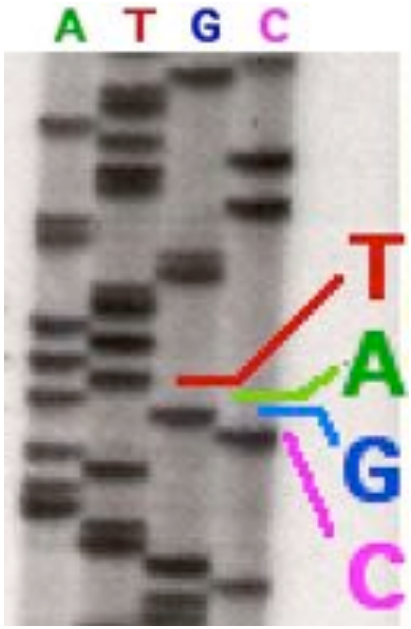
Sequencing Pitfalls

Michael Schatz

July 16, 2012
URP Bioinformatics

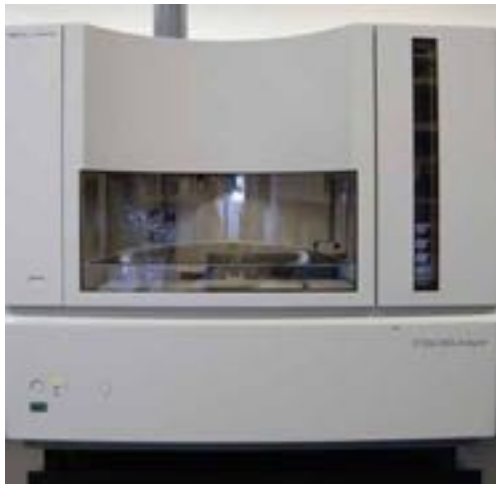
Advances in Sequencing: Zeroth, First, Second Generation



1970s: 0th Gen

Radioactive Chain Termination

5000bp / week



1980s-1990s: 1st Gen

Automated Capillary Sequencing

384kbp / day

2000s: 2nd Gen

Pyrosequencing, SOLiD Sequencing-by-Synthesis

1Gbp+ / day

Advances in Sequencing: Now Generation Sequencing

Illumina HiSeq 2000
Sequencing by Synthesis

>60Gbp / day
100bp reads

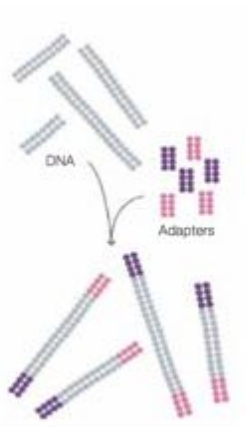
PacBio
SMRT-sequencing

~1Gbp / day
Long Reads

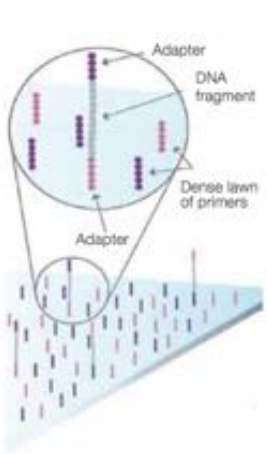
Oxford Nanopore
Nanopore sensing

Many GB / day?
Very Long Reads?

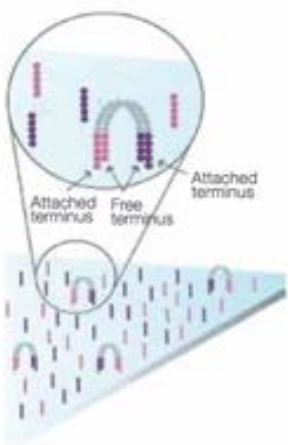
Illumina Sequencing by Synthesis



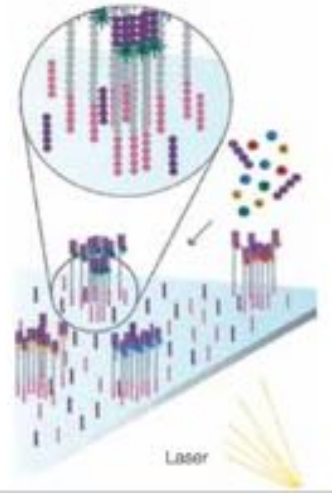
1. Prepare



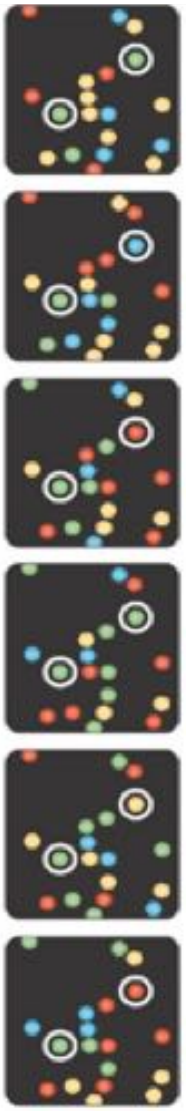
2. Attach



3. Amplify



4. Image



5. Basecall

Metzker (2010) Nature Reviews Genetics 11:31-46

http://www.youtube.com/watch?v=l99aKKHcxC4

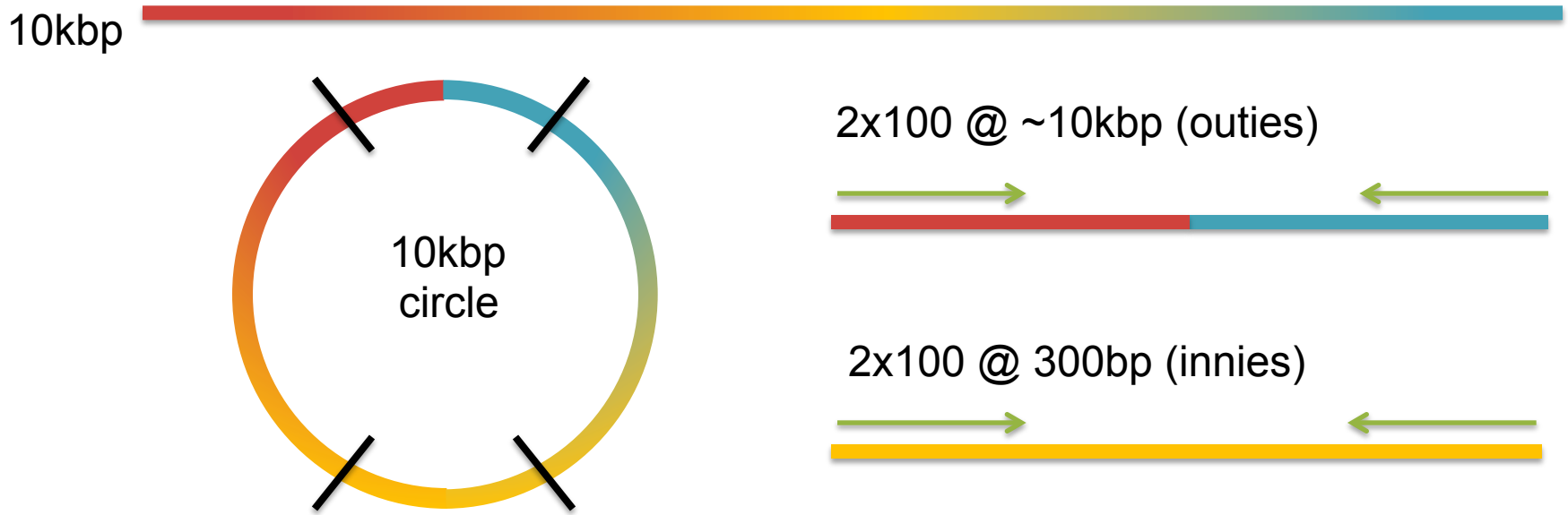
Paired-end and Mate-pairs

Paired-end sequencing

- Read one end of the molecule, flip, and read the other end
- Generate pair of reads separated by up to 500bp with inward orientation

Mate-pair sequencing

- Circularize long molecules (1-10kbp), shear into fragments, & sequence
- Mate failures create short paired-end reads



Short Read Mapping

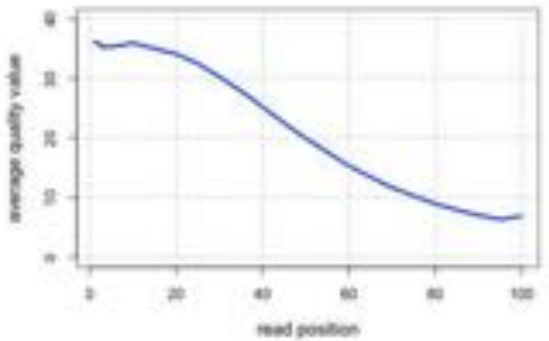
- Given a reference and many subject reads, report one or more "good" end-to-end alignments per alignable read
  - Fundamental computation to genotyping and many assays
    - RNA-seq
    - ChIP-seq
    - Methyl-seq
    - Dnase-seq
    - FAIRE-seq
    - Hi-C-seq

- Desperate need for scalable solutions
  - Single human requires >1,000 CPU hours / genome
  - **1000 hours * 1000 genomes = 1M CPU hours / project**

Illumina Quality

QV	p_{error}
40	1/10000
30	1/1000
20	1/100
10	1/10

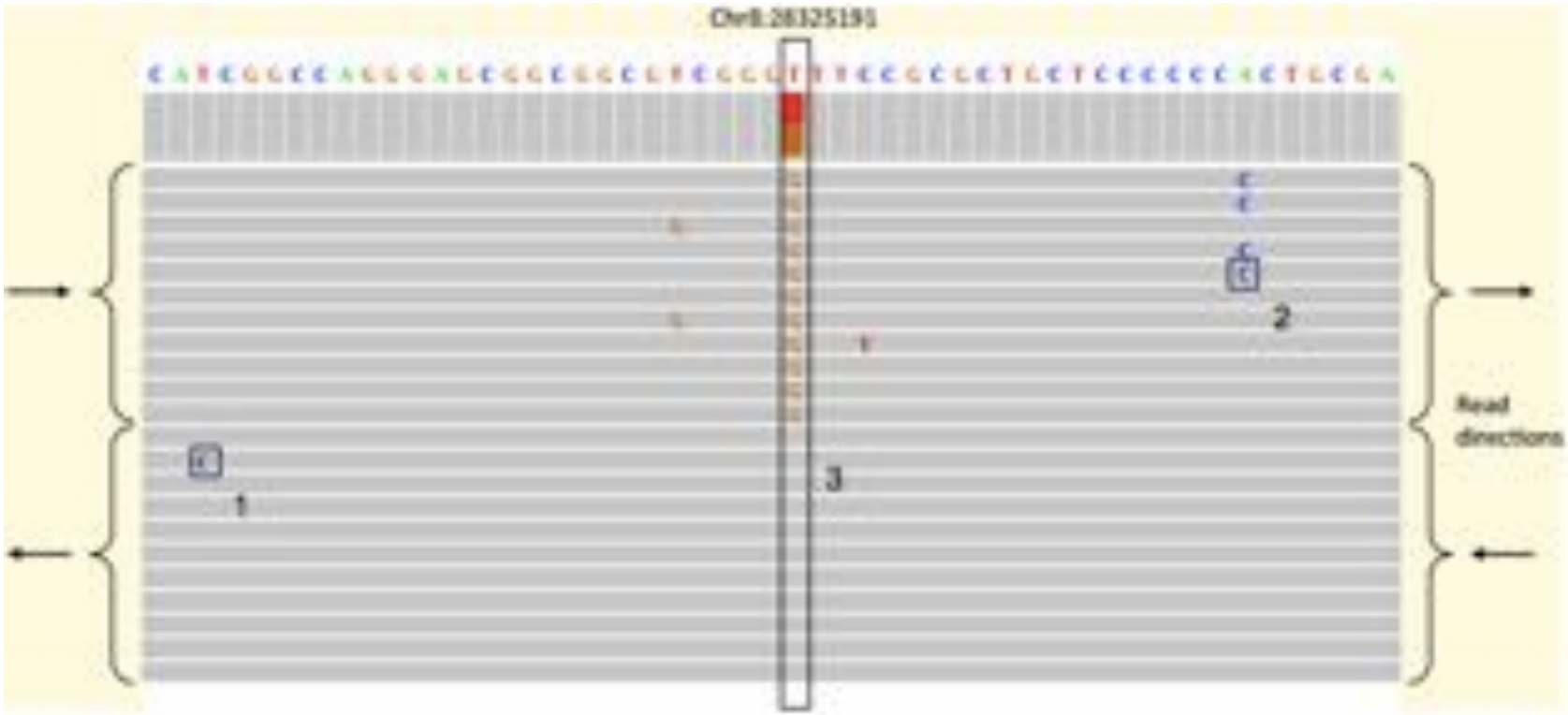
$$Q_{\text{sanger}} = -10 \log_{10} p$$




```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS..................................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX..............
  ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII..................
  ...............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ..................
  LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL..............................................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |        |                              |                     |
 33                        59   64       73                            104                   126

 S - Sanger        Phred+33,  raw reads typically (0, 40)
 X - Solexa        Solexa+64, raw reads typically (-5, 40)
 I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
 J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
     with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
     (Note: See discussion above).
 L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

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

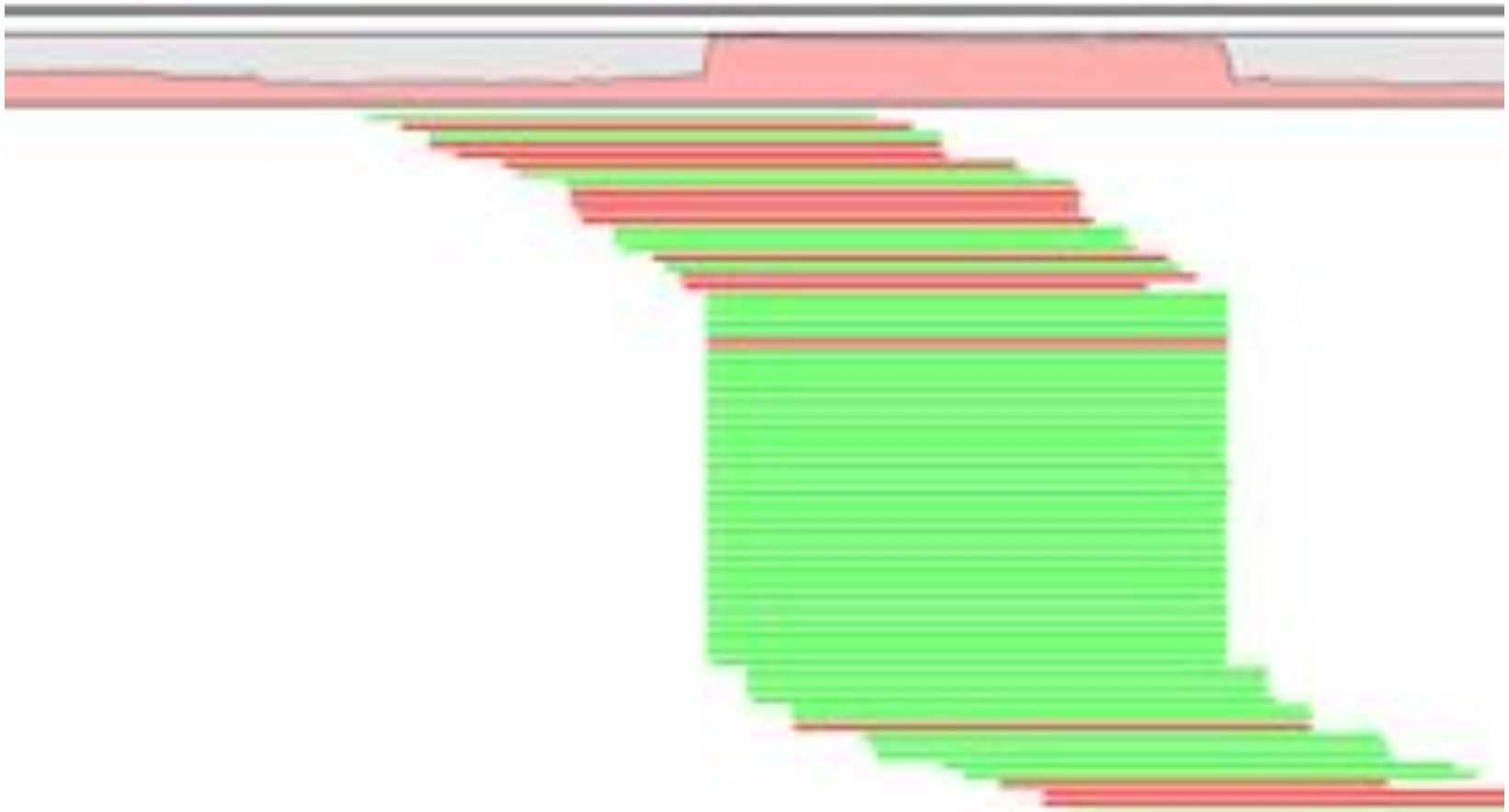
Beware of (Systematic) Errors



Identification and correction of systematic error in high-throughput sequence data
Meacham et al. (2011) *BMC Bioinformatics.* 12:451

A closer look at RNA editing.
Lior Pachter (2012) *Nature Biotechnology.* 30:246-247

Beware of Duplicate Reads



The Sequence alignment/map (SAM) format and SAMtools.
Li et al. (2009) *Bioinformatics.* 25:2078-9

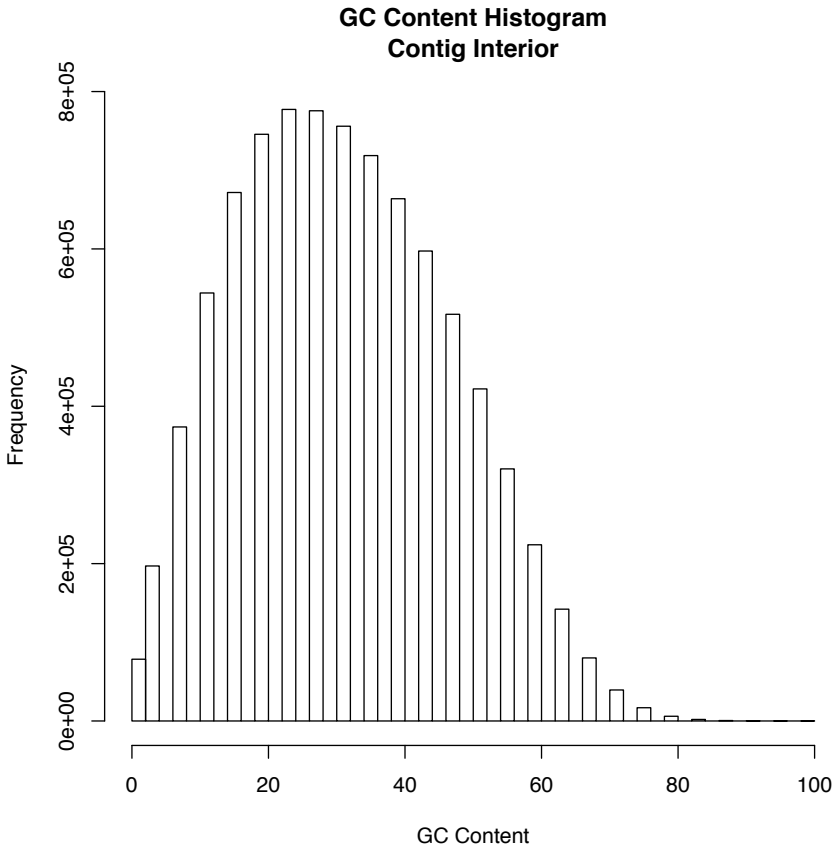
Picard: http://picard.sourceforge.net

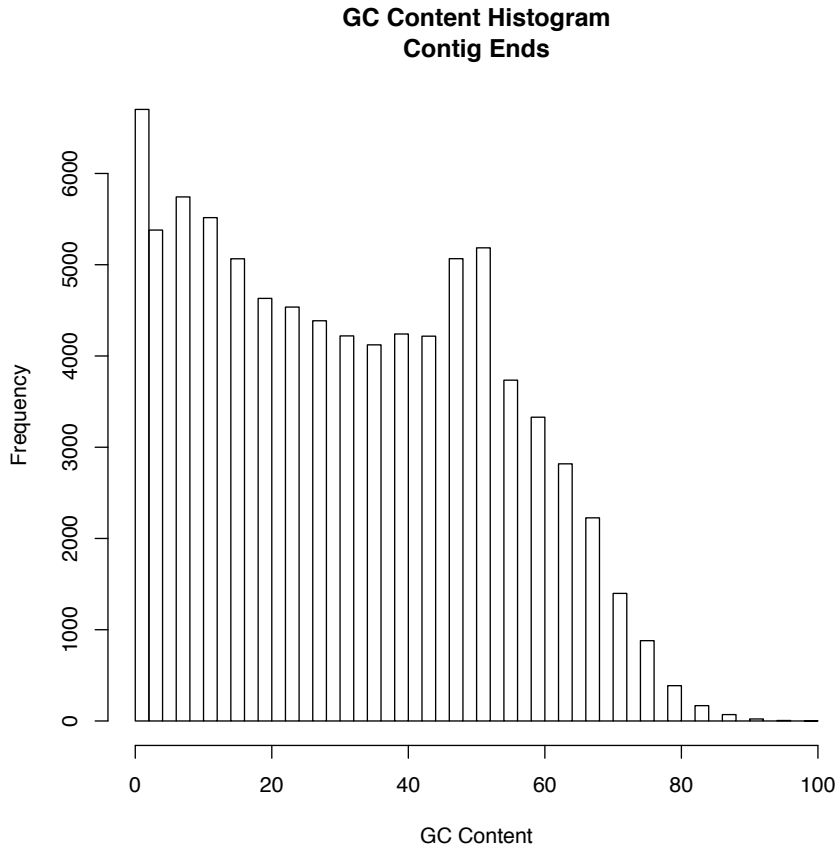
Beware of GC Biases

Apis dorsata (236Mbp)

2x500bp, 2x1.2kbp, 2x3kb, 2x5kbp

714kbp Scaffold N50, 8.3kbp Contig N50





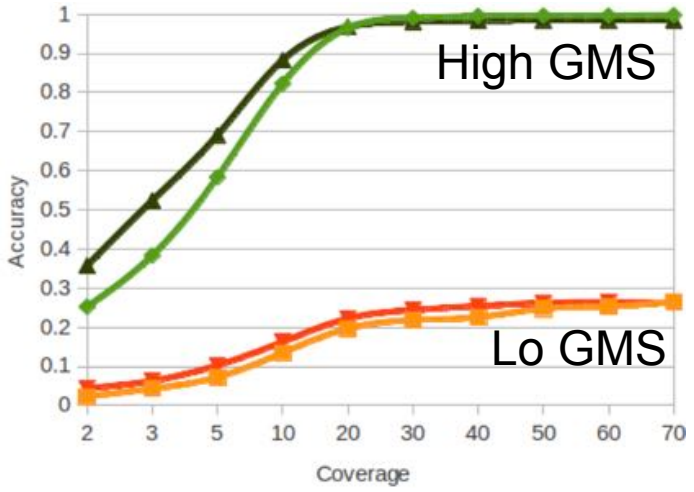
Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries.
Aird et al. (2011) *Genome Biology.* 12:R18.

Beware of Mapping Errors

- Short read mapping is a essential for identifying mutations in the genome
  - Not every base of the genome can mapped equally well, especially because of repeats

- Introduced a new probabilistic metric - the Genome Mappability Score - that quantifies how reliably reads can be mapped to every position in the genome
  - We have little power to measure 11-13% of the human genome, including of known clinically relevant variations
  - Errors in variation discovery are dominated by errors in low GMS regions

Species (build)	size	paired/single	whole (%)	transcription (%)
yeast (sc2)	12 Mbp	paired	94.85	95.04
		single	94.25	94.62
fly (dm3)	130 Mbp	paired	90.52	96.14
		single	89.70	95.94
mouse (mm9)	2.7 Gbp	paired	89.39	96.03
		single	87.47	94.75
human (hg19)	3.0 Gbp	paired	89.02	97.40
		single	87.79	96.38

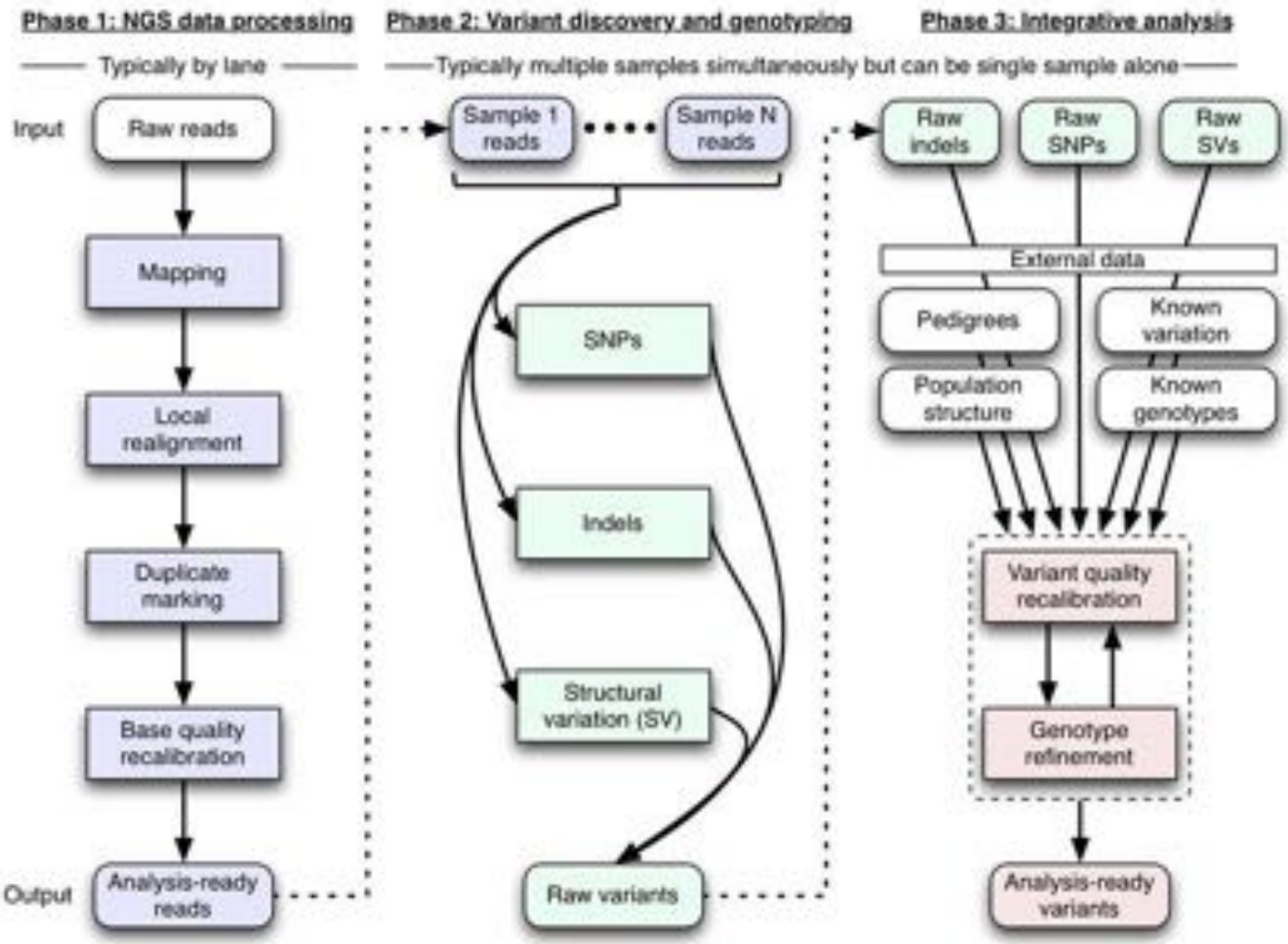


Genomic Dark Matter: The reliability of short read mapping illustrated by the GMS.
Lee, H., Schatz, M.C. (2012) *Bioinformatics. doi: 10.1093/bioinformatics/bts330*

Beware of Indels

Indel Cleaning and Calling
http://www.broadinstitute.org/files/shared/mpg/nextgen2010/nextgen_sivachenko.pdf

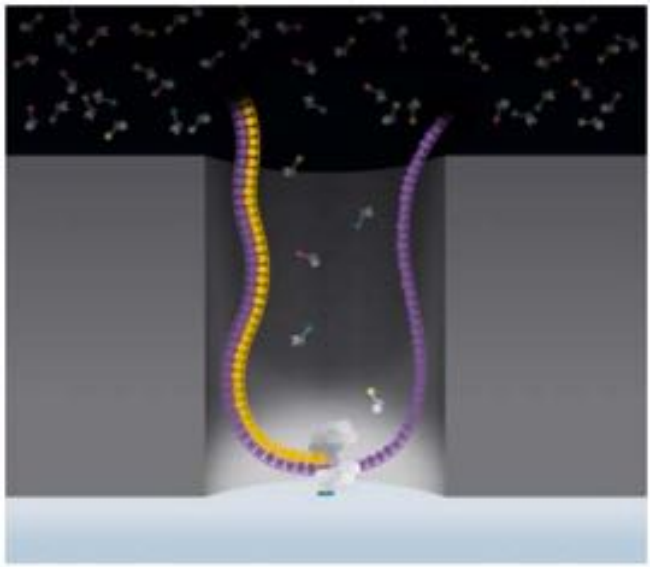
Recommendation: GATK

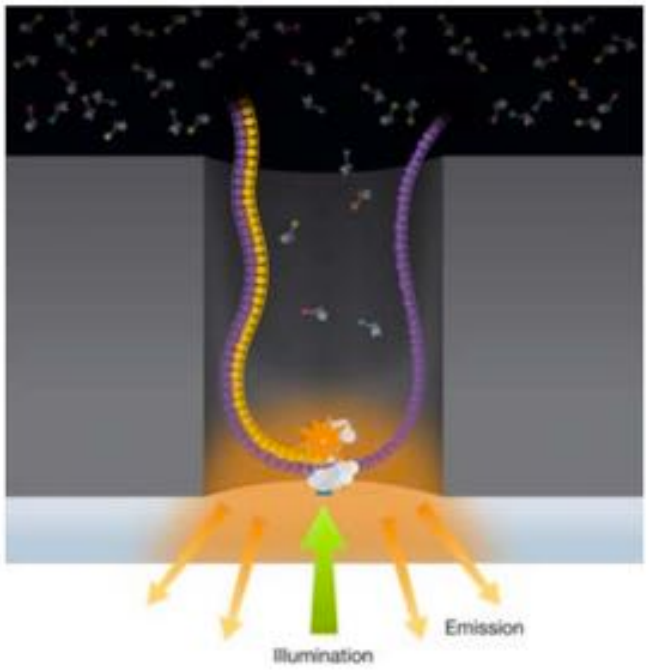


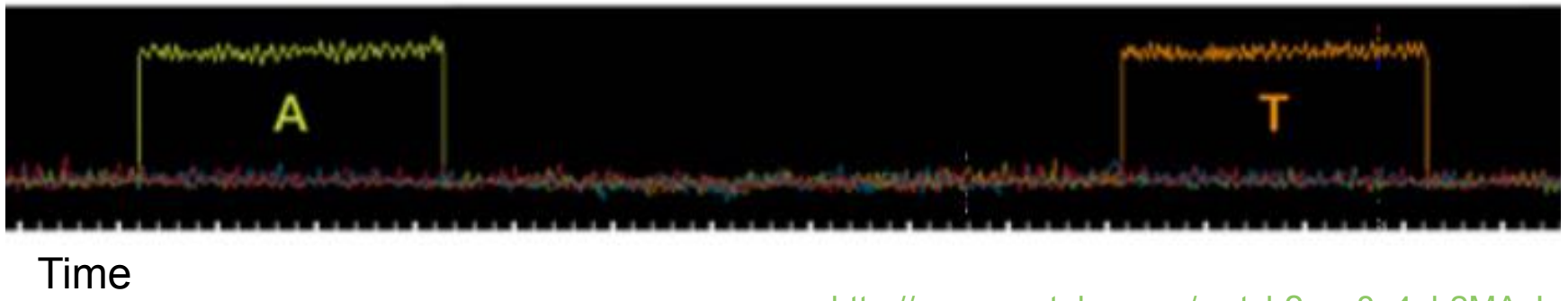
The Genome Analysis Toolkit:
A MapReduce framework for analyzing next-generation DNA sequencing data.
McKenna et al. (2010) *Genome Research.* (9):1297-303.

PacBio: SMRT Sequencing

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





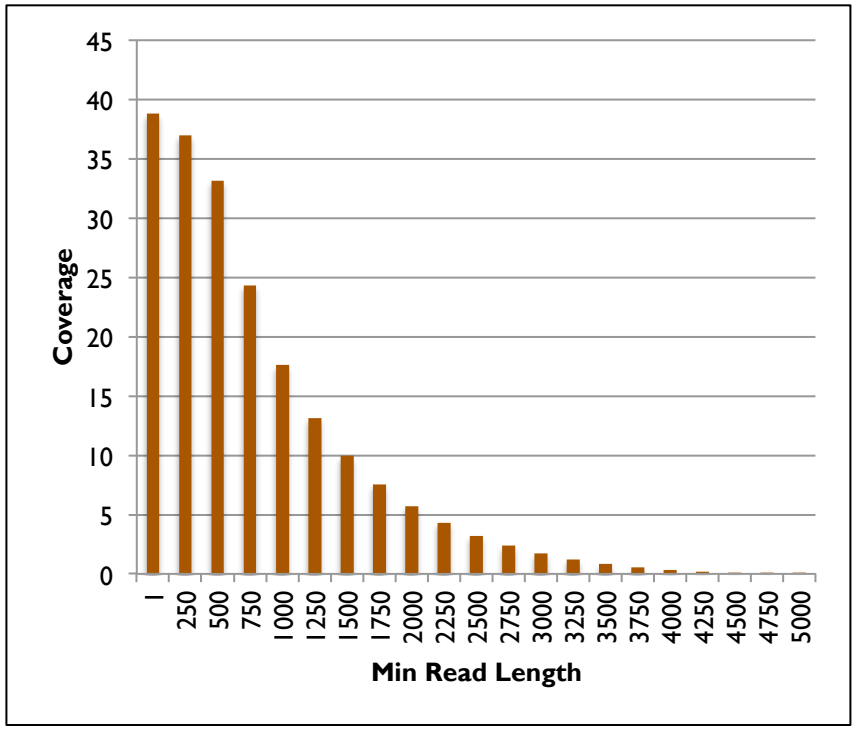


http://www.youtube.com/watch?v=v8p4ph2MAvI

SMRT Sequencing Data

Yeast
(12 Mbp genome)

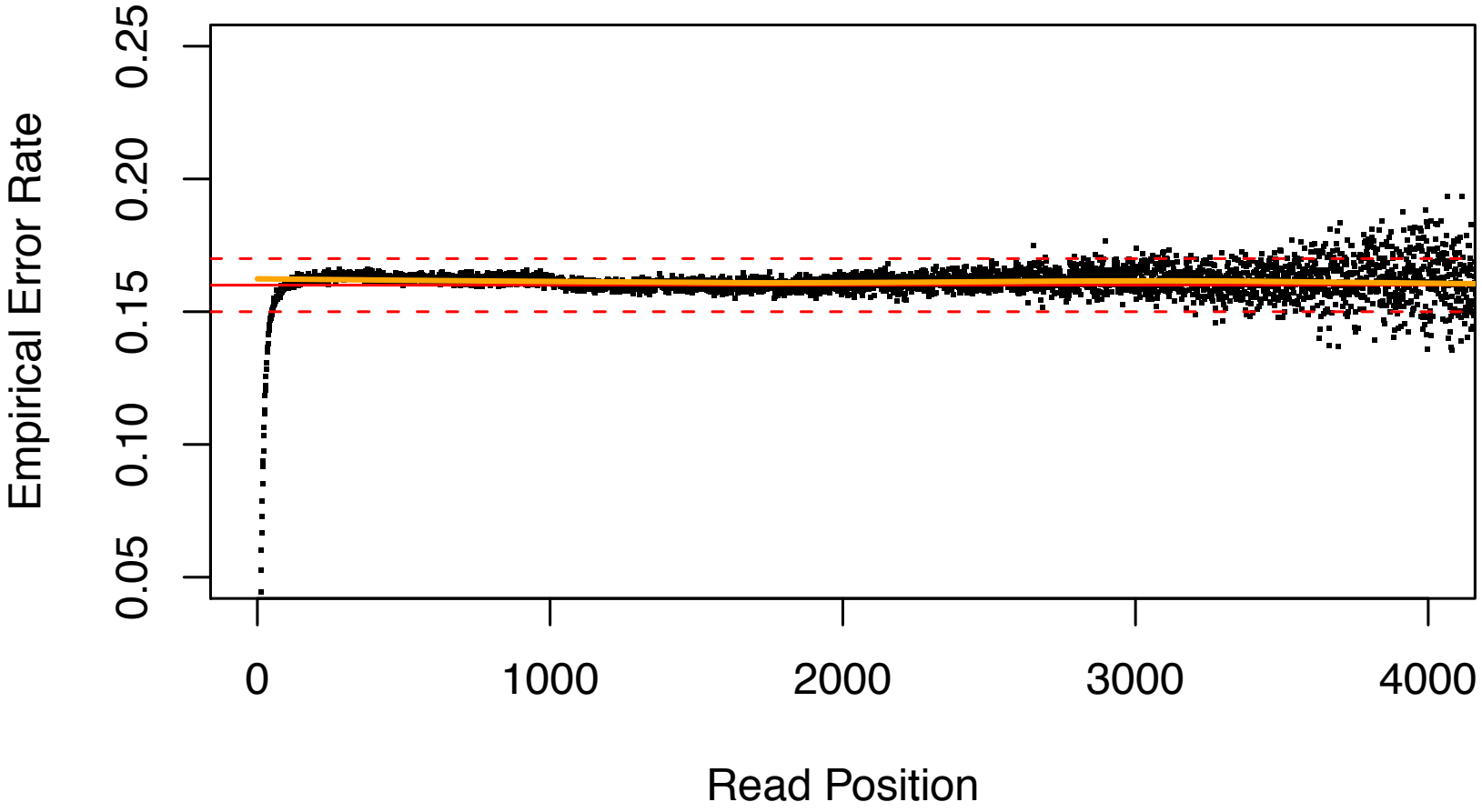
65 SMRT cells
734,151 reads after filtering
Mean: 642.3 +/- 587.3
Median: 553 Max: 8,495




```
TTGTAAGCAGTTGAAAACTATGTGTGGATTTAGAATAAAGAACATGAAAG
|||||||||||||||||||||||||| ||||||| |||||||||||| |||
TTGTAAGCAGTTGAAAACTATGTGT-GATTTAG-ATAAAGAACATGGAAG

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

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

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

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

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

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

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

Sample of 100k reads aligned with BLASR requiring >100bp alignment
Average overall accuracy: 83.7%, 11.5% insertions, 3.4% deletions, 1.4% mismatch

Read Quality

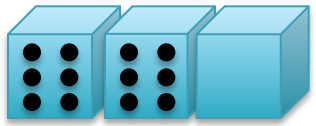
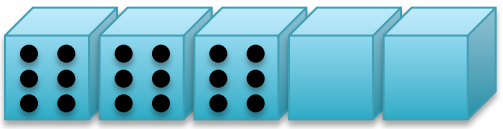


Consistent quality across the entire read

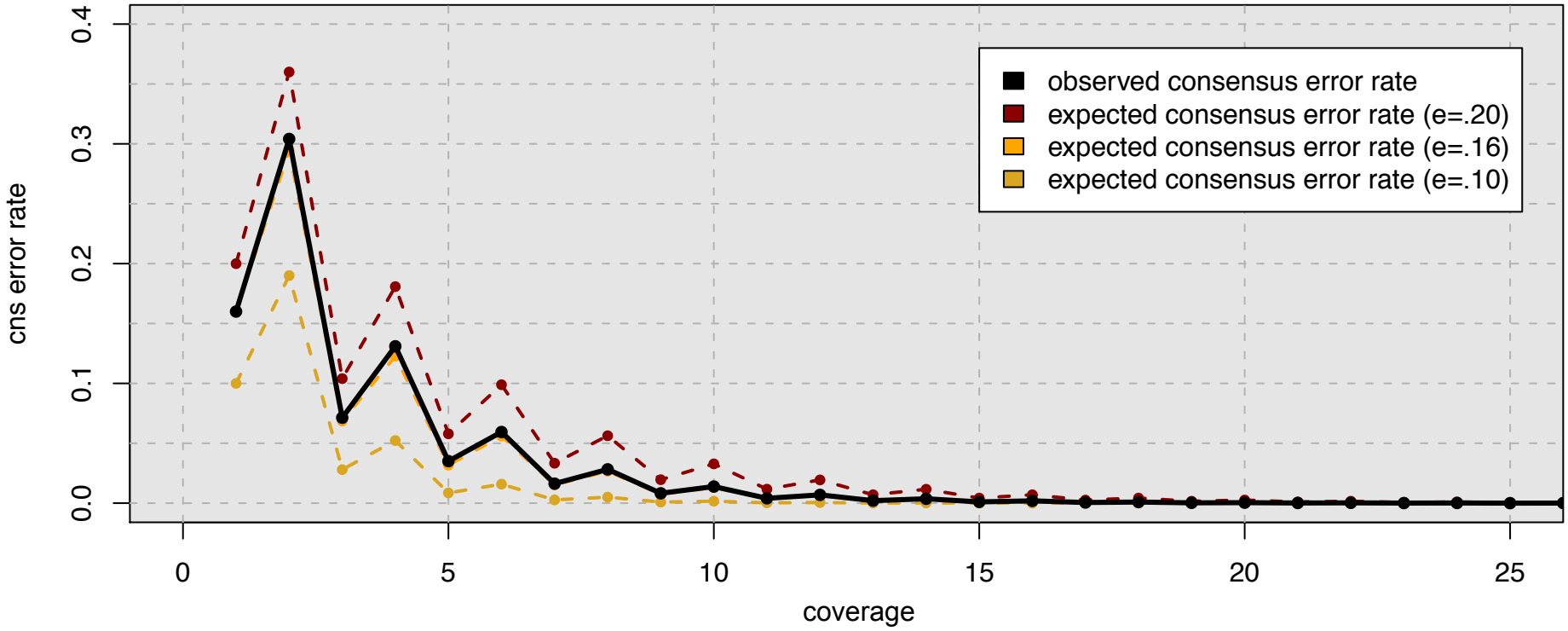
- Uniform error rate, no apparent biases for GC/motifs
- Sampling artifacts at beginning and ends of alignments

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)*	$\sum_{i=\lceil n/2 \rceil}^{n} P(i \ of \ n) = \sum_{i=\lceil n/2 \rceil}^{n} \binom{n}{i} (p)^{i} (1-p)^{n-i}$	

Consensus Accuracy and Coverage



Coverage can overcome random errors

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

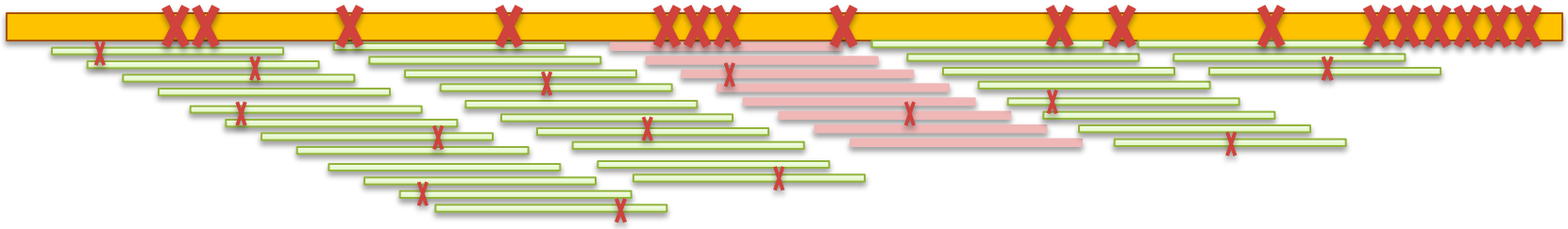
$$CNS\,Error = \sum_{i=\lceil c/2 \rceil}^{c} \binom{c}{i} (e)^{i} (1-e)^{n-i}$$

PacBio Error Correction

http://wgs-assembler.sf.net

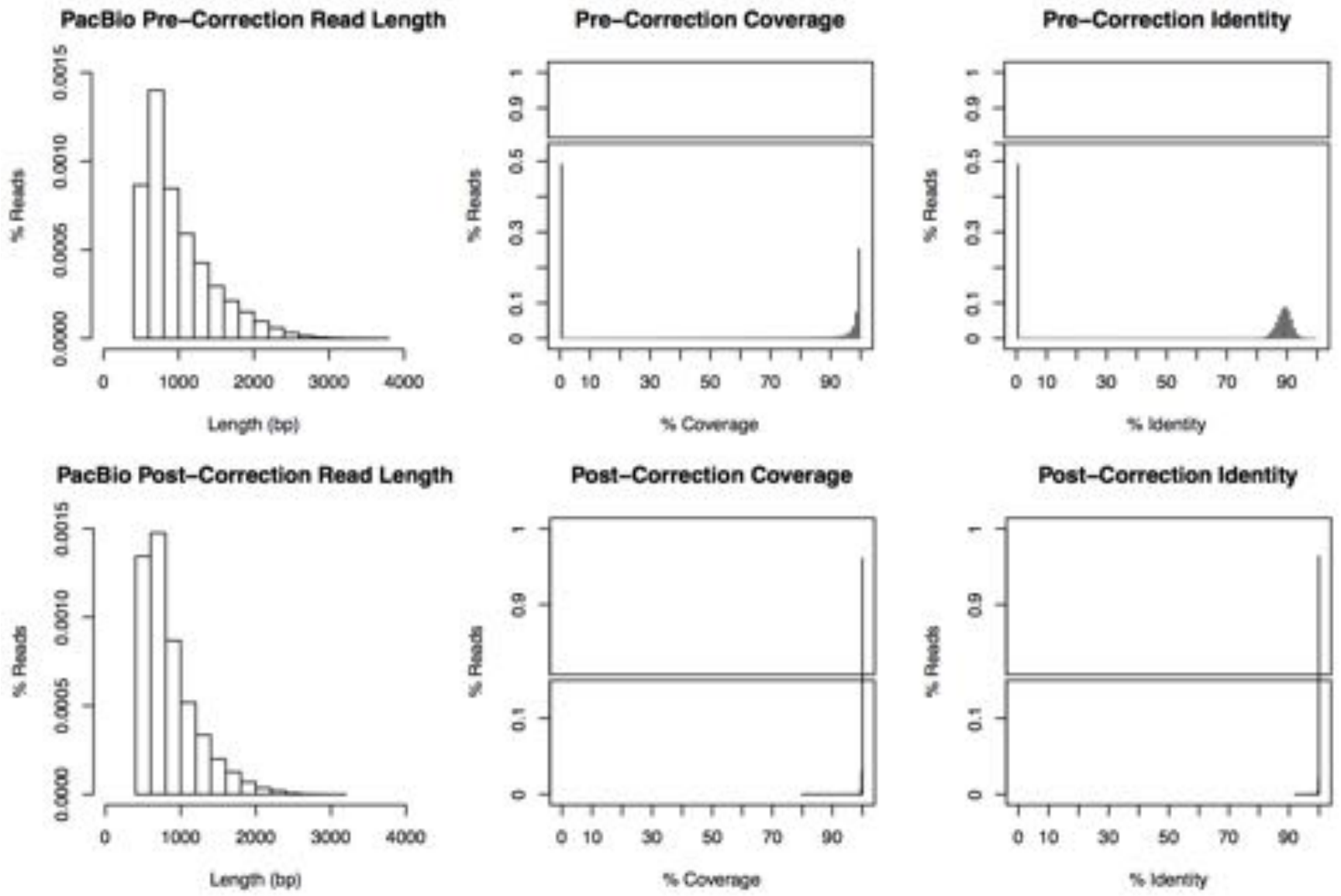
1. Correction Pipeline
    1. Map short reads (SR) to long reads (LR)
    2. Trim LRs at coverage gaps
    3. Compute consensus for each LR

2. Error corrected reads can be easily assembled, aligned



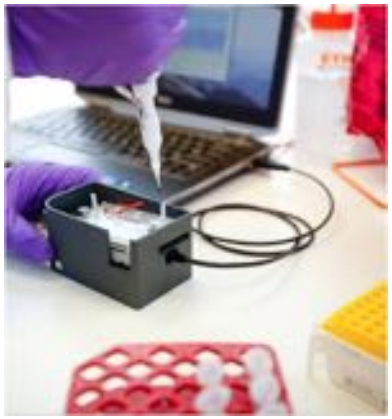
Hybrid error correction and de novo assembly of single-molecule sequencing reads.
Koren, S, Schatz, MC, Walenz, BP, Martin, J, Howard, J, Ganapathy, G, Wang, Z, Rasko, DA, McCombie, WR, Jarvis, ED, Phillippy, AM. (2012) *Nature Biotechnology.* 30: 693–700.

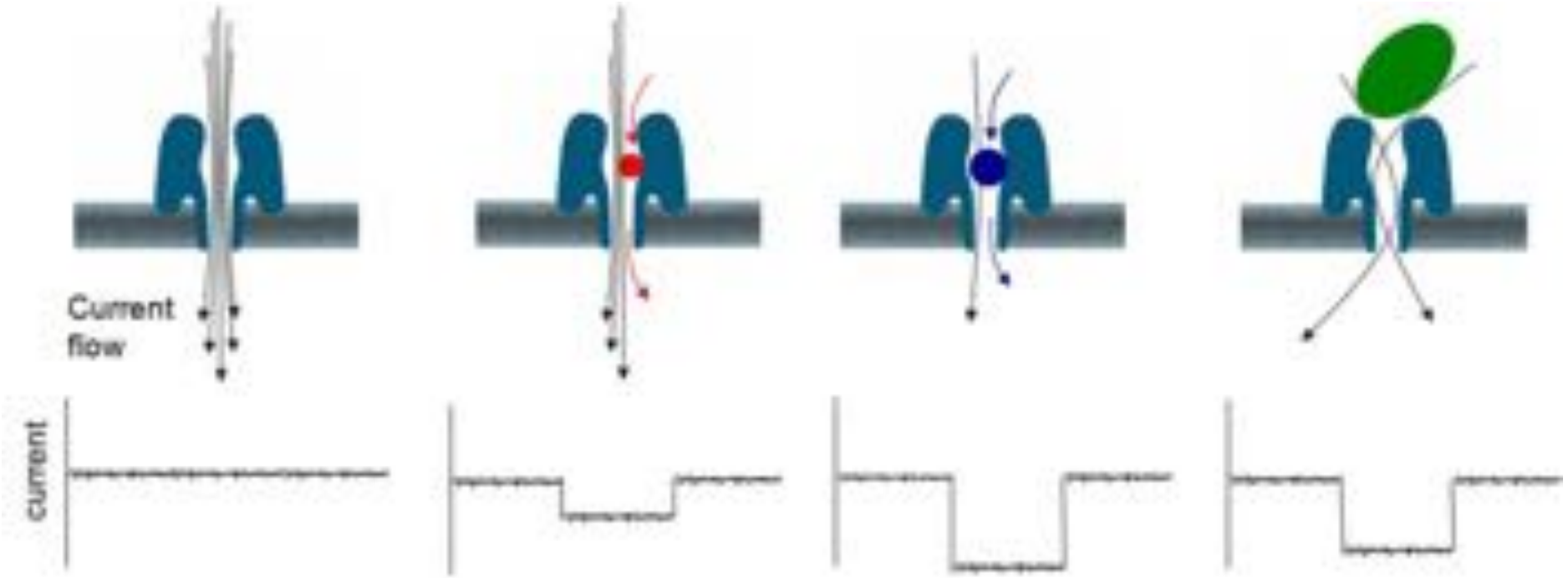
Error Correction Results



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

Oxford Nanopore: Nanosensing





http://www.nanoporetech.com/news/movies#movie-24-nanopore-dna-sequencing

Oxford Nanopore: Data Quality

As of AGBT in February

- Sequencing 40kbp lambda phage in one read
- Accuracy around 90-96%
- Costs, throughput, distribution on read length unknown

Thank You!

http://schatzlab.cshl.edu
@mike_schatz