Applications of micro-, mega-, and meta- assembly

Michael Schatz

Nov. 3, 2011

Genome Informatics

micro-

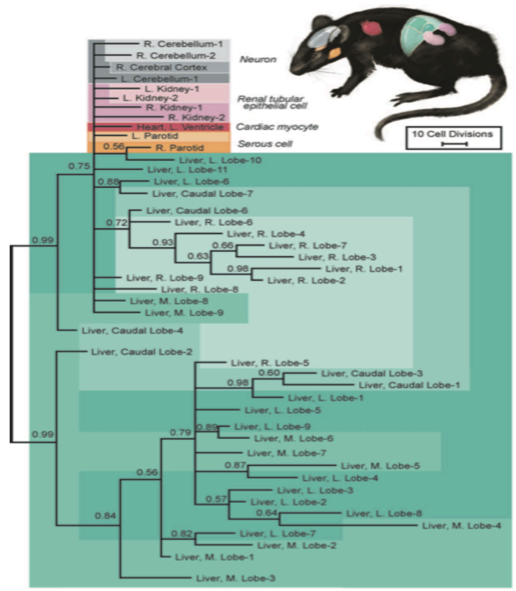
MicroSeq: high-throughput microsatellite genotyping

Mitch Bekritsky, Jennifer Troge, Dan Levy, Michael Wigler, Michael Schatz

- Highly variable simple sequence repeats
  - …GCACACACACAT… = …G(CA)$_5$T…
  - Created and mutate primarily through slippage during replication

- Genotyping with MicroSeq:
  1. Rapidly detect MS sequences
  2. Map reads using a new MS-mapper
  3. Analyze profiles in across cells & populations
     - Loss of heterozygosity, de novo mutations
     - Development of somatic & cancer cells
     - Relations across strains, across species
     - etc…

- Currently being applied to look for de novo mutations associated with autism

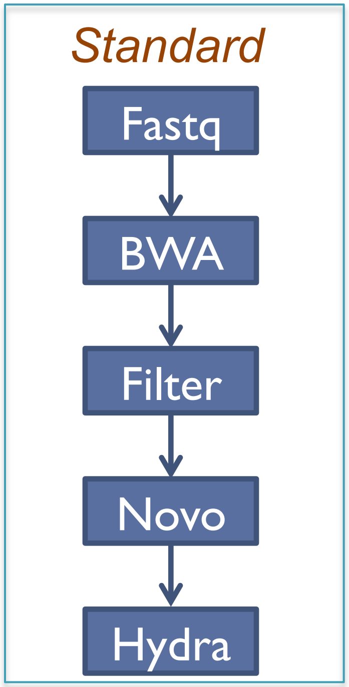


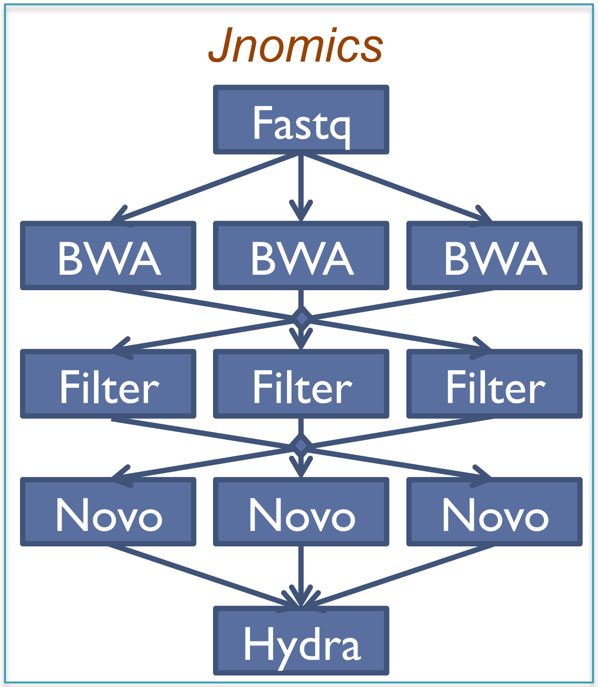
(Salipante *et al.* 2006)

mega- (x2)

Jnomics: Cloud-scale genomics

Matt Titmus, James Gurtowski, Michael Schatz





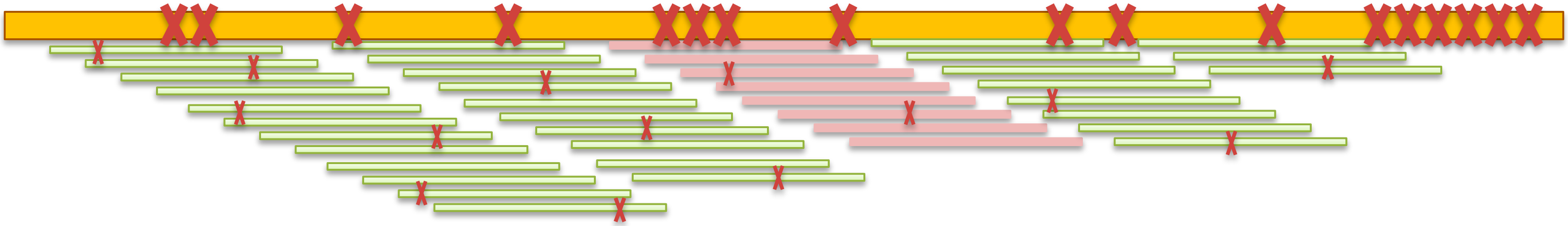
Poster 173

- Rapid parallel execution of NGS analysis pipelines
  - FASTX, BWA, Novoalign, SAMTools, Hydra
- Seamless read/write of common formats
  - BAM, SAM, BED, fastq, fasta
  - Sorting, merging, filtering, selection, etc

PacBio Error Correction & Assembly

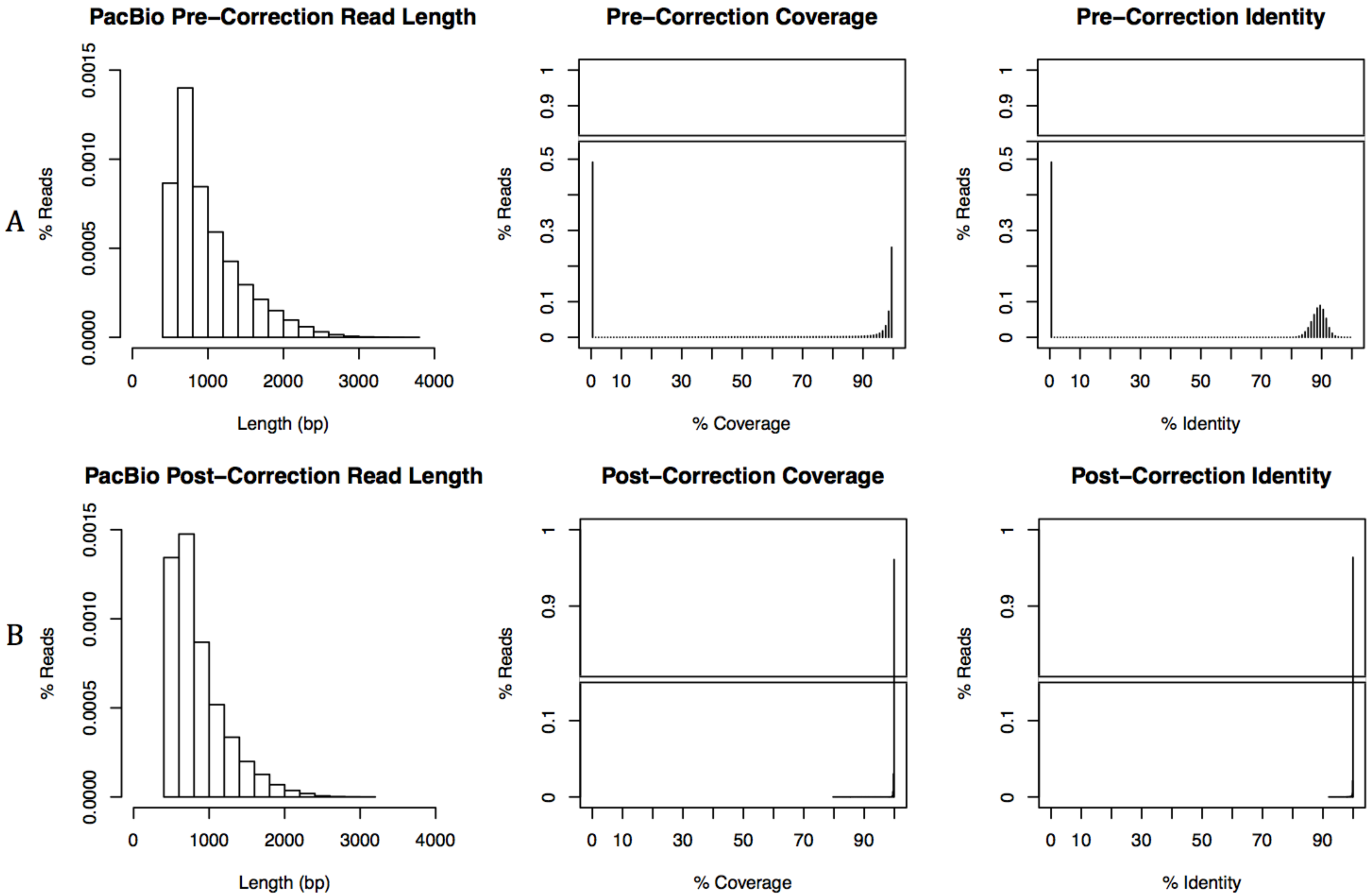
http://wgs-assembler.sf.net

1. Correction Pipeline
   1. Map short reads (SR) to long reads (LR)
   2. Trim LRs at coverage gaps
   3. Compute consensus for each LR
2. Co-assemble corrected LRs and SRs
   – Celera Assembler enhanced to support 32 Kbp reads
3. Assemblies substantially improve with longer reads



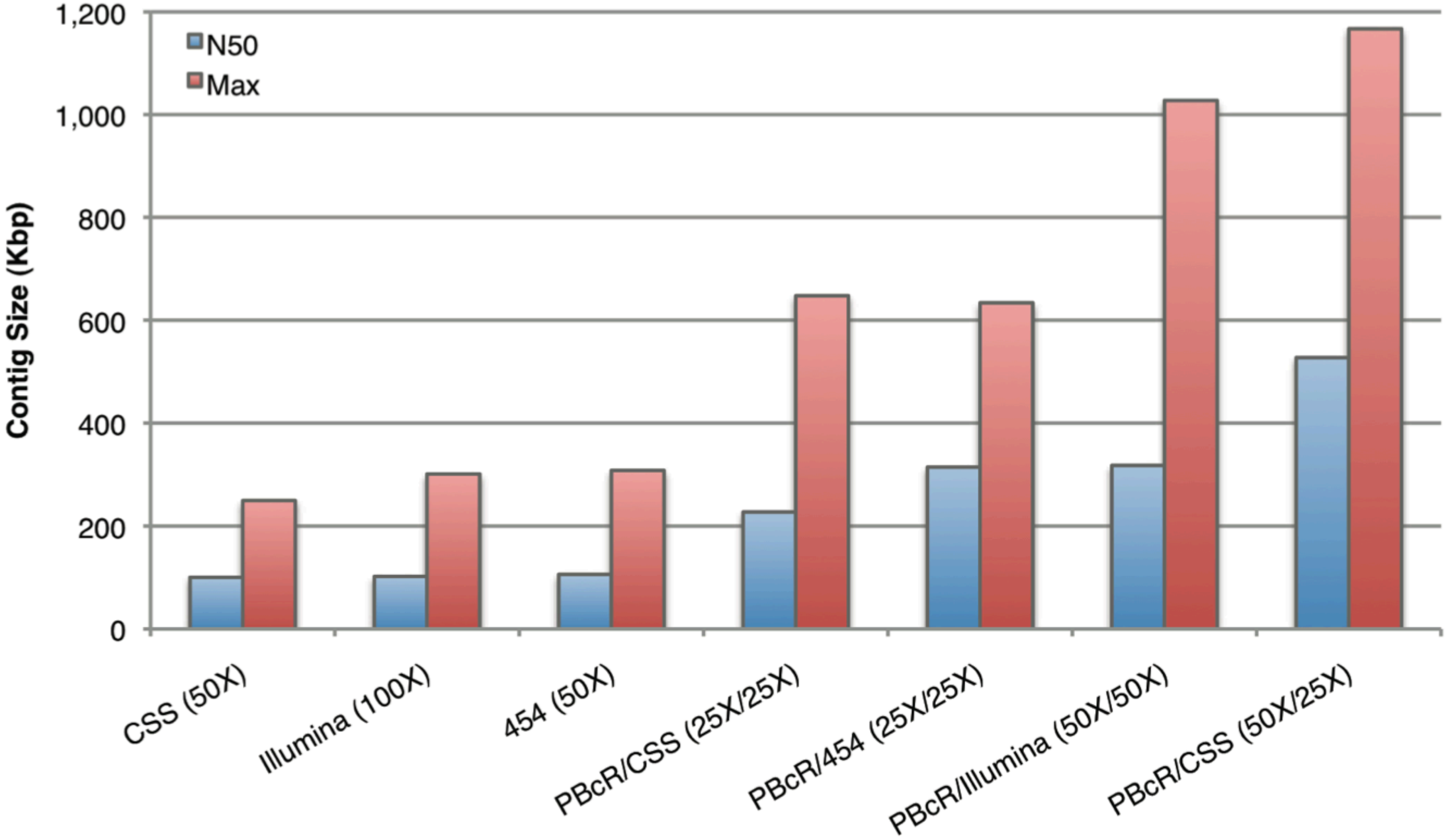
Hybrid error correction and de novo assembly of single-molecule sequencing reads.
Koren, S, Schatz, MC, Walenz, BP, Martin, J, Howard, J, Ganapathy, G, Wang, Z, Rasko, DA, McCombie, WR, Jarvis, ED, Phillippy, AM. (2011) *Under Review*

Error Correction Results



Correction results of 20x PacBio coverage of E. coli K12 corrected using 50x Illumina

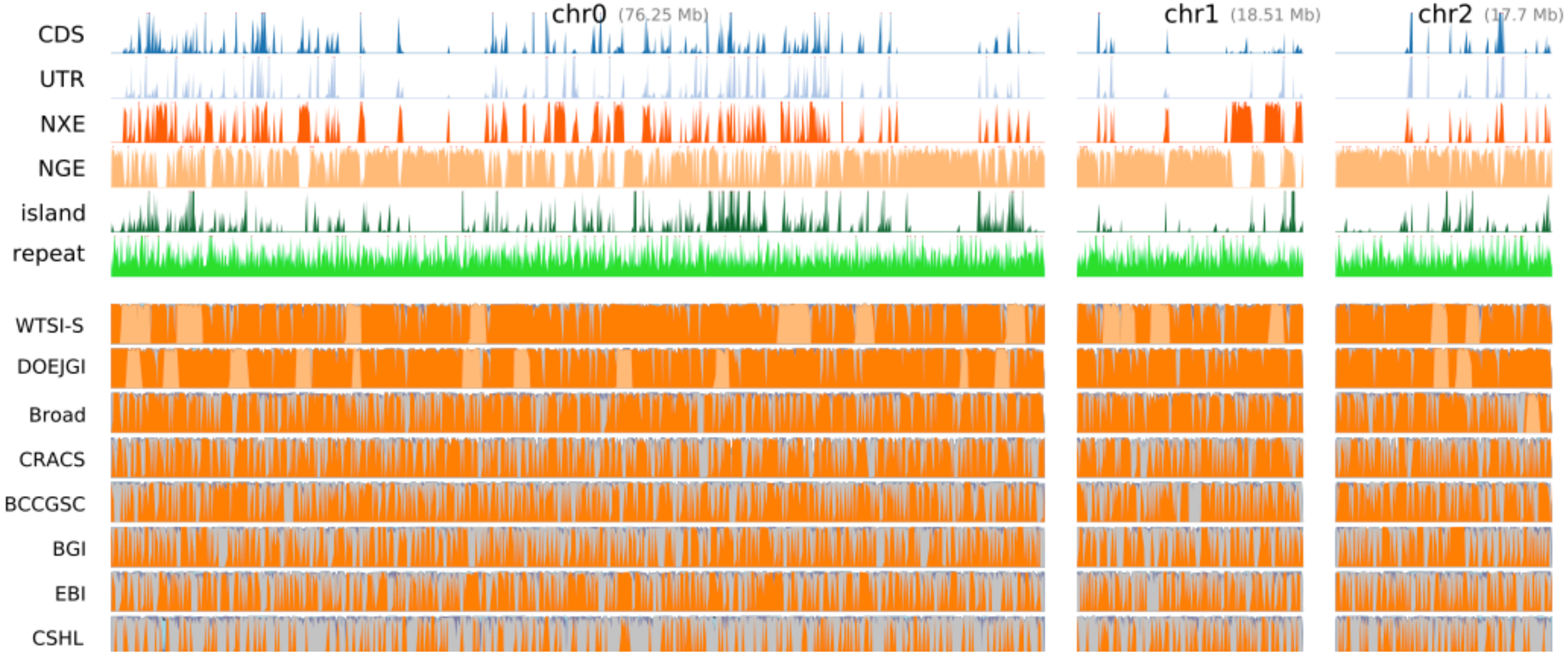
Assembly Results



SMRT-hybrid assembly results of 50x PacBio corrected coverage of E. coli K12
Long reads lead to ***contigs*** over 1Mbp

meta-

Assemblathon 1



- Assembly competition with a known reference genome enables base-by-base comparison to the truth
  - But evaluating an assembly in absence of a reference is difficult
  - Once we identify differences, what can we do about them?

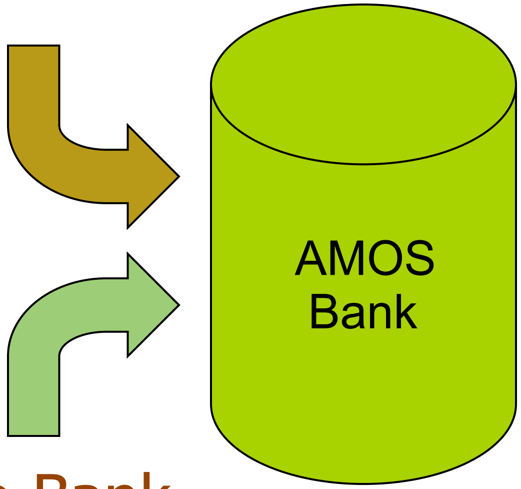
Forensics Pipeline

Computationally scan an assembly for mis-assemblies.

- Data inconsistencies are indicators for mis-assembly
- Some inconsistencies are merely statistical variations

AMOSvalidate

1. Load Assembly Data into Bank
2. Analyze Mate Pairs & Libraries
3. Analyze Depth of Coverage
4. Analyze Read Alignments
5. Analyze Read Breakpoints
6. Load Mis-assembly Signatures into Bank

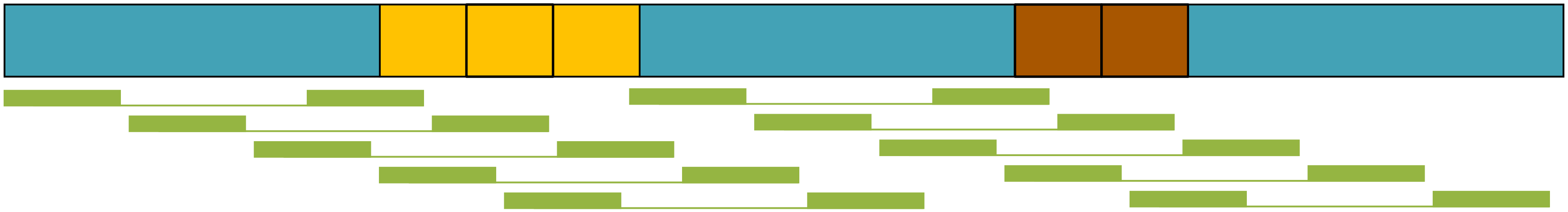


Genome Assembly forensics: finding the elusive mis-assembly.
Phillippy, AM, Schatz, MC, Pop, M. (2008) Genome Biology 9:R55.

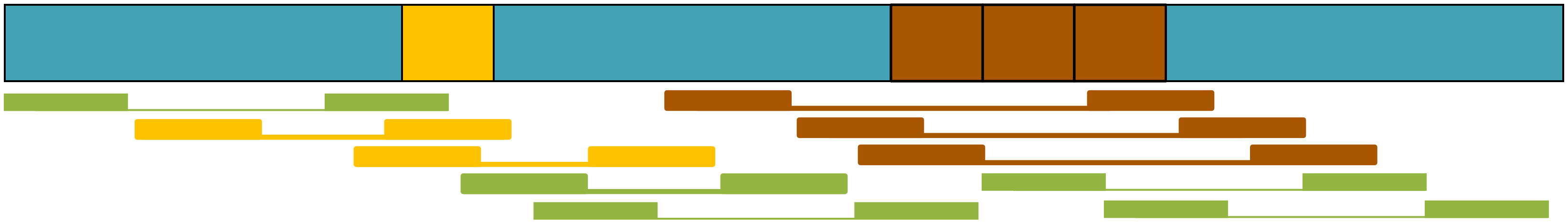
Forensics

Mate Evaluation

- Correct: mates have expected orientation and separation



- Mis-assembled: mates have incorrect orientation and separation



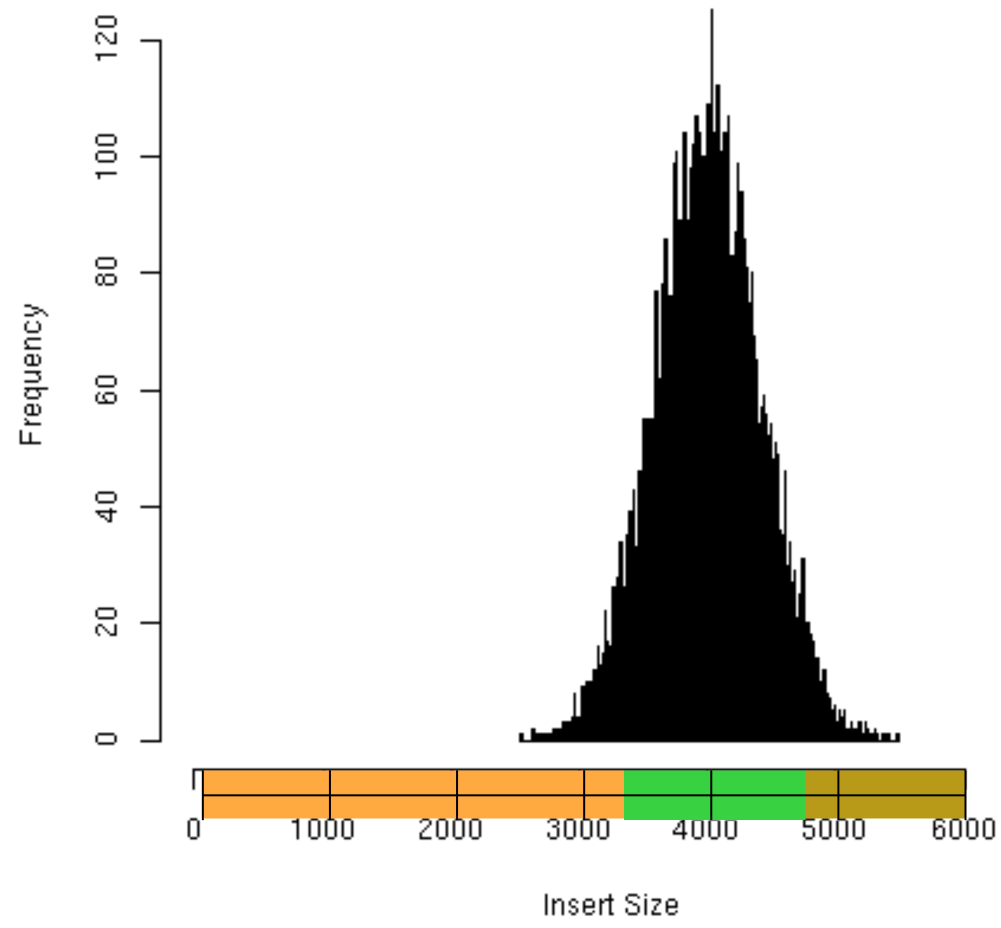
- Slightly compressed/expanded mates are expected because mates are sampled from a distribution of fragments

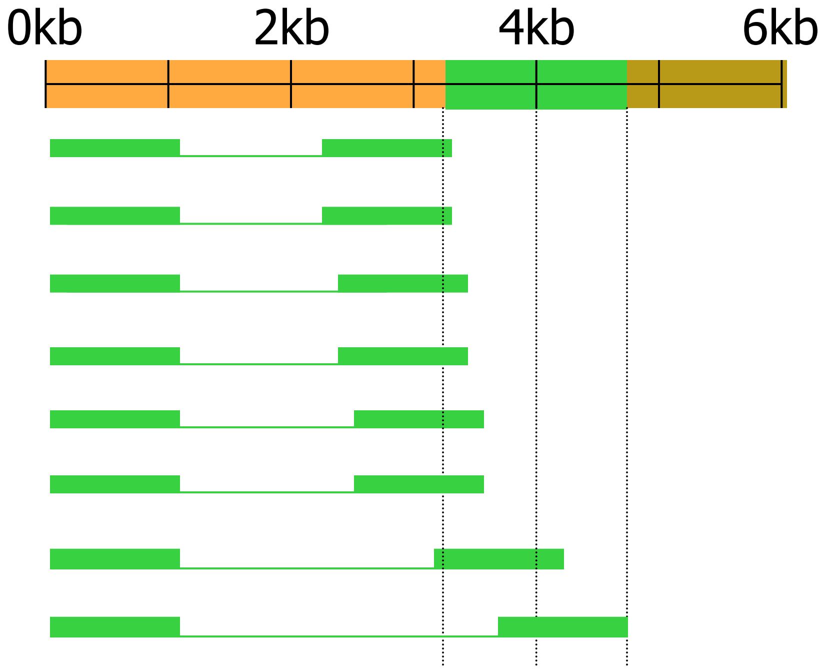
Forensics

Hidden Compression

Library size distribution

Mean: 4000, SD: 400





8 inserts: 3.2 kb-4.8kb

Local Mean: 3488

C/E Stat: $\frac{(3488-4000)}{(400\ /\ \sqrt{8})} = -3.62$

C/E Stat ≤ -3.0 indicates Compression

Assemblathon 2: Metassembly

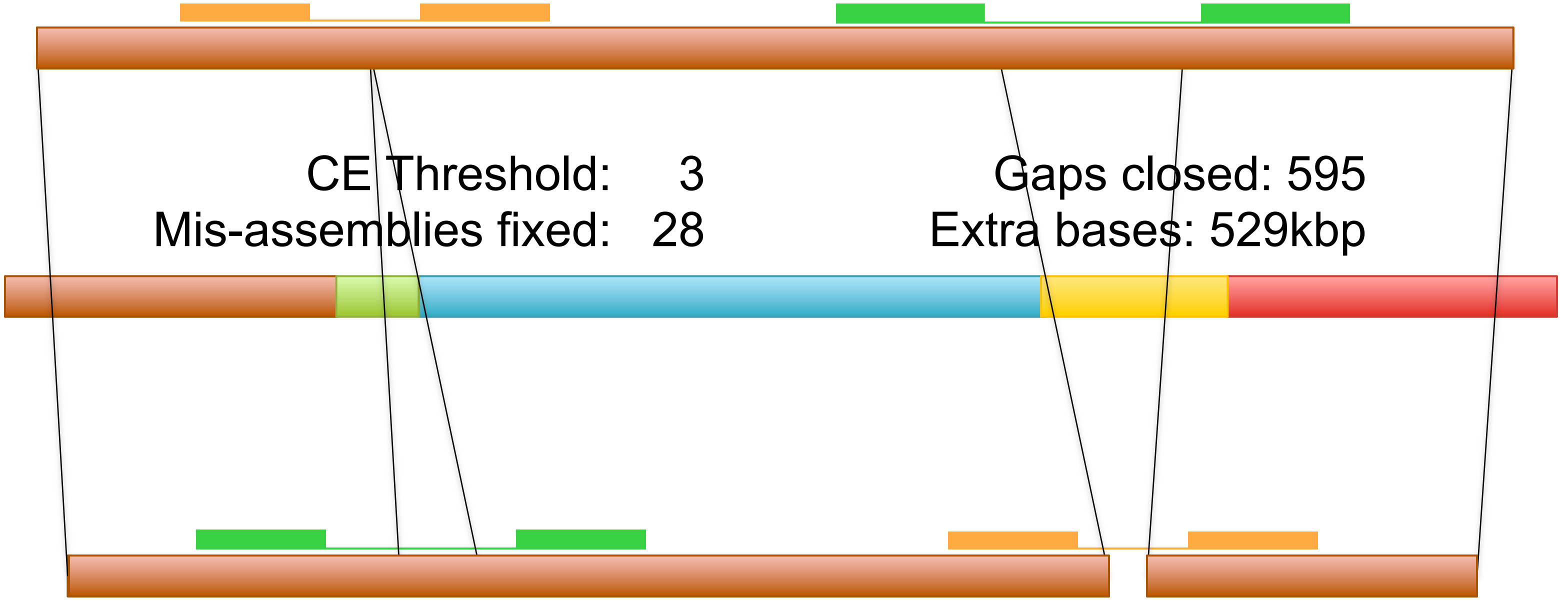
Paul Baranay, Scott Emrich, Michael Schatz

Poster 30

ALLPATHS-LG

Scaffold N50: 3,710,017
#>1000: 2,791

Contig N50: 20,183
#>1000: 68,591



CE Threshold: 3
Mis-assemblies fixed: 28

Gaps closed: 595
Extra bases: 529kbp

SOAPdenovo
+ FLASH
+ Quake
+ AMOS

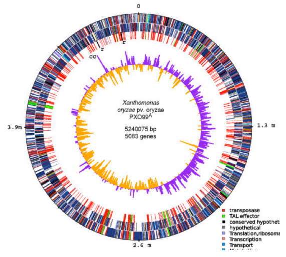
Scaffold N50: 285,413
#>1000: 29,119

Contig N50: 1,607
#>1000: 218,643

Inspired by Zimin *et al*. (2007) *Assembly Reconciliation*. Bioinformatics. 42(1) 42-45

Summary

- Assembly is moving to increasingly more complex and more diverse data types and organisms
  - PacBio error correction is my 3rd iteration of this problem
  - Assembly is useful in many different contexts, requires specialization and tuning
- There is a fundamental tension between connectivity and correctness
  - N50 is useful for evaluating connectivity but says nothing about correctness
  - CE can measure correctness at "gene-length" scale
- Metassembly is very promising for advancing assembly
  - Allows one to construct a consensus superior to the individual submissions
  - Enables one to select a locally optimal threshold



Acknowledgements

Schatzlab
Mitch Bekritsky
Matt Titmus
Hayan Lee
James Gurtowski
Giuseppe Narzisi
Rohith Menon
Goutham Bhat

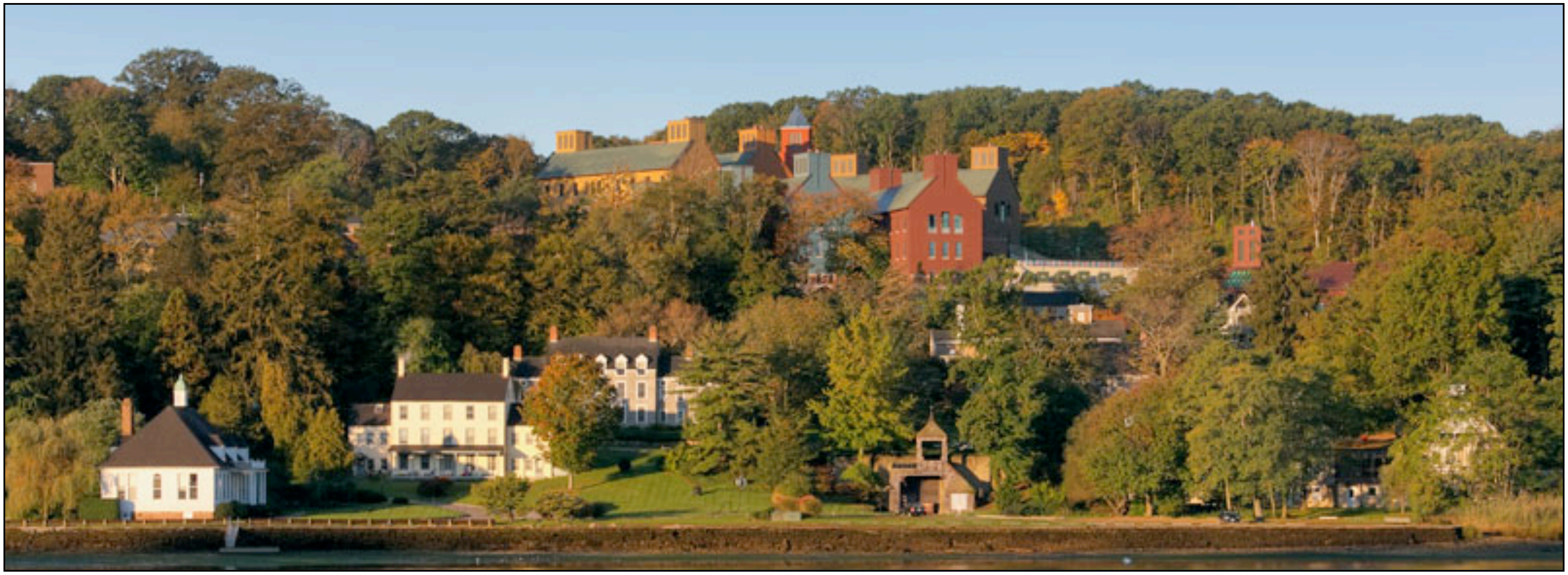
CSHL
Dick McCombie
Melissa Kramer
Eric Antonio
Mike Wigler
Zach Lippman
Doreen Ware
Ivan Iossifov

JHU
Steven Salzberg
Ben Langmead
Daniela Puiu

NBACC
Adam Phillipy
Sergey Koren

Univ. of Maryland
Mihai Pop
Art Delcher
David Kelley
Aleksey Zimin

ALLPATHS team
SOAPdenovo team

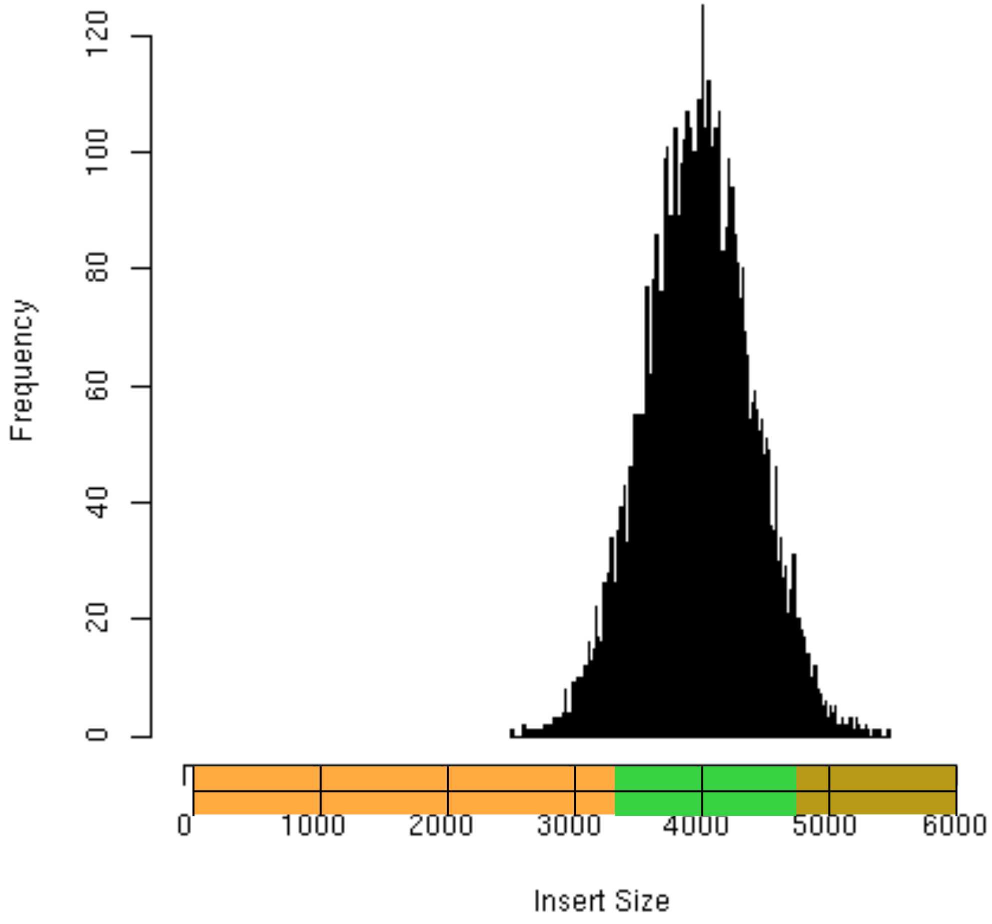


Thank You!

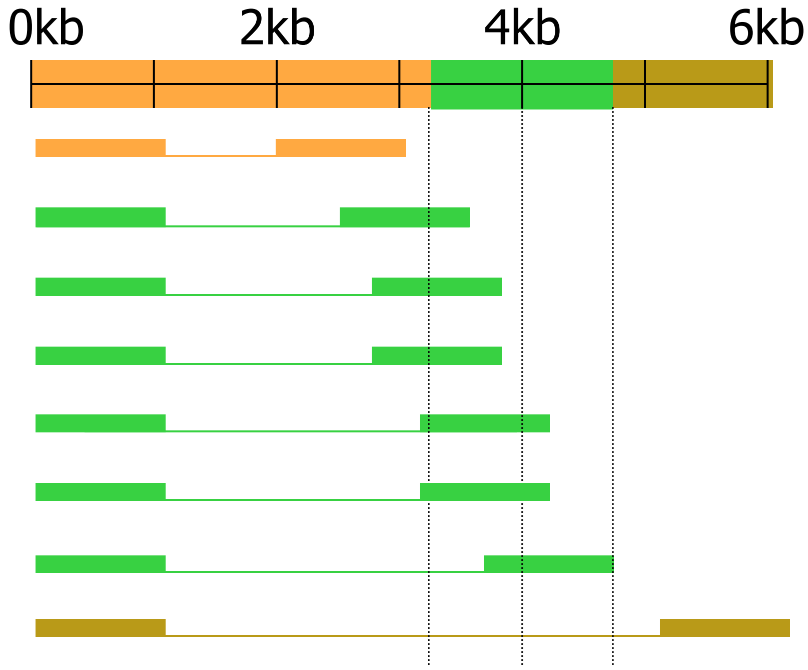
http://schatzlab.cshl.edu
@mike_schatz / #GI2011

Forensics

Compression/Expansion Statistic


Library size distribution

Mean: 4000, SD: 400



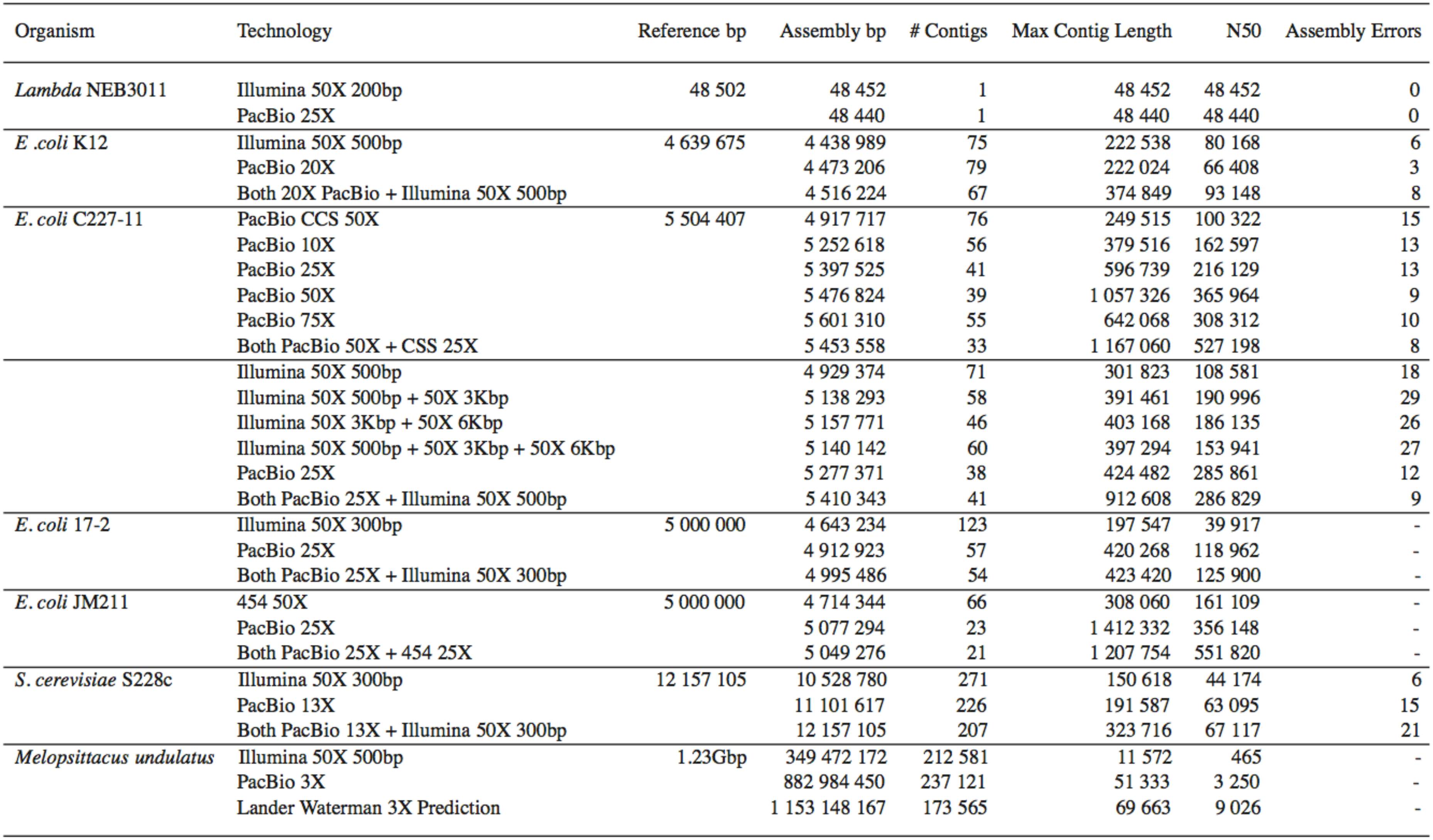
8 inserts: 3kb-6kb

Local Mean: 4048

C/E Stat: $\frac{(4048-4000)}{(400 / \sqrt{8})} = +0.33$

Near 0 indicates overall happiness

Hybrid Assembly Results



Organism	Technology	Reference bp	Assembly bp	# Contigs	Max Contig Length	N50	Assembly Errors
Lambda NEB3011	Illumina 50X 200bp	48 502	48 452	1	48 452	48 452	0
	PacBio 25X		48 440	1	48 440	48 440	0
E .coli K12	Illumina 50X 500bp	4 639 675	4 438 989	75	222 538	80 168	6
	PacBio 20X		4 473 206	79	222 024	66 408	3
	Both 20X PacBio + Illumina 50X 500bp		4 516 224	67	374 849	93 148	8
E. coli C227-11	PacBio CCS 50X	5 504 407	4 917 717	76	249 515	100 322	15
	PacBio 10X		5 252 618	56	379 516	162 597	13
	PacBio 25X		5 397 525	41	596 739	216 129	13
	PacBio 50X		5 476 824	39	1 057 326	365 964	9
	PacBio 75X		5 601 310	55	642 068	308 312	10
	Both PacBio 50X + CSS 25X		5 453 558	33	1 167 060	527 198	8
	Illumina 50X 500bp		4 929 374	71	301 823	108 581	18
	Illumina 50X 500bp + 50X 3Kbp		5 138 293	58	391 461	190 996	29
	Illumina 50X 3Kbp + 50X 6Kbp		5 157 771	46	403 168	186 135	26
	Illumina 50X 500bp + 50X 3Kbp + 50X 6Kbp		5 140 142	60	397 294	153 941	27
	PacBio 25X		5 277 371	38	424 482	285 861	12
	Both PacBio 25X + Illumina 50X 500bp		5 410 343	41	912 608	286 829	9
E. coli 17-2	Illumina 50X 300bp	5 000 000	4 643 234	123	197 547	39 917	-
	PacBio 25X		4 912 923	57	420 268	118 962	-
	Both PacBio 25X + Illumina 50X 300bp		4 995 486	54	423 420	125 900	-
E. coli JM211	454 50X	5 000 000	4 714 344	66	308 060	161 109	-
	PacBio 25X		5 077 294	23	1 412 332	356 148	-
	Both PacBio 25X + 454 25X		5 049 276	21	1 207 754	551 820	-
S. cerevisiae S228c	Illumina 50X 300bp	12 157 105	10 528 780	271	150 618	44 174	6
	PacBio 13X		11 101 617	226	191 587	63 095	15
	Both PacBio 13X + Illumina 50X 300bp		12 157 105	207	323 716	67 117	21
Melopsittacus undulatus	Illumina 50X 500bp	1.23Gbp	349 472 172	212 581	11 572	465	-
	PacBio 3X		882 984 450	237 121	51 333	3 250	-
	Lander Waterman 3X Prediction		1 153 148 167	173 565	69 663	9 026	-

Hybrid assembly results using error corrected PacBio reads
Meets or beats Illumina-only or 454-only assembly in every case