



Current Advances in Sequencing Technology

James Gurtowski
Schatz Lab

Outline

1. Assembly Review

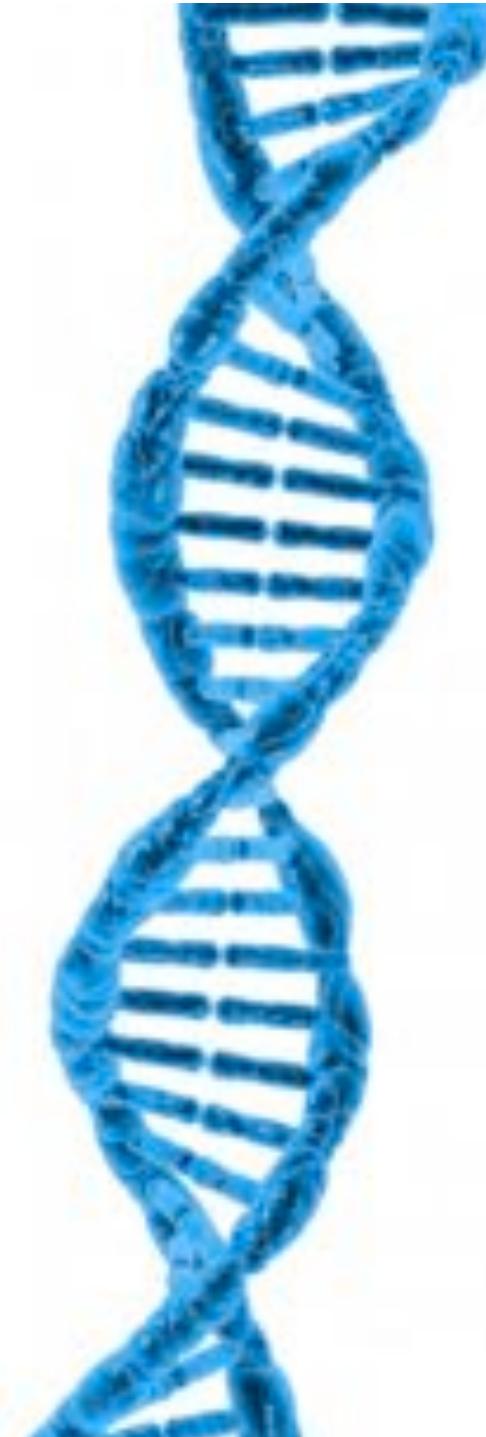
2. Pacbio

- Technology Overview
- Data Characteristics
- Algorithms
- Results – Assemblies

3. Oxford Nanopore

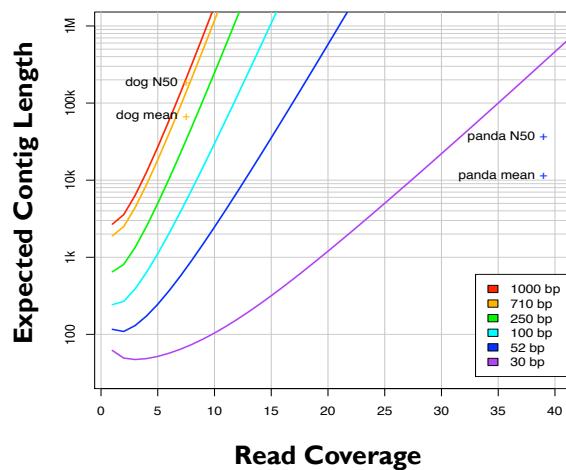
- Technology Overview
- Data Characteristics
- Algorithms
- Results – Assemblies

4. Summary



Ingredients for a good assembly

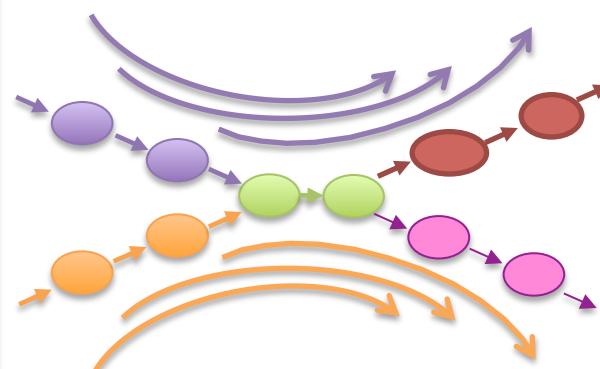
Coverage



High coverage is required

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

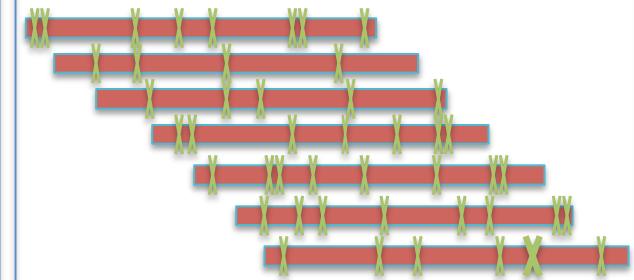
Read Length



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

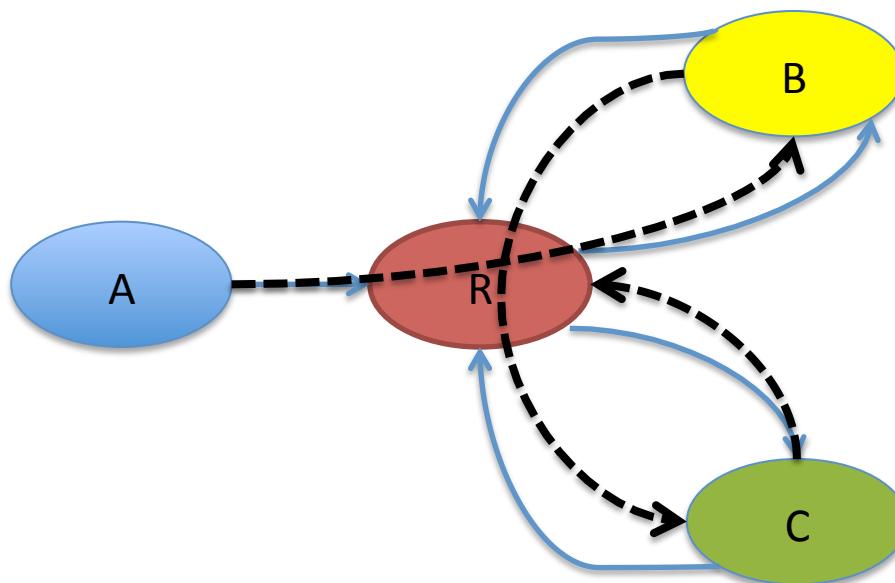
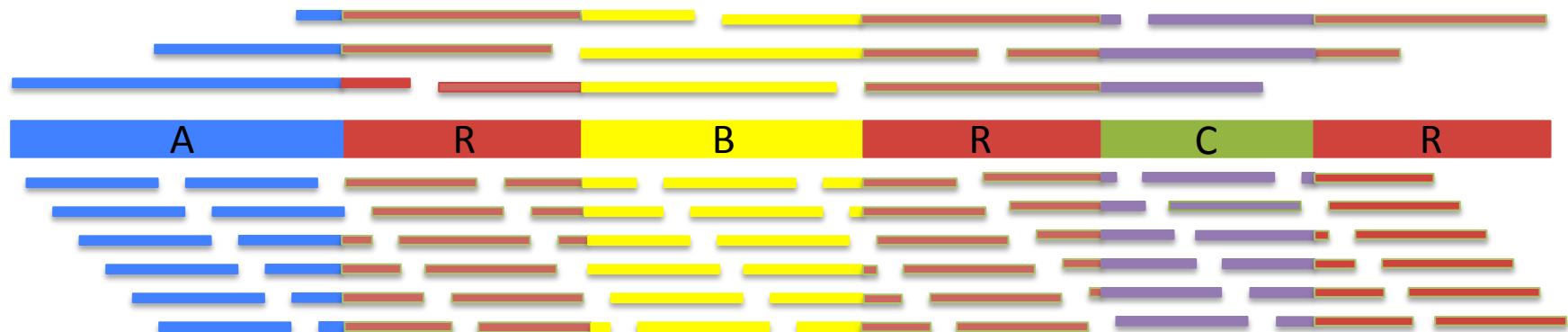


Errors obscure overlaps

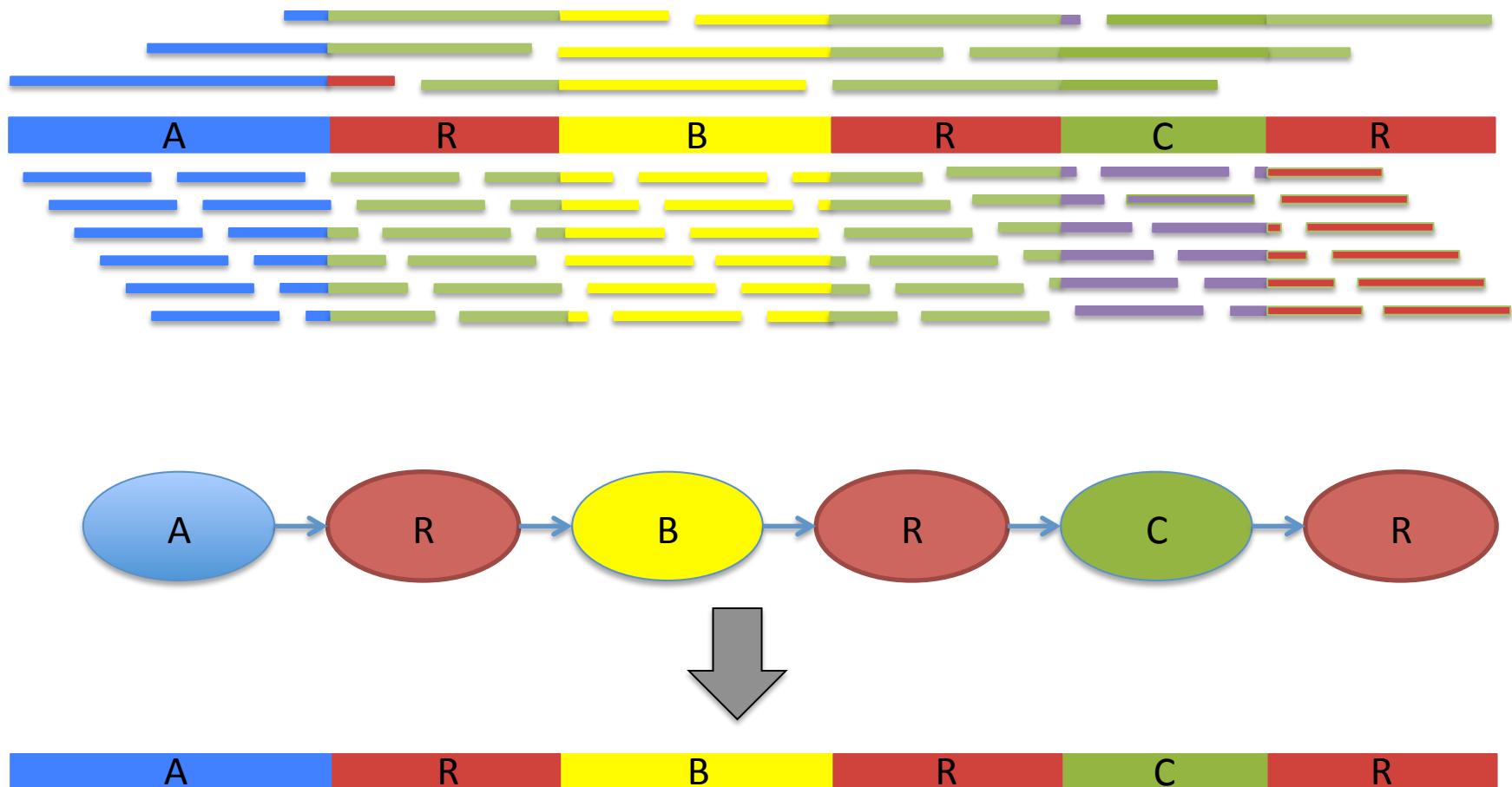
- Reads are assembled by finding kmers shared in pair of reads
- High error rate requires very short seeds, increasing complexity and forming assembly hairballs

Current challenges in *de novo* plant genome sequencing and assembly
Schatz MC, Witkowski, McCombie, WR (2012) *Genome Biology*. 12:243

Assembly Complexity



Assembly Complexity



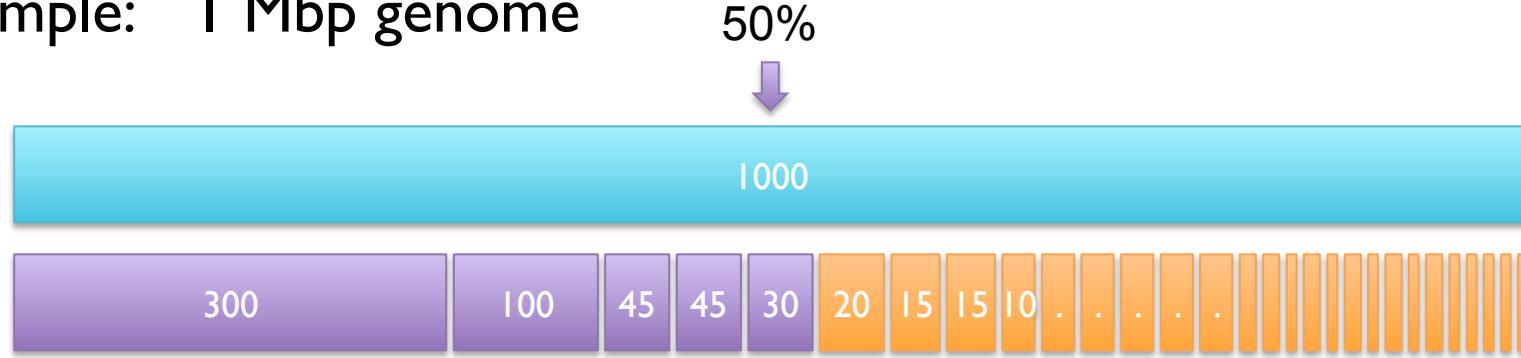
The advantages of SMRT sequencing

Roberts, RJ, Carneiro, MO, Schatz, MC (2013) *Genome Biology*. 14:405

N50 size

Def: 50% of the genome is in contigs as large as the N50 value

Example: 1 Mbp genome



N50 size = 30 kbp

$$(300k + 100k + 45k + 45k + 30k = 520k \geq 500\text{kbp})$$

A greater N50 is indicative of improvement in every dimension:

- Better resolution of genes and flanking regulatory regions
- Better resolution of transposons and other complex sequences
- Better resolution of chromosome organization
- Better sequence for all downstream analysis

Outline

1. Assembly Review

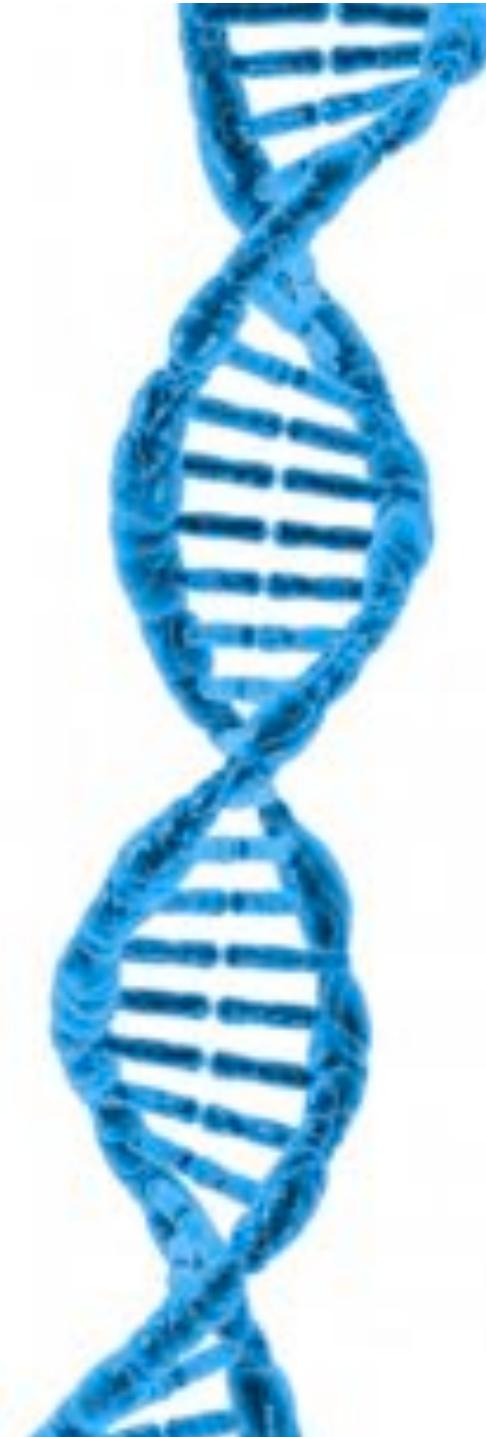
2. Pacbio

- Technology Overview
- Data Characteristics
- Algorithms
- Results – Assemblies

3. Oxford Nanopore

- Technology Overview
- Data Characteristics
- Algorithms
- Results – Assemblies

4. Summary



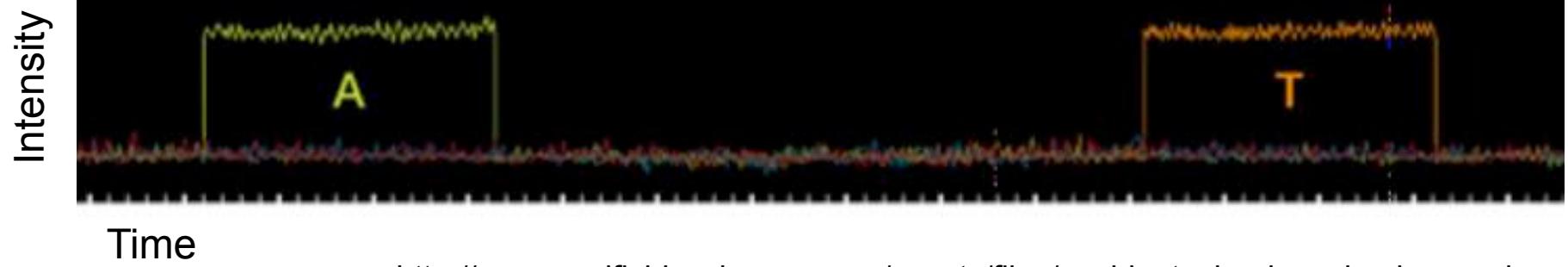
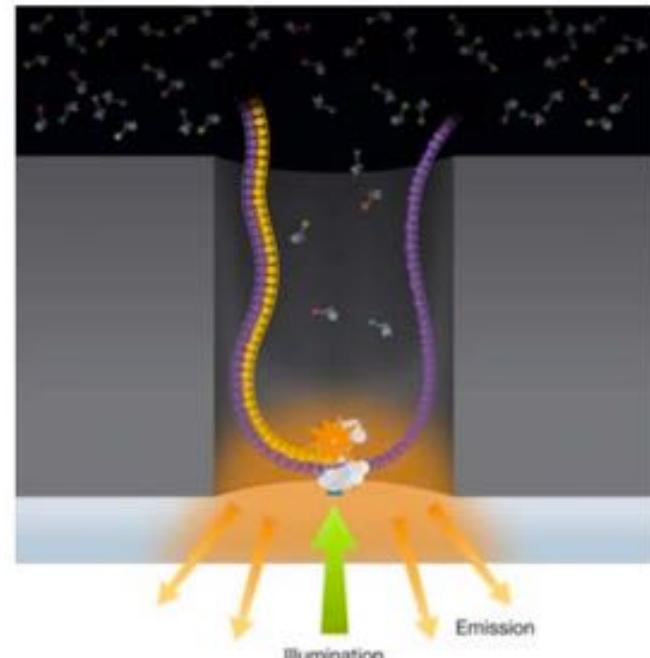
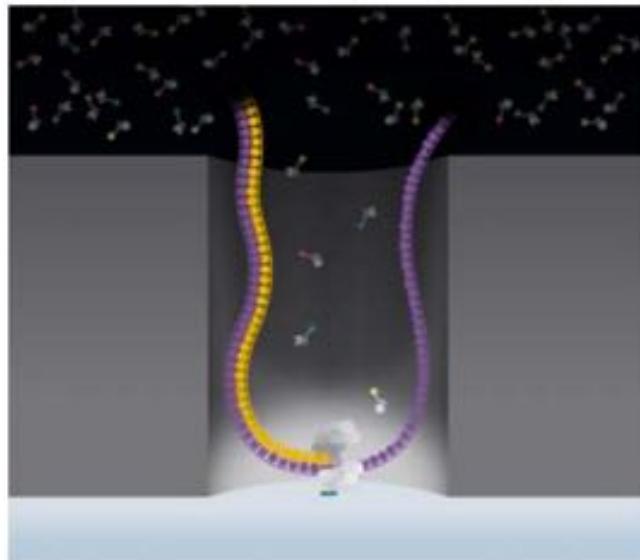


PACIFIC
BIOSCIENCES™

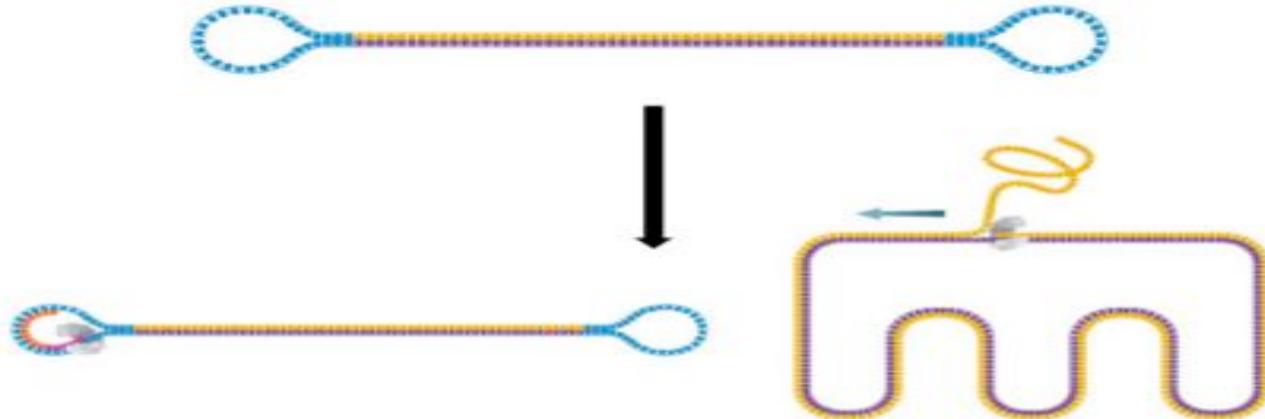


SMRT Sequencing

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

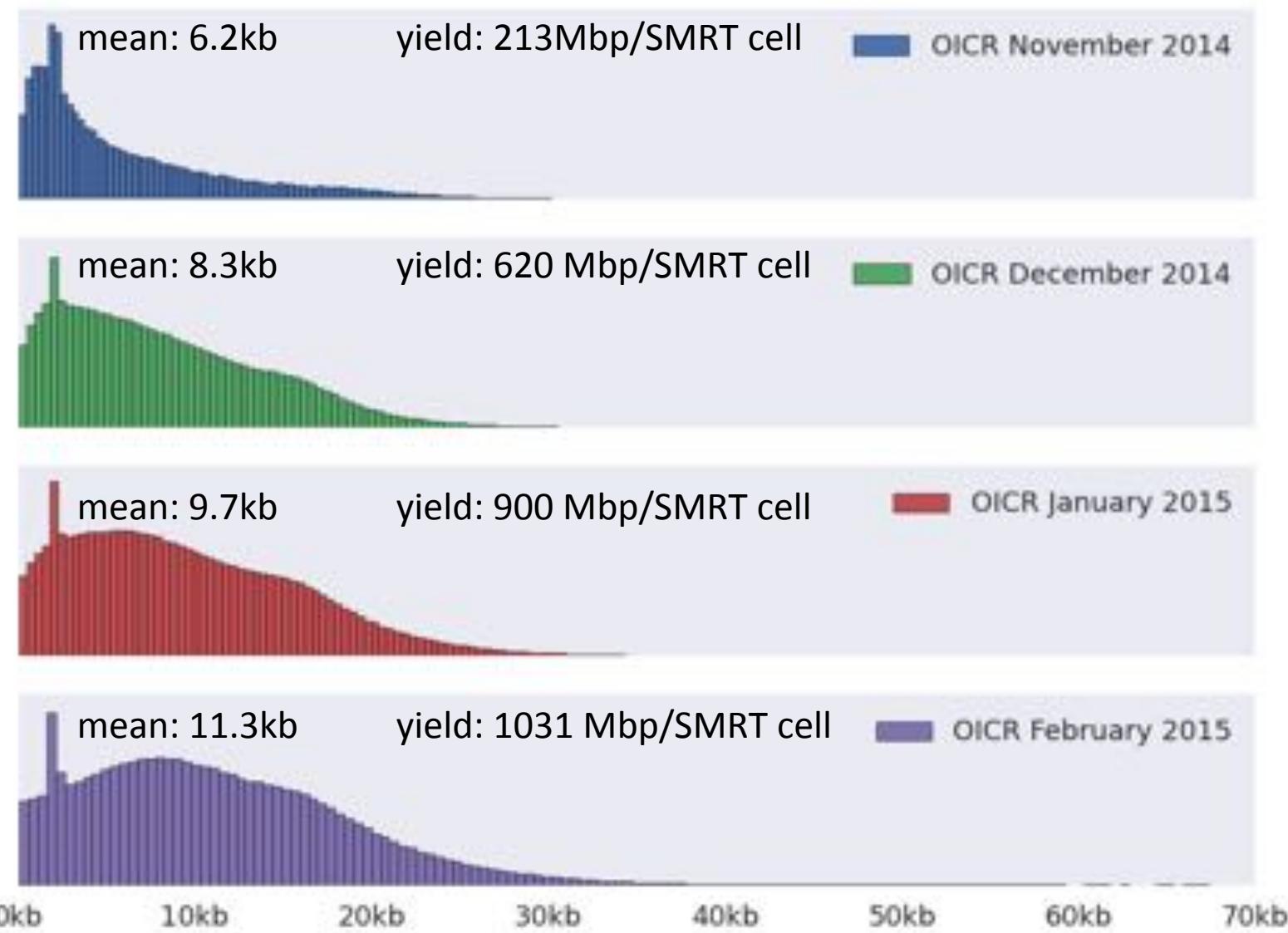


SMRT Read Types



- **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.
 - Barbell sequence: ATCTCTCTCttttcctcctccgttgttGAGAGAGAT

Dramatic changes just by experimenting with library preparation



Average Error Rate	16%
Mismatches	30.0%
Deletions	11.5%
Insertions	58.5%

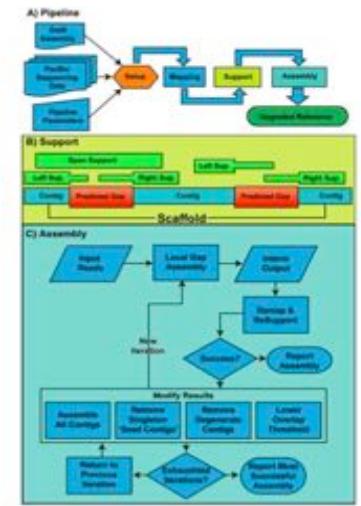
> chr7
Length=1898948

Score = 238 bits (1872), Expect = 3e-68
Identities = 612/889 (70%), Gaps = 172/889 (21%)
Strand=Plus/Minus

Error Profile Dominated by **Insertions**

Long Read Correction Algorithms

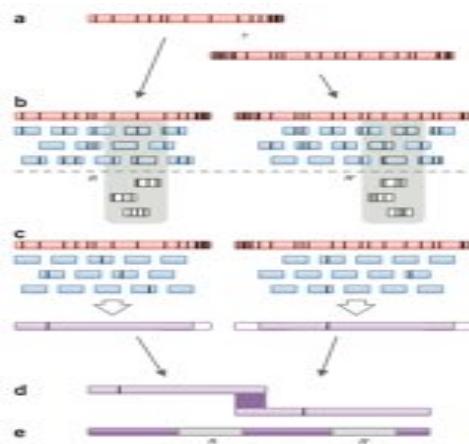
PBJelly



Gap Filling and Assembly Upgrade

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

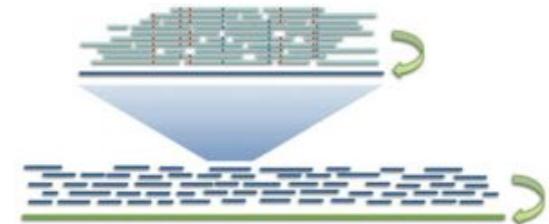
PacBioToCA & ECTools



Hybrid Error Correction

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

HGAP & Quiver



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

A probability tree diagram where node T branches into nodes R_1 , R_2 , ..., R_N .

Quiver Performance Results Comparison to Reference Genome (<i>M. ruber</i> ; 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

< 5x

Long Read Coverage

> 50x

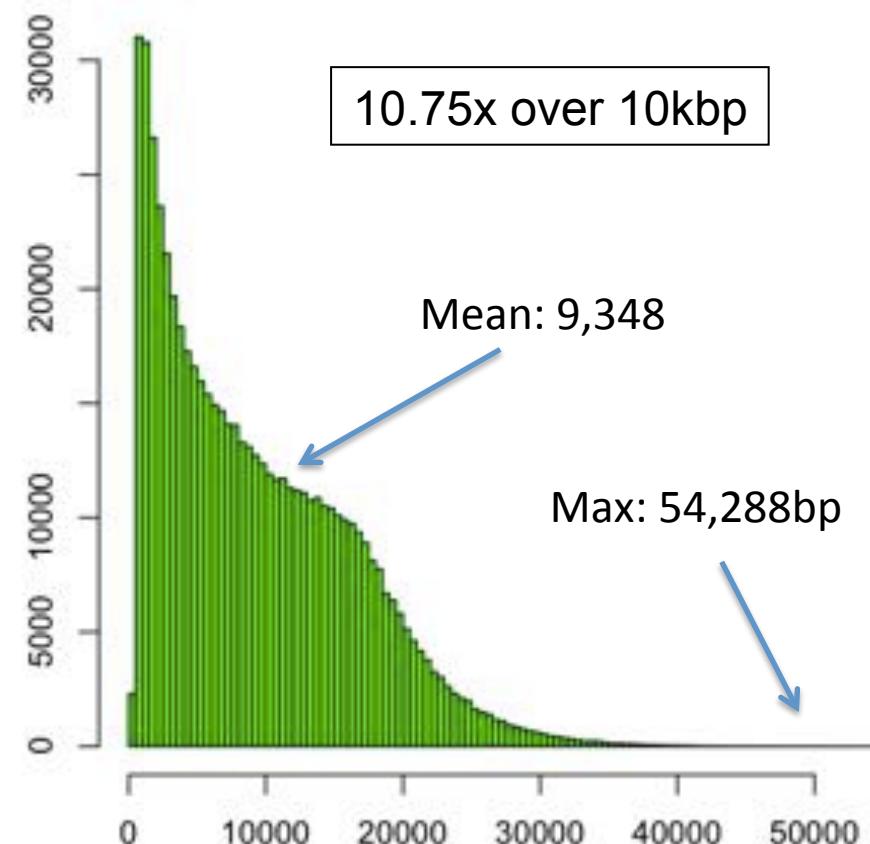
O. sativa pv Indica (IR64)

Genome size: ~370 Mb

Chromosome N50: ~29.7 Mbp

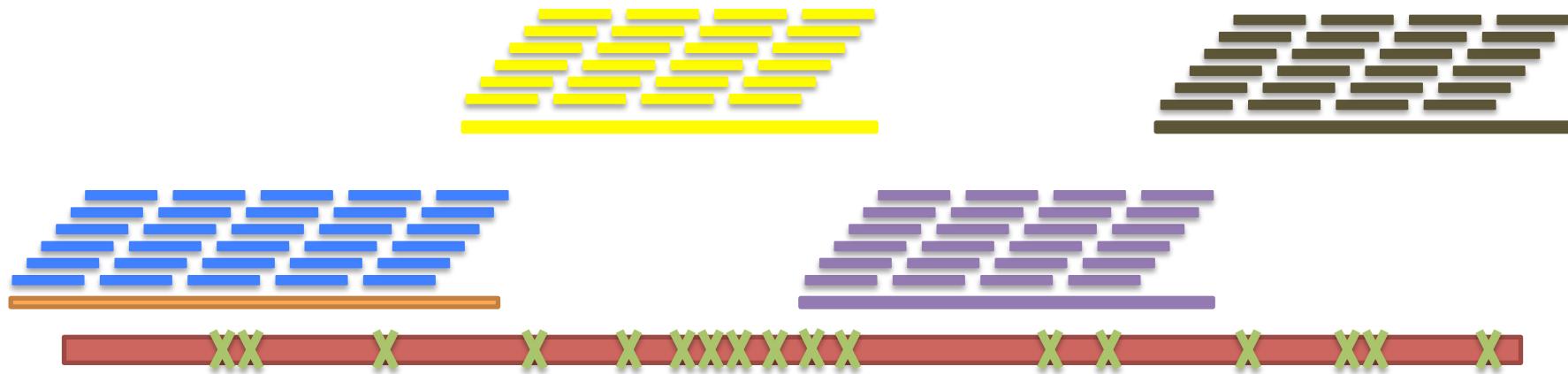


Assembly	Contig NG50
“ALLPATHS-recipe” 50x 2x100bp @ 180 36x 2x50bp @ 2100 51x 2x50bp @ 4800	18,450
MiSeq Fragments 25x 456bp (3 runs 2x300 @ 450 FLASH)	19,078
PacbioToCA – 47 SMRTCells 10.7x @ 10kbp	144,042
ECTools - 47 SMRTCells 10.7x @ 10kbp	????



ECTools: Error Correction with pre-assembled reads

<https://github.com/jgurtowski/ectools>



Short Reads -> Assemble Unitigs -> Align & Select -> Error Correct

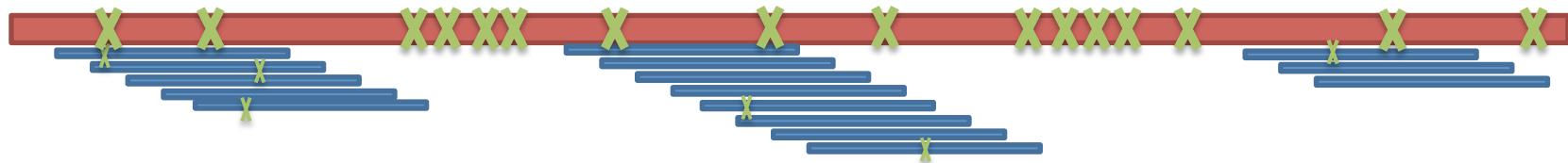
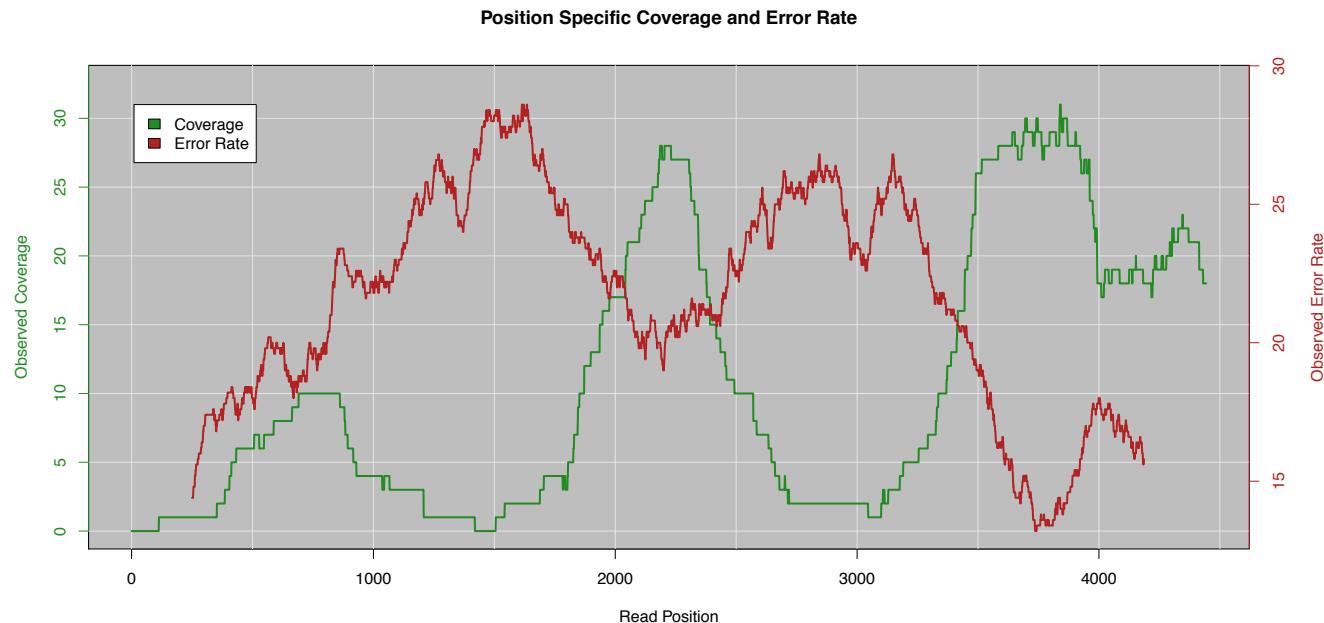
Can Help us overcome:

1. Error Dense Regions – Longer sequences have more seeds to match
2. Simple Repeats – Longer sequences easier to resolve

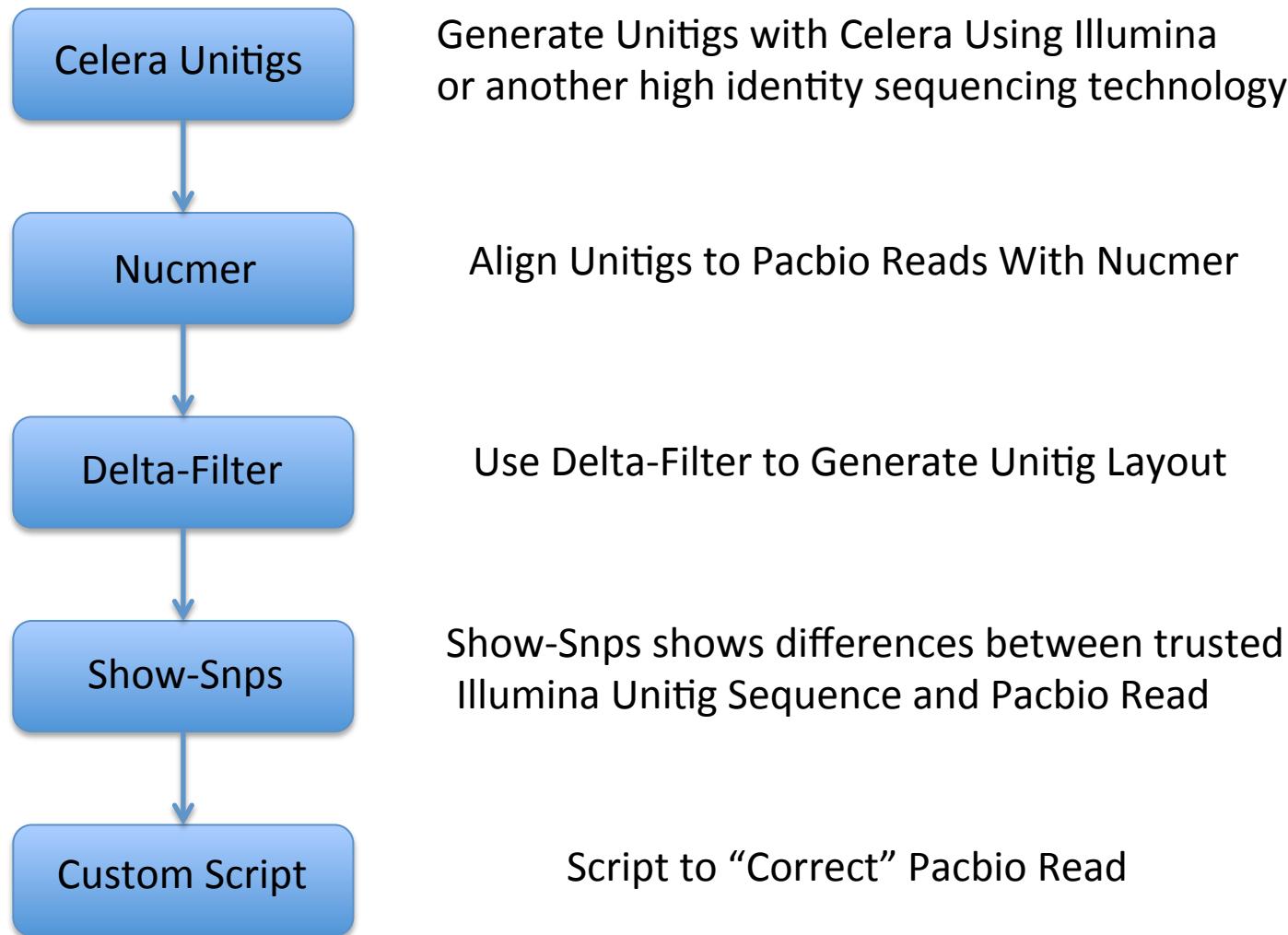
However, cannot overcome Illumina coverage gaps & other biases

Low Coverage Regions

1. Simple Repeats – Kmer Frequency Too High to Seed Overlaps
2. GC Rich Regions – Known Illumina Bias
3. **Error Dense Regions – Difficult to compute overlaps with many errors**



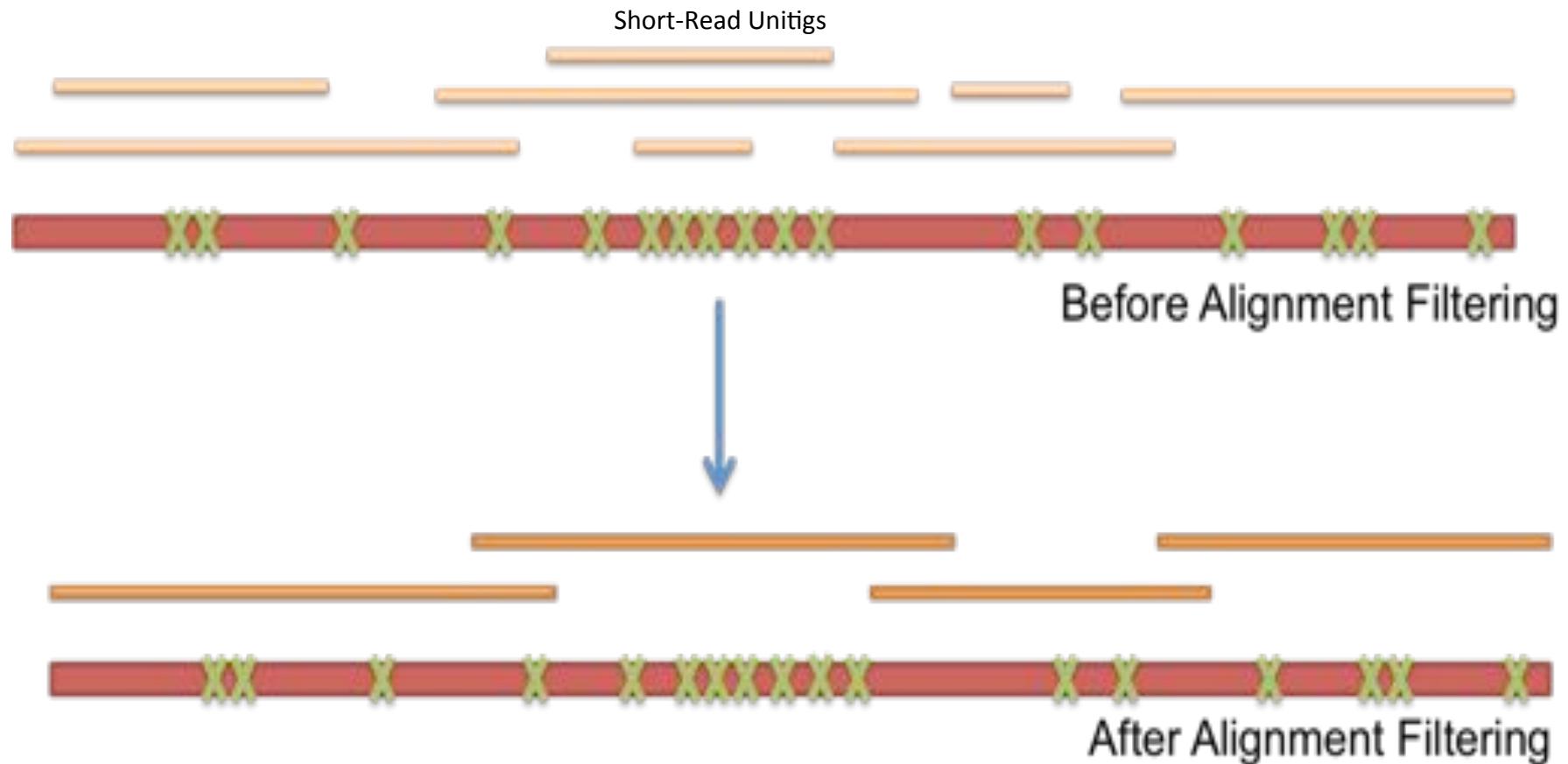
ECTools Pipeline



Note: Reads are never split or trimmed

Delta-Filter Alignment filtering

Uses Dynamic Programming (Longest Increasing Subset) to find the longest mutually consistent subset of unitigs with respect to the Pacbio Read



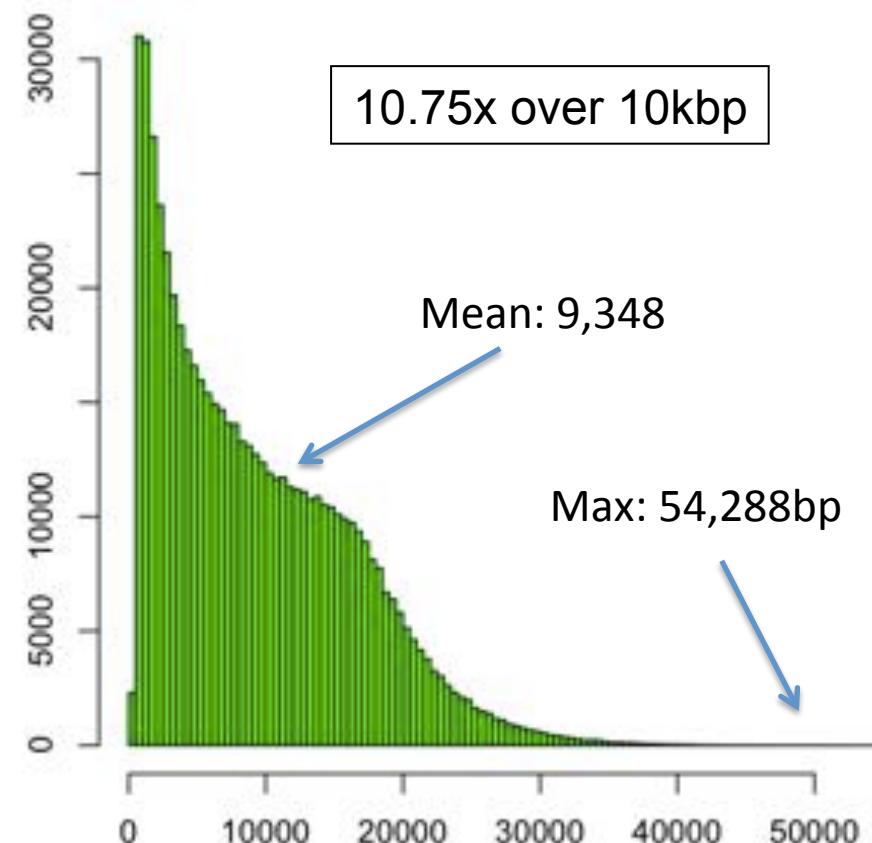
O. Sativa Indica (IR64)

Genome size: ~370 Mb

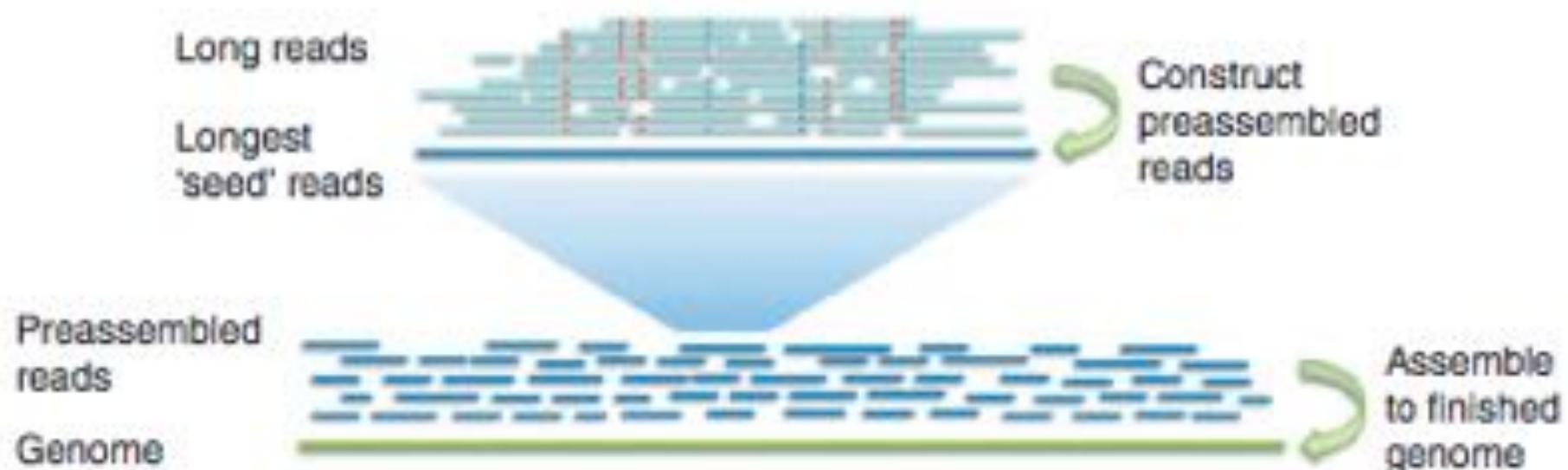
Chromosome N50: ~29.7 Mbp



Assembly	Contig NG50
“ALLPATHS-recipe” 50x 2x100bp @ 180 36x 2x50bp @ 2100 51x 2x50bp @ 4800	18,450
MiSeq Fragments 25x 456bp (3 runs 2x300 @ 450 FLASH)	19,078
PacbioToCA – 47 SMRTCells 10.7x @ 10kbp	144,042
ECTools - 47 SMRTCells 10.7x @ 10kbp	272,137



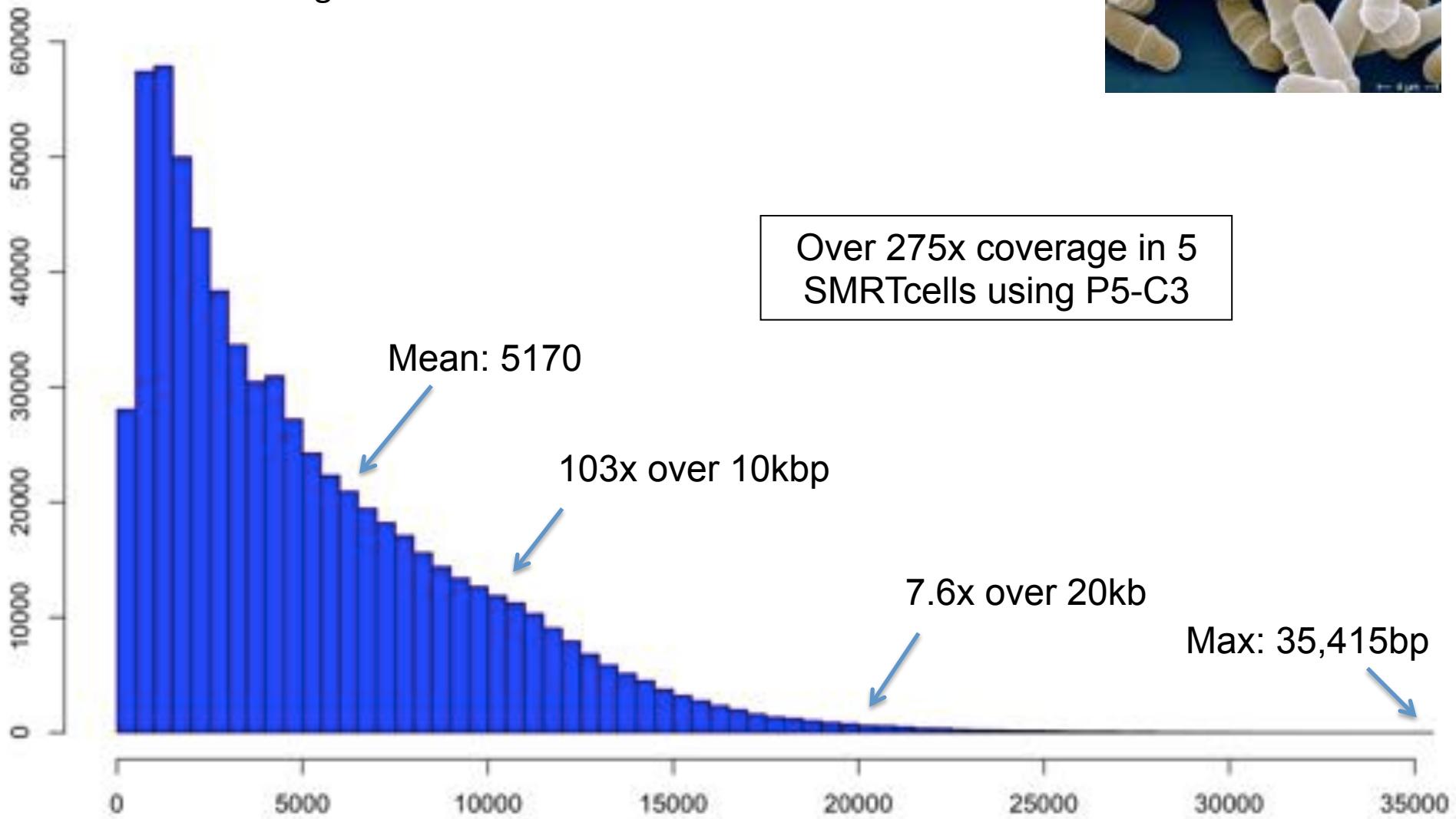
HGAP Error Correction



S. pombe dg21

PacBio RS II sequencing at CSHL

- Size selection using an 7 Kb elution window on a BluePippin™ device from Sage Science



S. pombe dg21

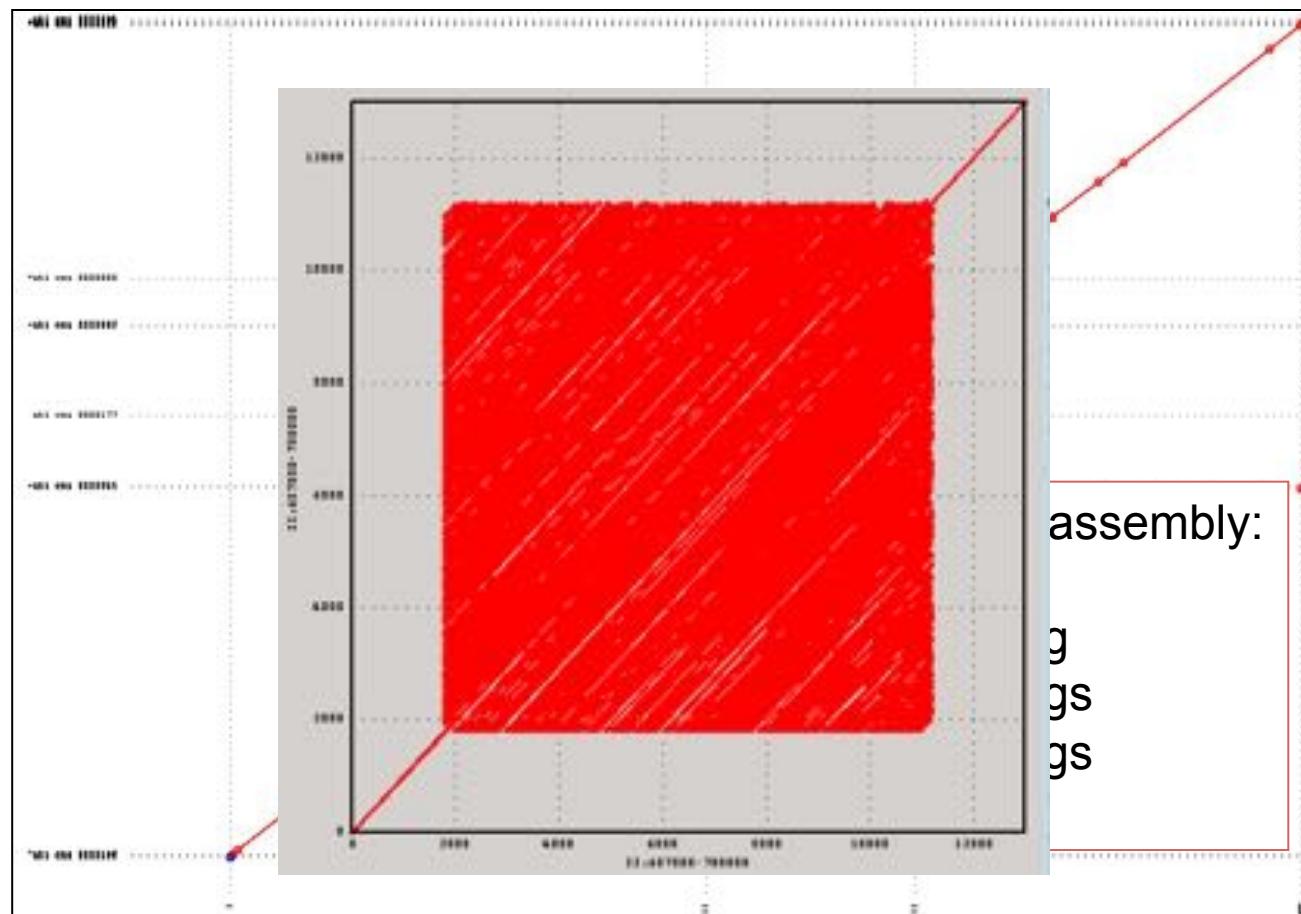
ASM294 Reference sequence

- 12.6Mbp; 3 chromo + mitochondria; N50: 4.53Mbp



PacBio assembly using HGAP + Celera Assembler

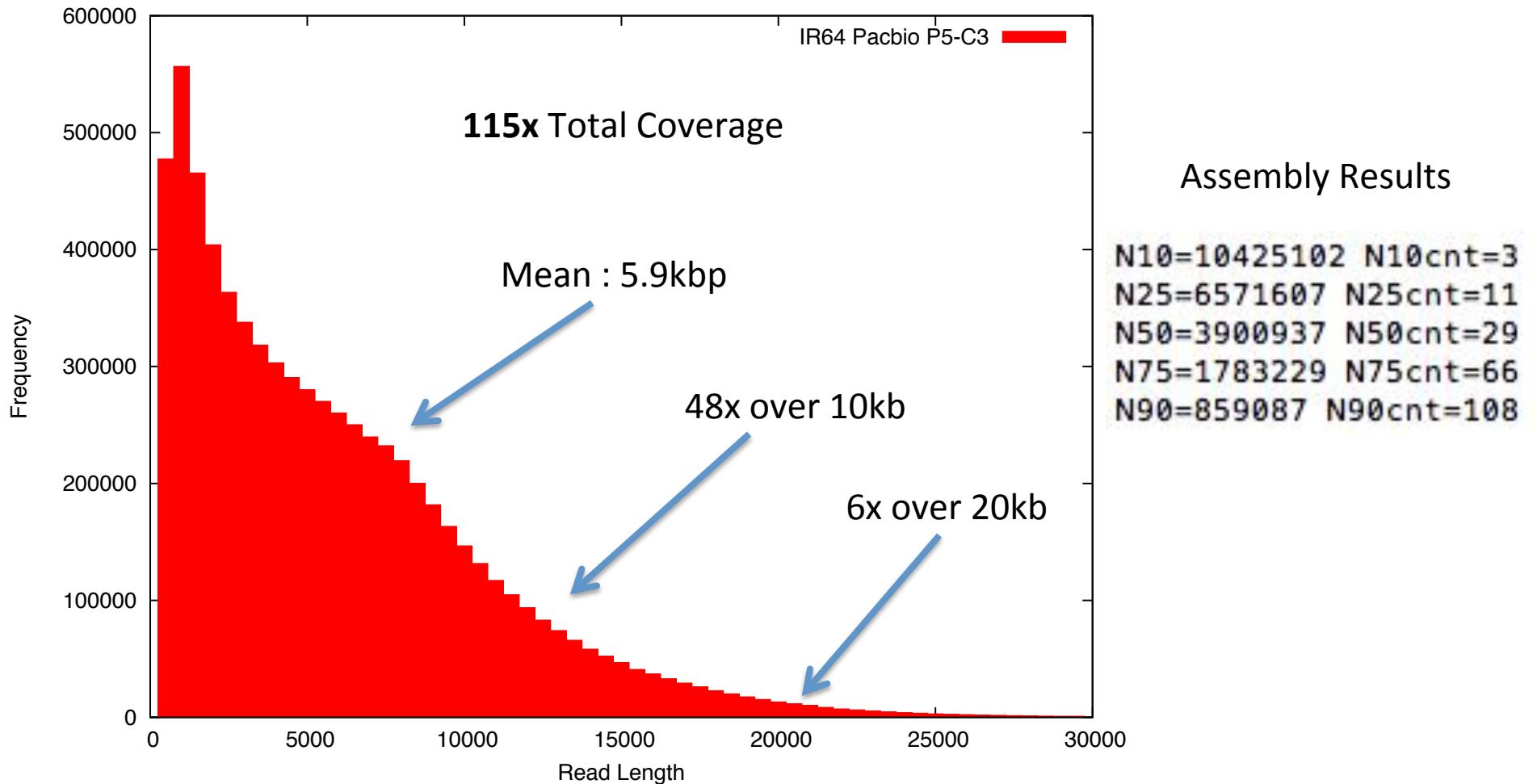
- 12.7Mbp; 13 non-redundant contigs; N50: 3.83Mbp; >99.98% id



O. sativa pv Indica (IR64)

Genome size: ~370 Mb

Chromosome N50: ~29.7 Mbp



Dazcon



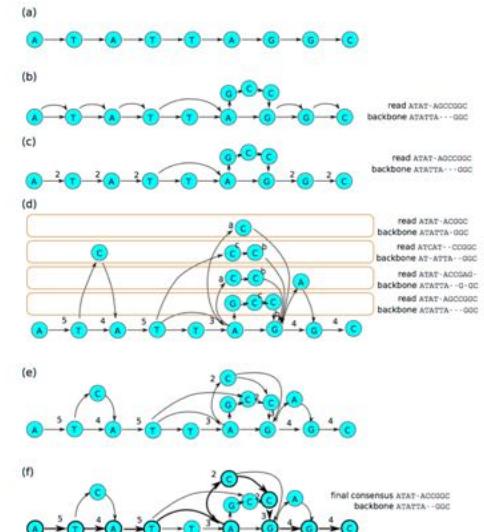
thegenemyers / DALIGNER

A -----+====>
B <=====+-----
9017

dazcon -ox -j 4 -s subreads.db -a subreads.las > corrected.fasta



PacificBiosciences / pbdagcon

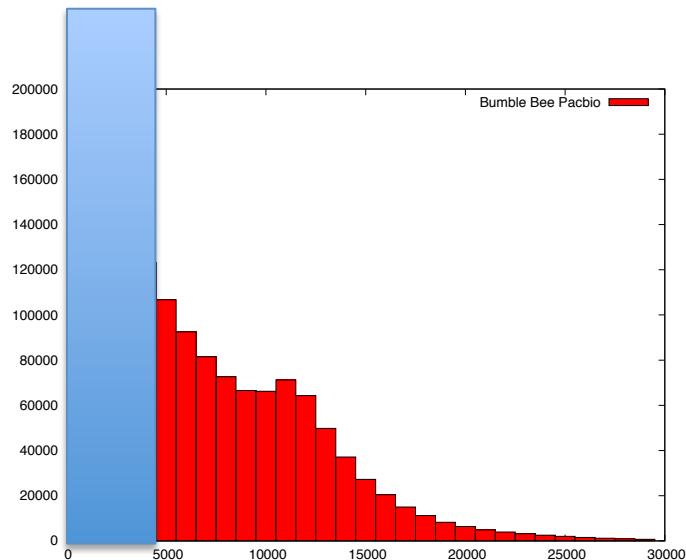


Bumble Bee Genome Assembly



n=369356 [26, 453616] 767.79 +/- 5960.50 sum=283588978 cov=1.13

Discard reads <5kb



N10=124463 N10cnt=145
N25=72986 N25cnt=547
N50=35935 N50cnt=1792
N75=13011 N75cnt=4670
N90=3553 N90cnt=9793



n=26131 [7976, 2171155] 21683.57 +/- 67351.74 sum=566613279 cov=2.27

N10=1219658 N10cnt=17
N25=669835 N25cnt=60
N50=359646 N50cnt=191
N75=183176 N75cnt=436
N90=81409 N90cnt=725

Outline

1. Assembly Review

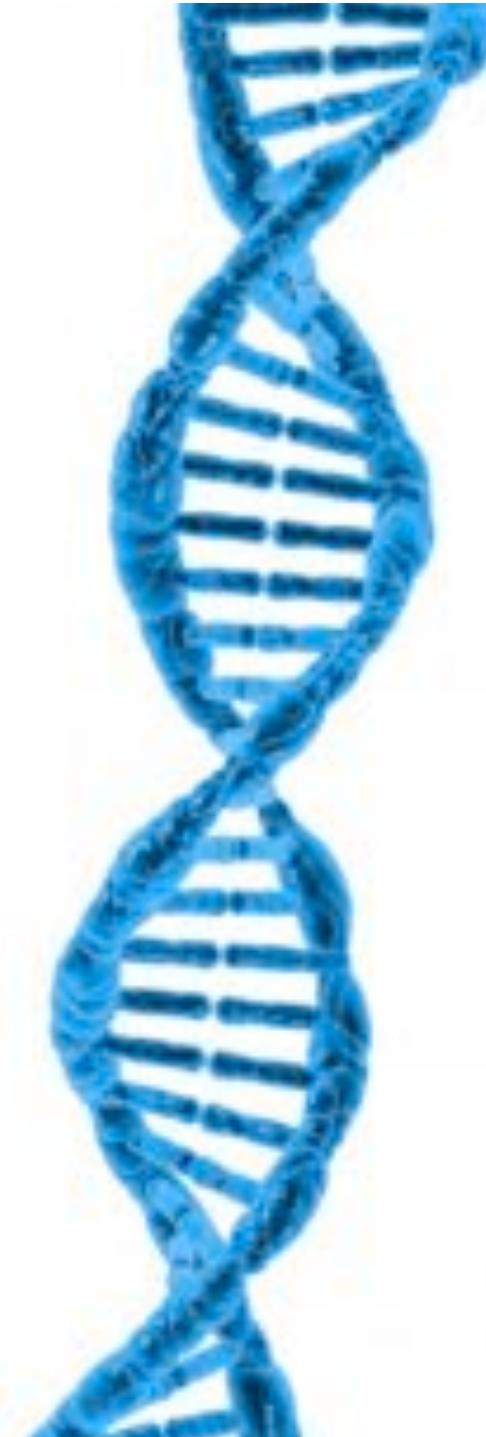
2. Pacbio

- Technology Overview
- Data Characteristics
- Algorithms
- Results – Assemblies

3. Oxford Nanopore

- Technology Overview
- Data Characteristics
- Algorithms
- Results – Assemblies

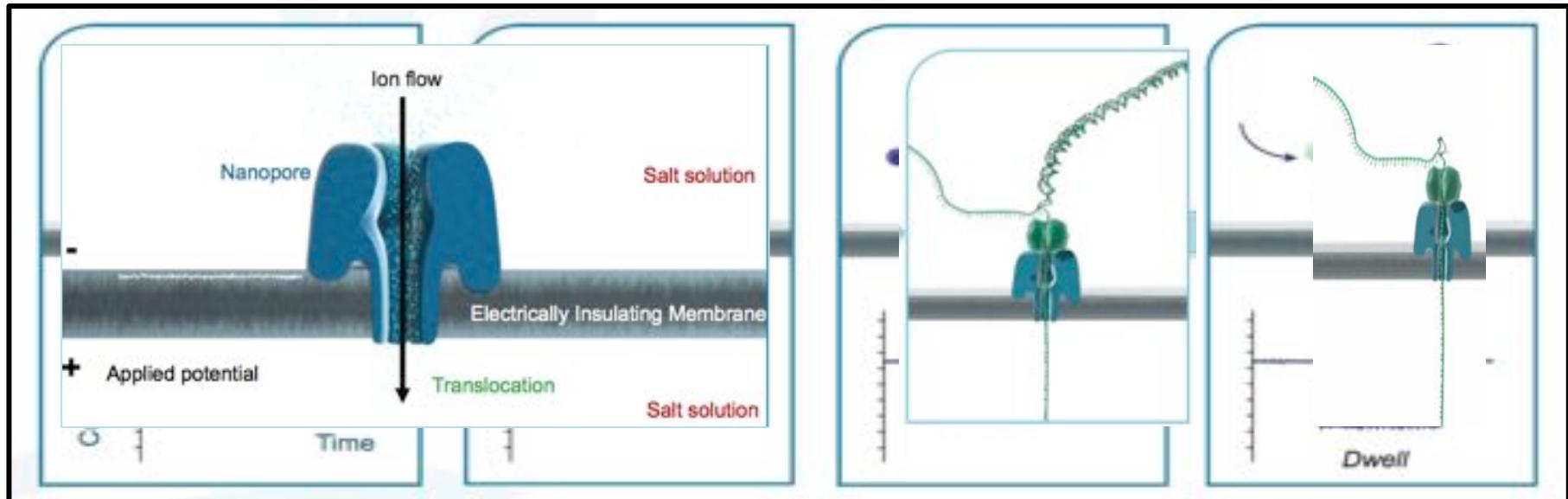
4. Summary



Oxford Nanopore MinION



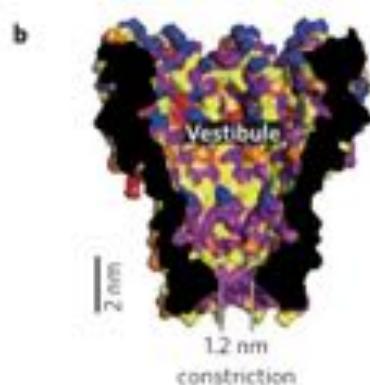
- MAP Program
- Thumb drive sized sequencer powered over USB
- Senses DNA by measuring changes to ion flow
- Reads both DNA Strands (2D)



Advantages And Challenges of Nanopore DNA Sequencing

Advantages

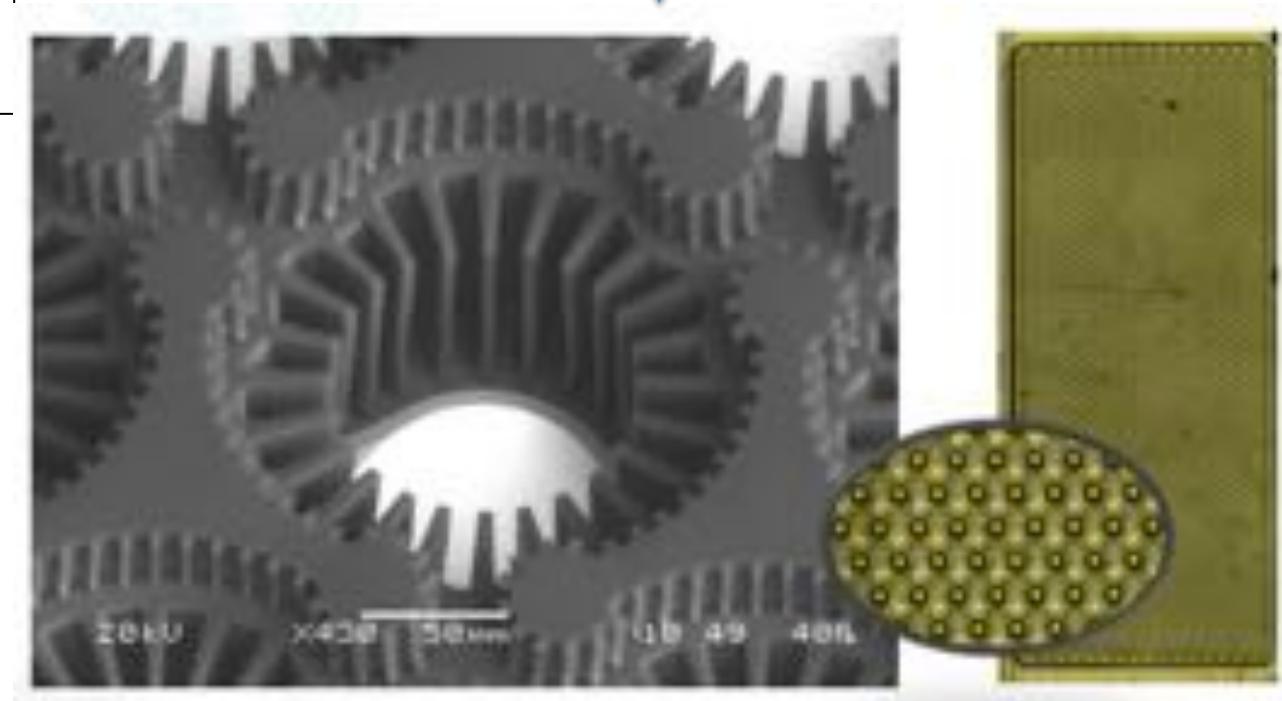
- Label-free
- Amplification-free
- Single-Molecule
- High-Throughput
- Inexpensive Instrument
- Simple/quick Sample Prep
- Produces Long Reads



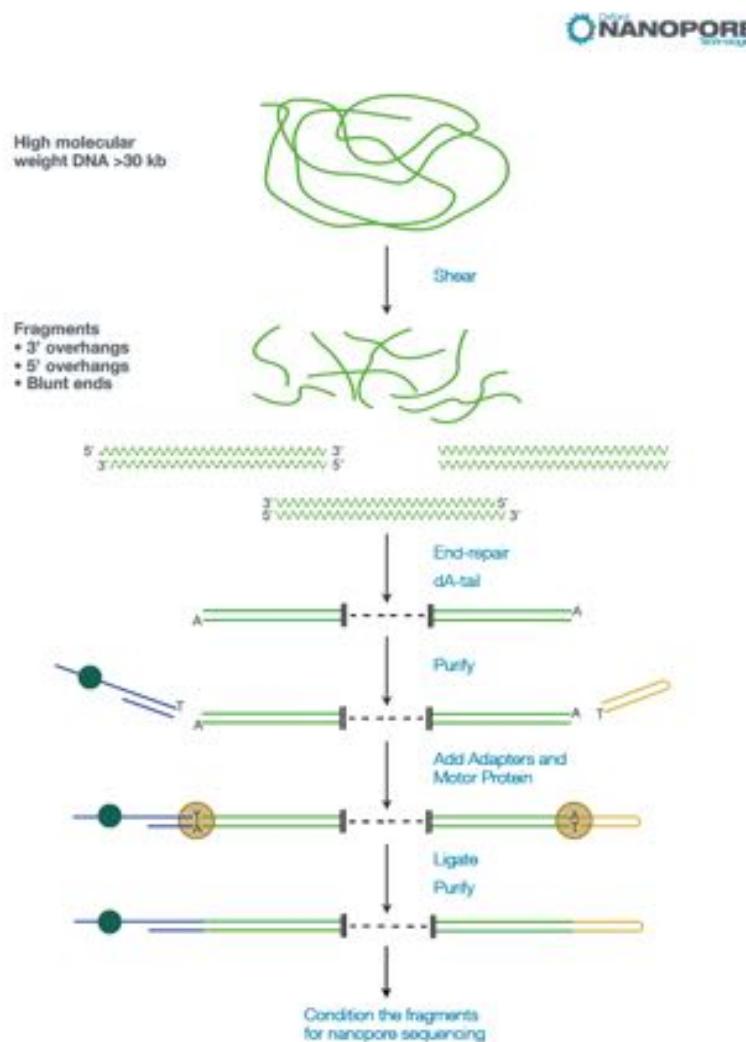
Challenges

- Controlling rate at which DNA translocates through the pore (1base/microsecond too fast to accurately measure current change)
- Pore does not have single base resolution (complicates basecalling and makes it hard to deal with modified bases)
- Commercially: Biological pores are sensitive to pH, temperature, salt concentration

Under the hood of the MinION

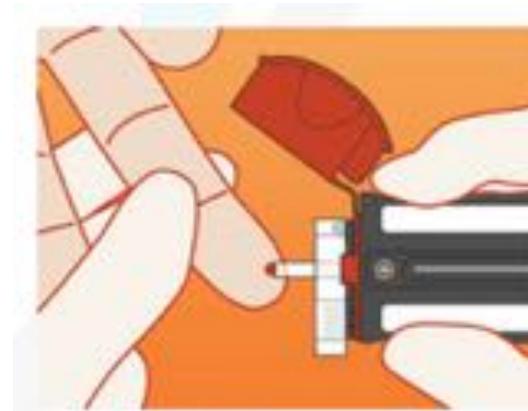


Oxford Nanopore DNA Prep

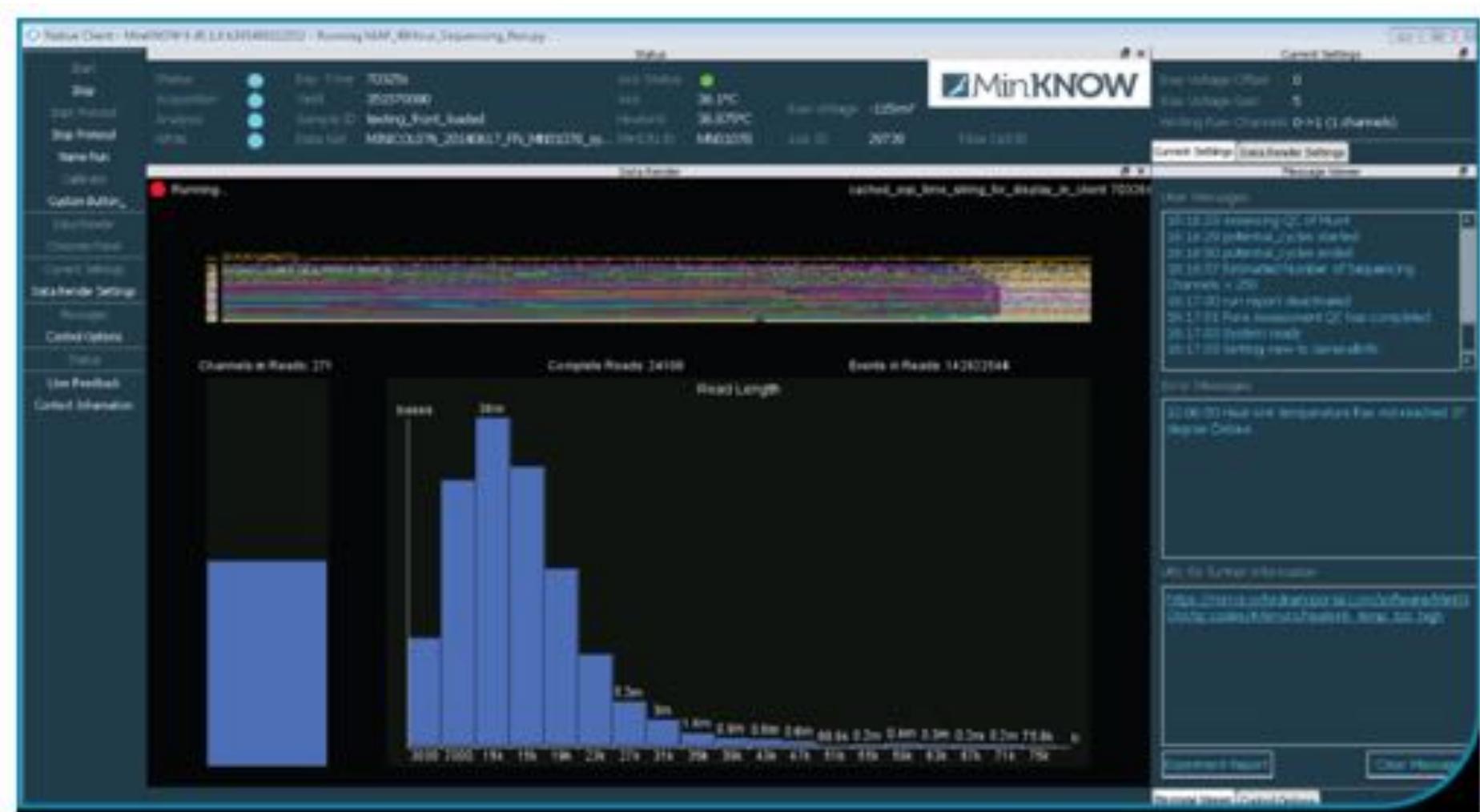


Simple DNA prep

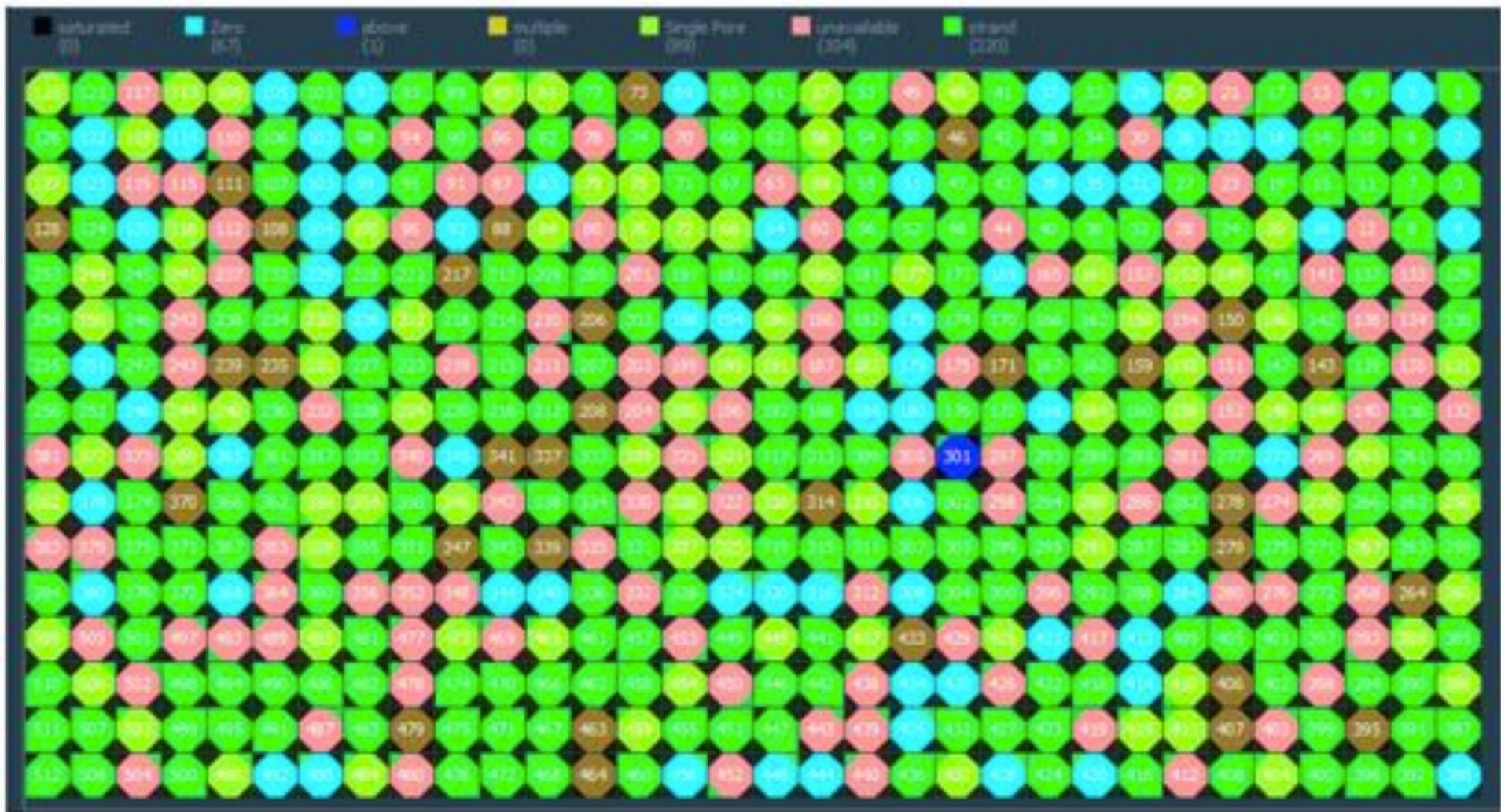
Can do it in the field?



Nanopore Desktop Software



View of Pore Activity

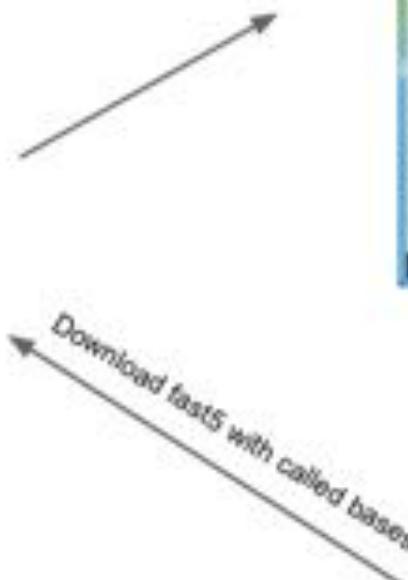


Base Calling

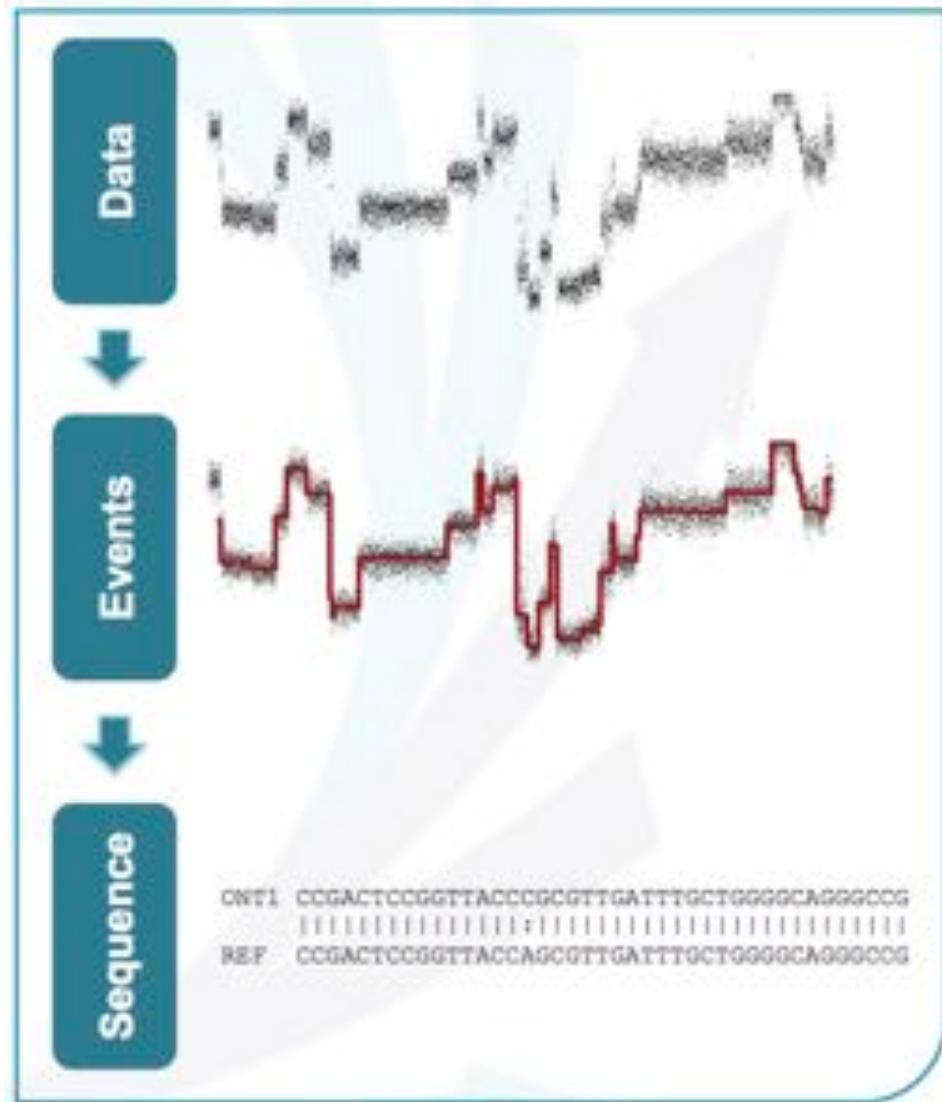
Raw Signal data contained in fast5 (hdf5) files on local machine

Name	Date modified	Type	Size
TEST12345_ch0_fna1	18/04/2014 13:38	Fast5 file	257 KB
TEST12345_ch0_fna2	18/04/2014 13:38	Fast5 file	259 KB
TEST12345_ch0_fna3	18/04/2014 13:38	Fast5 file	259 KB
TEST12345_ch0_fna4	18/04/2014 13:38	Fast5 file	242 KB
TEST12345_ch0_fna5	18/04/2014 13:38	Fast5 file	254 KB
TEST12345_ch0_fna6	18/04/2014 13:38	Fast5 file	243 KB
TEST12345_ch0_fna7	18/04/2014 13:38	Fast5 file	246 KB
TEST12345_ch0_fna8	18/04/2014 13:38	Fast5 file	251 KB
TEST12345_ch0_fna9	18/04/2014 13:38	Fast5 file	369 KB
TEST12345_ch0_fna10	18/04/2014 13:40	Fast5 file	369 KB
TEST12345_ch0_fna11	18/04/2014 13:40	Fast5 file	363 KB
TEST12345_ch0_fna12	18/04/2014 13:40	Fast5 file	353 KB
TEST12345_ch0_fna13	18/04/2014 13:40	Fast5 file	664 KB
TEST12345_ch0_fna14	18/04/2014 13:40	Fast5 file	243 KB
TEST12345_ch0_fna15	18/04/2014 13:40	Fast5 file	259 KB
TEST12345_ch0_fna16	18/04/2014 13:40	Fast5 file	249 KB
TEST12345_ch0_fna17	18/04/2014 13:40	Fast5 file	248 KB
TEST12345_ch0_fna18	18/04/2014 13:40	Fast5 file	275 KB
TEST12345_ch0_fna19	18/04/2014 13:40	Fast5 file	256 KB
TEST12345_ch0_fna20	18/04/2014 13:40	Fast5 file	187 KB

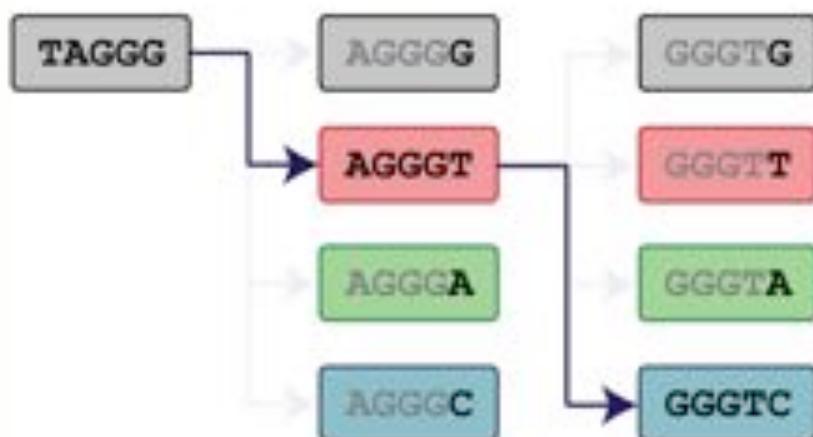
Local Software Agent facilitates data transaction with cloud basecaller



Nanopore Basecalling



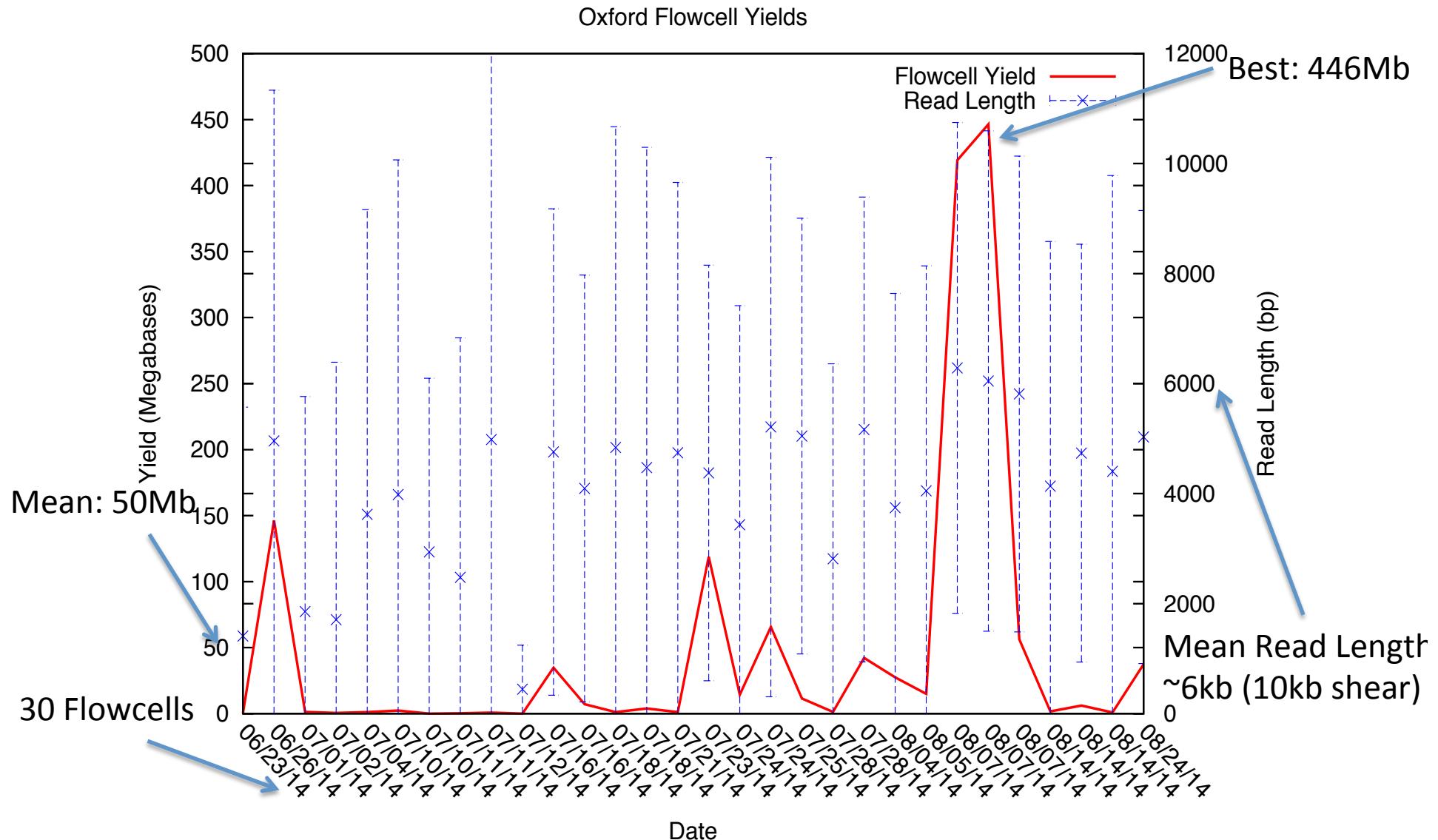
- Hidden Markov model
- Only four options per transition
- Pore type = distinct kmer length



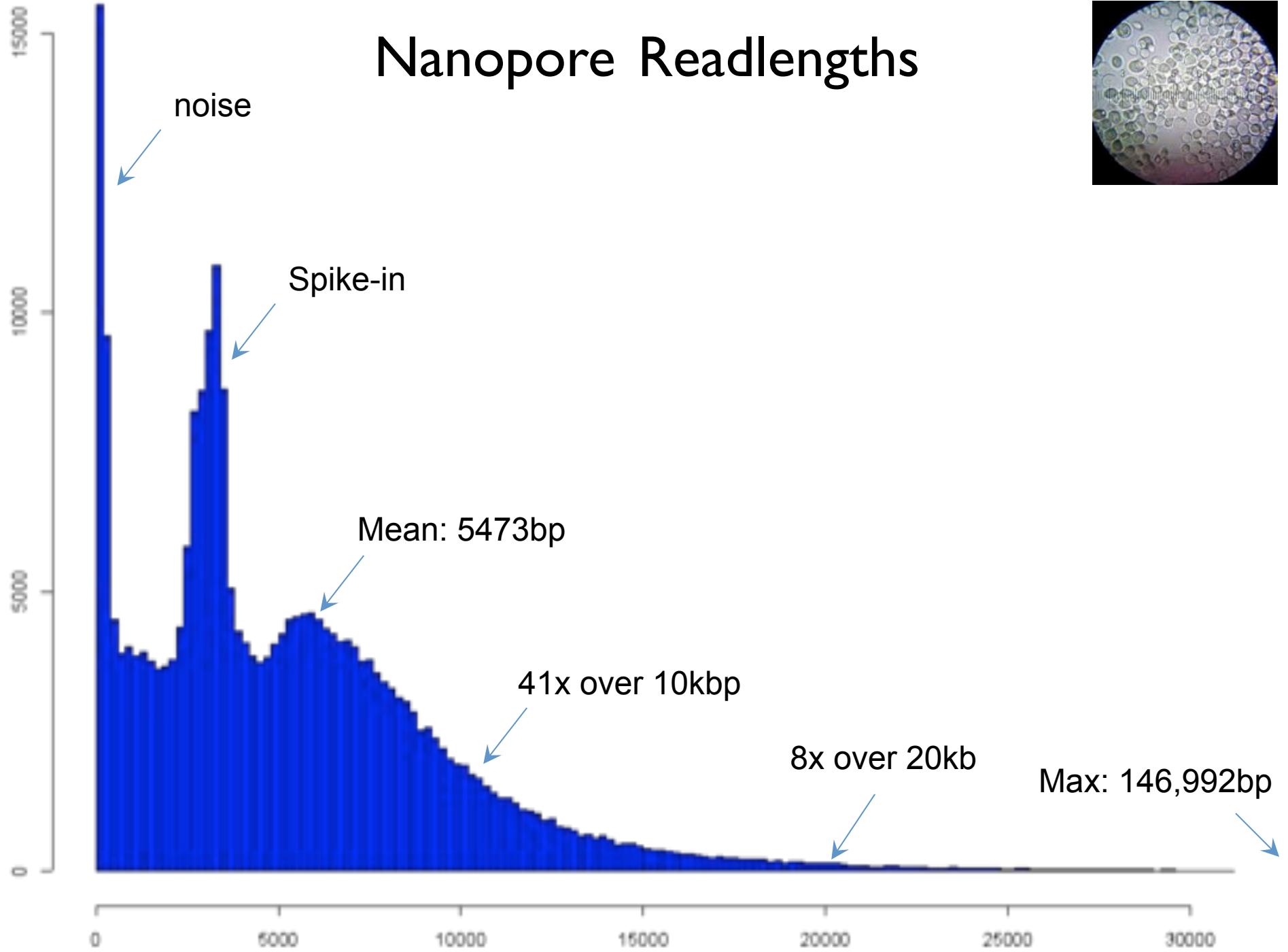
- Form probabilistic path through measured states currents and transitions
 - e.g. Viterbi algorithm

Basecalling currently performed at Amazon with frequent updates to algorithm

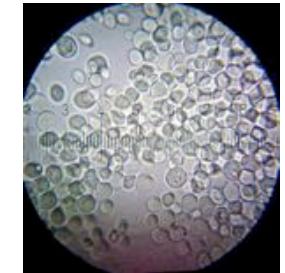
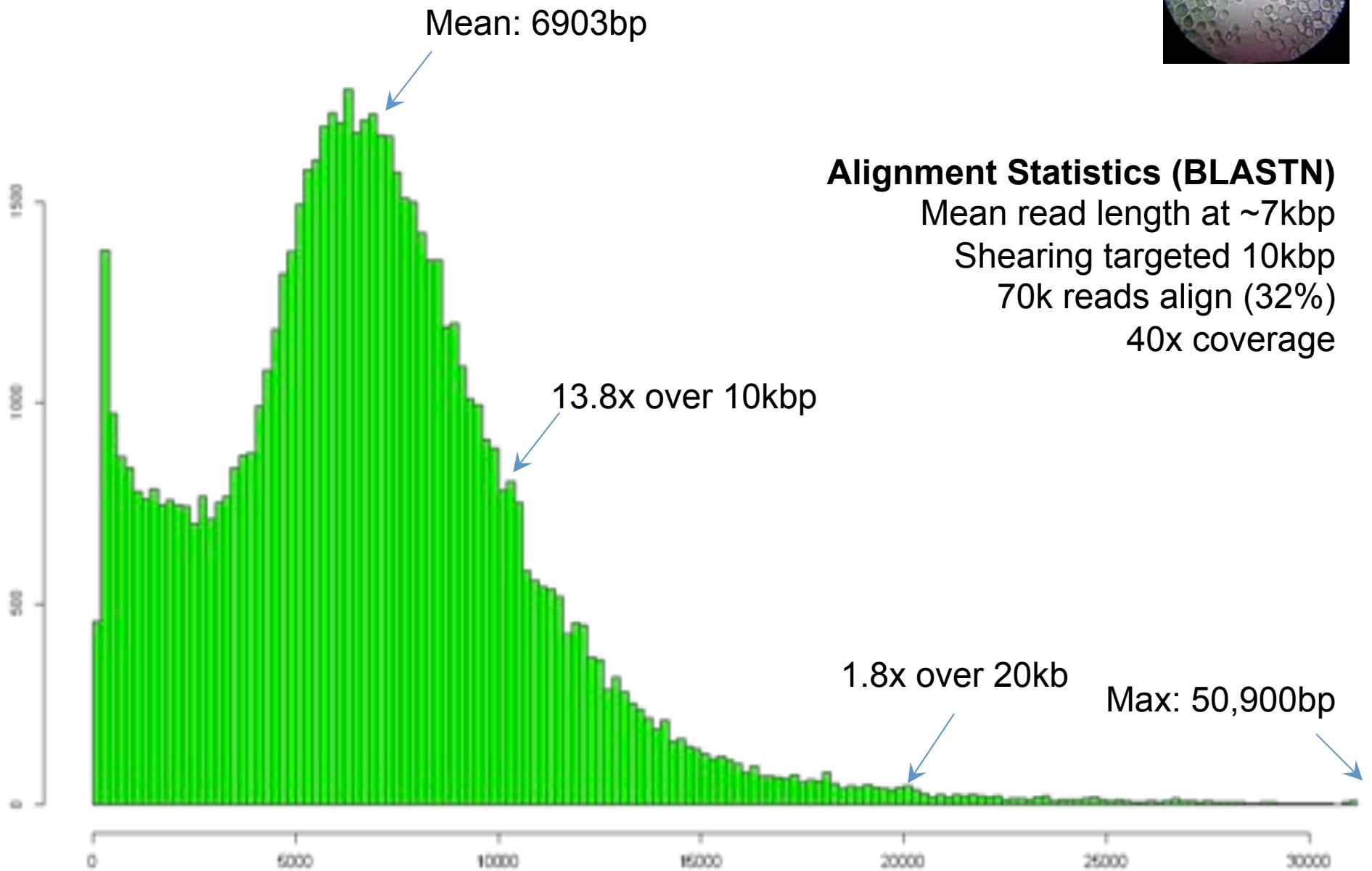
Our Data - Yeast W303



Nanopore Readlengths



Nanopore Alignments



Nanopore Accuracy

Alignment Quality (BLASTN)

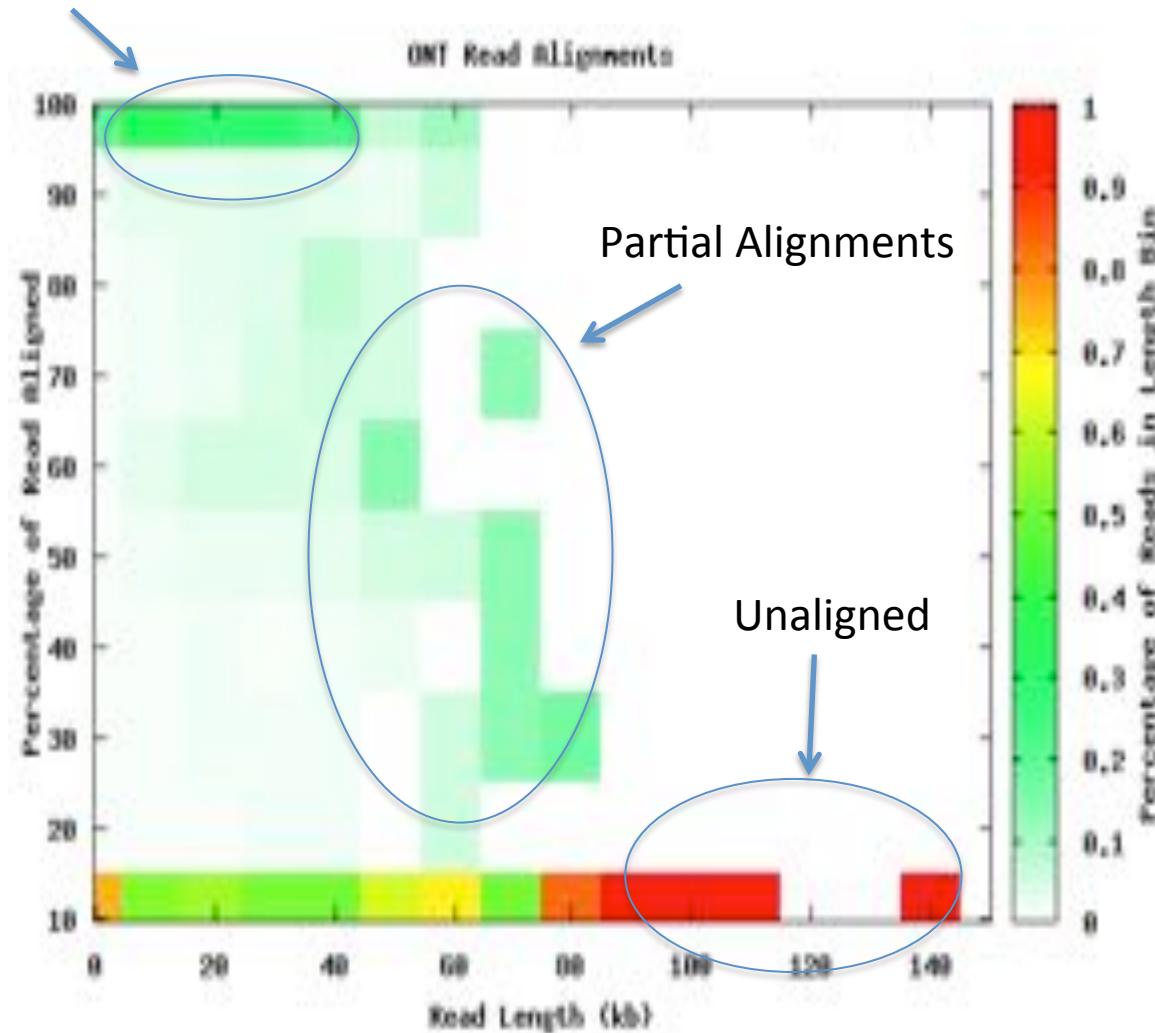
Of reads that align, average ~64% identity
“2D base-calling” improves to ~70% identity



Nanopore Alignment Summary

32% of the data map using BLASTN

Full Length Alignments

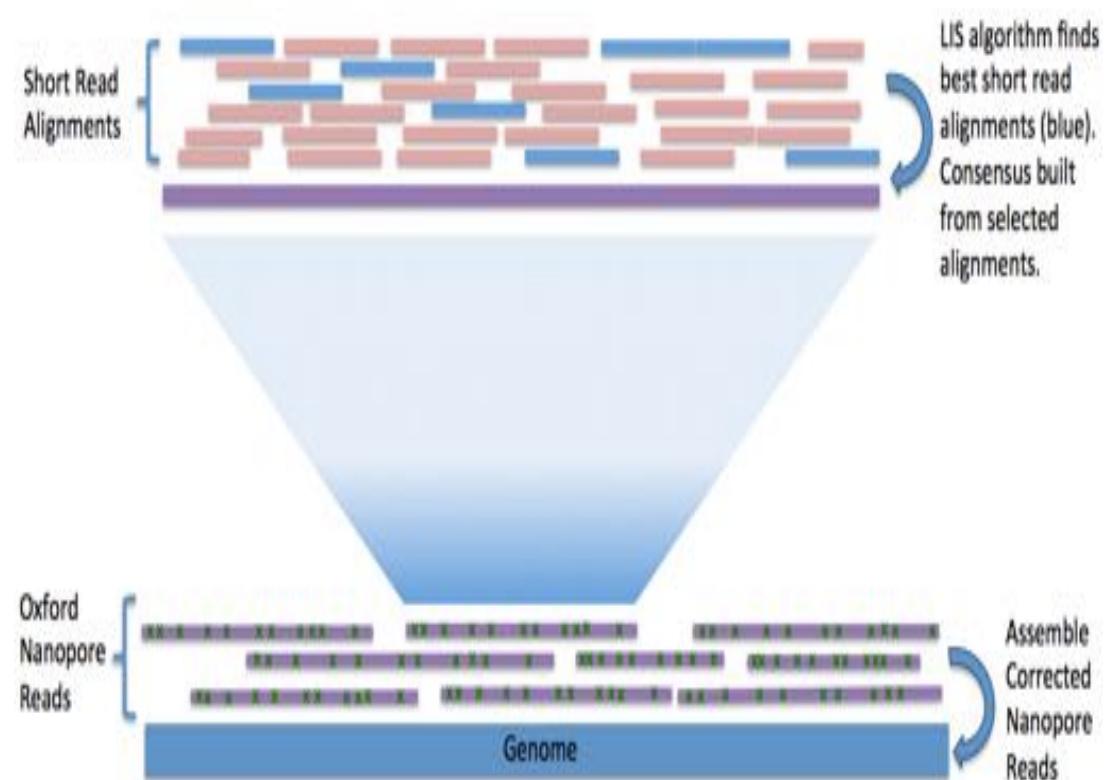


NanoCorr: Nanopore-Illumina Hybrid Error Correction

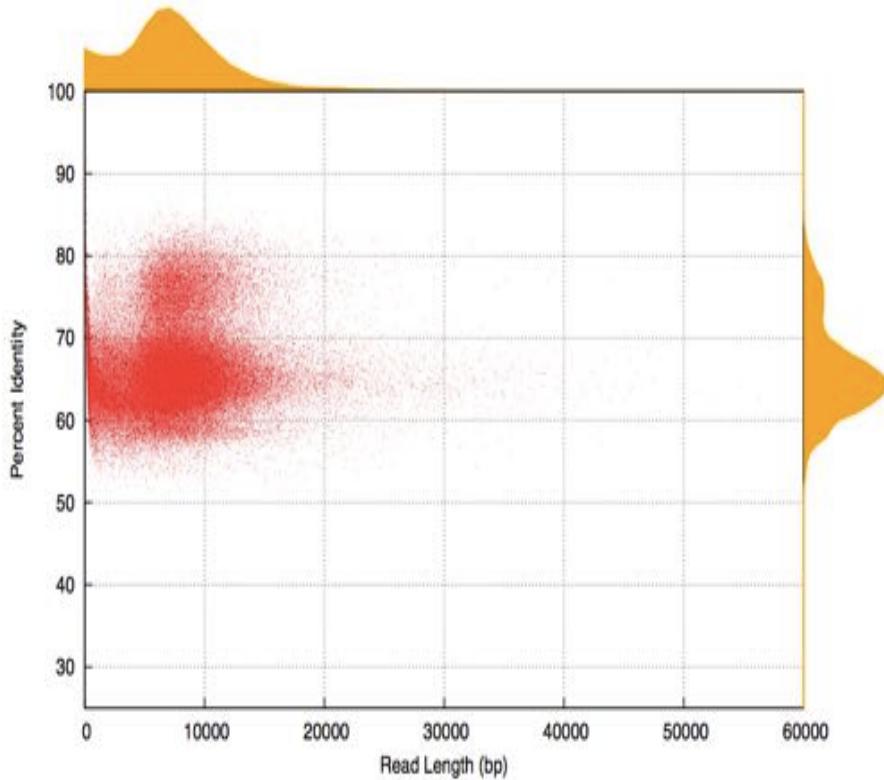


<https://github.com/jgurtowski/nanocorr>

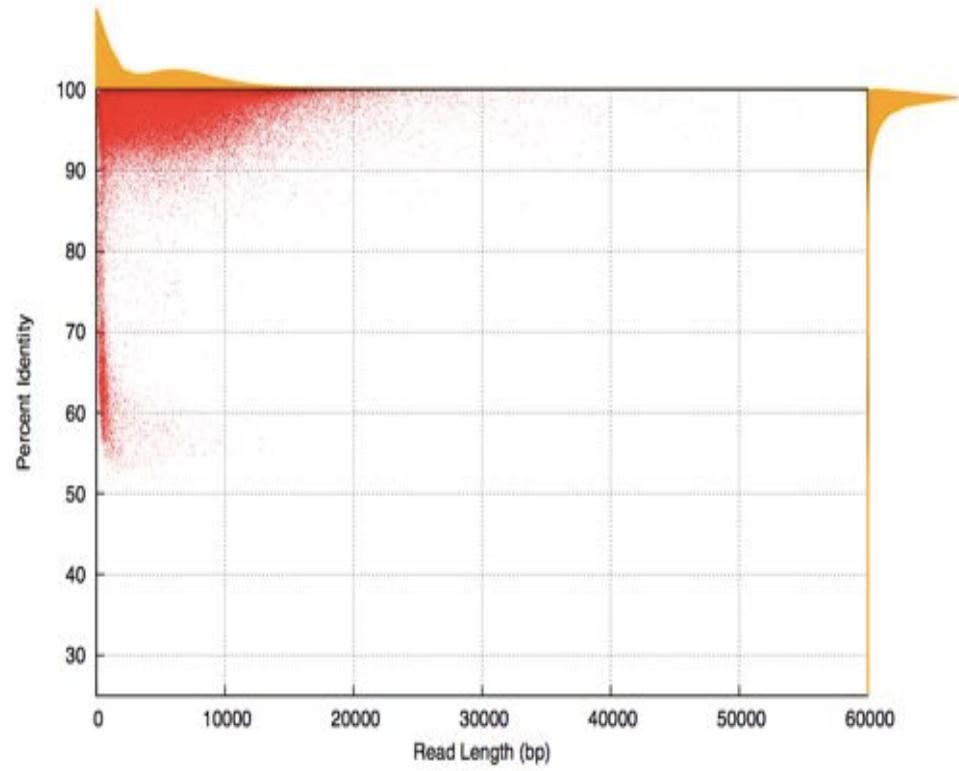
1. BLAST Miseq reads to all raw Oxford Nanopore reads
2. Select non-repetitive alignments
 1. First pass scans to remove “contained” alignments
 2. Second pass uses Dynamic Programming (LIS) to select set of high-identity alignments with minimal overlaps
3. Compute consensus of each Oxford Nanopore read
 1. Currently using Pacbio’s pbpdagcon



Nanocorr correction pipeline significantly improves read identity



Before



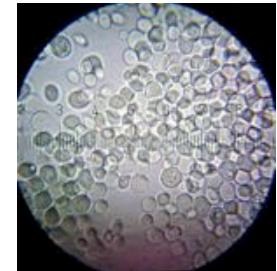
After

Percent identity versus read length before and after nanocorr correction

Long Read Assembly

S288C Reference sequence

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

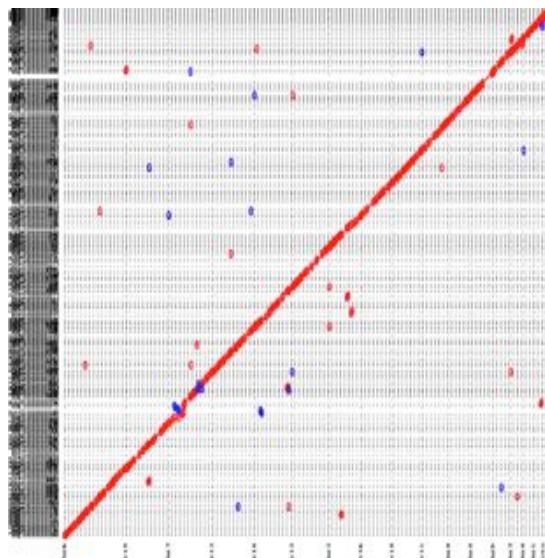


Illumina MiSeq

30x, 300bp PE (Flashed)

Celera Assembler

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

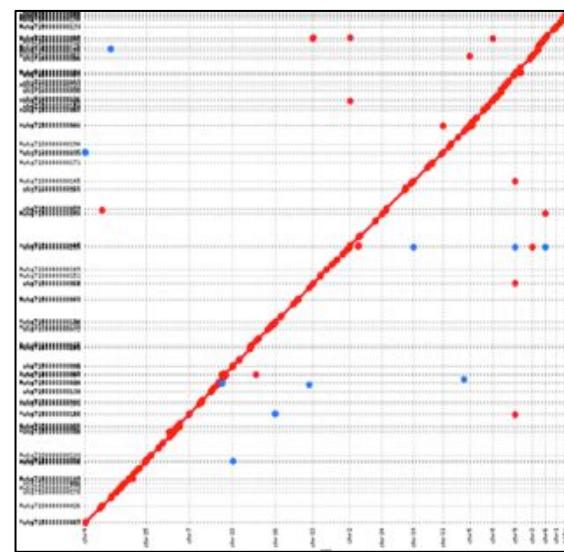


Oxford Nanopore

30x corrected reads > 6kb

NanoCorr + Celera Assembler

- 214 non-redundant contigs
- N50: 472kbp >99.78% id

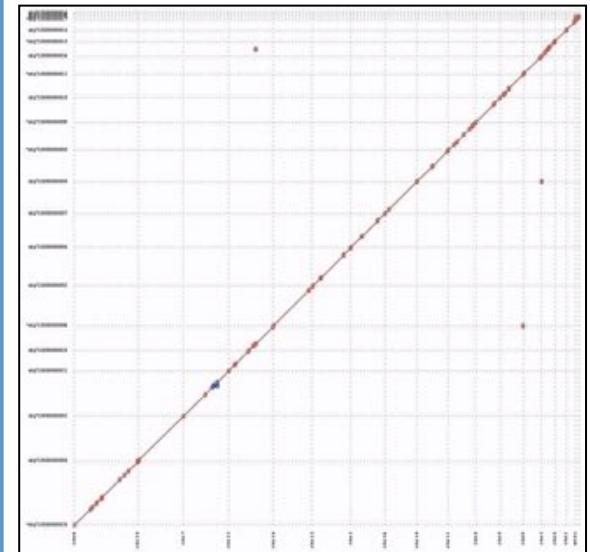


Pacific Biosciences

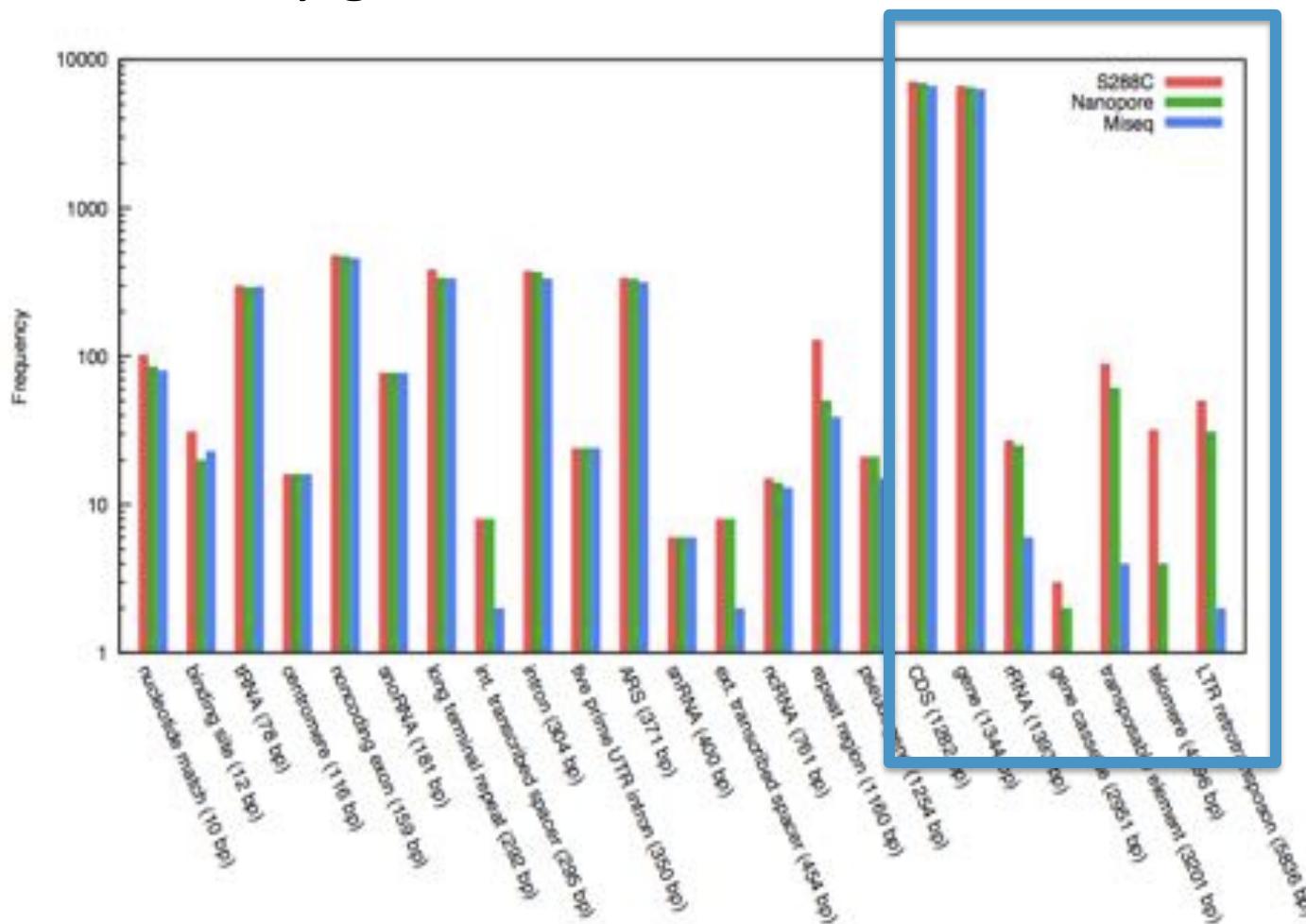
25x corrected reads > 10kb

HGAP + Celera Assembler

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

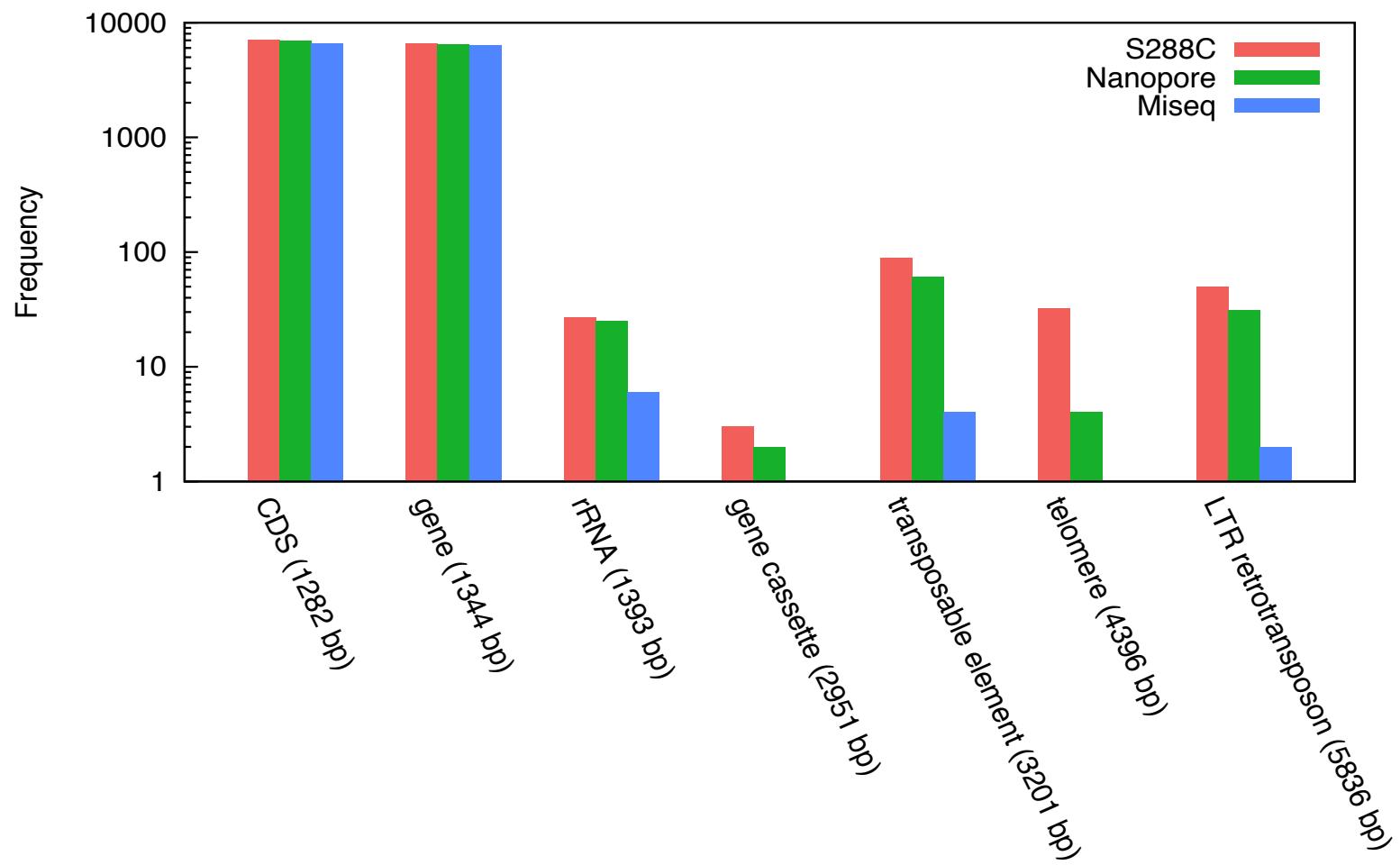


An assembly generated from Oxford Nanopore long reads is better able to identify genomic features



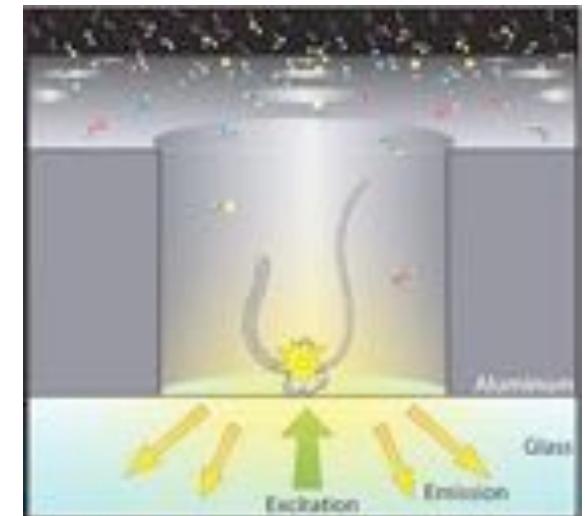
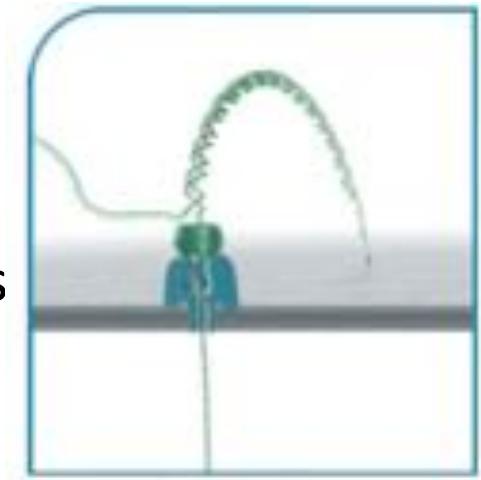
- S288C is an extremely high quality reference
 - In virtually all cases the Oxford estimate of the frequency of a genomic feature is closer to S288C than data generated by miSeq
 - In some cases (gene cassette, telomere) the miSeq is completely unable to detect features

ONT Assembly Completeness



Summary

1. New Single Molecule Sequencing Technologies
2. Produce very long reads
3. Have High Error rate -> Error Correction
4. Long reads Produce great assemblies, far better than short read technologies -> repeat resolution



Future of Oxford Nanopore



PromethION Setup

PromethION is a standalone benchtop instrument that includes substantial on-board computing to enable very high throughput real-time analyses. It is compatible with the cloud-based analysis service from Metrichor.

PromethION contains docking for 48 flow cells. Each flow cell contains a nanopore sensor array enabling 3,000 nanopores, so a total of 144,000 on the instrument. The sensor array interfaces with an ASIC within the instrument for signal processing. Each flow cell also allows for multiple samples to be processed separately.

Future for Pacbio

Sequel System: high-throughput, cost-effective access to SMRT Sequencing



The Sequel System is ideal for projects such as rapidly and cost-effectively generating high-quality whole genome *de novo* assemblies.

[Learn more](#) about the Sequel System.

Acknowledgements



Cold Spring Harbor Laboratory

Michael Schatz

Dick McCombie

Sara Goodwin

Schatz Lab

Tyler Garvin

Han Fang

Hayan Lee

Maria Nattestad

Srividya Ramakrishnan

