SMRT-assembly approaches

Michael Schatz

Sept 7, 2011

PacBio Users Meeting

Outline

1. Assembly preliminaries
    1. De Bruijn and Overlap graph
    2. Coverage, read length, repeats, and errors

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

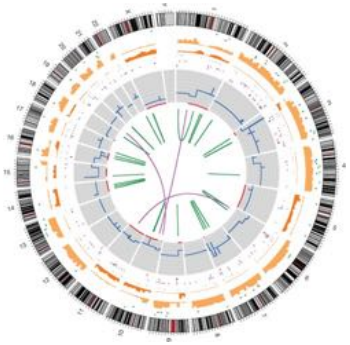
3. Review and best practices

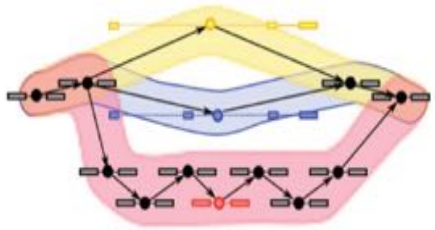
Assembly Applications

- Novel genomes

- Metagenomes

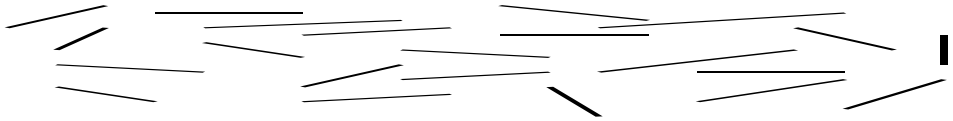
- Sequencing assays
  - Structural variations
  - Transcript assembly
  - …





Assembling a Genome

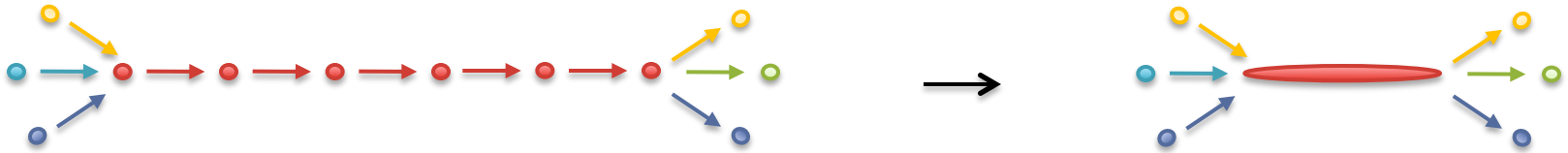
1. Shear & Sequence DNA



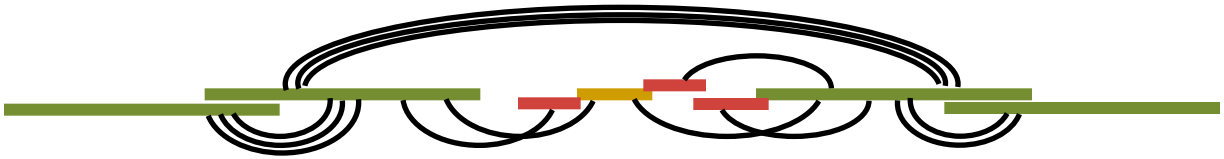
2. Construct assembly graph from overlapping reads

...AGCCTAGACCTACAGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGT...

3. Simplify assembly graph

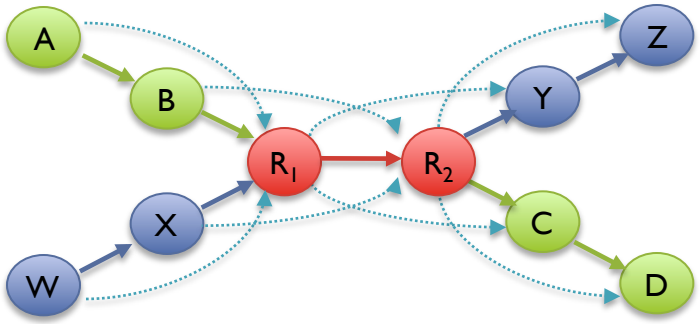


4. Detangle graph with long reads, mates, and other links



Two Paradigms for Assembly

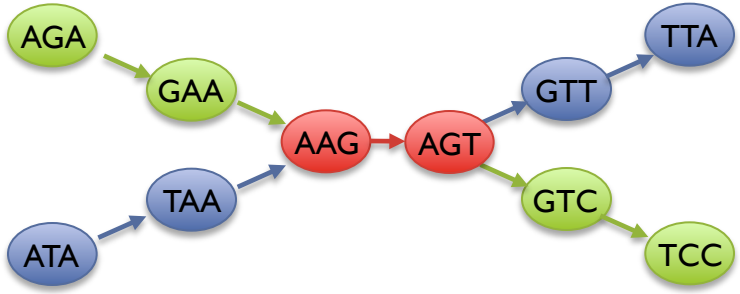
Overlap Graph



Long read assemblers

- Repeats depends on read length
- Read coherency, placements kept
- Tangled by high coverage

de Bruijn Graph



Short read assemblers

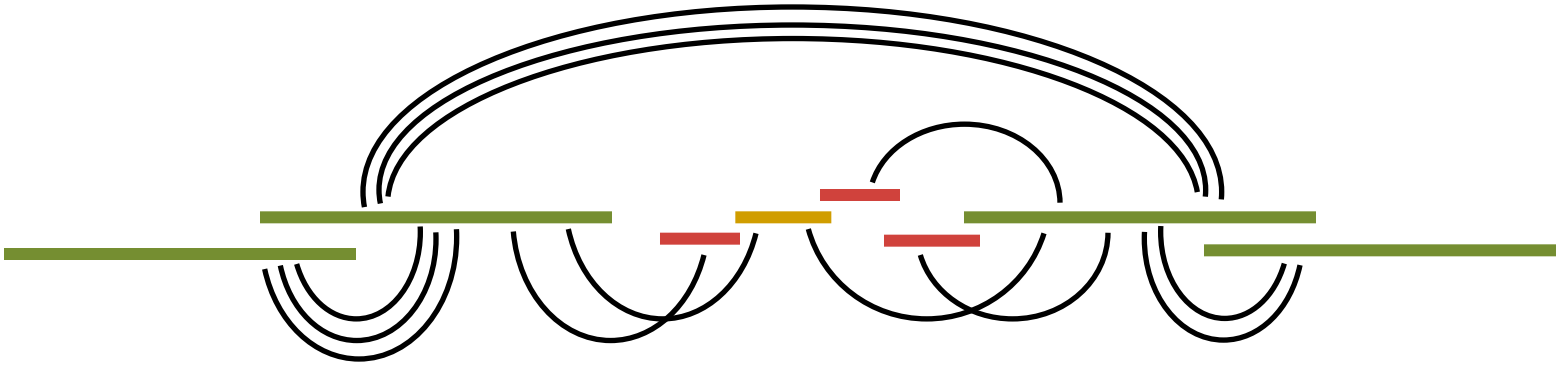
- Repeats depends on word length
- Read coherency, placements lost
- Robust to high coverage

Assembly of Large Genomes using Second Generation Sequencing
Schatz MC, Delcher AL, Salzberg SL (2010) *Genome Research.* 20:1165-1173.

Scaffolding

- Initial contigs (*aka* unipaths, unitigs) terminate at
  - *Coverage gaps*: especially extreme GC regions
  - *Conflicts*: sequencing errors, repeat boundaries

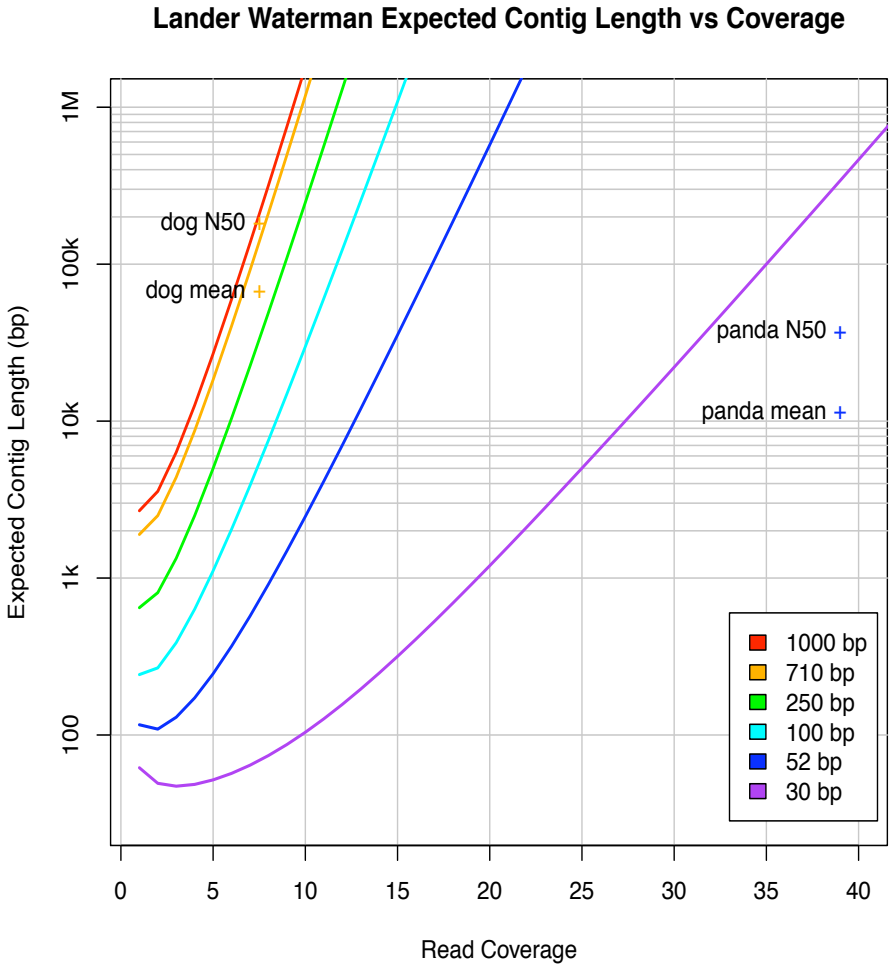
- Iteratively resolve longest, 'most unique' contigs
  - Both overlap graph and de Bruijn assemblers initially collapse repeats into single copies
  - Uniqueness measured by a statistical test on coverage



Coverage and Read Length

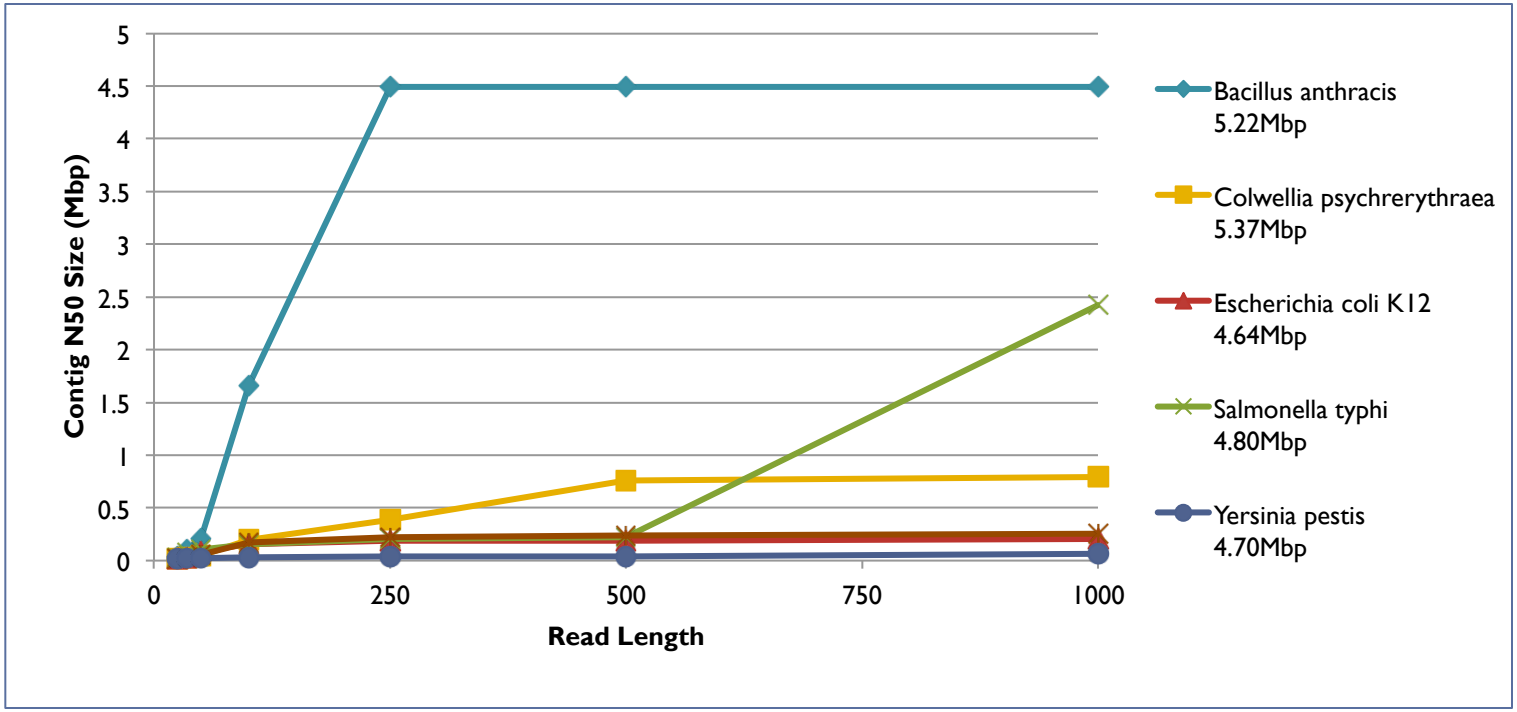
Idealized Lander-Waterman model

- Reads start at perfectly random positions
- Poisson distribution in coverage
  - Contigs end when there are no overlapping reads
- Contig length is a function of coverage and read length
  - Effective coverage reduced by *o/l*
  - Short reads require much higher coverage to reach same expected contig length



Assembly of Large Genomes using Second Generation Sequencing
Schatz MC, Delcher AL, Salzberg SL (2010) *Genome Research.* 20:1165-1173.

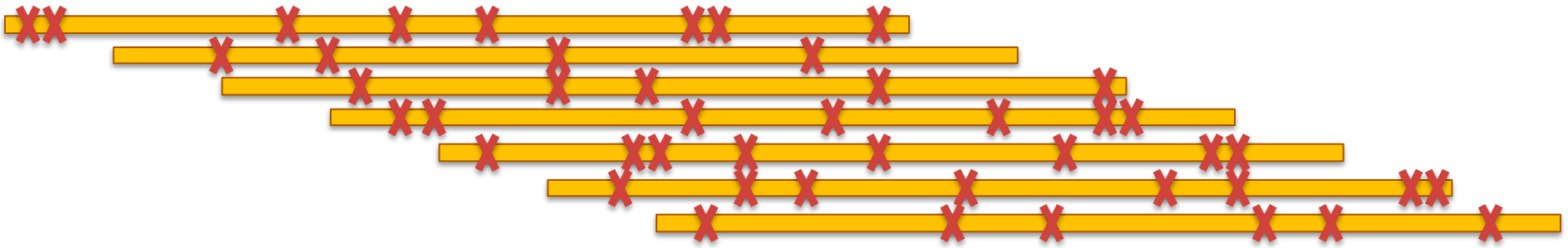
Repeats and Read Length



- Explore the relationship between read length and contig N50 size
  - Idealized assembly of read lengths: 25, 35, 50, 100, 250, 500, 1000
  - Contig/Read length relationship depends on specific repeat composition

Assembly Complexity of Prokaryotic Genomes using Short Reads.
Kingsford C, Schatz MC, Pop M (2010) *BMC Bioinformatics*. 11:21.

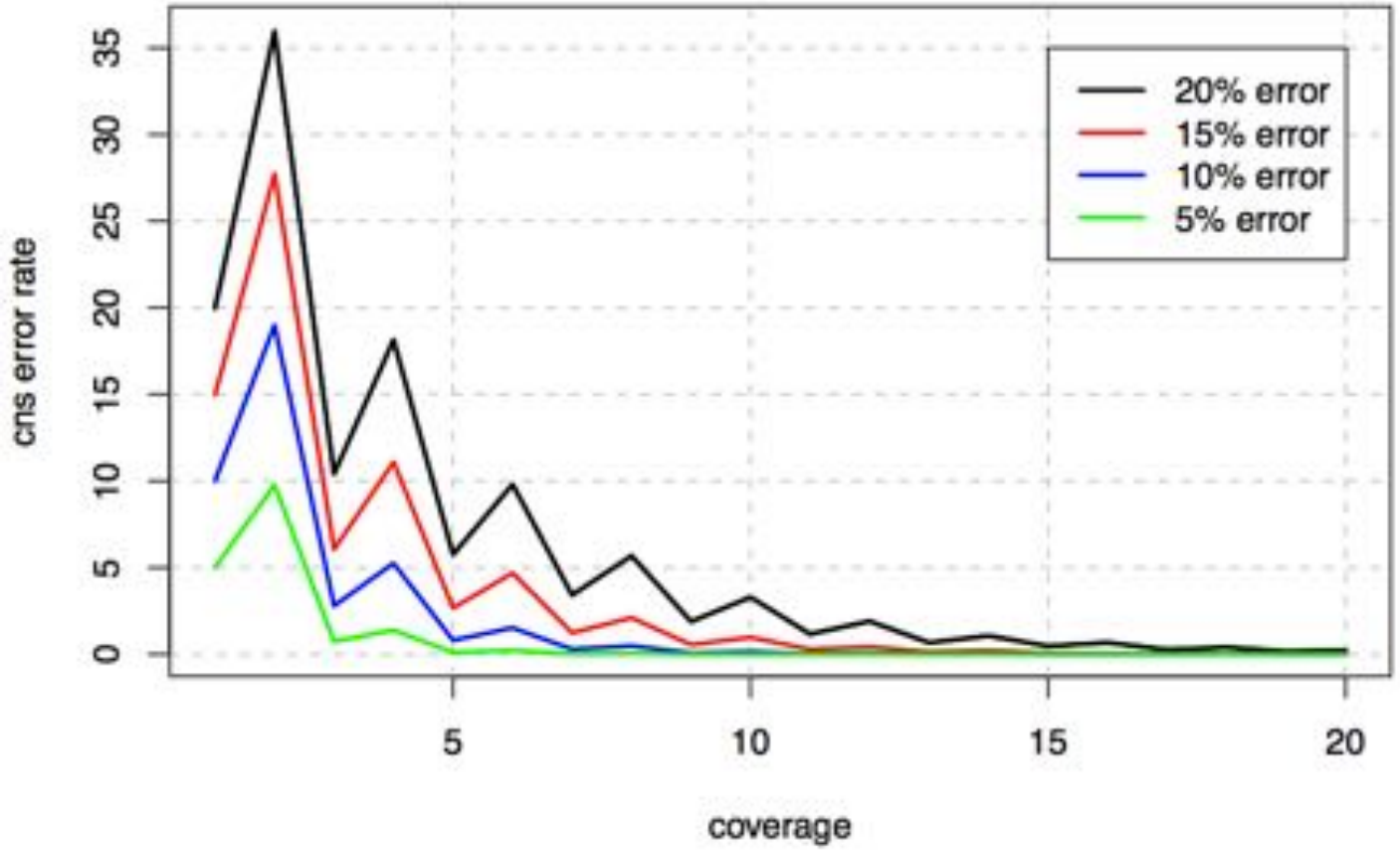
Sequencing Errors and Assembly



Sequencing errors add complexity to assembly

- Obscure overlaps, require shorter words
- Complicates graph by introducing spurs, bubbles, tips, etc
- Increases the effective repeat rate
- Potentially high error rate in consensus

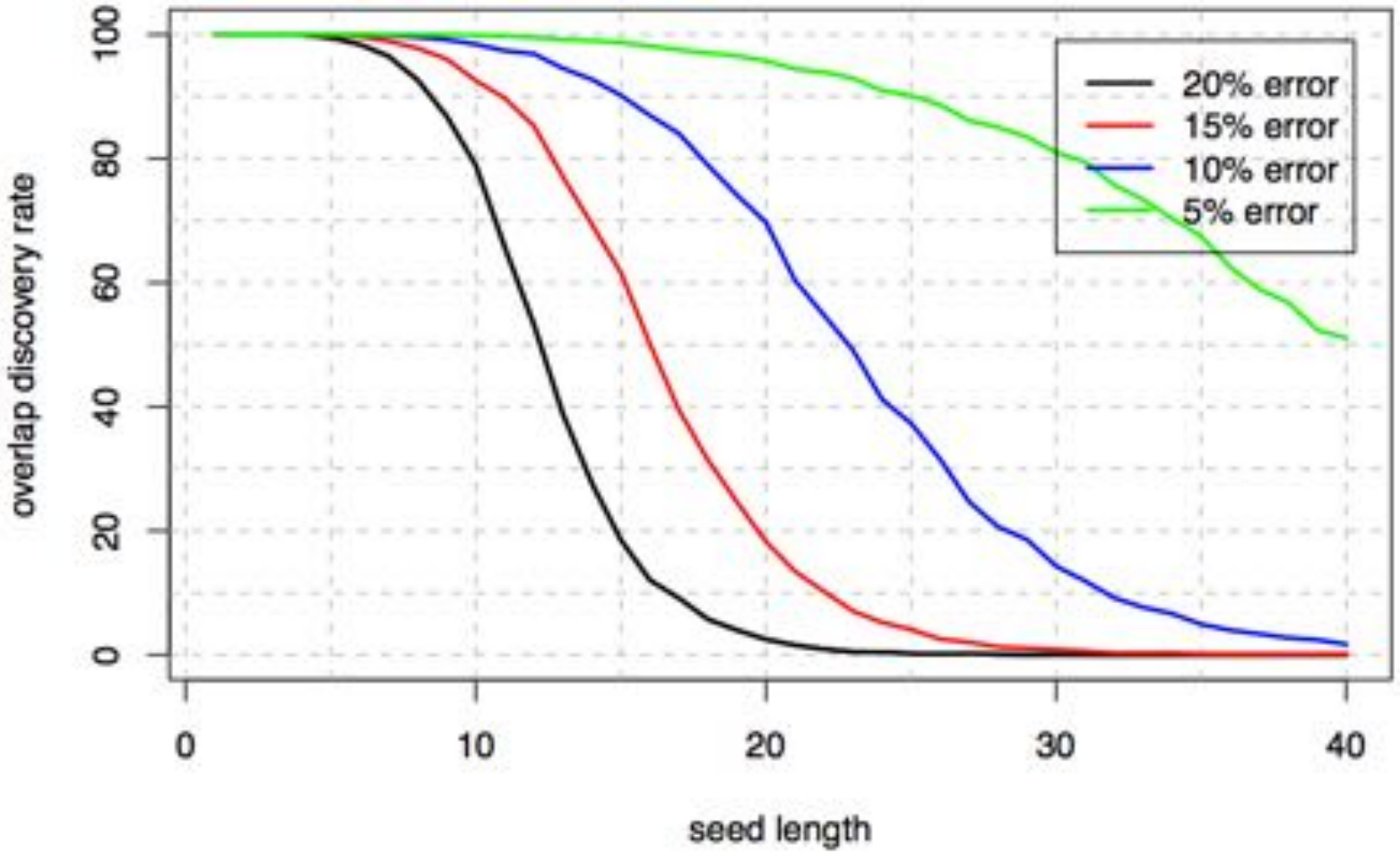
Consensus Accuracy and Coverage



Coverage can overcome most random errors

- Simulate layout of 1000bp reads with random errors
- Compute accuracy of consensus call

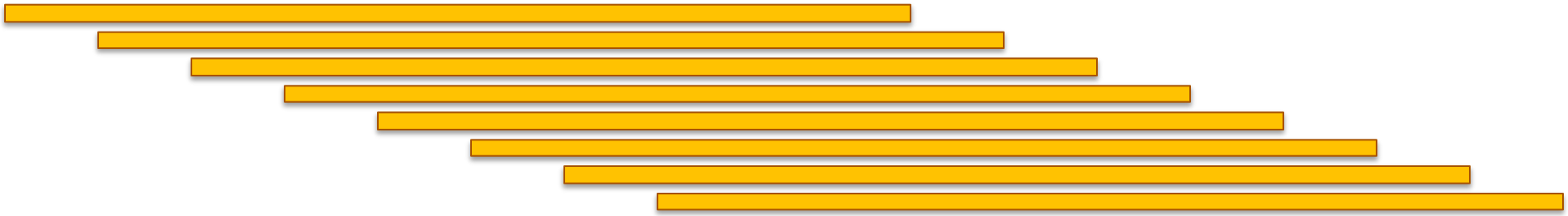
Overlap Seeds and Error Rate



Random errors obscure overlap seeds

- Simulate layout of 20x coverage of 1000bp reads
- What fraction of overlapping reads match for at least S bp?

Approach 1: SMRT-de novo



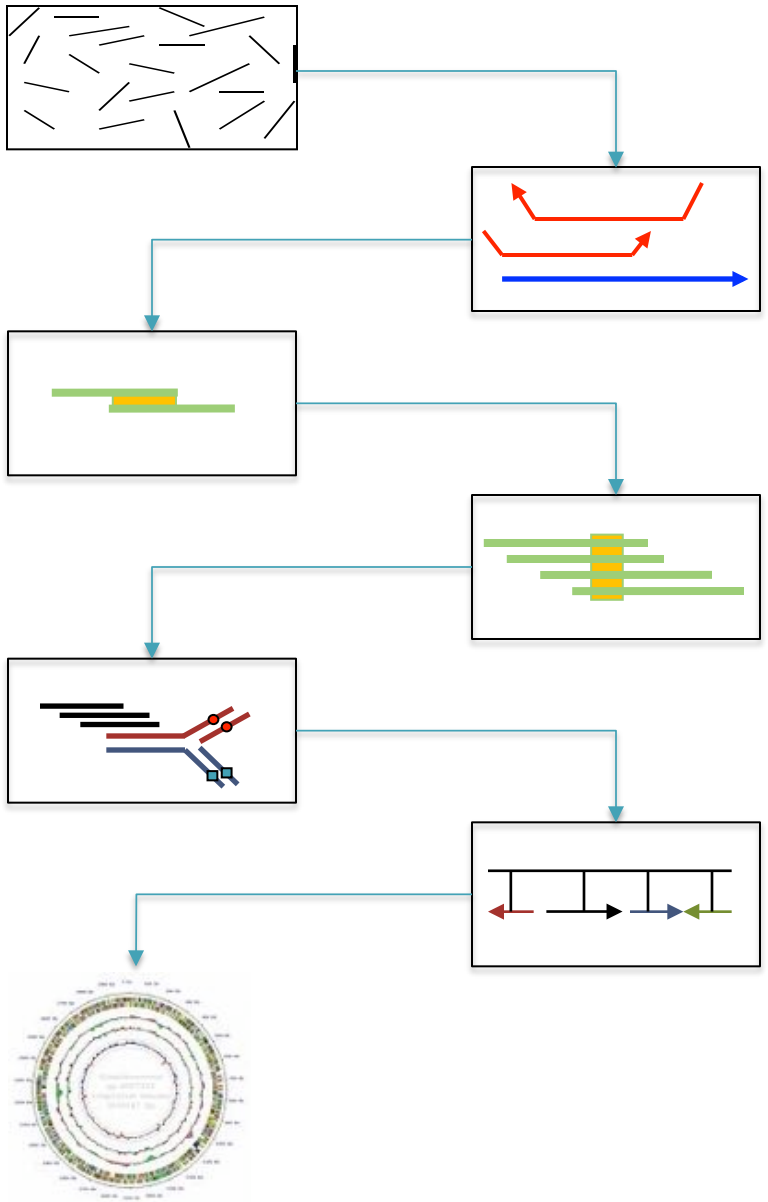
- De novo assembly of SMRT-reads
  - Rapid sequencing and assembly
  - Long reads to span repeats

- 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

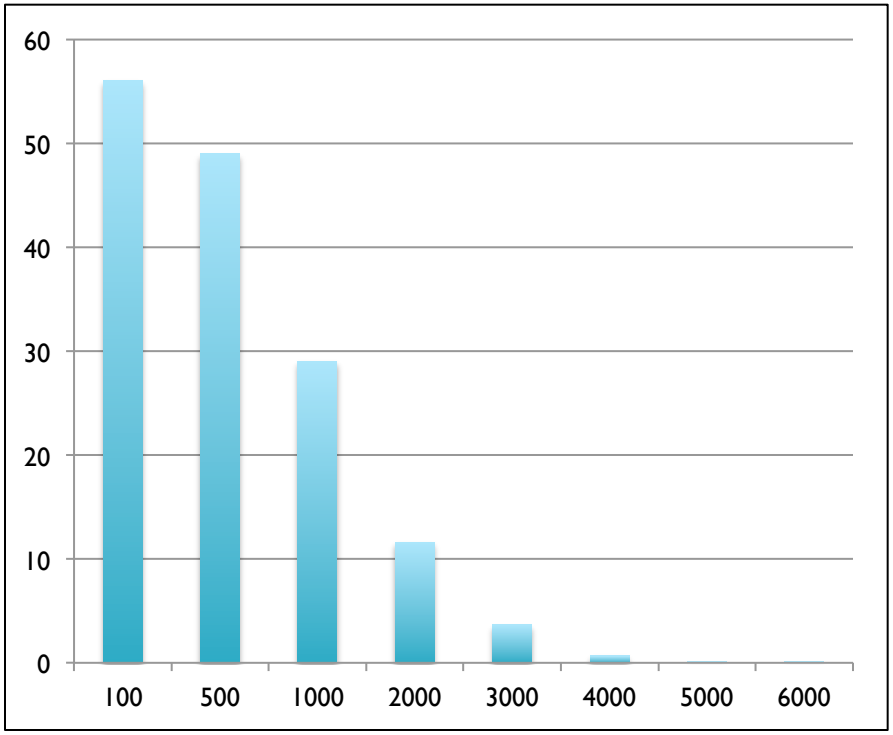
1. Pre-overlap
   - Consistency checks
2. Trimming
   - Quality trimming & partial overlaps
3. Compute Overlaps
   - Find high quality overlaps
4. Error Correction
   - Evaluate difference in context of overlapping reads
5. Unitigging
   - Merge consistent reads
6. Scaffolding
   - Bundle mates, Order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



SMRT Sequencing Runs

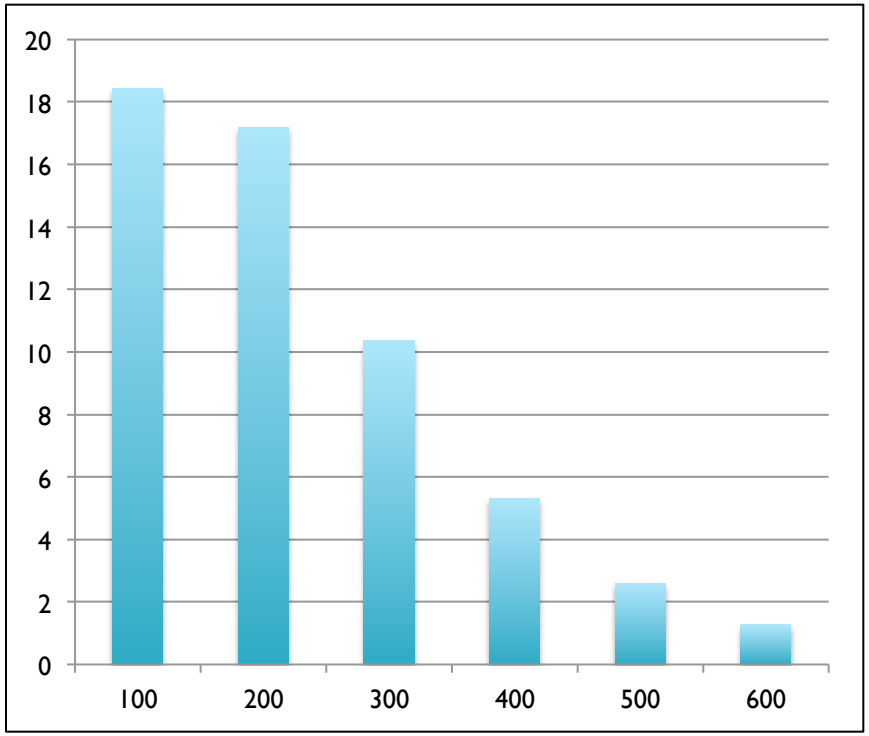
Yeast – Long reads

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



Yeast – CCS reads

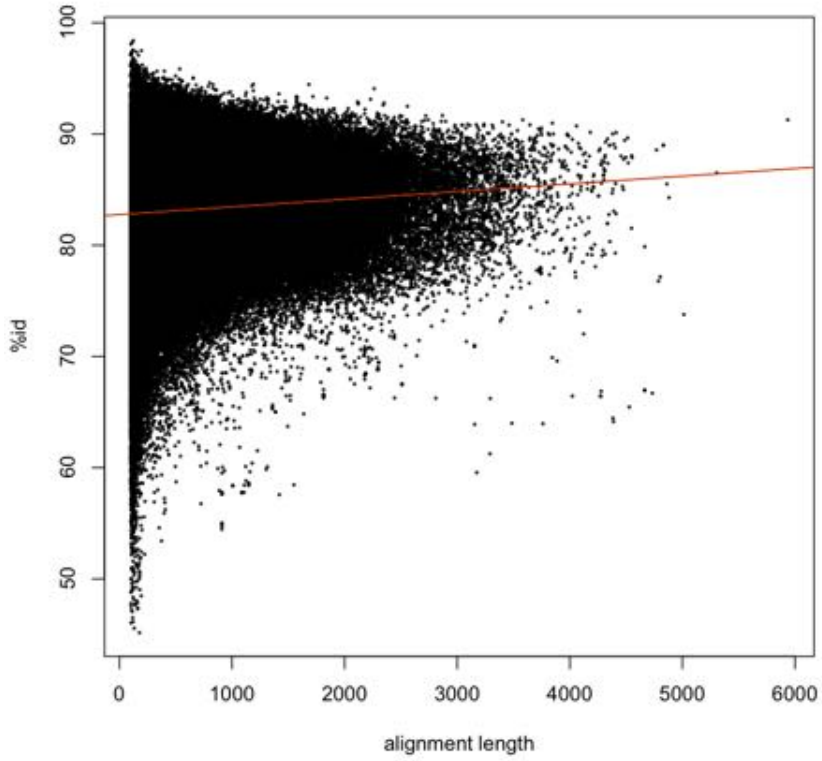
731,638 reads after filtering
Mean: 306 +/- 115
Median: 279 Max: 1,425



Read Accuracy

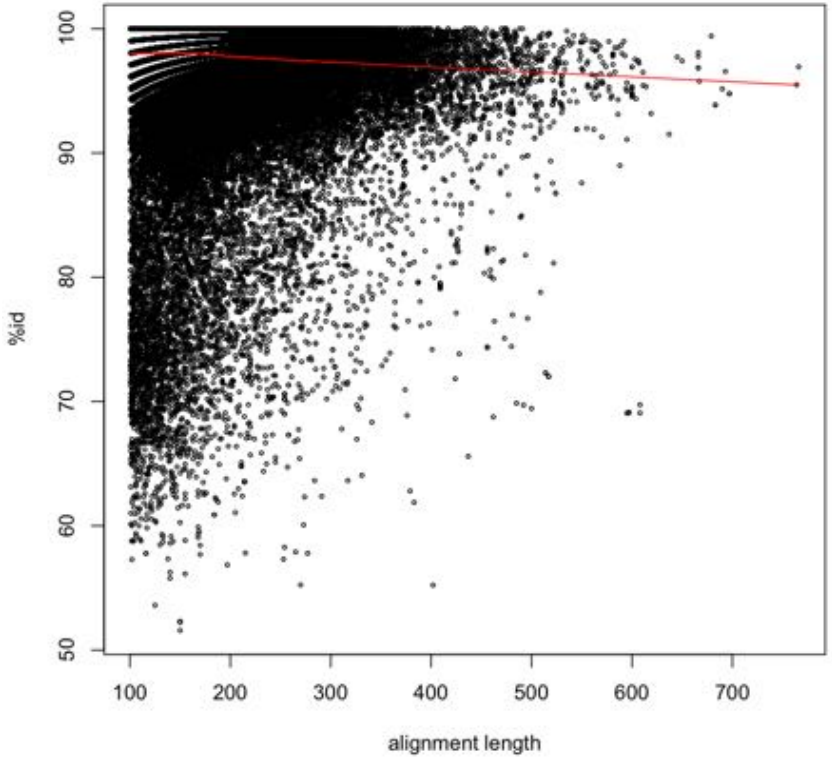
Yeast – Long reads

94% aligned reads
48% reads aligned >100bp
7% reads aligned >1kbp



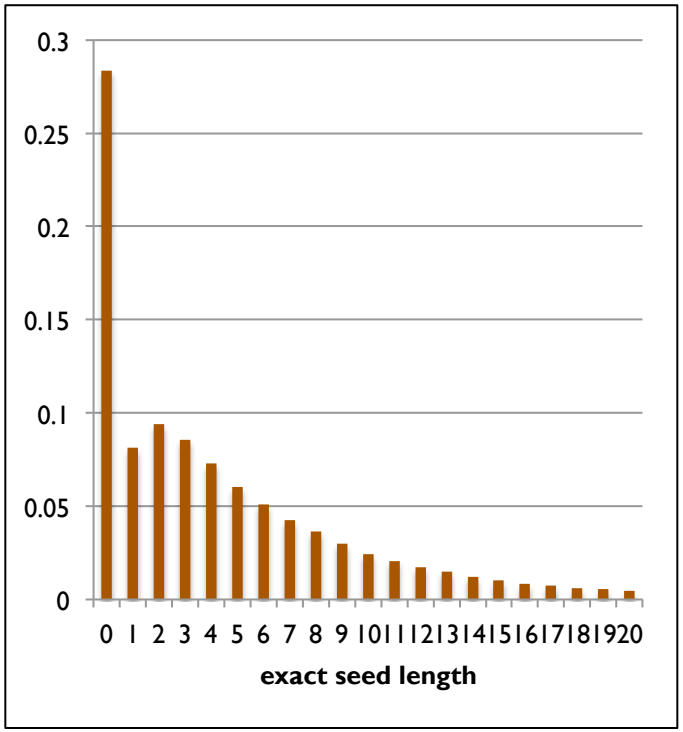
Yeast – CCS reads

99.93% aligned reads
98.2% reads aligned >100bp
38.8% reads aligned >300bp



Alignment Quality

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



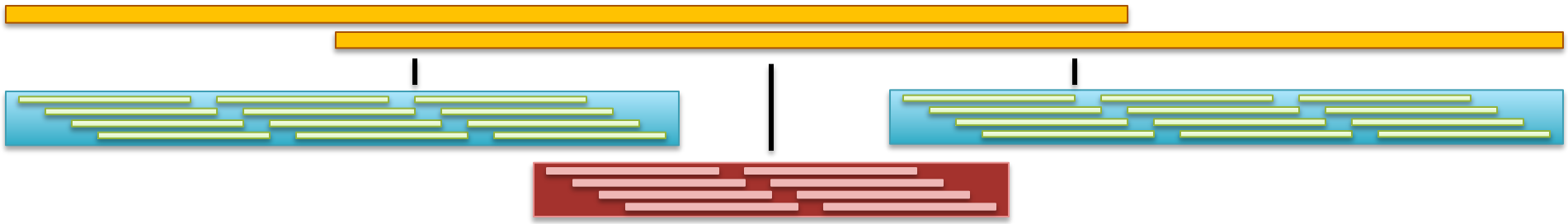

```
     4  TTGTAAGCAGTTGAAAACTATGTGTGGATTTAGAATAAAGAACATGAAAG
        ||||||||||||||||||||||||| ||||||| |||||||||||| |||
539752  TTGTAAGCAGTTGAAAACTATGTGT-GATTTAG-ATAAAGAACATGGAAG

    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  -AGGAGGGGAAGGGGGGAATATCT-ATAAAGATTACAAATTAGA-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 1M reads aligned with BLASR requiring >100bp alignment

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)

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

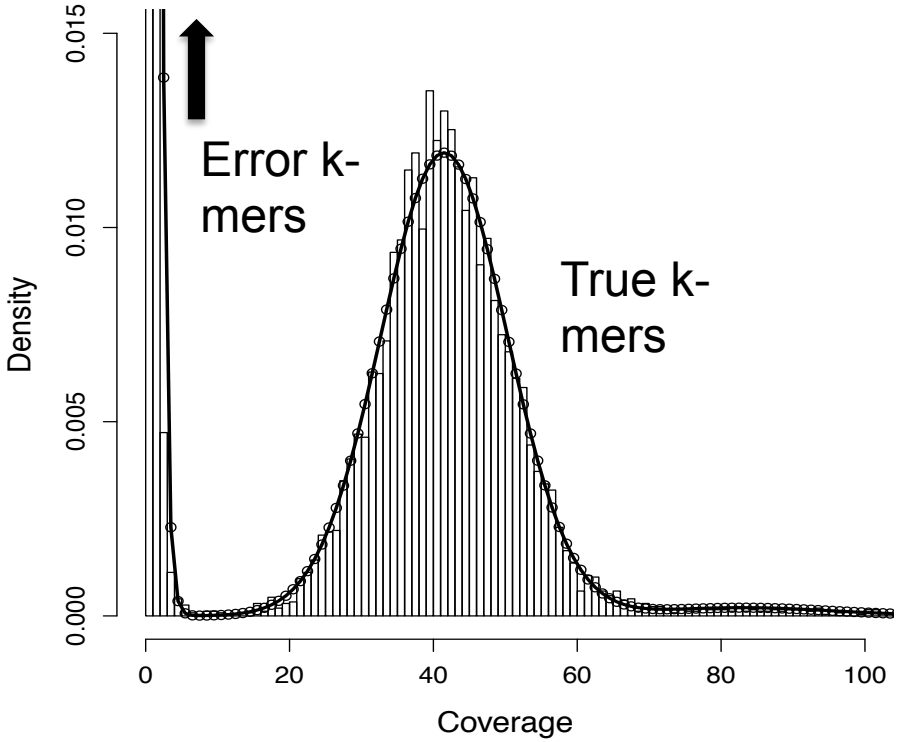
Error Correction with Quake

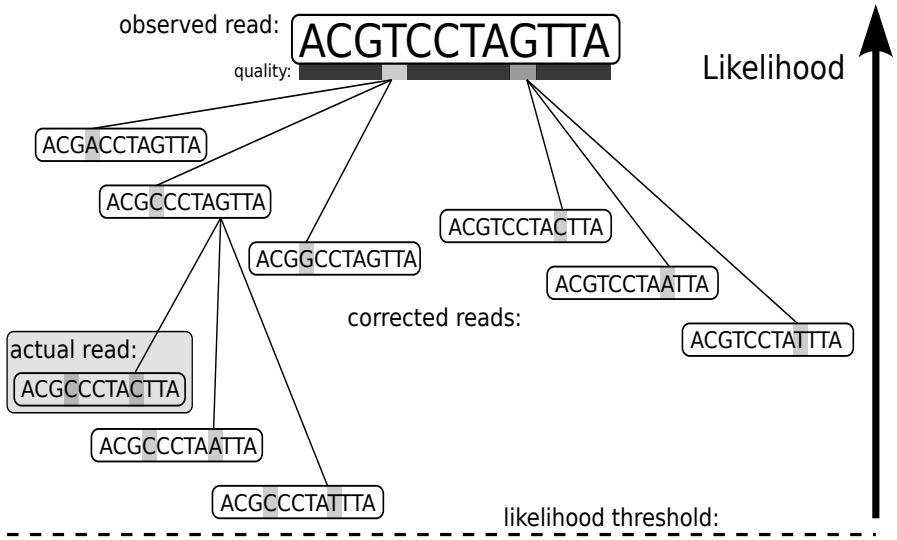
1. Count all "Q-mers" in reads

- Fit coverage distribution to mixture model of errors and regular coverage
- Automatically determines threshold for trusted k-mers

2. Correction Algorithm

- Considers editing erroneous kmers into trusted kmers in decreasing likelihood
- Includes quality values, nucleotide/nucleotide substitution rate



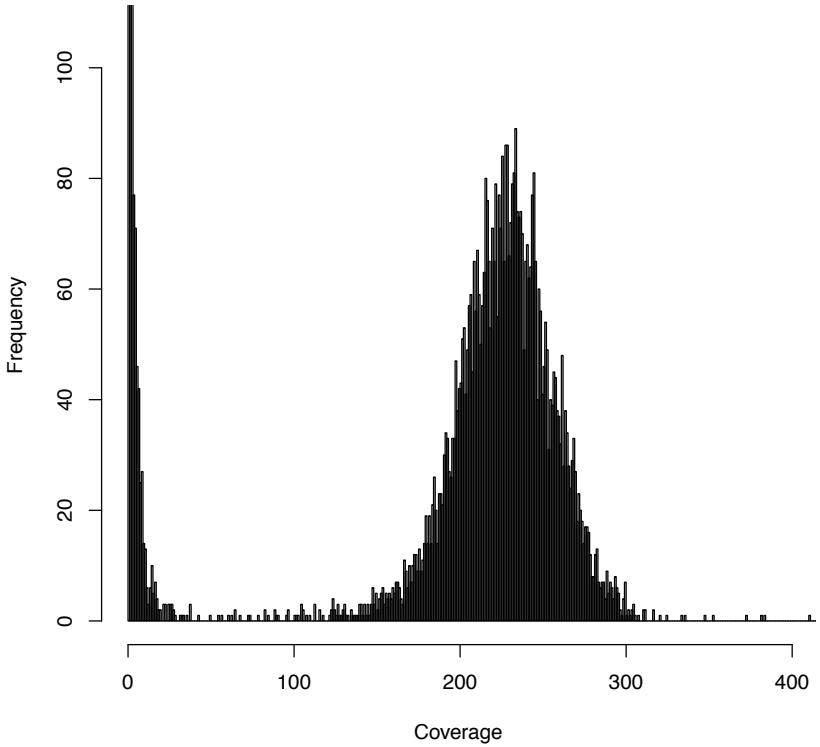


Quake: quality-aware detection and correction of sequencing reads.
Kelley, DR, Schatz, MC, Salzberg SL (2010) *Genome Biology*. 11:R116

Illumina Sequencing & Assembly

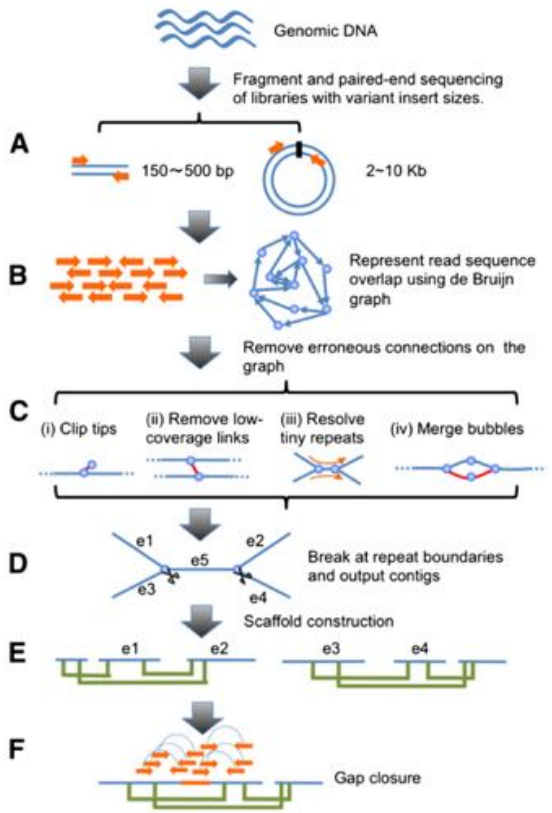
Quake Results

2x76bp @ 275bp
2x36bp @ 3400bp



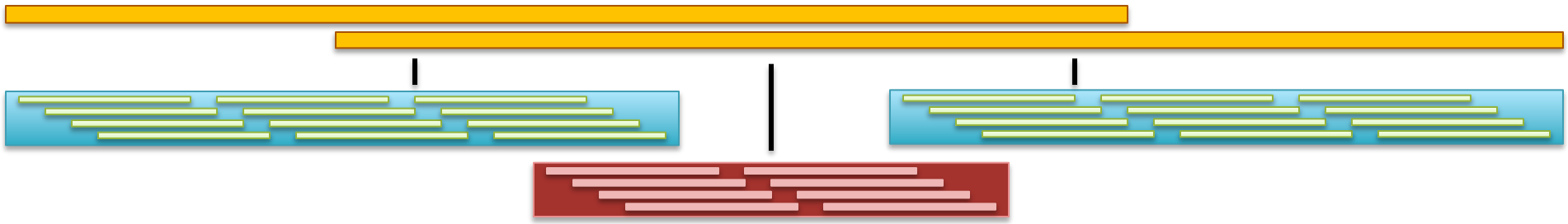
Validated	51,243,281	88.5%
Corrected	2,763,380	4.8%
Trim Only	3,273,428	5.6%
Removed	606,251	1.0%

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)

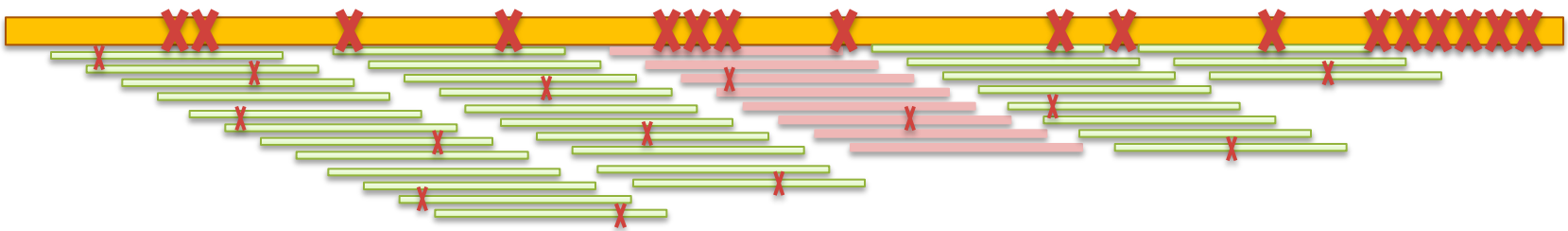
Approach 3: SMRT-hybrid

- Co-assemble long reads and short reads
  - Long reads 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

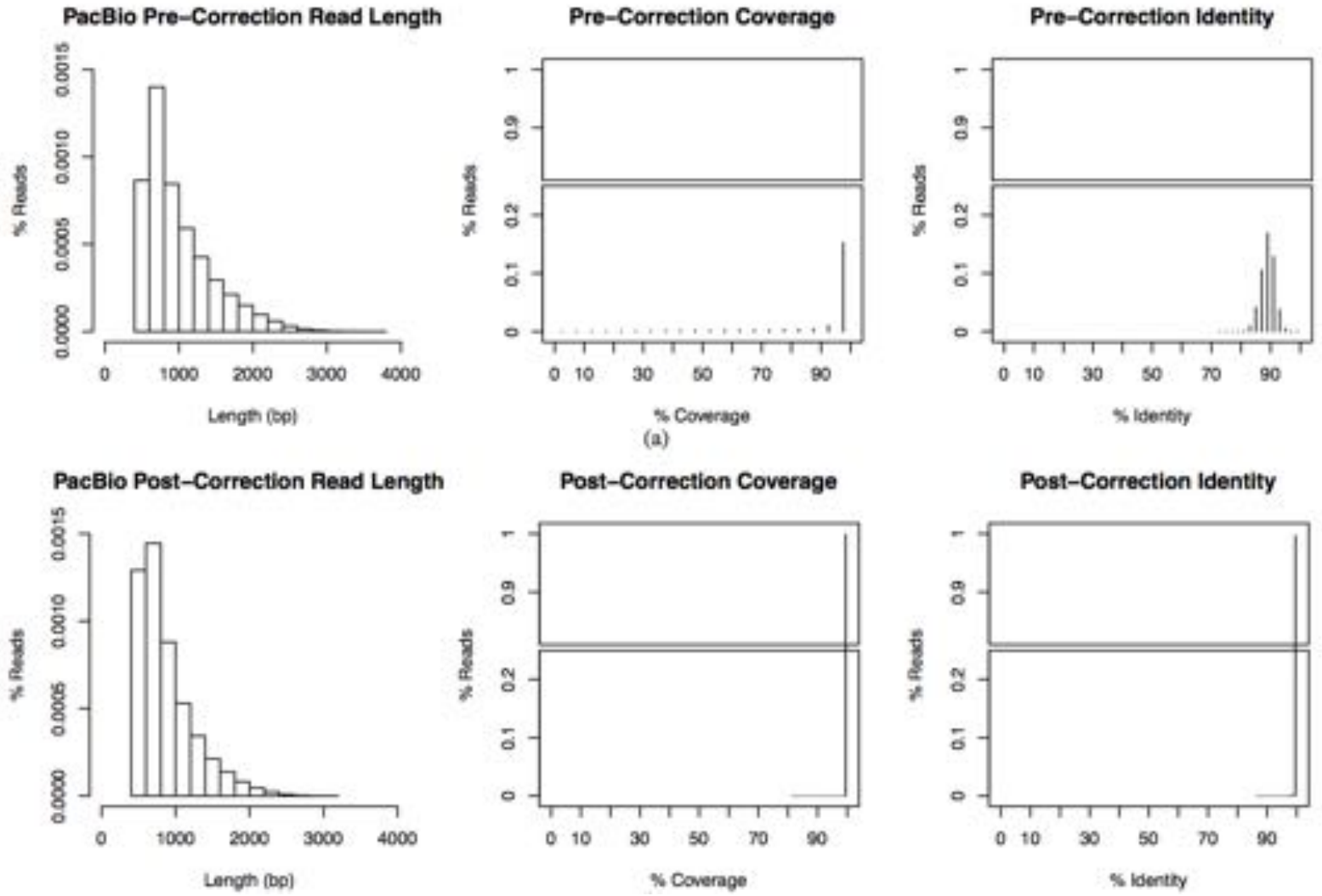
SMRT-hybrid Error Correction & Assembly

1. Trim/correct SR sequence
2. Compute an SR layout for each LR
   1. Map SRs to LRs
   2. Trim LRs at coverage gaps
   3. Compute consensus for each LR
3. Co-assemble corrected LRs and SRs
   – Celera Assembler enhanced to support 16 Kbp reads



A hybrid strategy for utilizing single-molecule sequencing data for genome assembly and RNA-Seq. Koren, S, Walenz, BP, Martin, J, Jarvis, ED, Rasko, DA, Schatz, MC, McCombie, WR, Phillippy, AM. (2011) *In preparation.*

Error Correction Results



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

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

Conclusions

- SMRT-sequencing extremely promising for de novo assembly
  - Compute high quality consensus sequence from error prone reads
  - Long reads are challenging to use alone, but very effective when combined with high quality CCS or short reads
- Error correction is key to unlocking potential of SMRT-sequencing
  - The leading second generation sequence assemblers aggressively compensate for the platform specific error model
  - Easy prediction: same will be true for 3rd generation assemblers
- Significant challenges ahead
  - Technology: Throughput, accuracy, read length, cost
  - Protocols: Sample preparation, library construction, read types
  - Informatics: Robust computational analysis methods

Acknowledgements

CSHL
Dick McCombie
Melissa Kramer
Eric Antoniou
Laura Gelley
Stephanie Muller

NBACC
Adam Phillippy
Sergey Koren

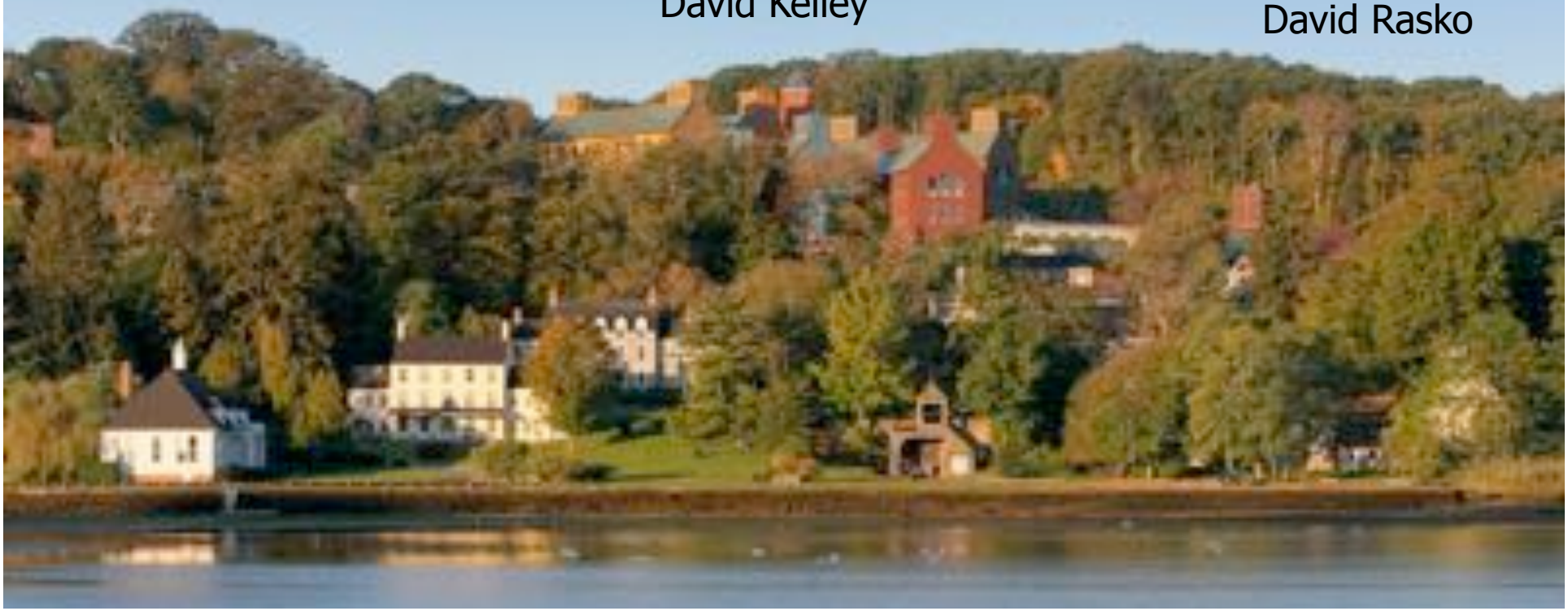
UMD
Steven Salzberg
Mihai Pop
David Kelley

JCVI
Brian Walenz

JGI
Jeffrey Martin

Duke
Erich Jarvis

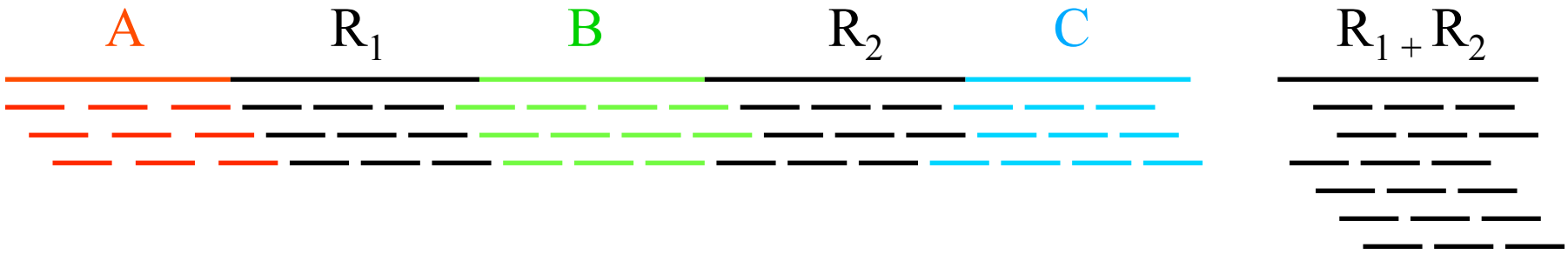
UMD SOM
David Rasko



Thank You

http://schatzlab.cshl.edu
@mike_schatz

Repeats and Coverage Statistics

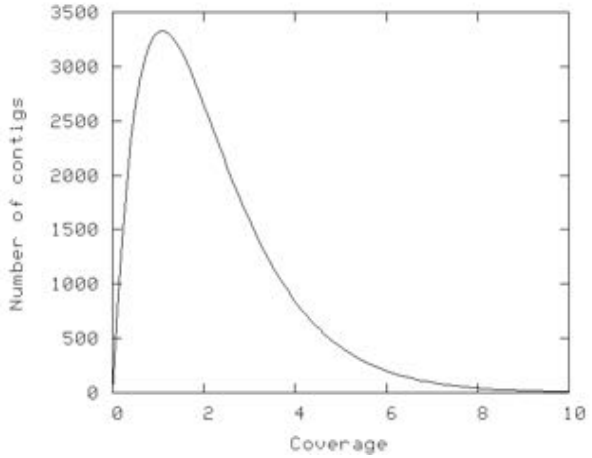


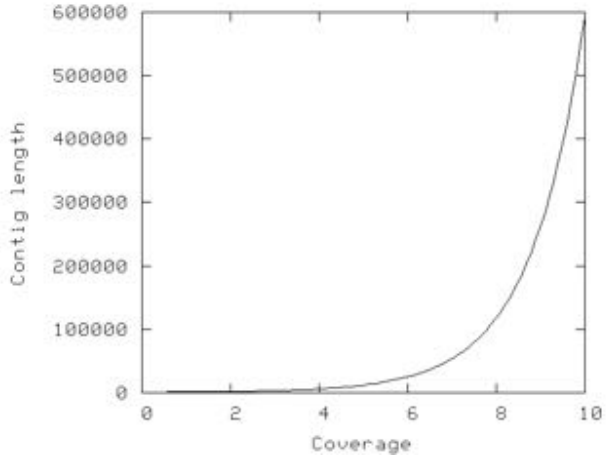
- If *n* reads are a uniform random sample of the genome of length *G*, we expect $k=n\Delta/G$ reads to start in a region of length Δ.
  - If we see many more reads than k (if the arrival rate is > A) , it is likely to be a collapsed repeat
  - Requires an accurate genome size estimate

$$\Pr(X - copy) = \binom{n}{k}\left(\frac{X\Delta}{G}\right)^{k}\left(\frac{G - X\Delta}{G}\right)^{n-k}$$

$$A(\Delta,k) = \ln\left(\frac{\Pr(1-copy)}{\Pr(2-copy)}\right) = \ln\left(\frac{\frac{(\Delta n/G)^{k}}{k!}e^{\frac{-\Delta n}{G}}}{\frac{(2\Delta n/G)^{k}}{k!}e^{\frac{-2\Delta n}{G}}}\right) = \frac{n\Delta}{G} - k\ln 2$$

Lander-Waterman statistics





L = read length

T = minimum overlap

G = genome size

N = number of reads

c = coverage (NL / G)

$\sigma = 1 - T/L$

$E(\#islands) = Ne^{-c\sigma}$

$E(island\ size) = L(e^{c\sigma} - 1) / c + 1 - \sigma$

contig = island with 2 or more reads