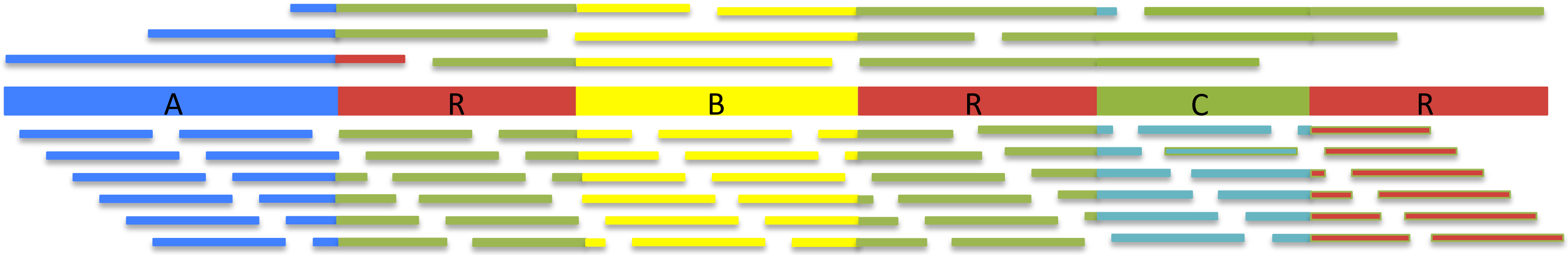
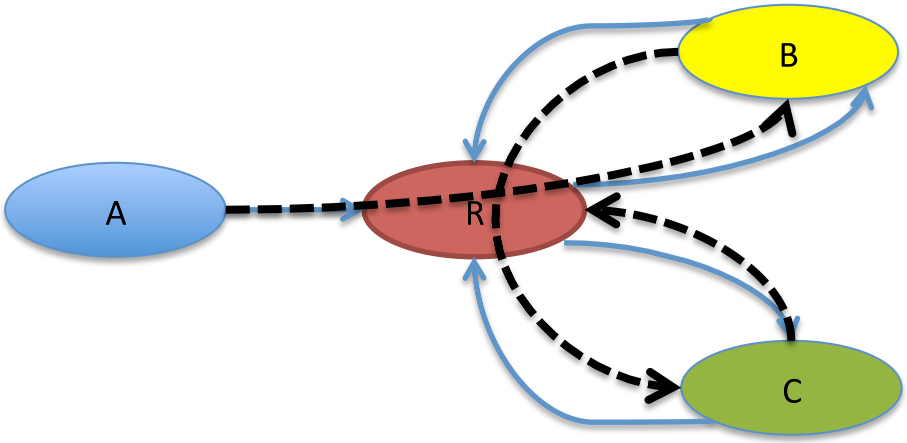
CSH Cold Spring Harbor Laboratory

Error Correction and Assembly of Single Molecule Sequencing Data

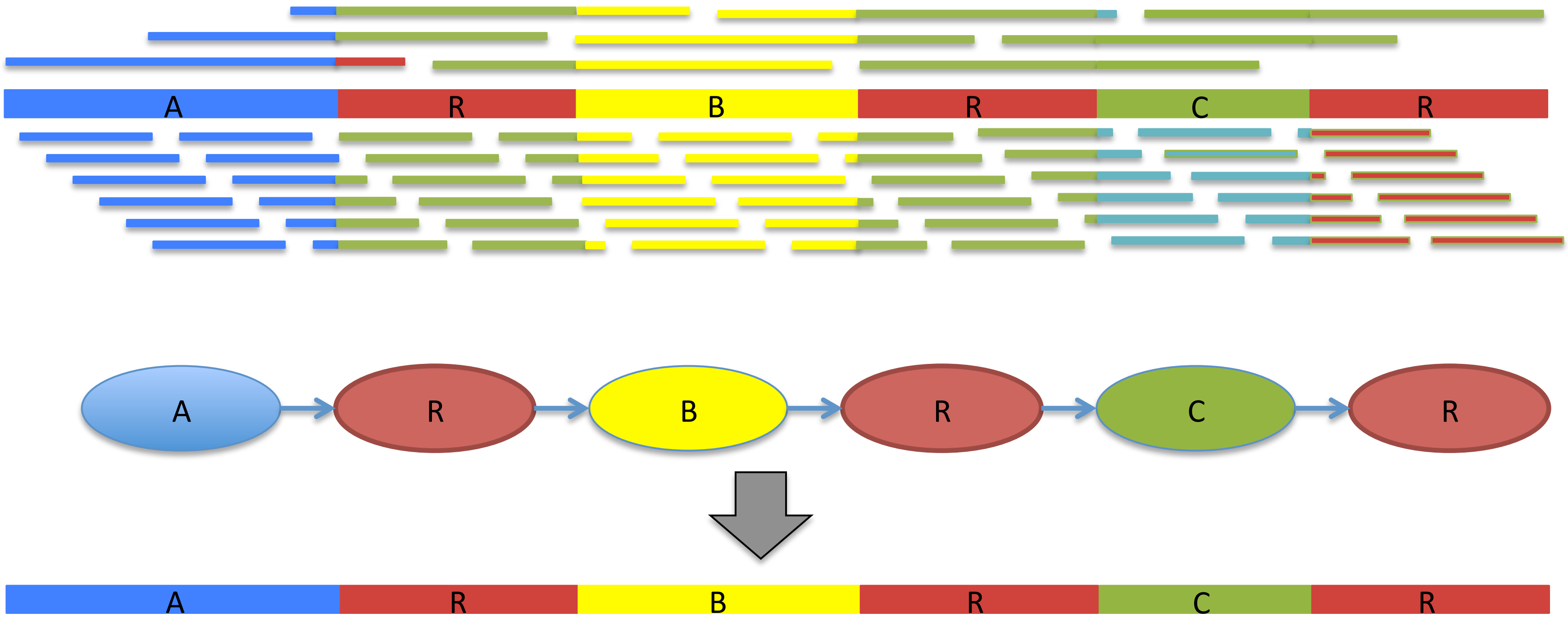
James Gurtowski

Assembly Complexity





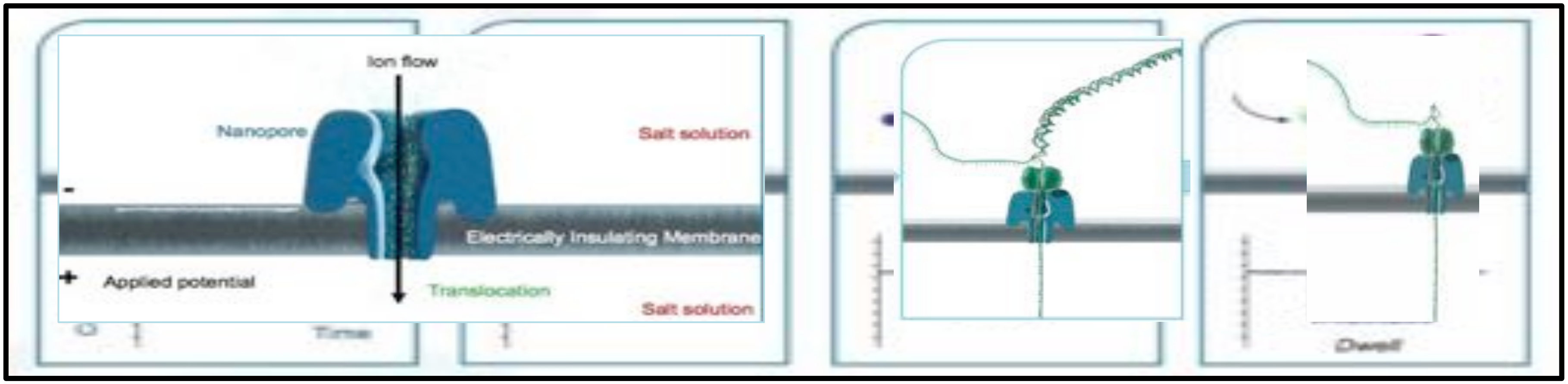
Assembly Complexity



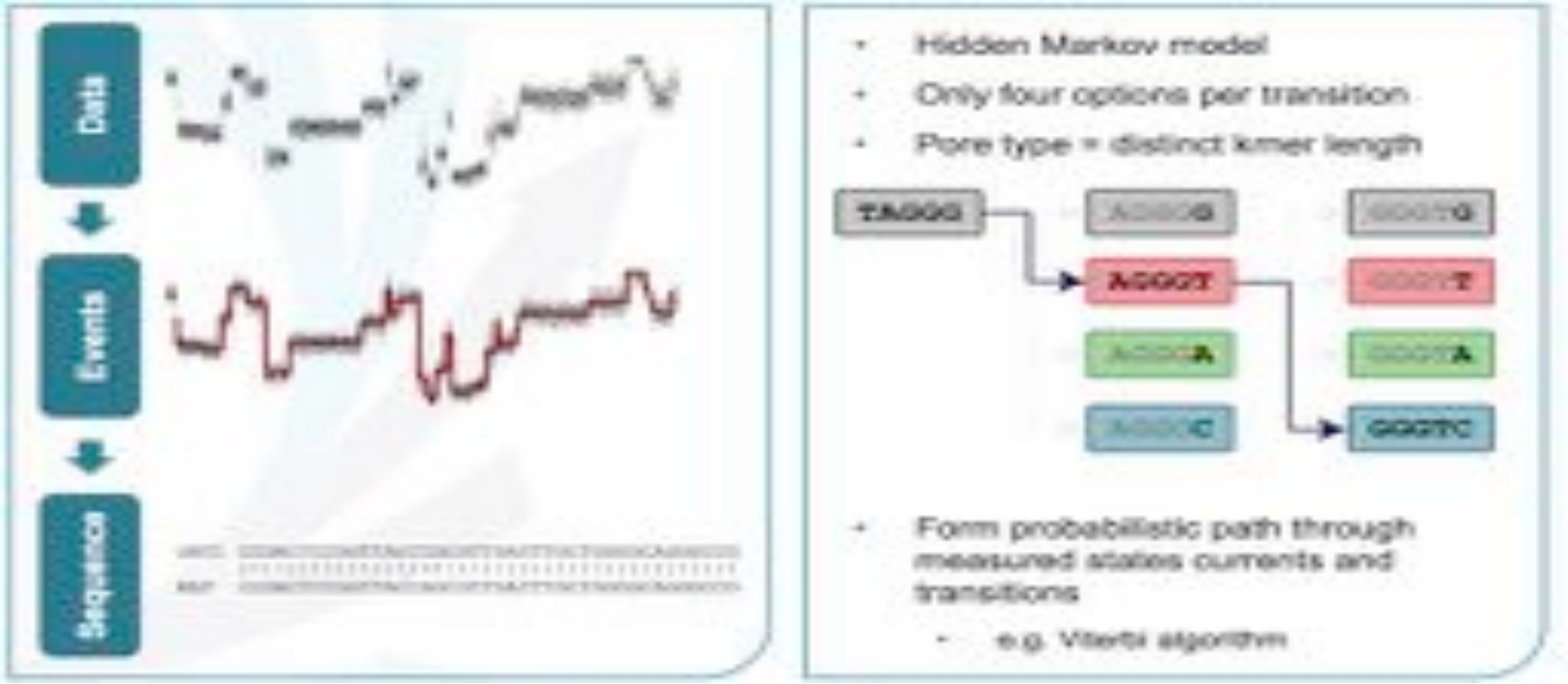
The advantages of SMRT sequencing
Roberts, RJ, Carneiro, MO, Schatz, MC (2013) *Genome Biology.* 14:405

Oxford Nanopore MinION

- Thumb drive sized sequencer powered over USB
- Senses DNA by measuring changes to ion flow
- Reads both DNA Strands (2D)



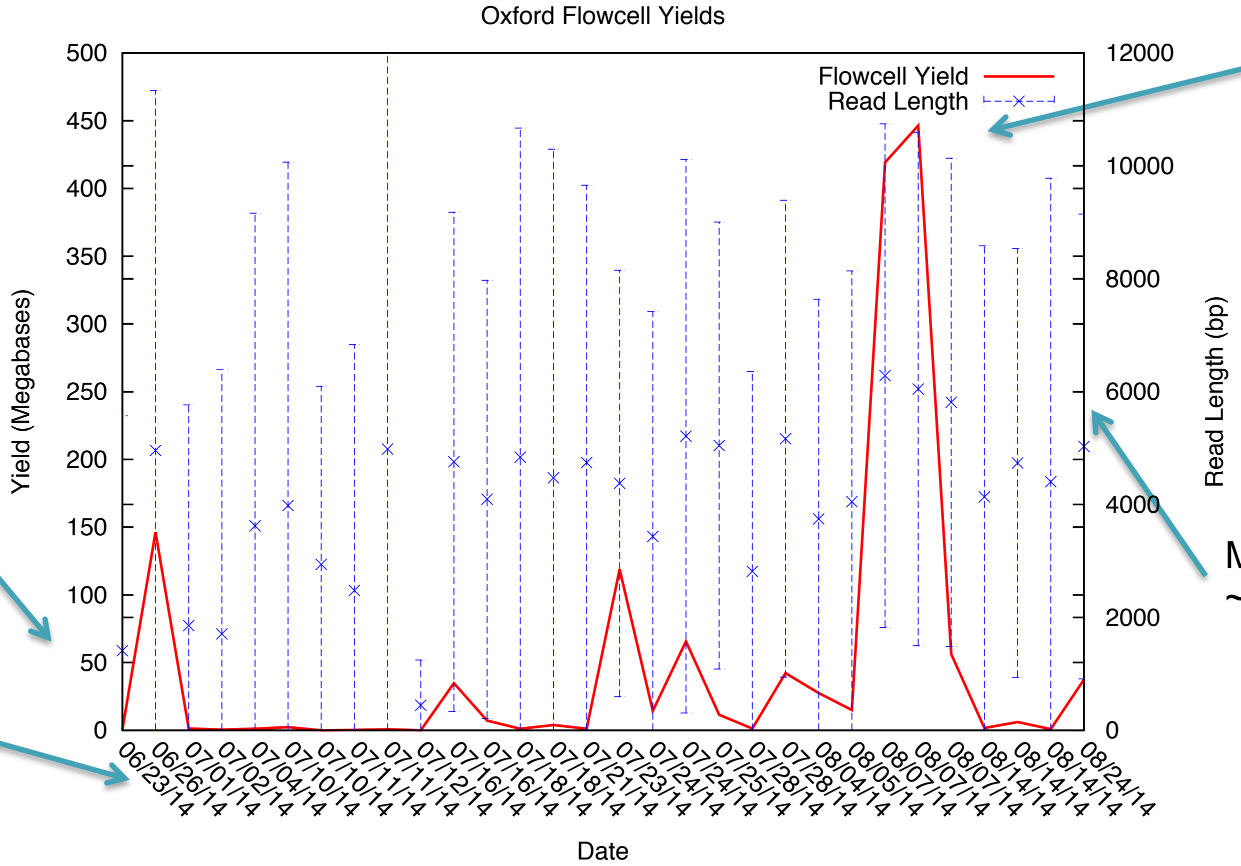
Nanopore Basecalling



- Hidden Markov model
- Only four options per transition
- Pore type = distinct kmer length
- Form probabilistic path through measured states currents and transitions
  - e.g. Viterbi algorithm

Basecalling currently performed at Amazon with frequent updates to algorithm

Our Data - Yeast W303



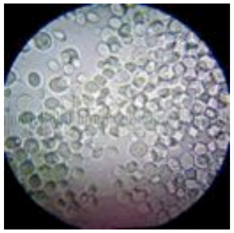
Best: 446Mb

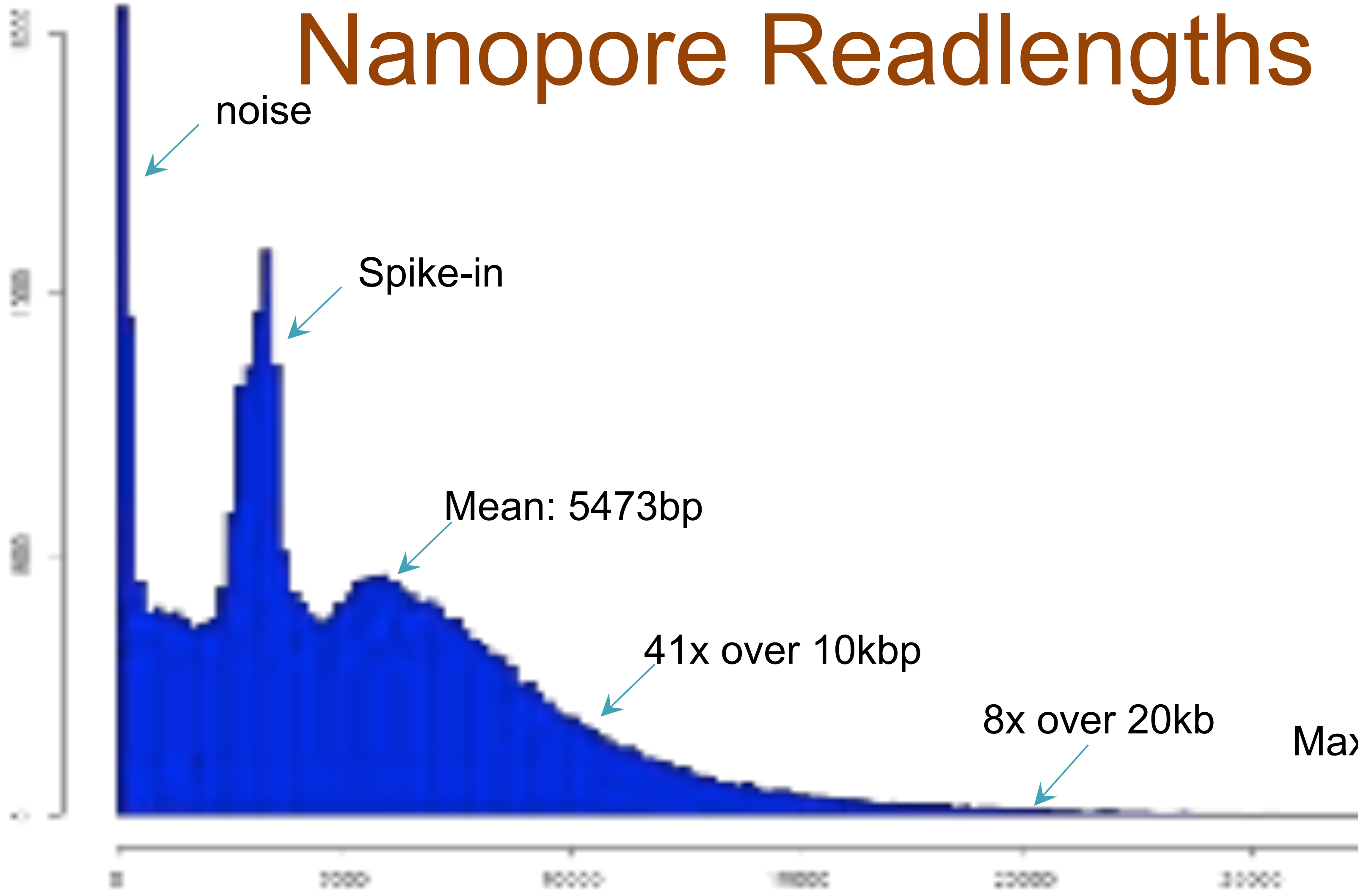
Mean: 50Mb

Mean Read Length:
~6kb (10kb shear)

30 Flowcells

Nanopore Readlengths





noise

Spike-in

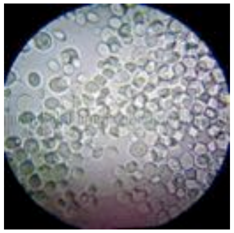
Mean: 5473bp

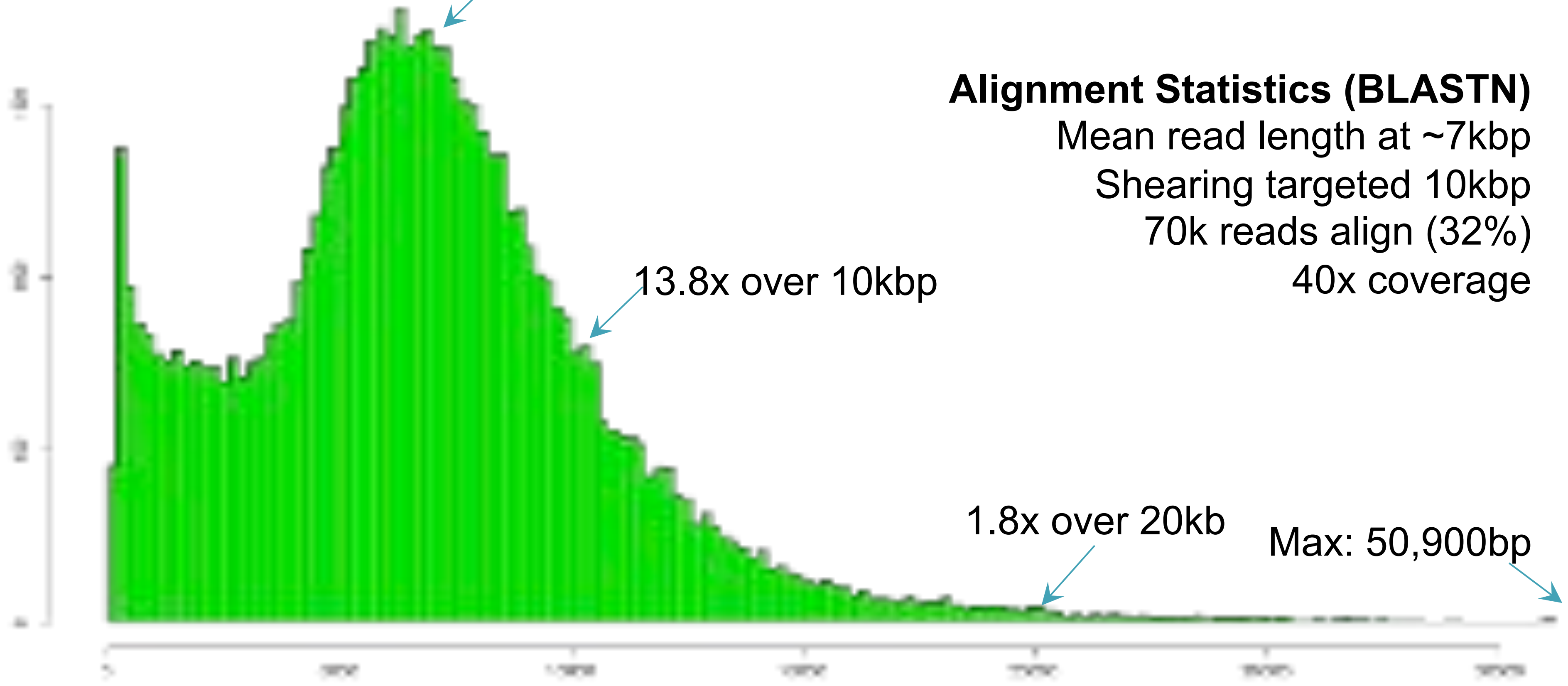
41x over 10kbp

8x over 20kb

Max: 146,992bp

Nanopore Alignments





Mean: 6903bp

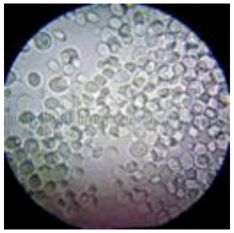
13.8x over 10kbp

1.8x over 20kb

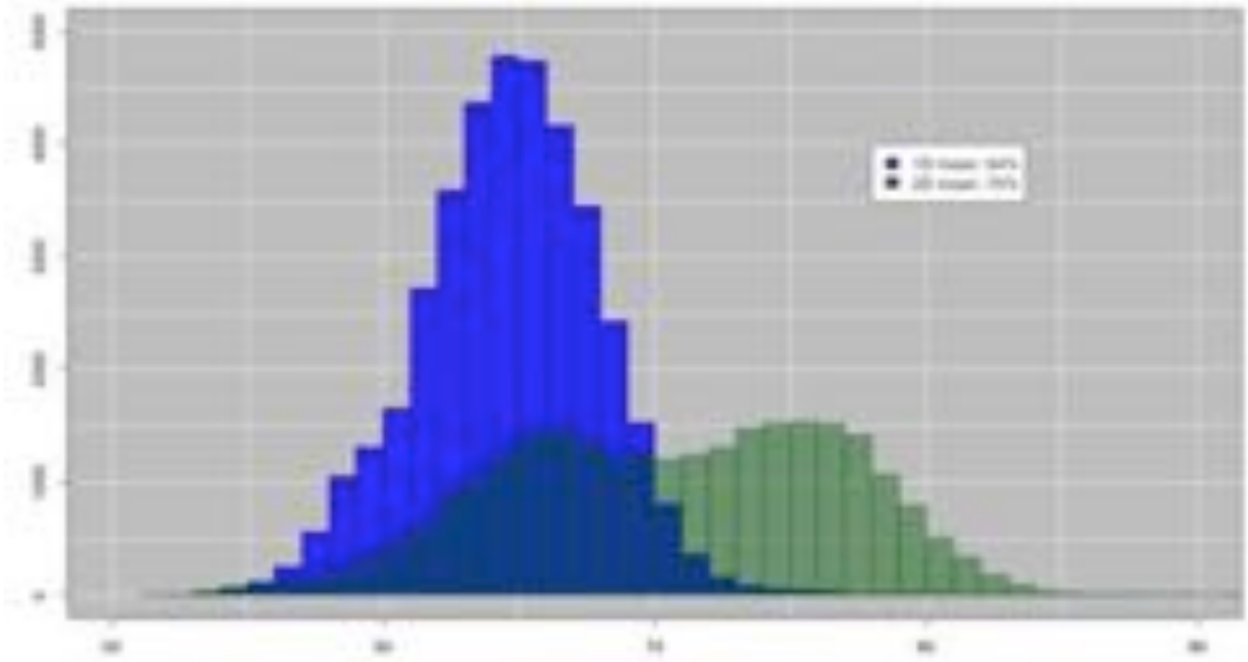
Max: 50,900bp

Alignment Statistics (BLASTN)
Mean read length at ~7kbp
Shearing targeted 10kbp
70k reads align (32%)
40x coverage

Nanopore Accuracy



Alignment Quality (BLASTN)
Of reads that align, average ~64% identity
“2D base-calling” improves to ~70% identity

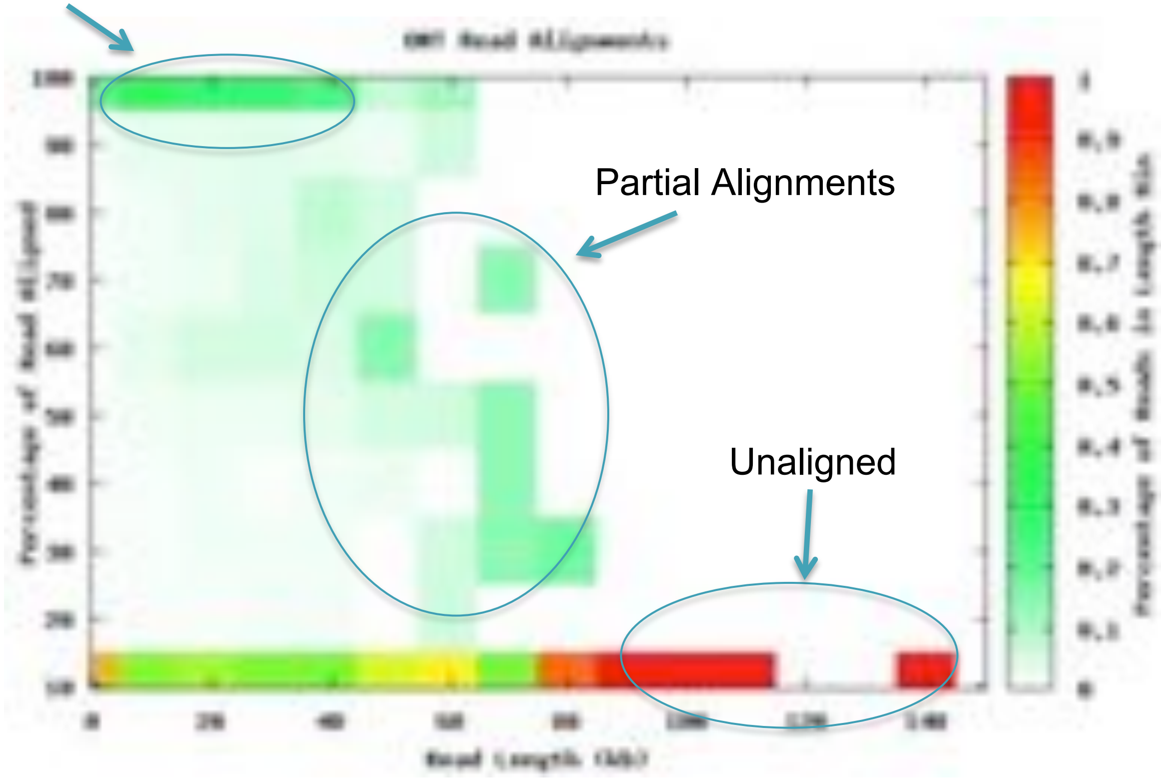


57% Mismatches
32% Deletions
11% Insertions

Nanopore Alignment Summary

32% of the data map using BLASTN

Full Length Alignments



Partial Alignments

Unaligned

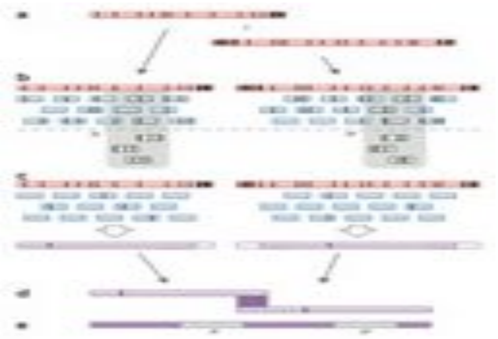
Long Read Correction Algorithms

PBJelly

Gap Filling and Assembly Upgrade

English *et al* (2012)
PLOS One. 7(11): e47768

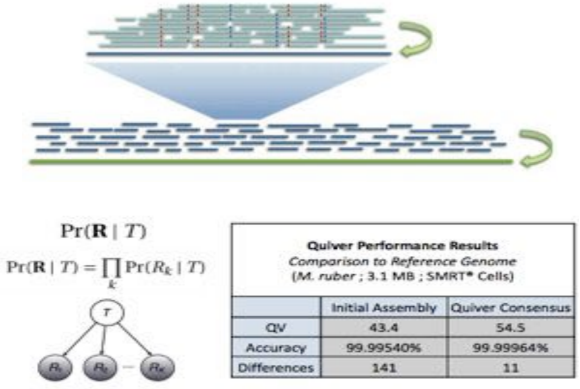
PacBioToCA & ECTools



Hybrid Error Correction

Koren, Schatz, *et al* (2012)
Nature Biotechnology. 30:693–700

HGAP & Quiver



$$\Pr(\mathbf{R} \mid T)$$

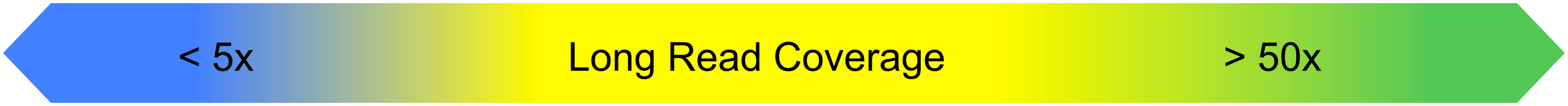
$$\Pr(\mathbf{R} \mid T) = \prod_k \Pr(R_k \mid T)$$

Quiver Performance Results
Comparison to Reference Genome
(*M. ruber* ; 3.1 MB ; SMRT® Cells)

	Initial Assembly	Quiver Consensus
QV	43.4	54.5
Accuracy	99.99540%	99.99964%
Differences	141	11

LR-only Correction & Polishing

Chin *et al* (2013)
Nature Methods. 10:563–569

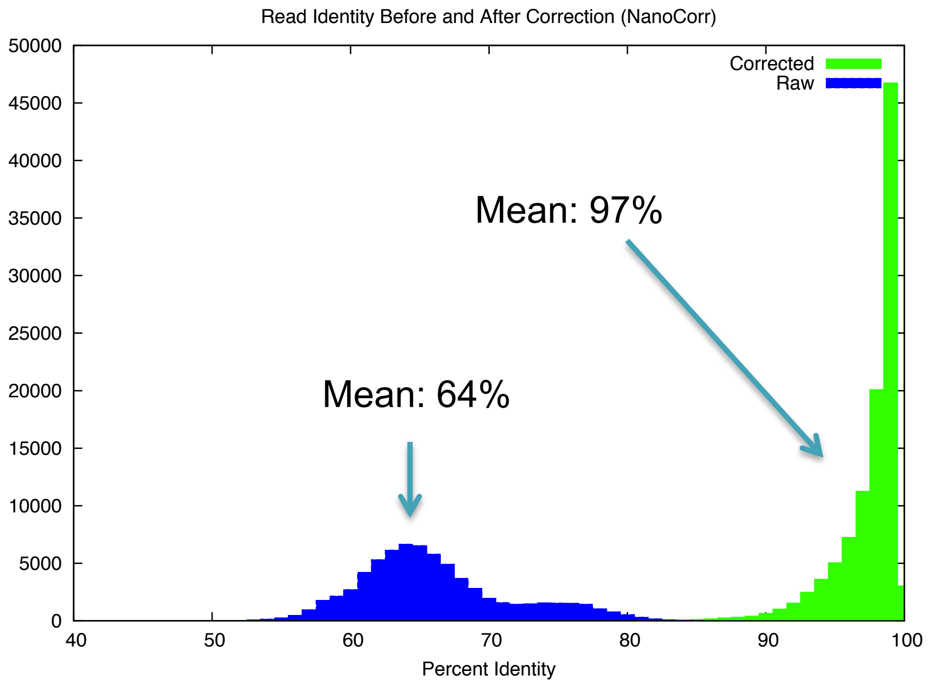


NanoCorr: Nanopore-Illumina Hybrid Error Correction

1. BLAST Miseq reads to all raw Oxford Nanopore reads

2. Select non-repetitive alignments
   - First pass scans to remove “contained” alignments
   - Second pass uses Dynamic Programming (LIS) to select set of high-identity alignments with minimal overlaps

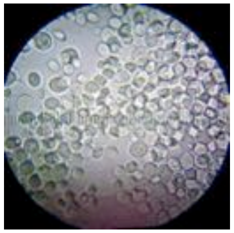
3. Compute consensus of each Oxford Nanopore read
   - Currently using Pacbio’s pbdagcon


Read Identity Before and After Correction (NanoCorr)

Mean: 97%

Mean: 64%

Long Read Assembly

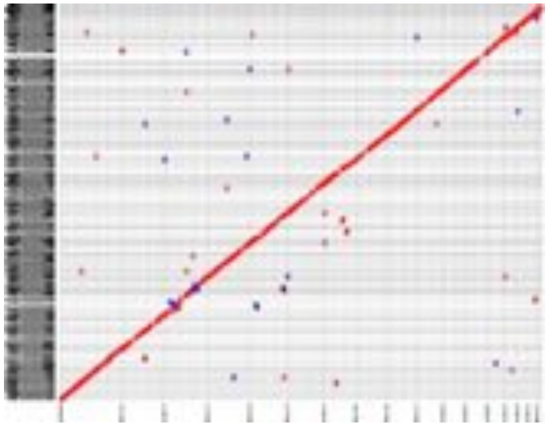


S288C Reference sequence

- 12.1Mbp; 16 chromo + mitochondria; N50: 924kbp

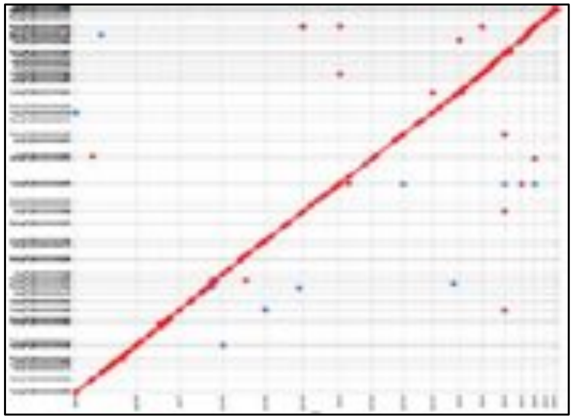
Illumina MiSeq
30x, 300bp PE (Flashed)
Celera Assembler

- 6953 non-redundant contigs
- N50: 59kb >99.9% id



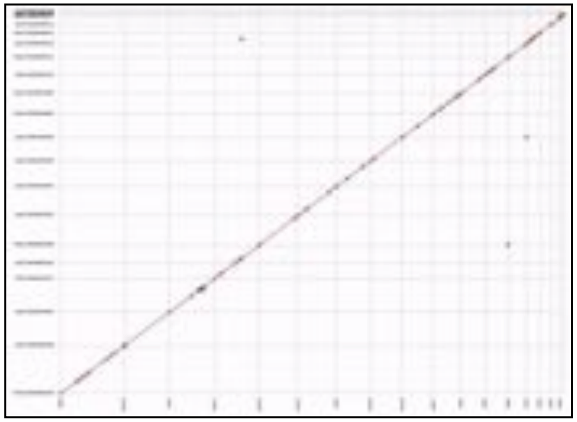
Oxford Nanopore
30x corrected reads > 6kb
NanoCorr + Celera Assembler

- 234 non-redundant contigs
- N50: 362kbp >99.78% id



Pacific Biosciences
25x corrected reads > 10kb
HGAP + Celera Assembler

- 21 non-redundant contigs
- N50: 811kb >99.8% id



Acknowledgements

CSH Cold Spring Harbor Laboratory

Michael Schatz

Dick McCombie

Sara Goodwin

Schatz Lab

Oxford NANOPORE Technologies

PACIFIC BIOSCIENCES