SMRT-assembly

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

Michael Schatz

Jan 17, 2012
PAG-XX: PacBio Workshop

@mike_schatz / #PAGXX

Outline

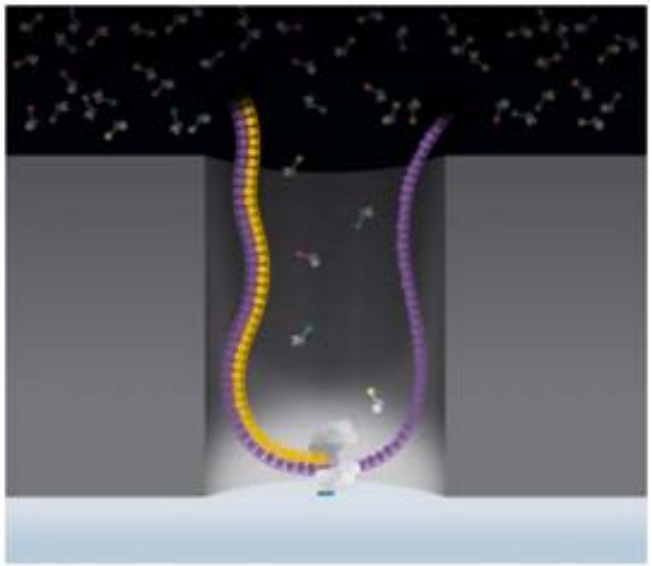
1. SMRT-sequencing
    1. Coverage, read length, and accuracy

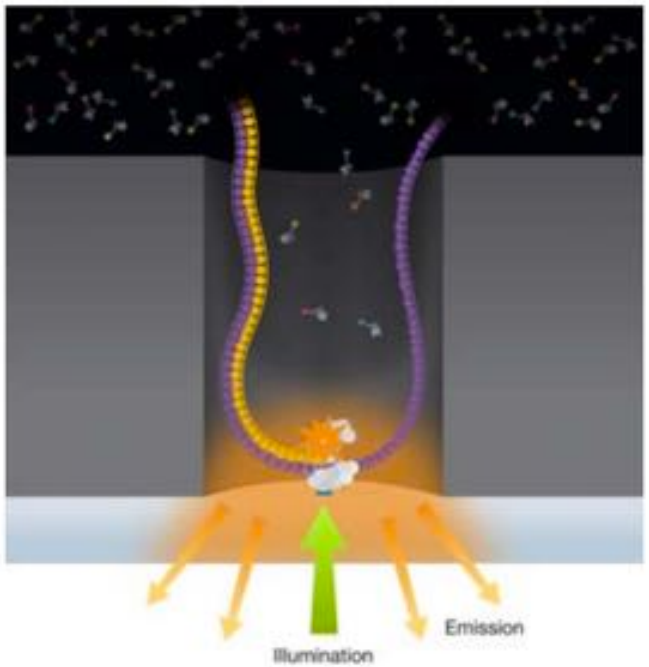
2. SMRT-assembly approaches
    1. SMRT-de novo: SMRT-only assembly
    2. SMRT-scaffolding: Long reads as links
    3. SMRT-hybrid: Short and long together

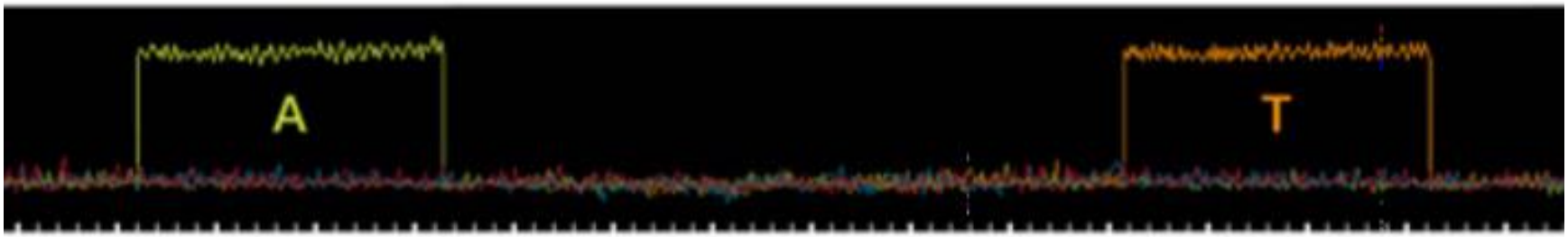
3. Review and best practices

SMRT Sequencing

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





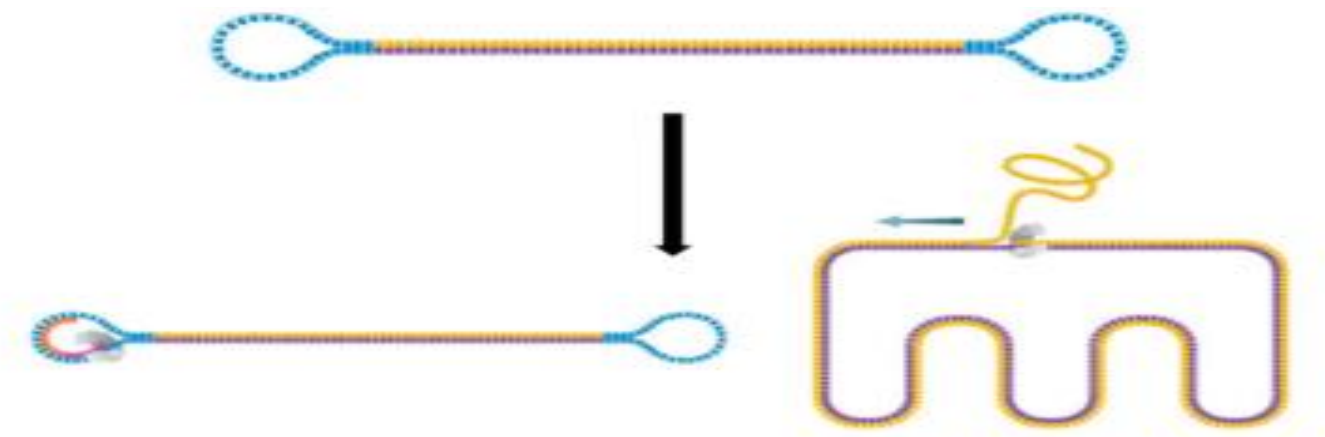


Intensity

Time

http://www.pacificbiosciences.com/assets/files/pacbio_technology_backgrounder.pdf

SMRT Read Types



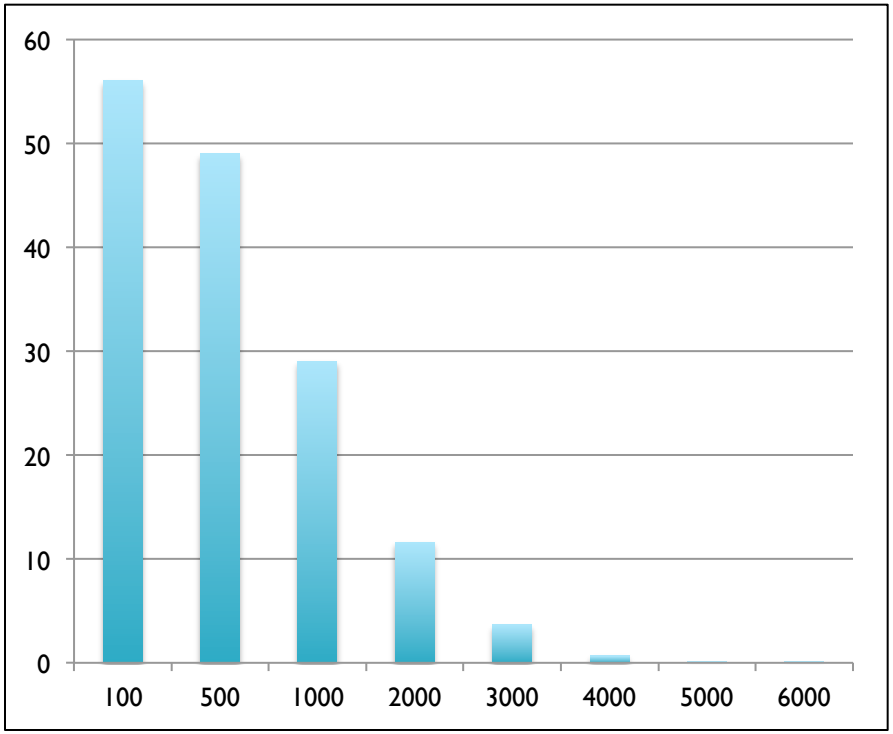
- ***Standard sequencing***
  - Long inserts so that the polymerase can synthesize along a single strand
- ***Circular consensus sequencing***
  - Short inserts, so polymerase can continue around the entire SMRTbell multiple times and generate multiple sub-reads from the same single molecule.
- ***Strobe sequencing***
  - Very long inserts, alternate the lasers in the instrument between on and off. On periods generate strobe sub-reads and the off periods determine the length of the spacing between, known as strobe advance

http://www.pacificbiosciences.com/assets/files/pacbio_technology_backgrounder.pdf

SMRT Sequencing Runs

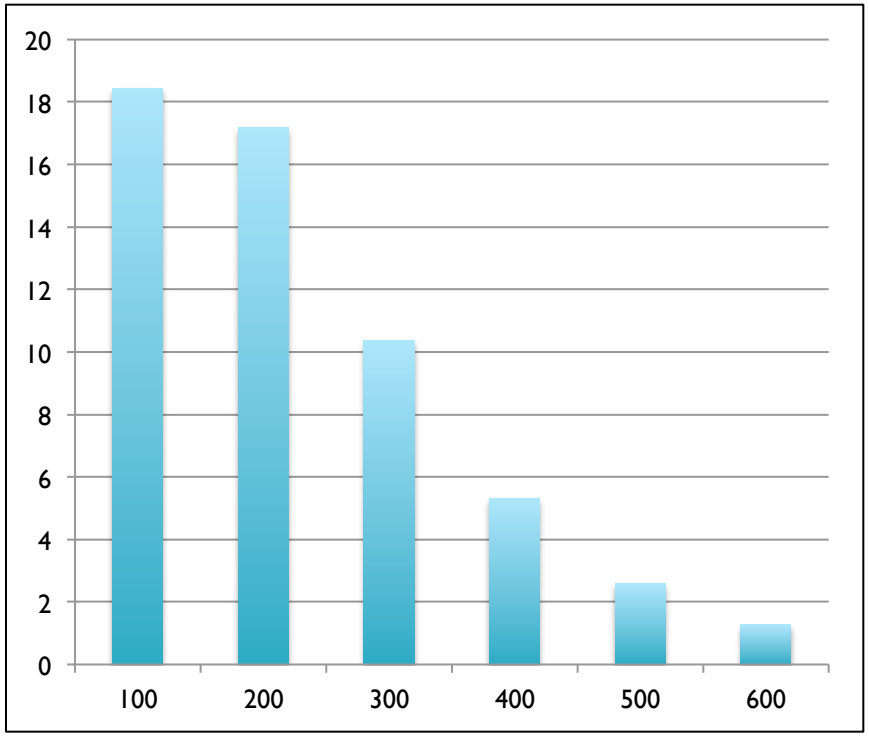
Yeast – Long reads

969,445 reads after filtering
Mean: 710 +/- 663
Median: 558 Max: 8,495

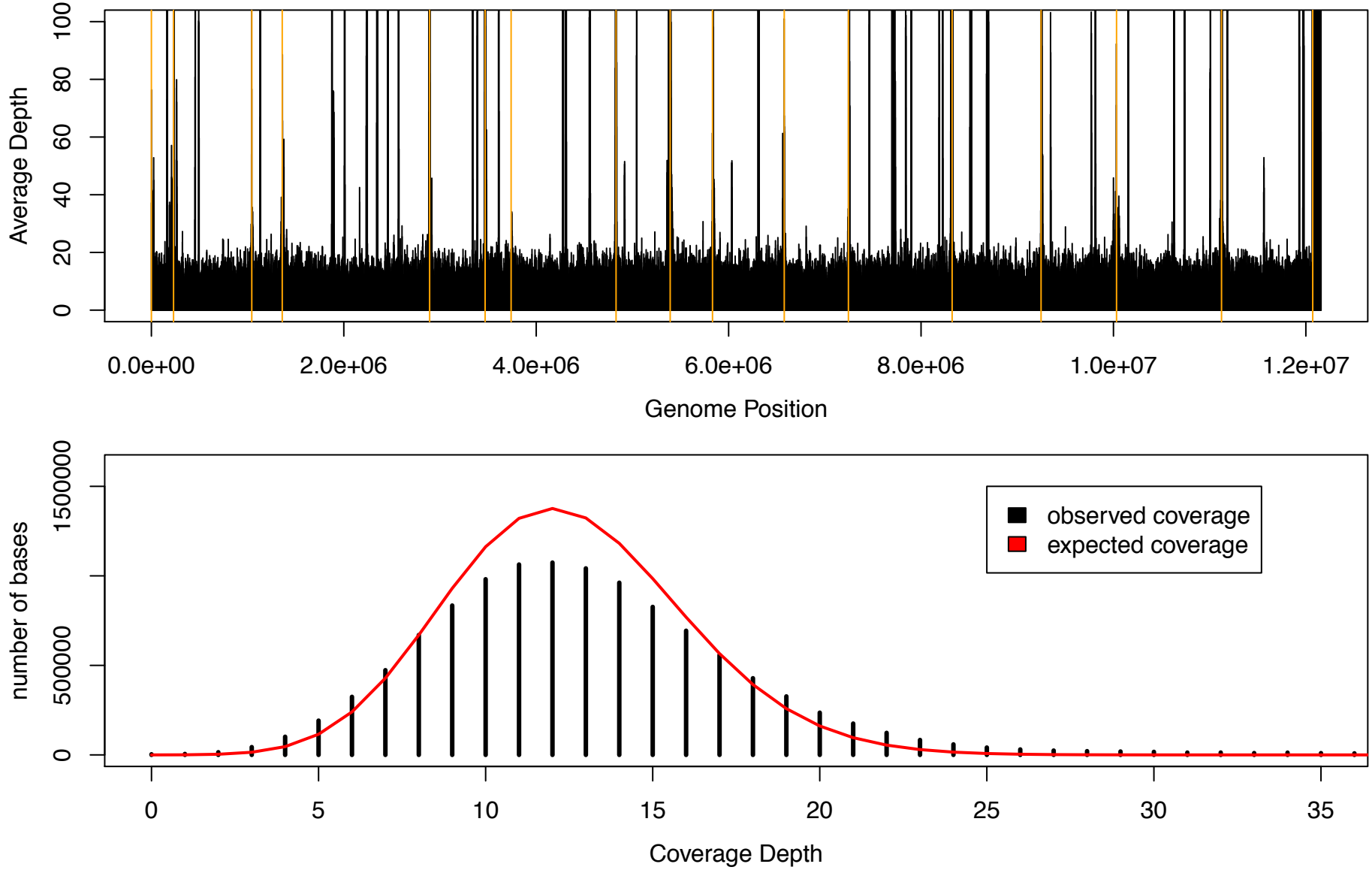


Yeast – CCS reads

731,638 reads after filtering
Mean: 306 +/- 115
Median: 279 Max: 1,425



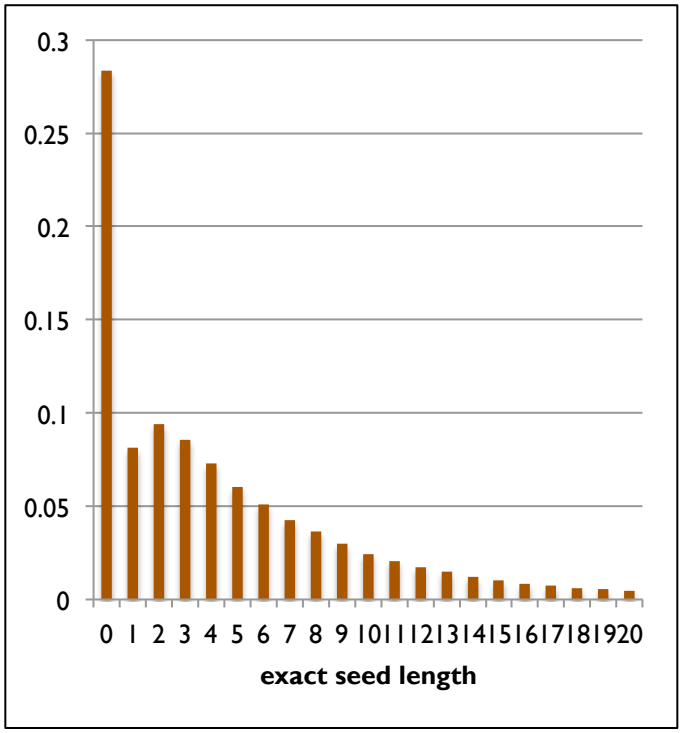
Genome Coverage



Coverage plots of long reads along yeast genome computed by BLASR

Alignment Quality

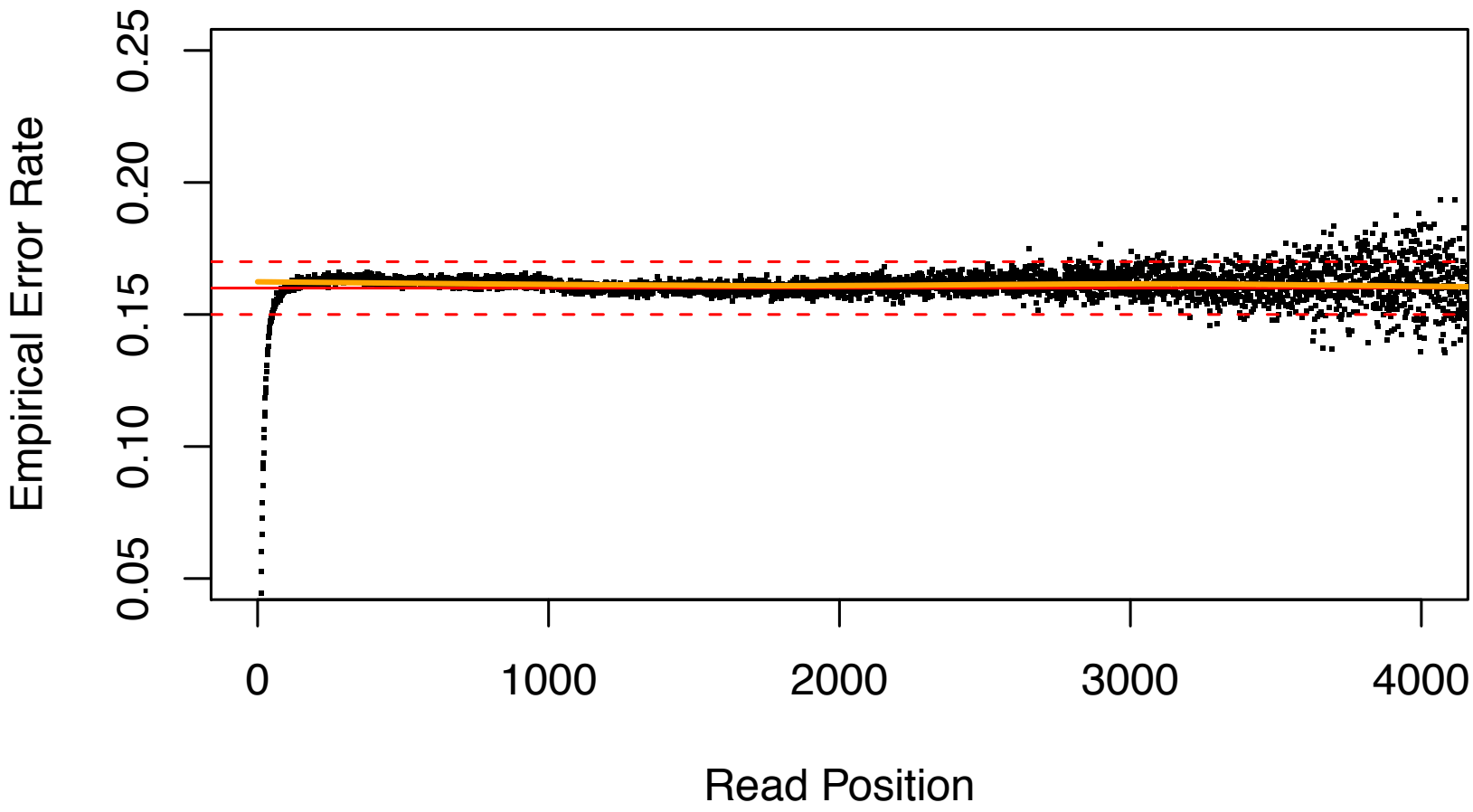
Match	83.7%
Mismatch	1.4%
Insertions	11.5%
Deletions	3.4%




```
     4  TTGTAAGCAGTTGAAAACTATGTGTGGATTTAGAATAAAGAACATGAAAG
        ||||||||||||||||||||||||| ||||||| ||||||||||||| |||
539752  TTGTAAGCAGTTGAAAACTATGTGT-GATTTAG-ATAAAGAACATGGAAG

    54  ATTATAAA-CAGTTGATCCATT-AGAAGA-AAACGCAAAAGGCGGCTAGG
        | |||||| ||||||||||||| |||| | |||||| |||||| ||||||
539800  A-TATAAATCAGTTGATCCATTAAGAA-AGAAACGC-AAAGGC-GCTAGG

   101  CAACCTTGAATGTAATCGCACTTGAAGAACAAGATTTTATTCCGCGCCCG
        | |||||| |||| ||  |||||||||||||||||||||||||||||||
539846  C-ACCTTG-ATGT-AT--CACTTGAAGAACAAGATTTTATTCCGCGCCCG

   151  TAACGAATCAAGATTCTGAAAACACAT-ATAACAACCTCCAAAA-CACAA
        | ||||||| |||||||||||||| || ||    |||||||||| |||||
539891  T-ACGAATC-AGATTCTGAAAACA-ATGAT----ACCTCCAAAAGCACAA

   199  -AGGAGGGGAAAGGGGGGAATATCT-ATAAAAGATTACAAATTAGA-TGA
         ||||||   ||     |||||||| || |||||||||||||| || |||
539934  GAGGAGG---AA-----GAATATCTGAT-AAAGATTACAAATT-GAGTGA

   246  ACT-AATTCACAATA-AATAACACTTTTA-ACAGAATTGAT-GGAA-GTT
        ||| ||||||||| | |||||||||||||| ||| ||||||| |||| |||
539974  ACTAAATTCACAA-ATAATAACACTTTTAGACAAAATTGATGGGAAGGTT

   291  TCGGAGAGATCCAAAACAATGGGC-ATCGCCTTTGA-GTTAC-AATCAAA
        || ||||||||| ||||||| ||| |||| |||||| ||||| |||||||
540023  TC-GAGAGATCC-AAACAAT-GGCGATCG-CTTTGACGTTACAAATCAAA

   338  ATCCAGTGGAAAATATAATTTATGCAATCCAGGAACTTATTCACAATTAG
        ||||||| |||||||||  |||||| ||||| |||||||||||||||||||
540069  ATCCAGT-GAAAATATA--TTATGC-ATCCA-GAACTTATTCACAATTAG
```

Sample of 100k reads aligned with BLASR requiring >100bp alignment

Read Quality

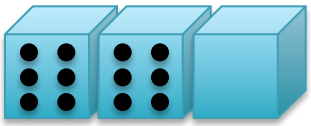
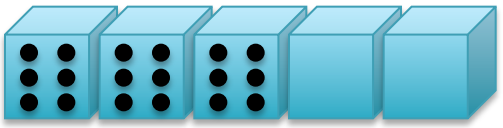


Consistent quality across the entire read

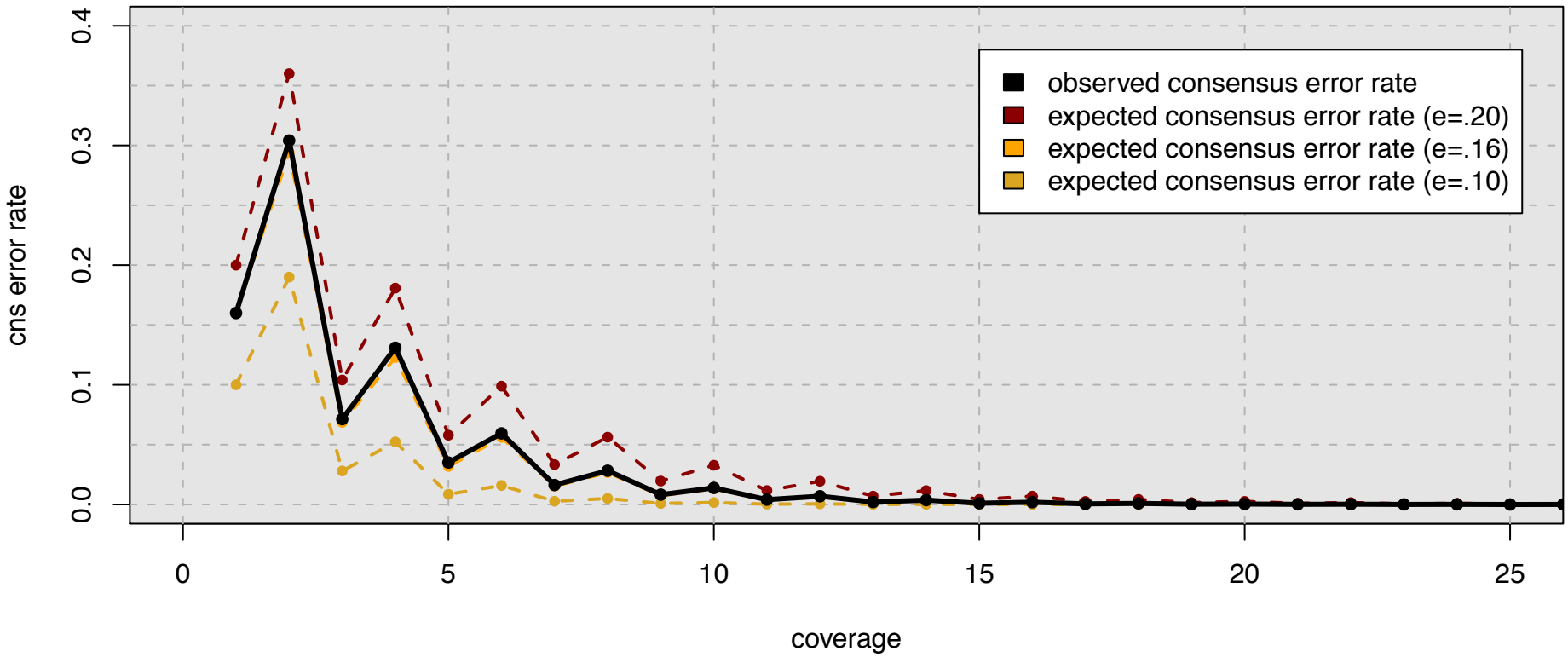
- Uniform error rate, no apparent biases for GC/motifs
- Sampling artifacts at beginning and ends of alignments

Consensus Quality: Probability Review

Roll *n* dice => What is the probability that at least half are 6's

n	*Min to Win*	*Winning Events*	*P(Win)*
1		*1/6*	*16.7%*
2		*P(1 of 2) + P(2 of 2)*	*30.5%*
3		*P(2 of 3) + P(3 of 3)*	*7.4%*
4		*P(2 of 4) + P(3 of 4) + P(4 of 4)*	*13.2%*
5		*P(3 of 5) + P(4 of 5) + P(5 of 5)*	*3.5%*
n	*ceil(n/2)*	$\sum_{i=\lceil n/2 \rceil}^{n} P(i \text{ of } n) = \sum_{i=\lceil n/2 \rceil}^{n} \binom{n}{i} (p)^i (1-p)^{n-i}$	

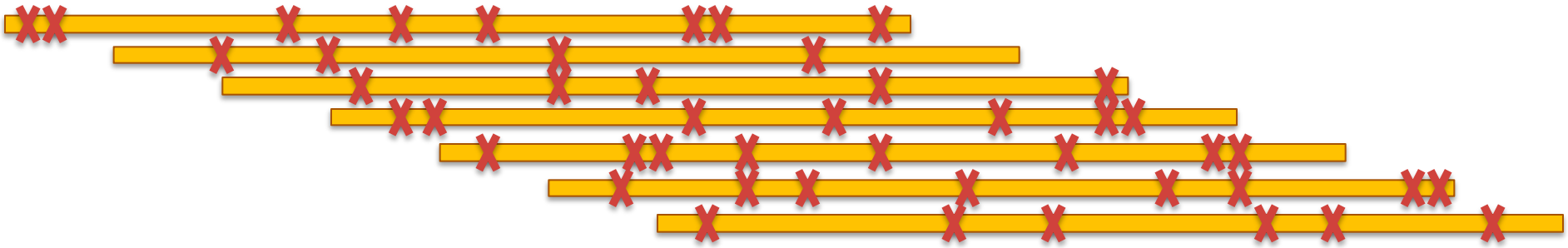
Consensus Accuracy and Coverage



Coverage can overcome random errors

- Dashed lines – accuracy model of binomial sampling
- Solid line – observed consensus error rate
- For same reason, CCS is extremely accurate when using 5+ subreads

Approach 1: SMRT-de novo



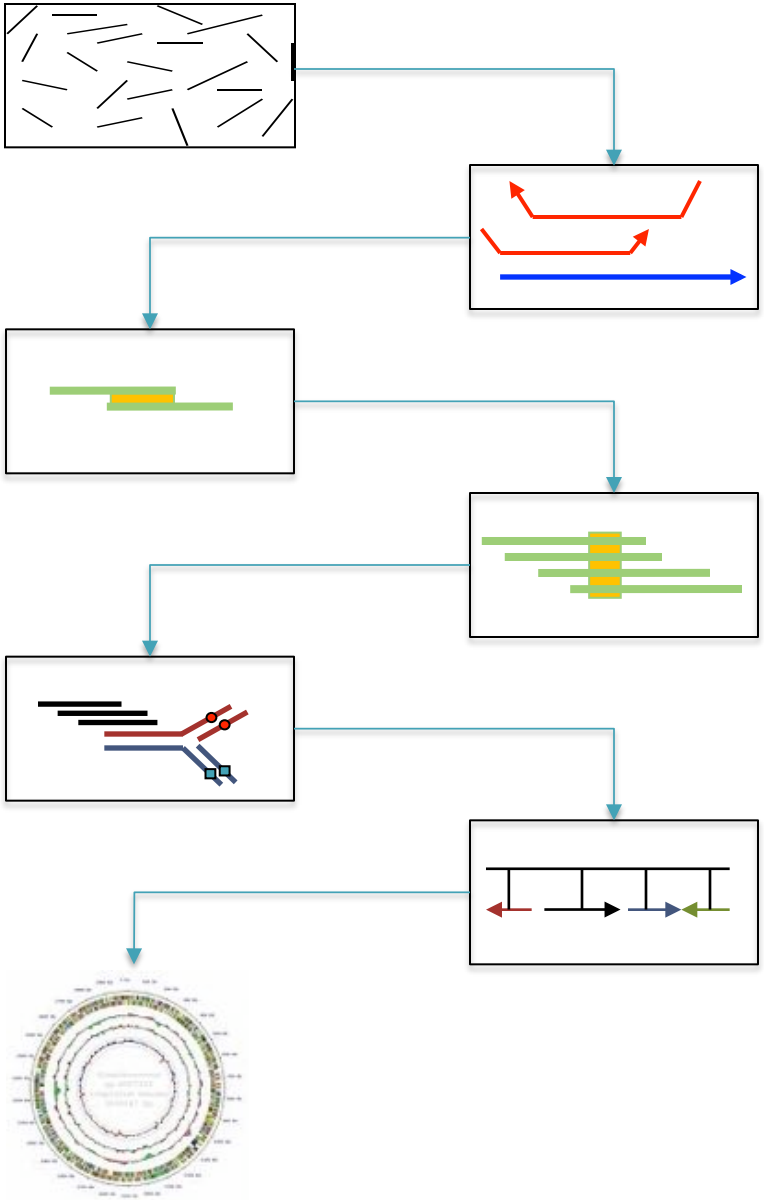
- De novo assembly of SMRT-reads
  - Rapid sequencing and assembly
  - Long reads to span repeats

- Challenges
  - 15% error rate per read equates to ~30% error rate per overlap
  - CCS reads as shorter, but higher quality reads

Celera Assembler

http://wgs-assembler.sf.net

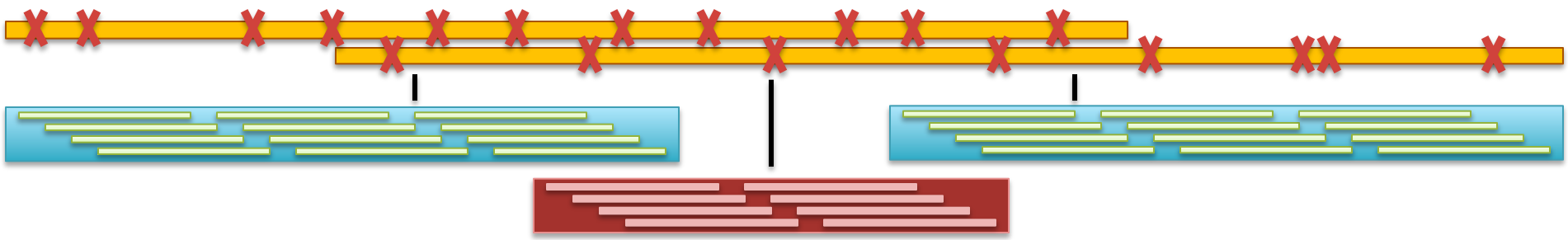
1. Pre-overlap
   - Consistency checks
2. Trimming
   - Quality trimming & partial overlaps
3. Compute Overlaps
   - Find high quality overlaps
4. Error Correction
   - Evaluate difference in context of overlapping reads
5. Unitigging
   - Merge consistent reads
6. Scaffolding
   - Bundle mates, Order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



SMRT-de novo Results

- De novo assembly of long reads
  - Experiments in progress
  - Very challenging to find good overlaps with very high error rate
- De novo assembly of CCS reads
  - Contig N50: 24,582bp
- De novo assembly of ref-corrected CCS
  - Contig N50: 65,119bp

Approach 2: SMRT-scaffolding



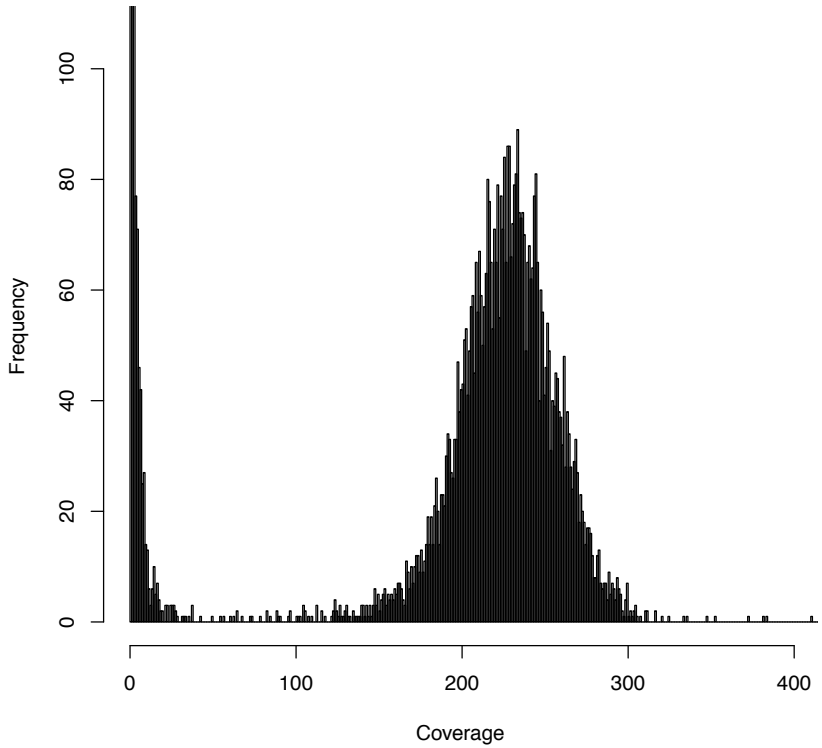
- Use long reads (or strobe reads) to link high quality contigs from short reads
  - Long reads (orange) span repetitive short-read contig (red)
  - Doesn't need very high coverage nor accuracy of long reads

- Challenges
  - Creating good short read assembly
  - Properly aligning reads to contigs
  - Untangling complex repeats

Illumina Sequencing & Assembly

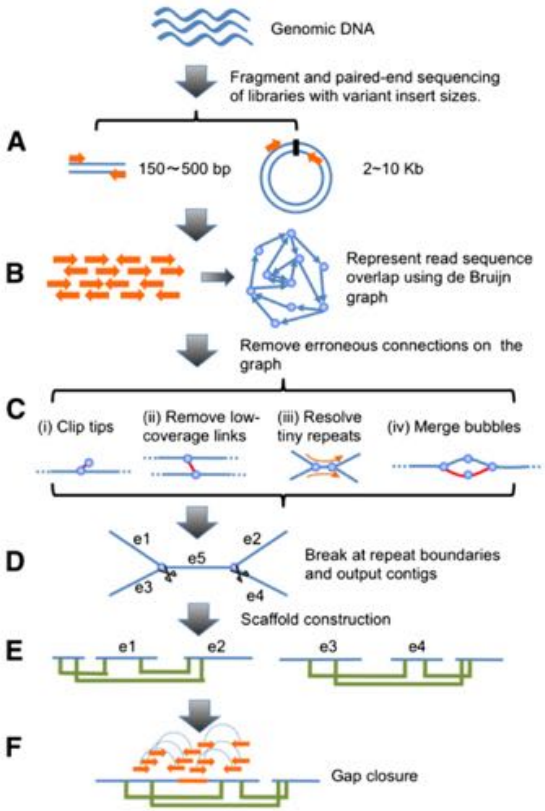
Quake Results

2x76bp @ 275bp
2x36bp @ 3400bp



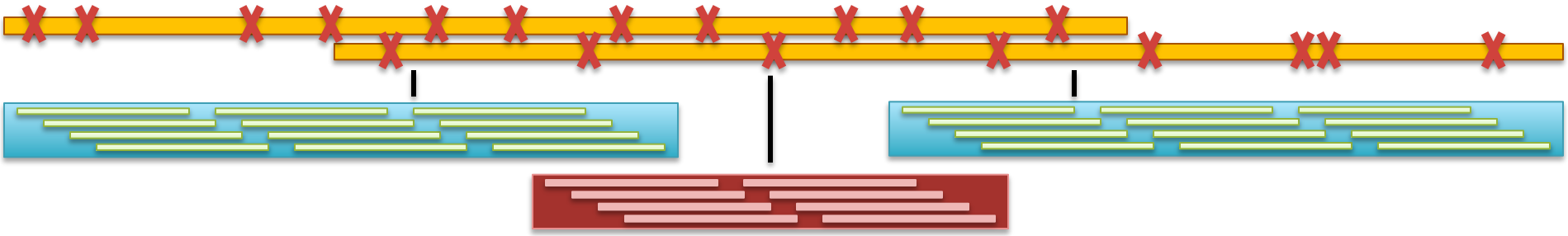
Validated	51,243,281	88.5%
Corrected	2,763,380	4.8%
Trim Only	3,273,428	5.6%
Removed	606,251	1.0%

SOAPdenovo Results



	# ≥ 100bp	N50 (bp)
Scaffolds	2,340	253,186
Contigs	2,782	56,374
Unitigs	4,151	20,772

SMRT-scaffolding results



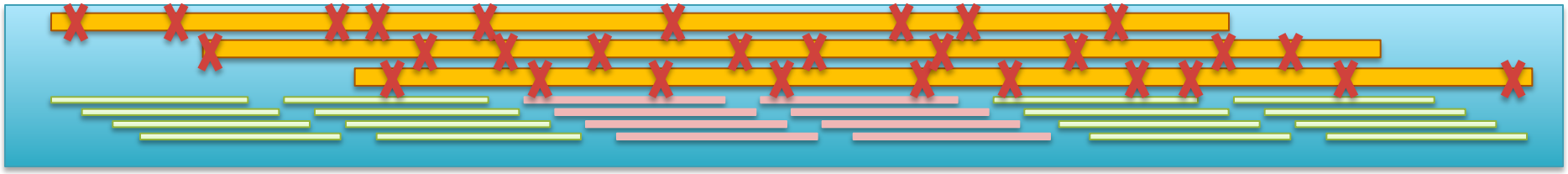
SMRTpipe hybrid scaffold of SOAPdenovo assembly + >2kbp long reads

- Scaffold N50: 310,246bp (+22% improvement)
- Scaffold cnt: 2246 (4% reduction)

SMRTpipe hybrid scaffold of ref-CCS assembly + >2kbp long reads

- Scaffold N50: 97,414bp (+50% improvement)
- Scaffold cnt: 6,610 (3% reduction)

Approach 3: SMRT-hybrid



- 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

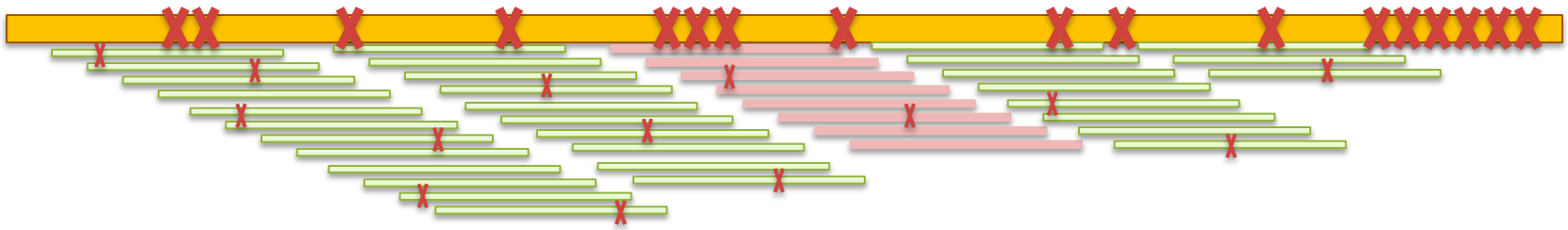
- Challenges
  - Long reads have too high of an error rate to assemble directly
  - Assembler must supports a wide mix of read lengths

PacBio Error Correction

http://wgs-assembler.sf.net

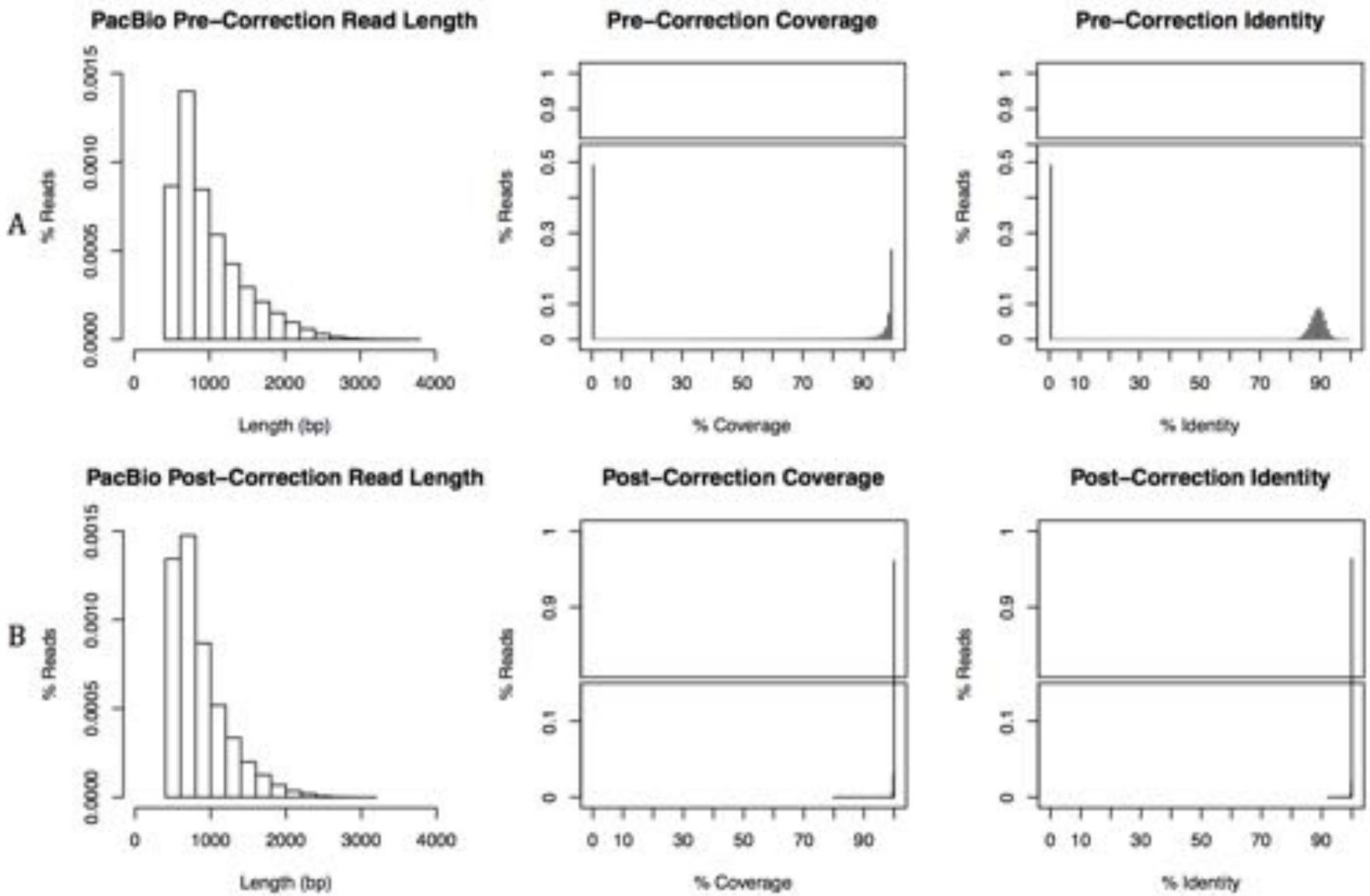
1. Correction Pipeline
   1. Map short reads (SR) to long reads (LR)
   2. Trim LRs at coverage gaps
   3. Compute consensus for each LR

2. Error corrected reads can be easily assembled, aligned



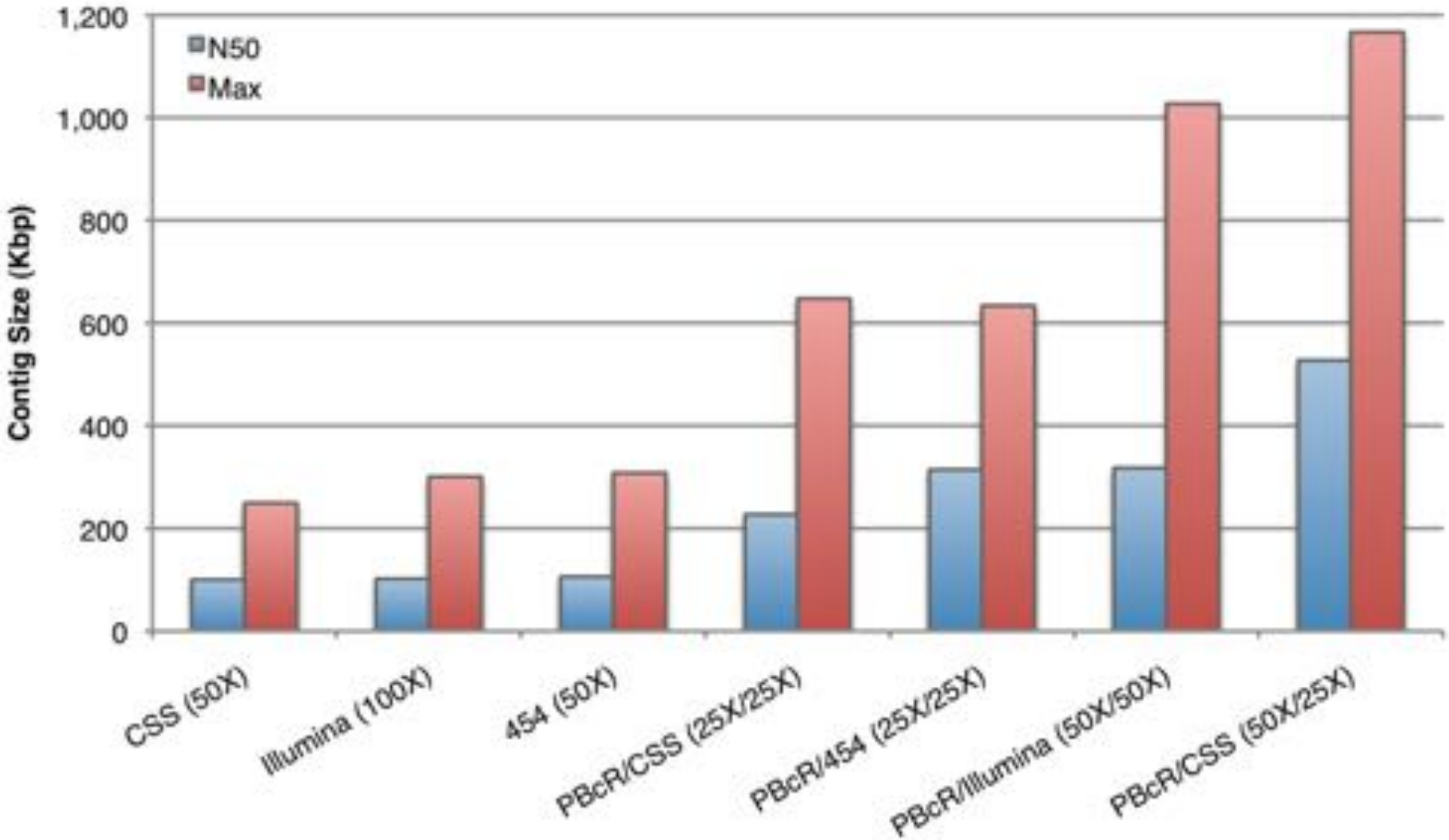
Hybrid error correction and de novo assembly of single-molecule sequencing reads.
Koren, S, Schatz, MC, Walenz, BP, Martin, J, Howard, J, Ganapathy, G, Wang, Z, Rasko, DA, McCombie, WR, Jarvis, ED, Phillippy, AM. (2012) *Under Review*

Error Correction Results



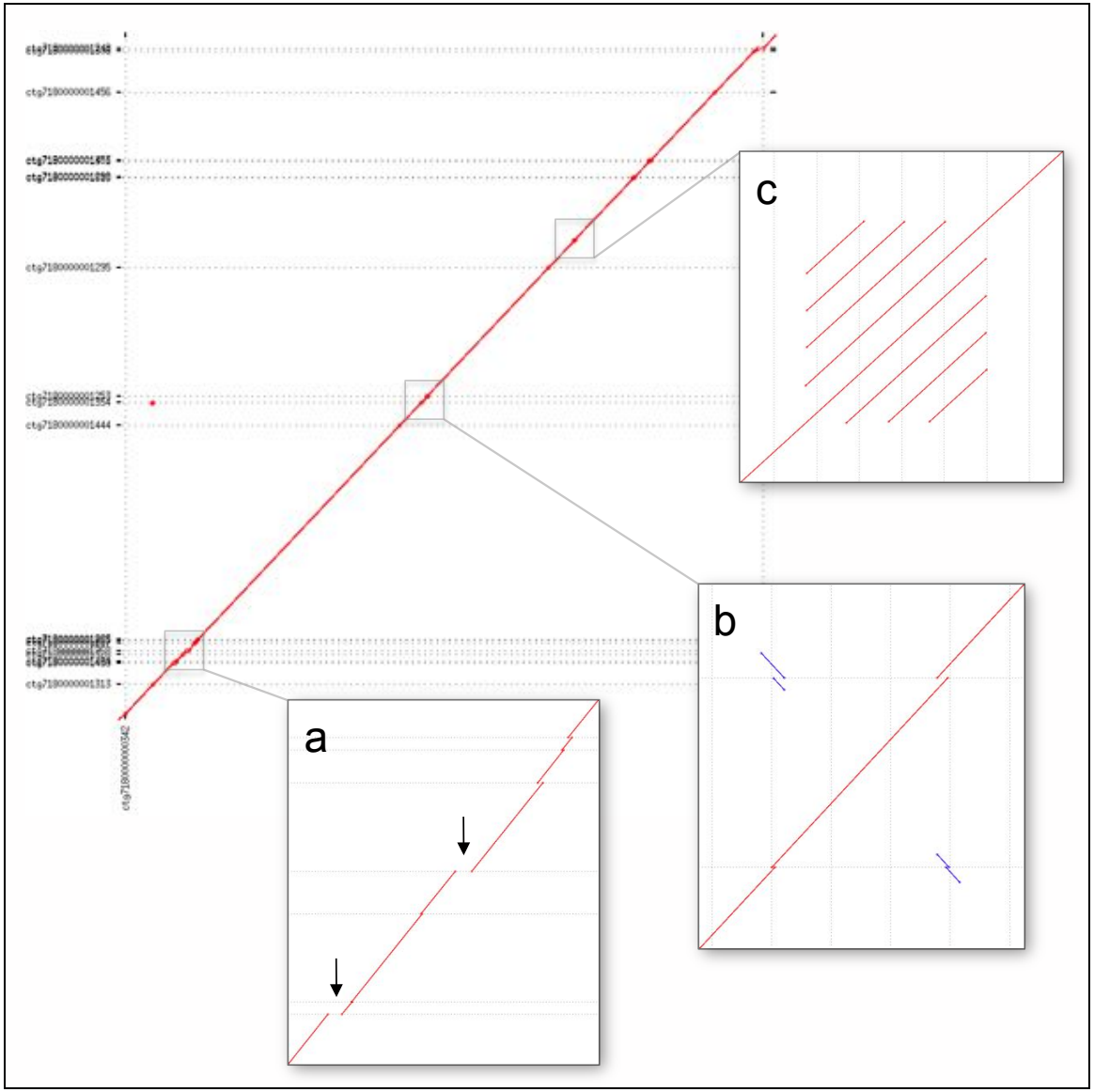
Correction results of 20x PacBio coverage of E. coli K12 corrected using 50x Illumina

Assembly Results



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

PacBio Advantages

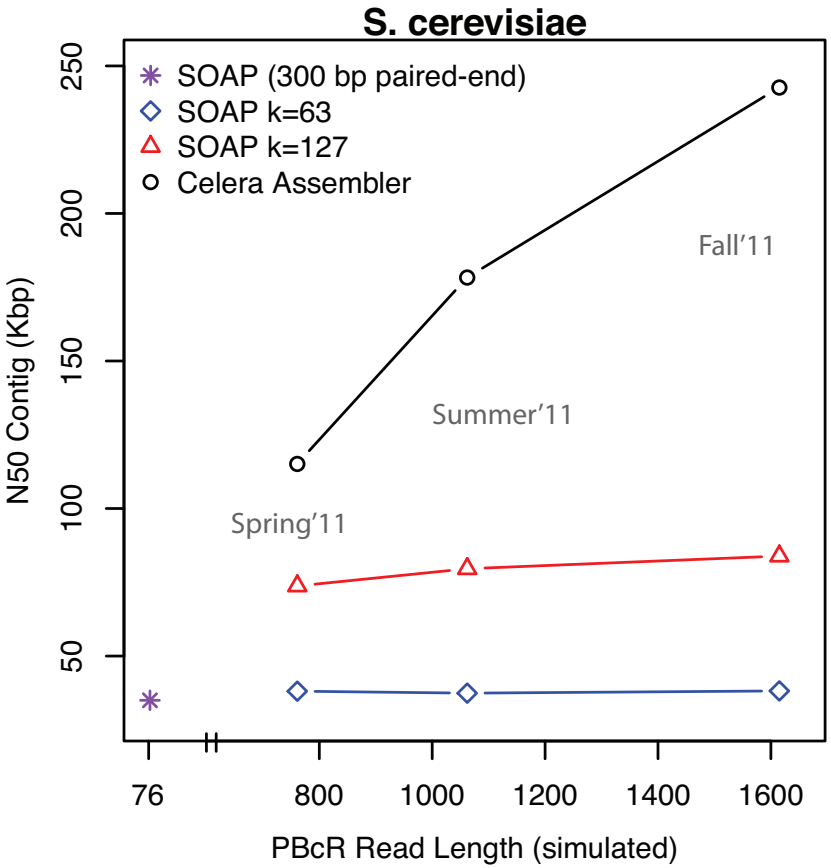


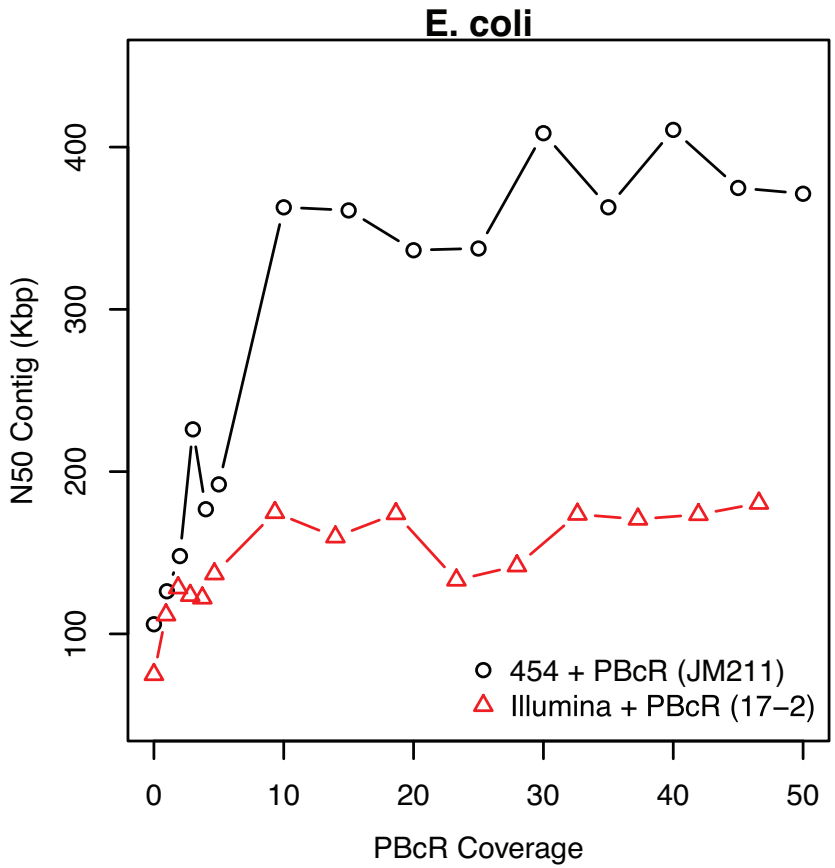
(a) Long reads close sequencing gaps

(b) Long reads assemble across long repeats

(c) Long reads span complex microsatellites

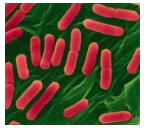
Assembly Results

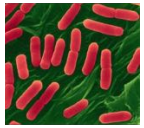


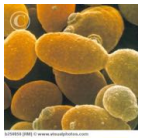


- The longer the read, the greater the improvement to assembly quality
- Best assemblies come from PacBio + 454 reads or PacBio + CCS
- Best assemblies need ~10x coverage PacBio long reads

Hybrid Assembly Results



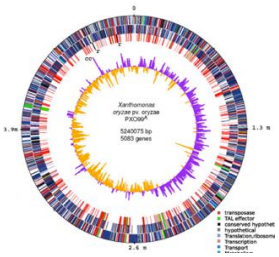




Organism	Technology	Reference bp	Assembly bp	# Contigs	Max Contig Length	N50
Lambda NEB3011	Illumina 100X 200bp	48 502	48 492	1	48 492 / 48 492	48 492 / 48 492 (100%) *
(median: 727 max: 3 280)	PacBio PBcR 25X		48 440	1	48 444 / 48 444	48 444 / 48 440 (100%) *
E. coli K12	Illumina 100X 500bp	4 639 675	4 462 836	61	221 615 / 221 553	100 338 / 83 037 (82.76%) *
(median: 747 max: 3 068)	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%) *
E. coli C227-11	PacBio CCS 50X	5 504 407	4 917 717	76	249 515	100 322
(median: 1 217 max: 14 901)	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[a]		5 452 251	23	653 382	402 041
S. cerevisiae S228c	Illumina 100X 300bp	12 157 105	11 034 156	192	266 528 / 227 714	73 871 / 49 254 (66.68%) *
(median: 674 max: 5 994)	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%) *
Melopsittacus undulatus	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
(median 997, max 13 079)	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

Assembly Summary



Assembly quality depends on

1. ***Coverage***: low coverage is mathematically hopeless
2. ***Read length***: longer reads resolve repeats and complex regions
3. ***Read Quality:*** need clean libraries, clean reads

- 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

- PacBio enables highest quality de novo assembly
  - Longer reads have fundamentally more information than shorter reads
  - Because the errors are random we can compensate for them informatically
  - Software available open source at http://wgs-assembler.sf.net

Acknowledgements

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

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

NBACC
Adam Phillipy
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

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