

SMRT-assembly

Error correction and de novo assembly of complex genomes using single molecule, real-time sequencing

Michael Schatz

May 10, 2012

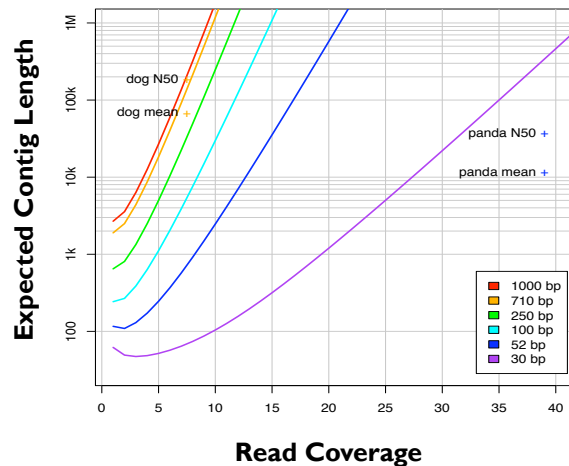
Biology of Genomes



@mike_schatz / #bog12

Ingredients for a good assembly

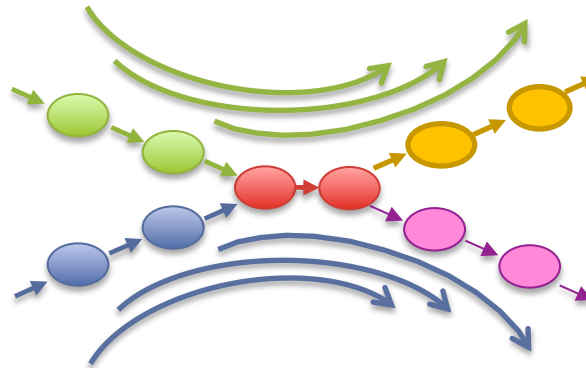
Coverage



High coverage is required

- Oversample the genome to ensure every base is sequenced with long overlaps between reads
- Biased coverage will also fragment assembly

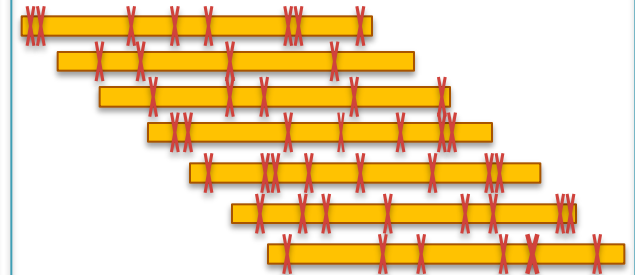
Read Length



Reads & mates must be longer than the repeats

- Short reads will have **false overlaps** forming hairball assembly graphs
- With long enough reads, assemble entire chromosomes into contigs

Quality



Errors obscure overlaps

- Reads are assembled by finding kmers shared in pair of reads
- High error rate requires very short seeds, increasing complexity and forming assembly hairballs

Current challenges in *de novo* plant genome sequencing and assembly

Schatz MC, Witkowski, McCombie, VWR (2012) *Genome Biology*. In Press.

Hybrid Sequencing



Illumina

Sequencing by Synthesis

High throughput (60Gbp/day)

High accuracy (~99%)

Short reads (~100bp)



Pacific Biosciences

SMRT Sequencing

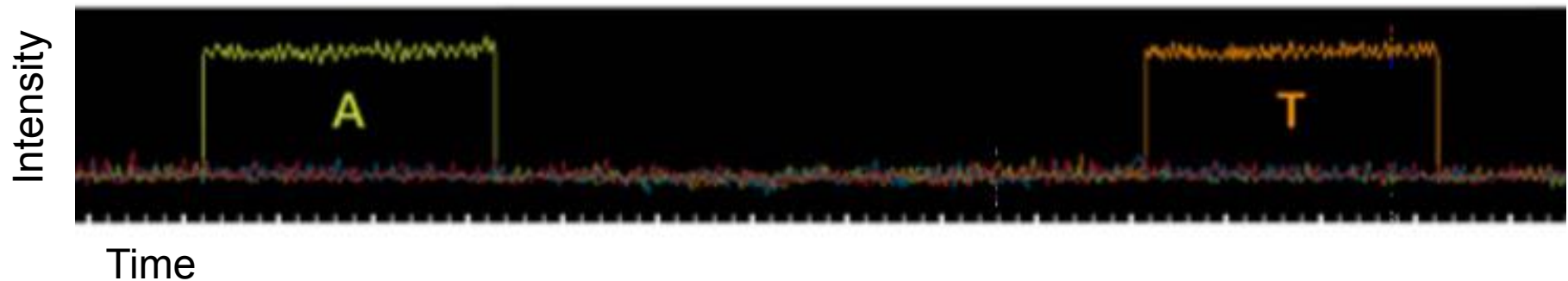
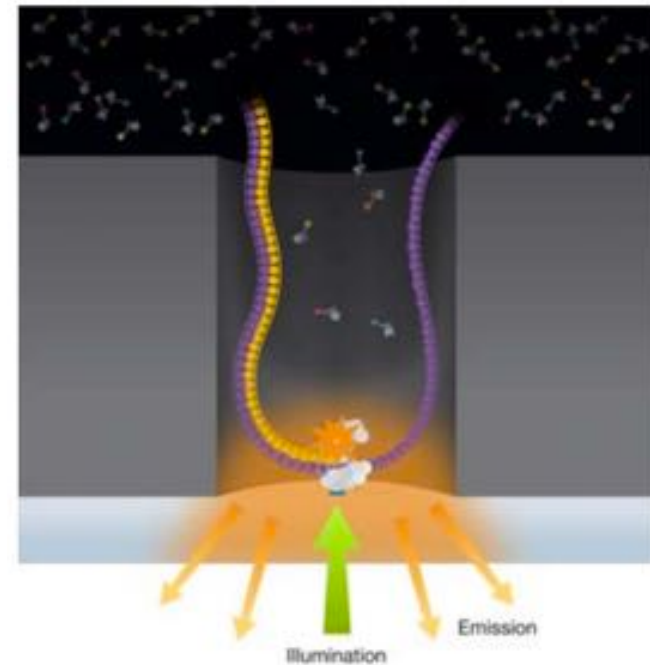
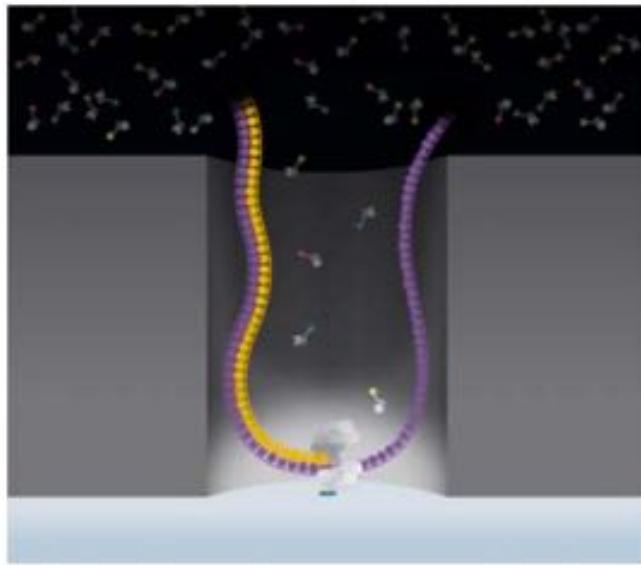
Lower throughput (600Mbp/day)

Lower accuracy (~85%)

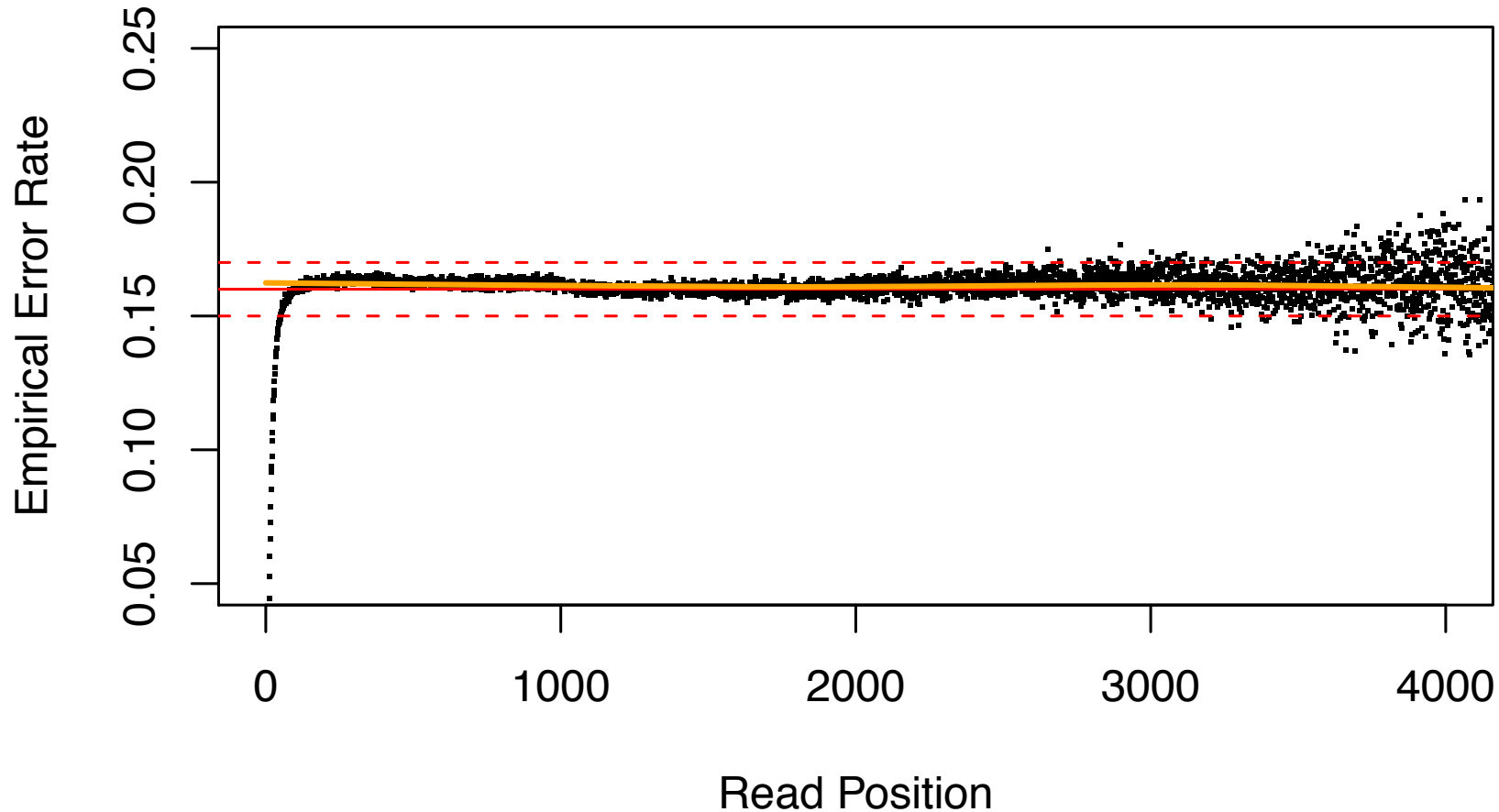
Long reads (10kbp+)

SMRT Sequencing

Imaging of fluorescent phospholinked labeled nucleotides as they are incorporated by a polymerase anchored to a Zero-Mode Waveguide (ZMW).



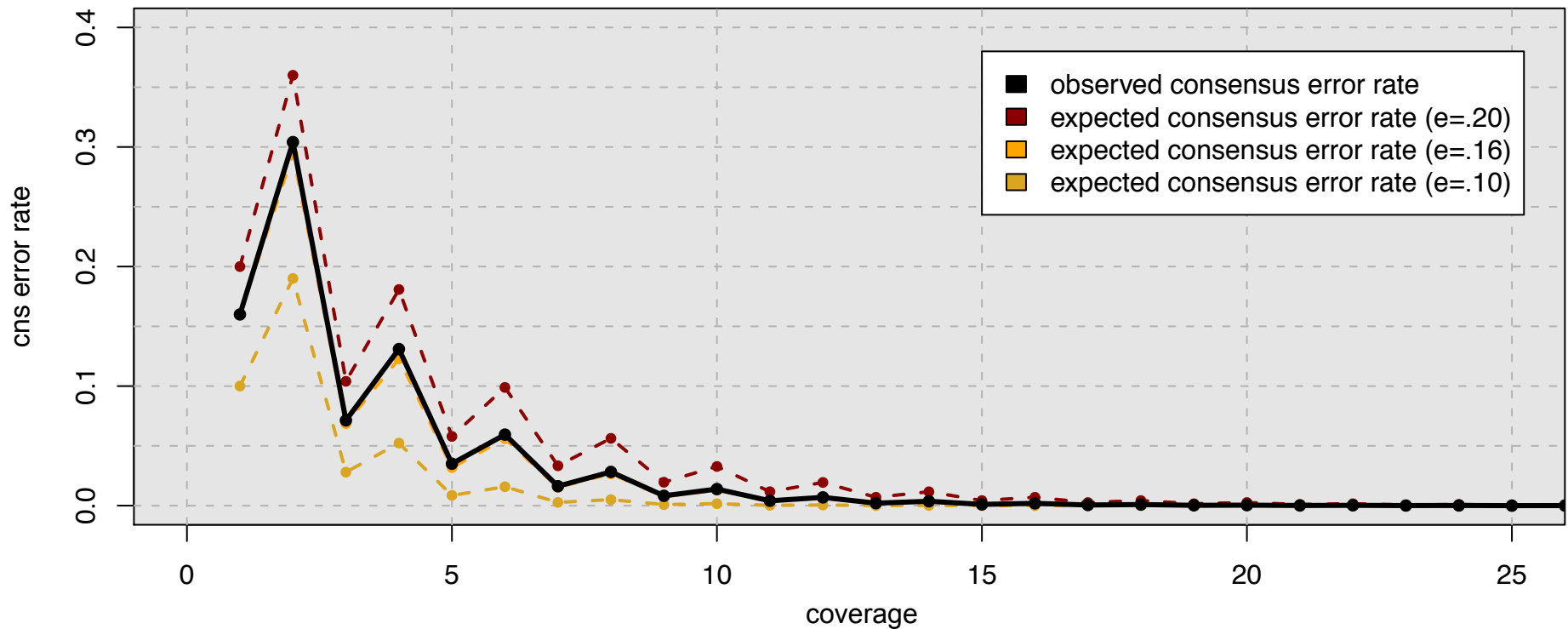
Read Quality



Consistent quality across the entire read

- Uniform error rate, no apparent biases for GC/motifs
- Sampling artifacts at beginning and ends of alignments

Consensus Accuracy and Coverage

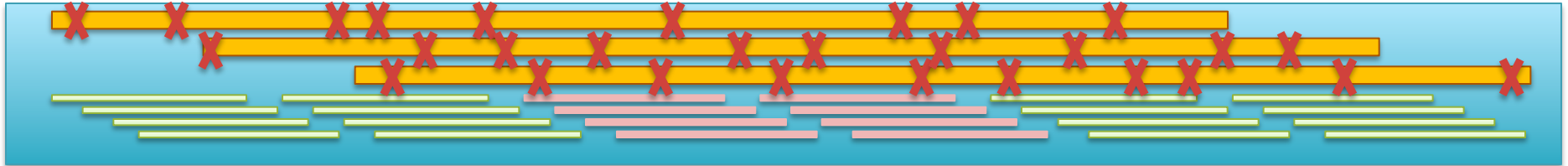


Coverage can overcome random errors

- Dashed: error model from binomial sampling; solid: observed accuracy
- For same reason, CCS is extremely accurate when using 5+ subreads

$$CNS\ Error = \sum_{i=\lceil c/2 \rceil}^c \binom{c}{i} (e)^i (1-e)^{n-i}$$

SMRT-hybrid Assembly



- Co-assemble long reads and short reads
 - Long reads (orange) natively span repeats (red)
 - Guards against mis-assemblies in draft assembly
 - Use all available data at once
- Challenges
 - Long reads have too high of an error rate to assemble directly
 - Assembler must support a wide mix of read lengths

PacBio Error Correction

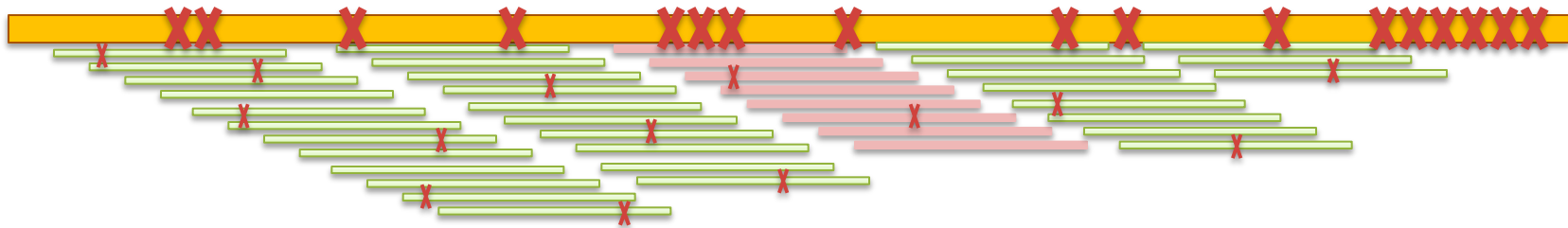
<http://wgs-assembler.sf.net>



I. Correction Pipeline

1. Map short reads (SR) to long reads (LR)
2. Trim LRs at coverage gaps
3. Compute consensus for each LR

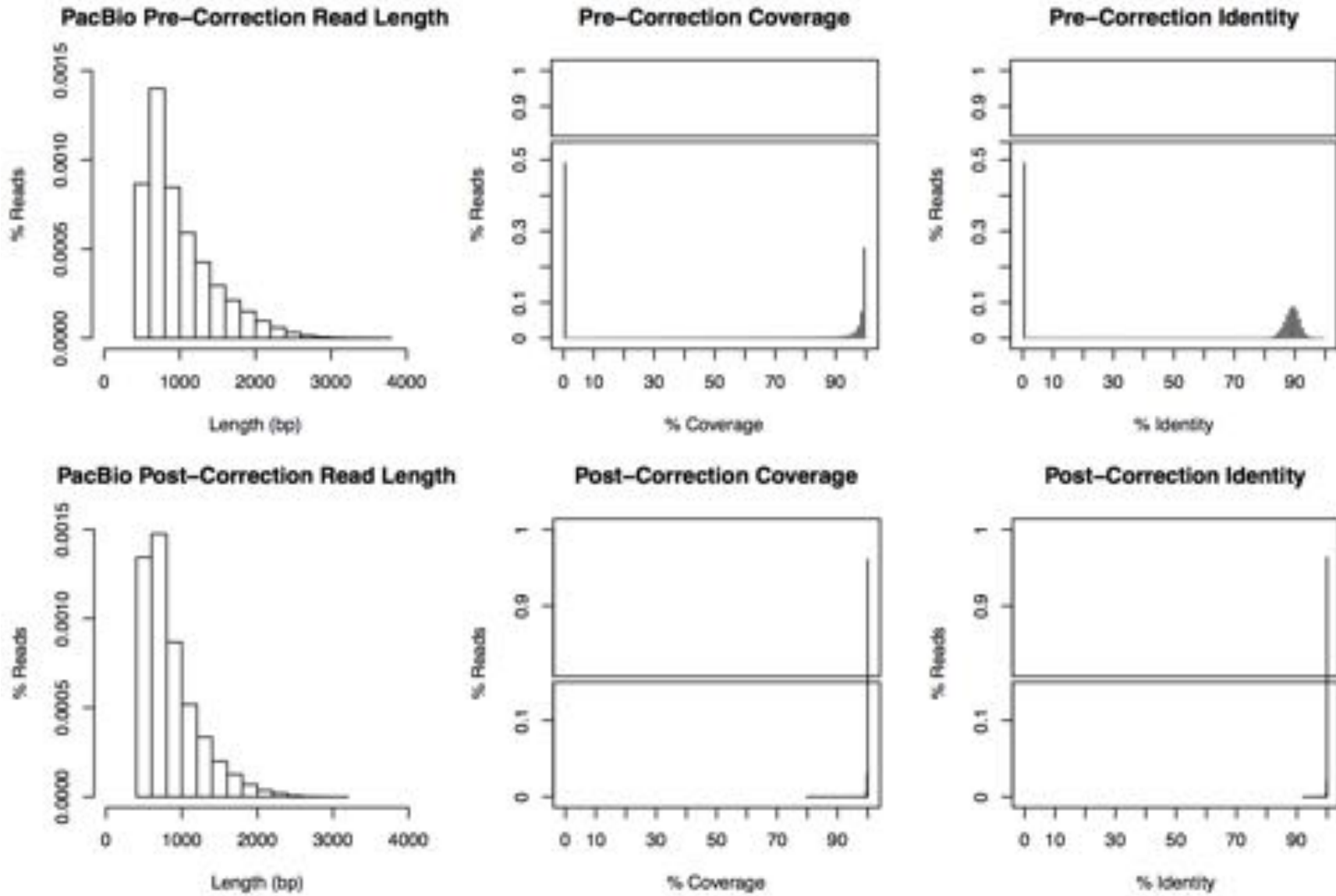
2. Error corrected reads can be easily assembled, aligned



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, VWR, Jarvis, ED, Phillippy, AM. (2012) *Under Review*

Error Correction Results

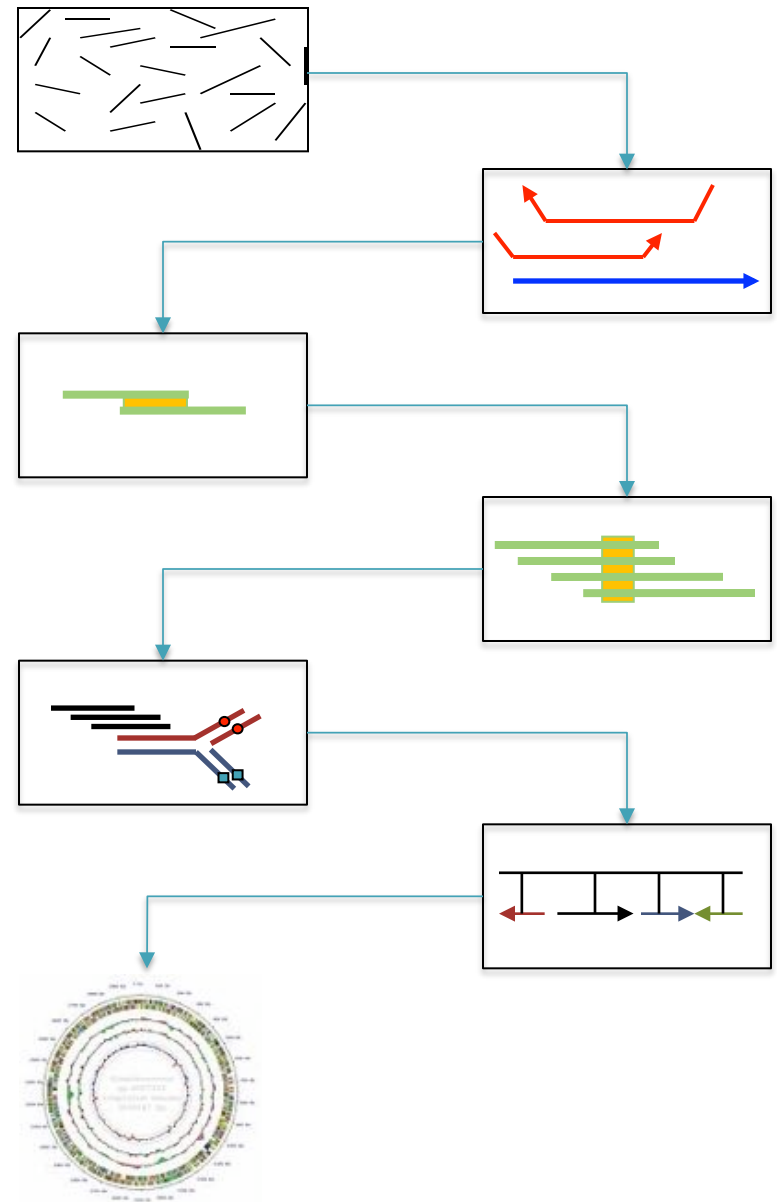


Correction results of 20x PacBio coverage of E. coli K12 corrected using 50x Illumina

Celera Assembler

<http://wgs-assembler.sf.net>

1. Pre-overlap
 - Consistency checks
2. Trimming
 - Quality trimming & partial overlaps
3. Compute Overlaps
 - Find high quality overlaps
4. Error Correction
 - Evaluate difference in context of overlapping reads
5. Unitigging
 - Merge consistent reads
6. Scaffolding
 - Bundle mates, Order & Orient
7. Finalize Data
 - Build final consensus sequences



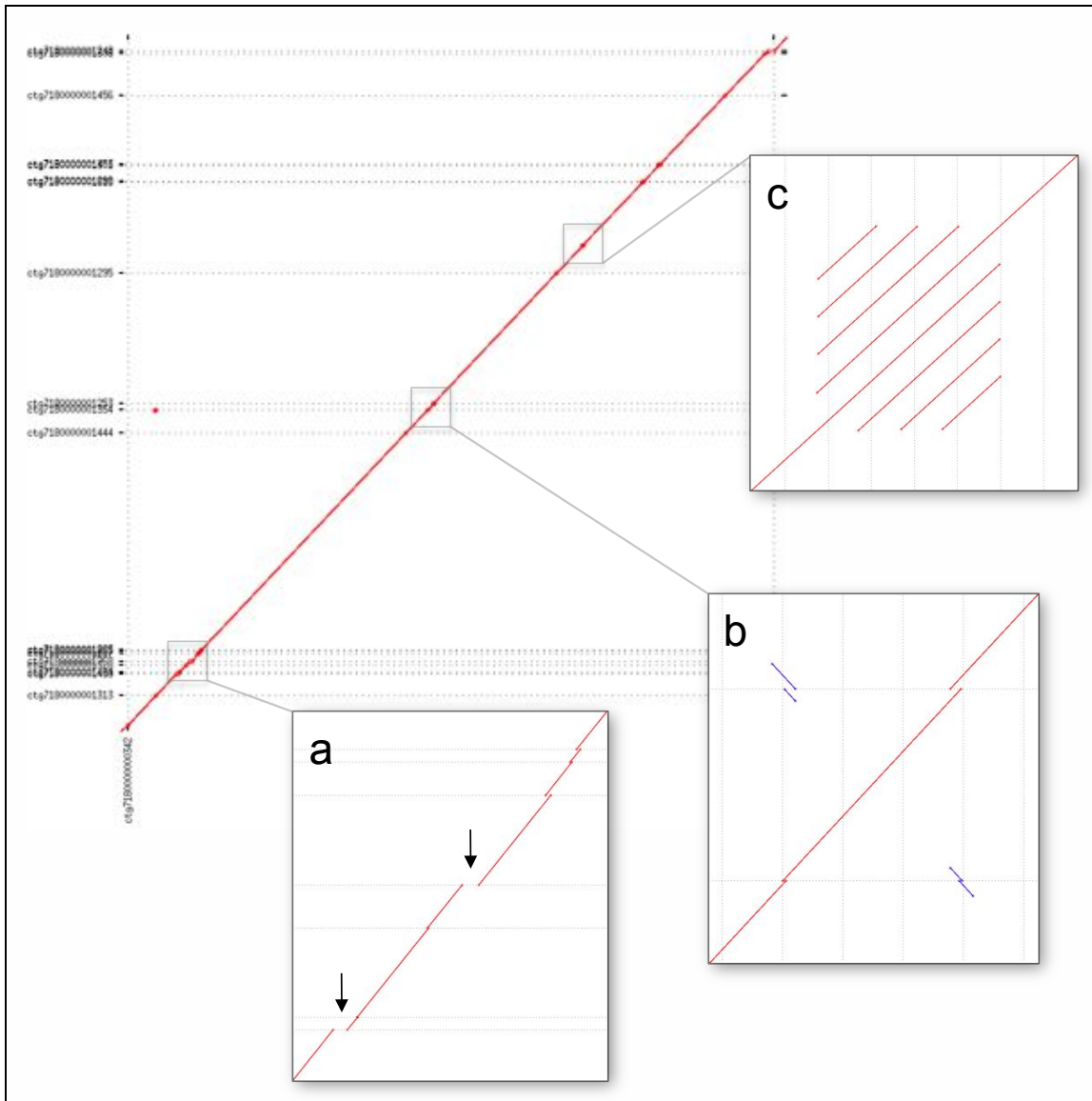
SMRT-Assembly Results



Organism	Technology	Reference bp	Assembly bp	# Contigs	Max Contig Length	N50
<i>Lambda</i> NEB3011 (median: 727 max: 3 280)	Illumina 100X 200bp	48 502	48 492	1	48 492 / 48 492	48 492 / 48 492 (100%) *
	PacBio PBcR 25X		48 440	1	48 444 / 48 444	48 444 / 48 440 (100%) *
<i>E. coli</i> K12 (median: 747 max: 3 068)	Illumina 100X 500bp	4 639 675	4 462 836	61	221 615 / 221 553	100 338 / 83 037 (82.36%) *
	PacBio PBcR 18X		4 465 533	77	239 058 / 238 224	71 479 / 68 309 (95.57%) *
	Both 18X PacBio PBcR + Illumina 50X 500bp		4 576 046	65	238 272 / 238 224	93 048 / 89 431 (96.11%) *
<i>E. coli</i> C227-11 (median: 1 217 max: 14 901)	PacBio CCS 50X	5 504 407	4 917 717	76	249 515	100 322
	PacBio 25X PBcR (corrected by 25X CCS)		5 207 946	80	357 234	98 774
	Both PacBio PBcR 25X + CCS 25X		5 269 158	39	647 362	227 302
	PacBio 50X PBcR (corrected by 50X CCS)		5 445 466	35	1 076 027	376 443
	Both PacBio PBcR 50X + CCS 25X		5 453 458	33	1 167 060	527 198
	Manually Corrected ALLORA Assembly ⁸		5 452 251	23	653 382	402 041
<i>S. cerevisiae</i> S228c (median: 674 max: 5 994)	Illumina 100X 300bp	12 157 105	11 034 156	192	266 528 / 227 714	73 871 / 49 254 (66.68%) *
	PacBio PBcR 13X		11 110 420	224	224 478 / 217 704	62 898 / 54 633 (86.86%) *
	Both PacBio PBcR 13X + Illumina 50X 300bp		11 286 932	177	262 846 / 260 794	82 543 / 59 792 (72.44%) *
<i>Melospiza arundinaria</i> (median 997, max 13 079)	Illumina 194X (220/500/800 paired-end 2/5/10Kb mate-pairs)	1.23 Gbp	1 023 532 850	24 181	1 050 202	47 383
	454 15.4X (FLX + FLX Plus + 3/8/20Kbp paired-ends)		999 168 029	16 574	751 729	75 178
	454 15.4X + PacBio PBcR 3.75X		1 071 356 415	15 081	1 238 843	99 573

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

PacBio Long Read Advantages

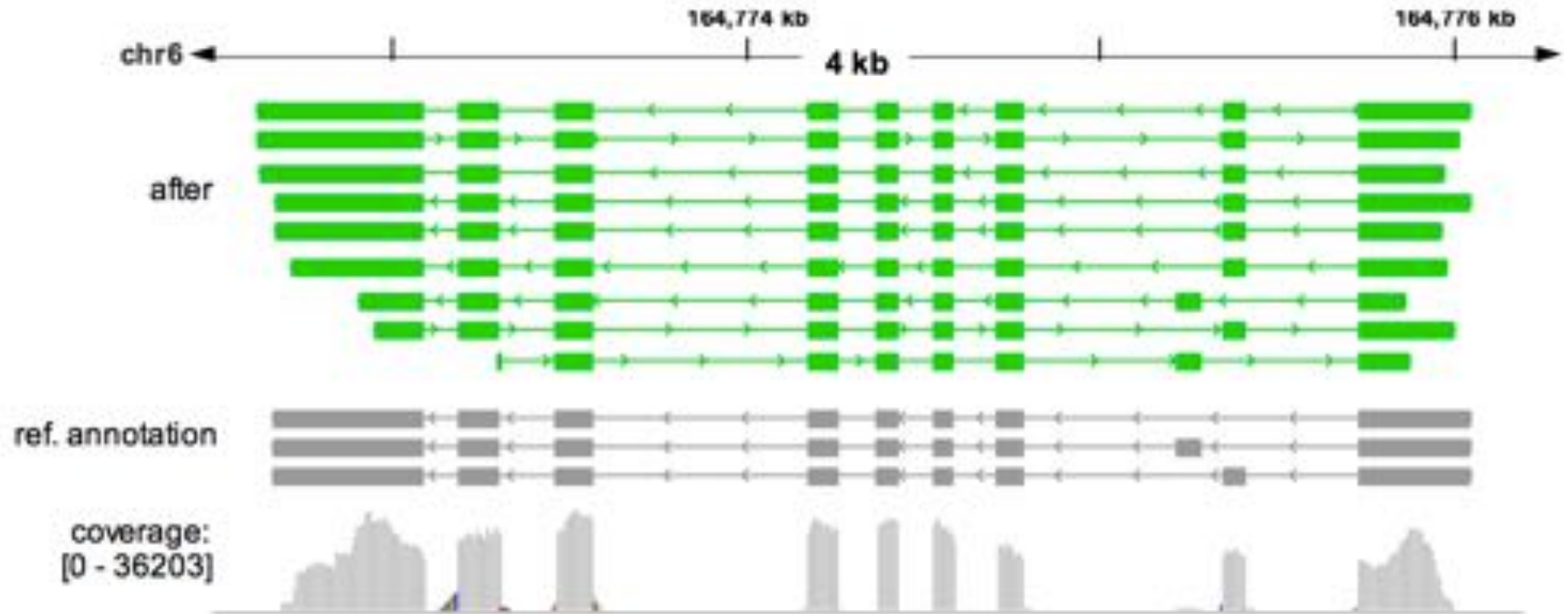


(a) Long reads close sequencing gaps

(b) Long reads assemble across long repeats

(c) Long reads span complex tandem repeats

Transcript Alignment



- Long-read single-molecule sequencing has potential to directly sequence full length transcripts
 - 11% mapped by BLAT pre-correction, 99% after correction
 - Directly observe alternative splicing events
- New collaboration with Gingeras Lab looking at splicing in human

Single Molecule Sequencing Summary

PacBio RS has capabilities not found in any other technology

- Substantially longer reads -> span repeats
- Unbiased sequence coverage -> close sequencing gaps
- Single molecule sequencing -> haplotype phasing, alternative splicing

Long reads enables highest quality de novo assembly

- Longer reads have more information than shorter reads
- Because the errors are random we can compensate for them
- One chromosome, one contig achieved in microbes

Exciting developments on the horizon

- Longer reads, higher throughput PacBio
- Nanopore Sequencing



Acknowledgements

Schatzlab

Giuseppe Narzisi
James Gurtowski
Mitch Bekritsky
Hayan Lee
Robert Aboukhalil
Charles Underwood

CSHL

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

NBACC

Adam Phillippy
Sergey Koren

JHU

Steven Salzberg
Ben Langmead
Jeff Leek

Univ. of Maryland

Mihai Pop
Art Delcher
Jimmy Lin
David Kelley
Dan Sommer
Cole Trapnell



Thank You

Poster 209: Giuseppe Narzisi *et al.*
Detection and validation of *de novo* mutations in exome-capture data using micro-assembly

**Want to push the limits
of biotechnology and bioinformatics?**

<http://schatzlab.cshl.edu/apply/>