
EULER-SR	1499	6141424	4097	6458	70724	0.97	71	1462	5294
Newbler	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
OpenAssembler	119	3594577	30206	65623	178094	0.98	1	22	6
Velvet	489	3598332	7358	11843	56529	0.98	70	1081	4886

Who survived?

- 454 is Newbler's ecological niche.
- OpenAssembler is not the winner on 454
- OpenAssembler's excels with Illumina data.
- Mixing is OpenAssembler's specialty.

A. baylyi					
	Genome coverage	**Reads**	**Contigs**	**Mismatches**	**Indels**
Newbler	98%	454	118	64	356
OpenAssembler	98%	Mixed	119	22	6

Closing remarks

- OpenAssembler runs on mixes -- not the others
- OpenAssembler improves the quality of genome drafts
- Quality is important
- One (easy-to-use) tool to rule them all
- Paper submitted

Genome project standards in a new era of sequencing.
Chain PS et al.
Science. 2009 Oct 9;326(5950):236-7.

Acknowledgments

- Jacques Corbeil is the Canada Research Chair in Medical Genomics
- François Laviolette is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC)
- Sébastien Boisvert has a Master's award from the Canadian Institutes of Health Research (CIHR)

Canada Foundation for Innovation
Fondation canadienne pour l'innovation

CRSNG
NSERC

Canada Research Chairs
Chaires de recherche du Canada

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