The Resurgence of Reference Quality Genomes

Michael Schatz

April 9, 2015

UMN-MSI: Advances in Genome Assembly

Outline

1. Assembly Fundamentals

2. PacBio Sequencing of Rice

3. Oxford Nanopore Sequencing of Yeast

Outline

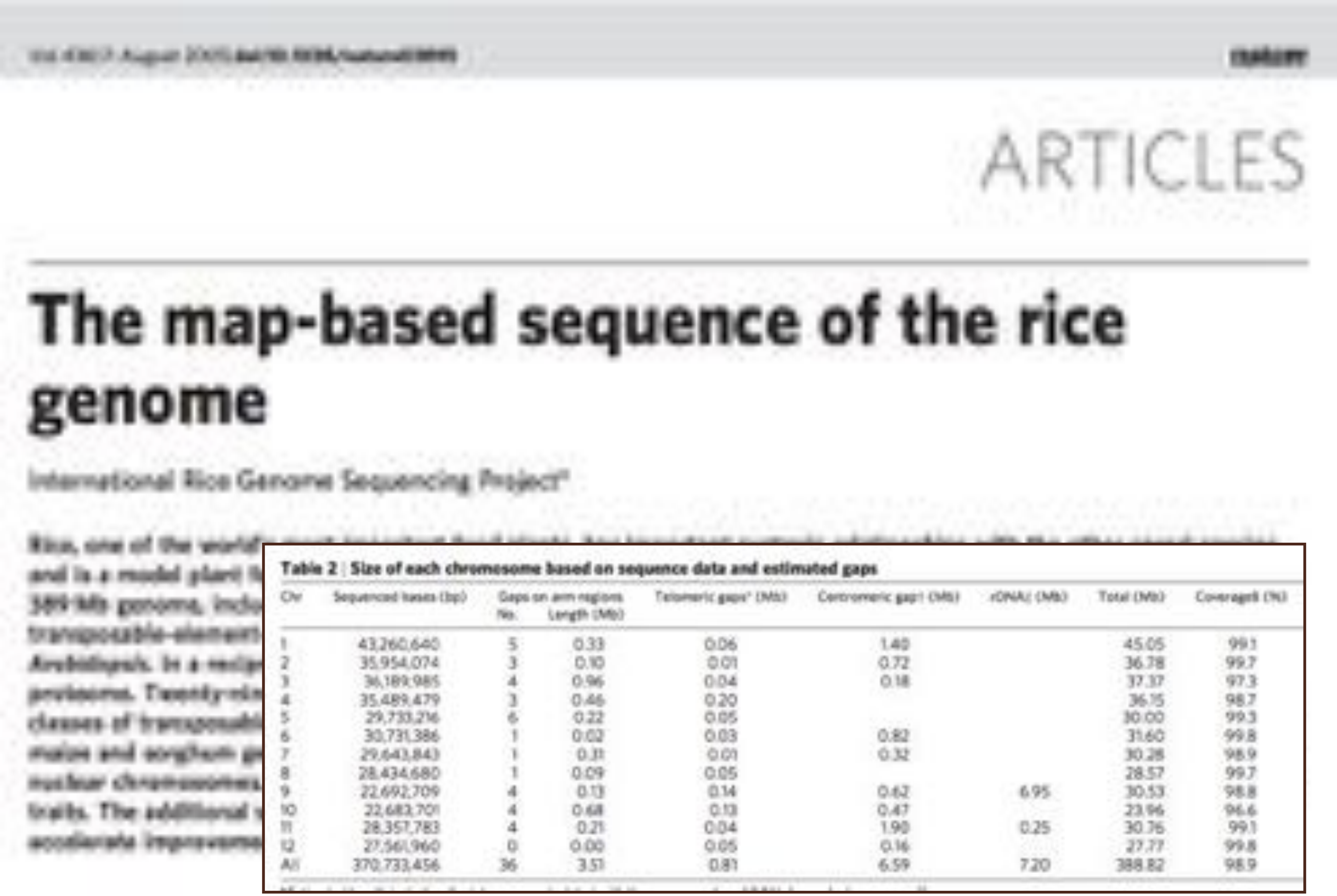
1. ~~Assembly Fundamentals~~

   Thanks Jason!

2. PacBio Sequencing of Rice

   and Human Cancer

3. Oxford Nanopore Sequencing of Yeast



Vol 436|11 August 2005|doi:10.1038/nature03895

nature

ARTICLES

The map-based sequence of the rice genome

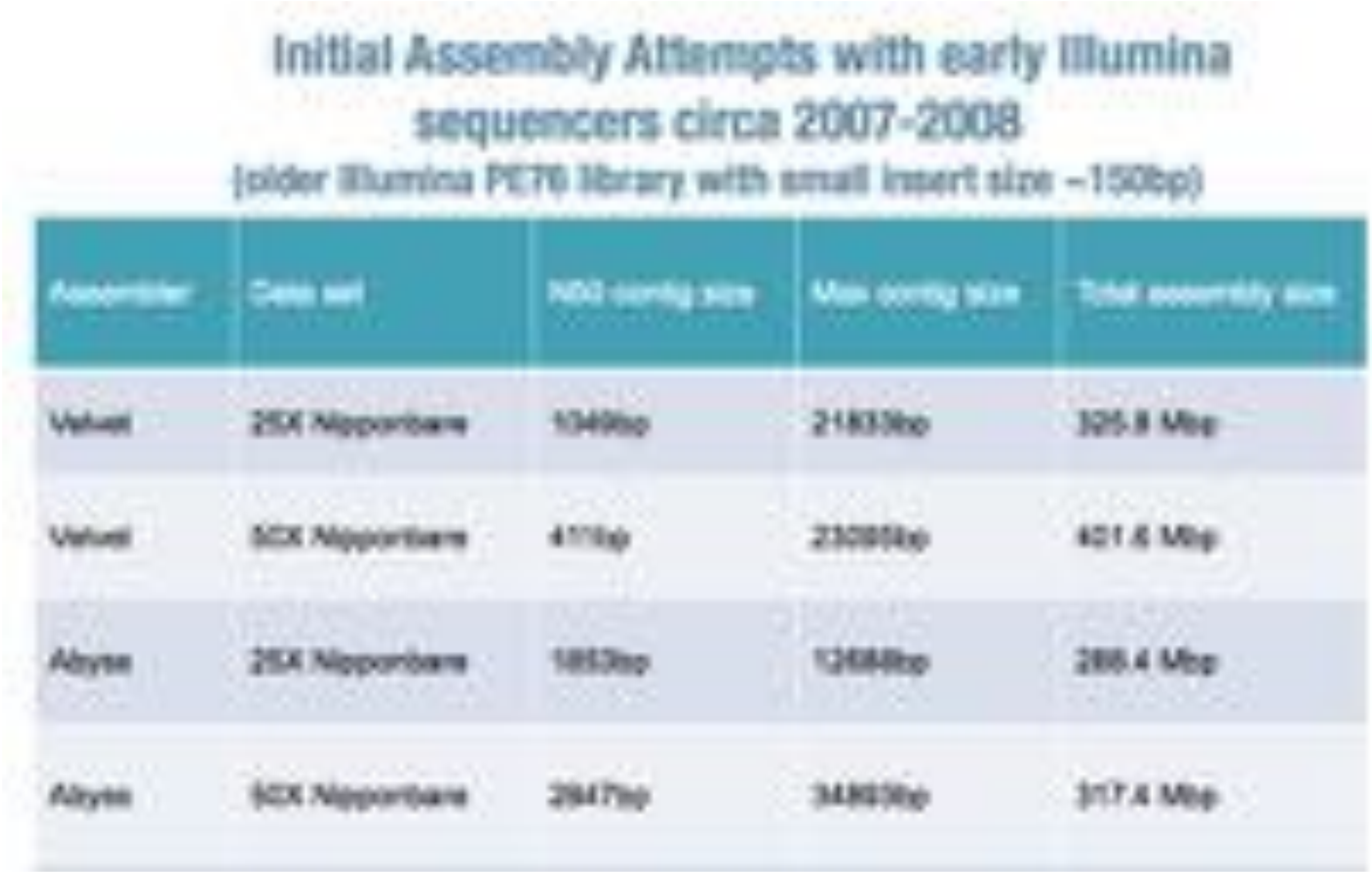
International Rice Genome Sequencing Project*

Rice, one of the world's most important food plants, has important syntenic relationships with the other cereal species and is a model plant
389-Mb genome, inclu
transposable-element
Arabidopsis. In a recip
proviruses. Twenty-nin
classes of transposable
maize and sorghum ge
nuclear chromosomes
traits. The additional
accelerate improveme

Table 2 | Size of each chromosome based on sequence data and estimated gaps

Chr	Sequenced bases (bp)	Gaps on arm regions		Telomeric gaps* (Mb)	Centromeric gap† (Mb)	rDNA‡ (Mb)	Total (Mb)	Coverage§ (%)
		No.	Length (Mb)					
1	43,260,640	5	0.33	0.06	1.40		45.05	99.1
2	35,954,074	3	0.10	0.01	0.72		36.78	99.7
3	36,189,985	4	0.96	0.04	0.18		37.37	97.3
4	35,489,479	3	0.46	0.20			36.15	98.7
5	29,733,216	6	0.22	0.05			30.00	99.3
6	30,731,386	1	0.02	0.03	0.82		31.60	99.8
7	29,643,843	1	0.31	0.01	0.32		30.28	98.9
8	28,434,680	1	0.09	0.05			28.57	99.7
9	22,692,709	4	0.13	0.14	0.62	6.95	30.53	98.8
10	22,683,701	4	0.68	0.13	0.47		23.96	96.6
11	28,357,783	4	0.21	0.04	1.90	0.25	30.76	99.1
12	27,561,960	0	0.00	0.05	0.16		27.77	99.8
All	370,733,456	36	3.51	0.81	6.59	7.20	388.82	98.9

Contig N50: 5.1Mbp
Total projects costs: >$100M



Initial Assembly Attempts with early Illumina sequencers circa 2007-2008

(older Illumina PE76 library with small insert size ~150bp)

Assembler	Data set	N50 contig size	Max contig size	Total assembly size
Velvet	25X Nipponbare	1049bp	21833bp	325.8 Mbp
Velvet	50X Nipponbare	411bp	23095bp	401.6 Mbp
Abyss	25X Nipponbare	1853bp	12684bp	288.4 Mbp
Abyss	50X Nipponbare	2947bp	34803bp	317.4 Mbp

Total costs: ~$10k
>1,000x times cheaper, but at what cost scientifically?

W.R. McCombie

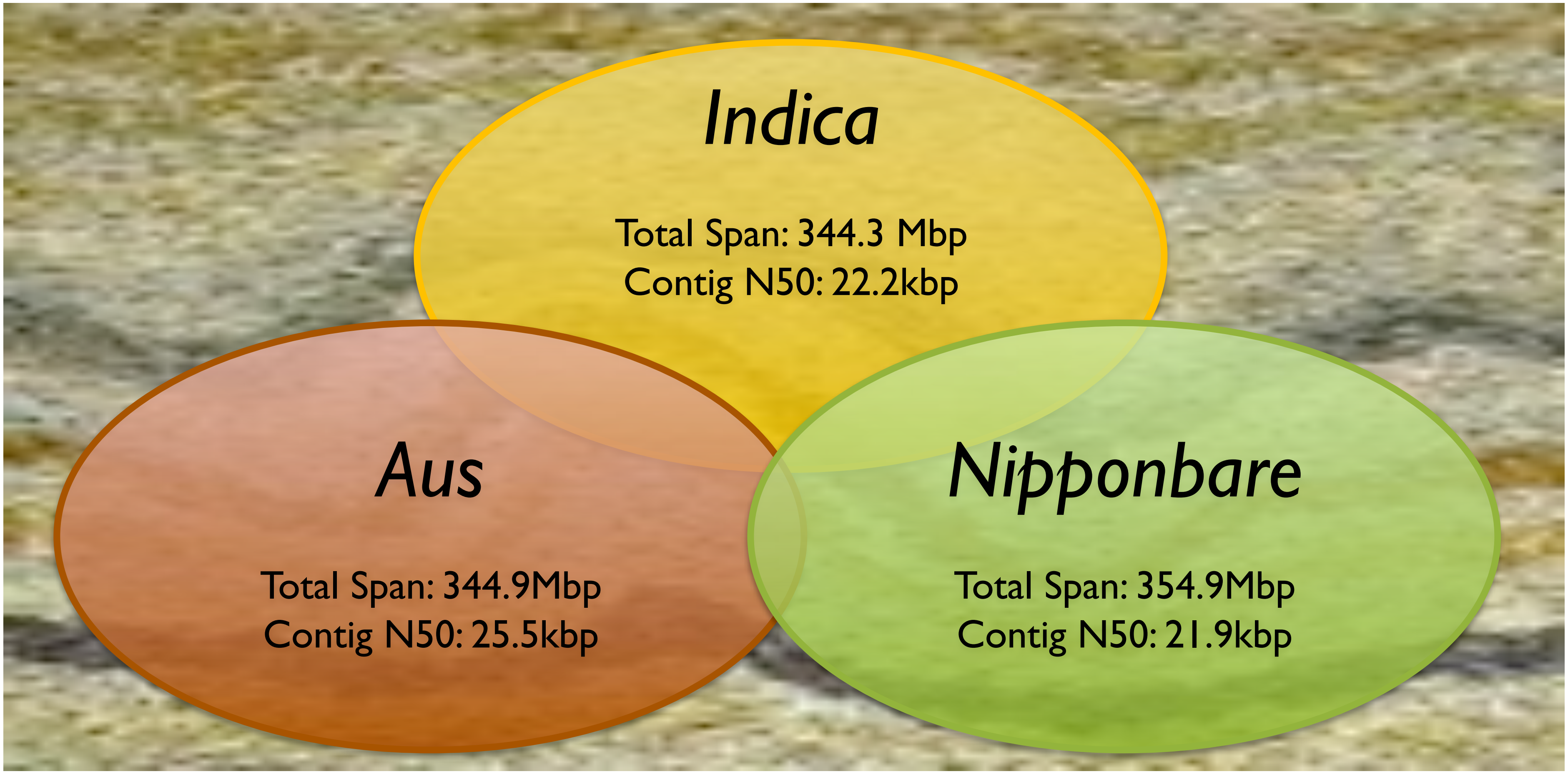
Genomics Arsenal in the year 2015

Sample Preparation

Sequencing

Chromosome Mapping

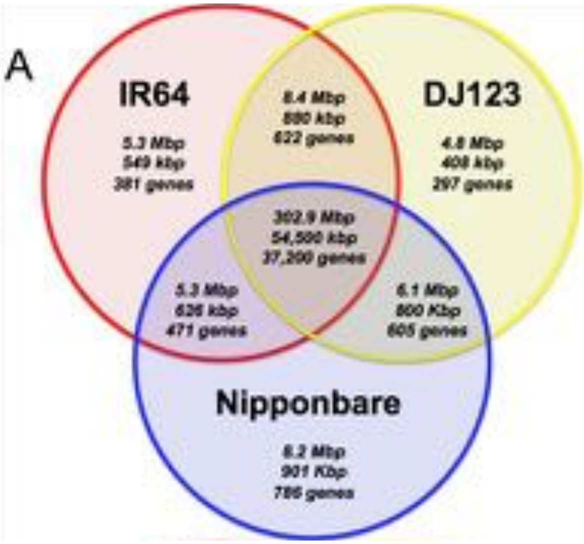
Population structure of *Oryza sativa*



Whole genome de novo assemblies of three divergent strains of rice (*O. sativa*) documents novel gene space of *aus* and *indica*
Schatz, Maron, Stein et a*l* (2014) *Genome Biology.* 15:506 doi:10.1186/s13059-014-0506-z

Oryza sativa Gene Diversity

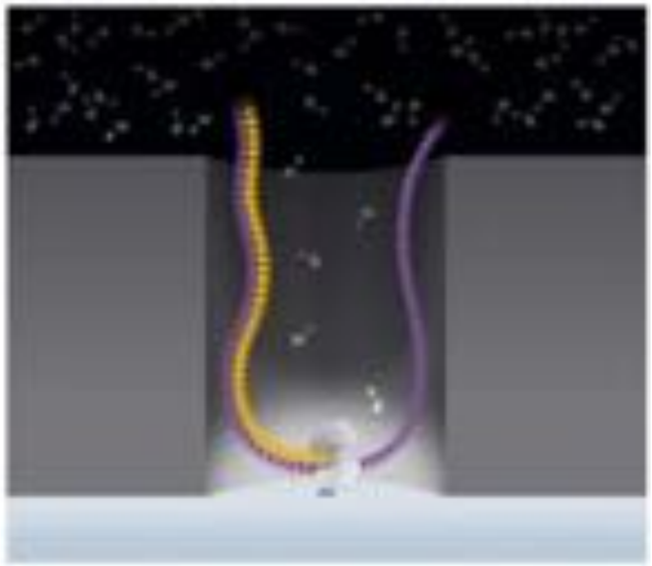
- Very high quality representation of the "gene-space"
  - Overall identity ~99.9%
  - Less than 1% of exonic bases missing
- Genome-specific genes enriched for disease resistance
  - Reflects their geographic and environmental diversity
- Assemblies fragmented at (high copy) repeats
  - Difficult to identify full length gene models and regulatory features

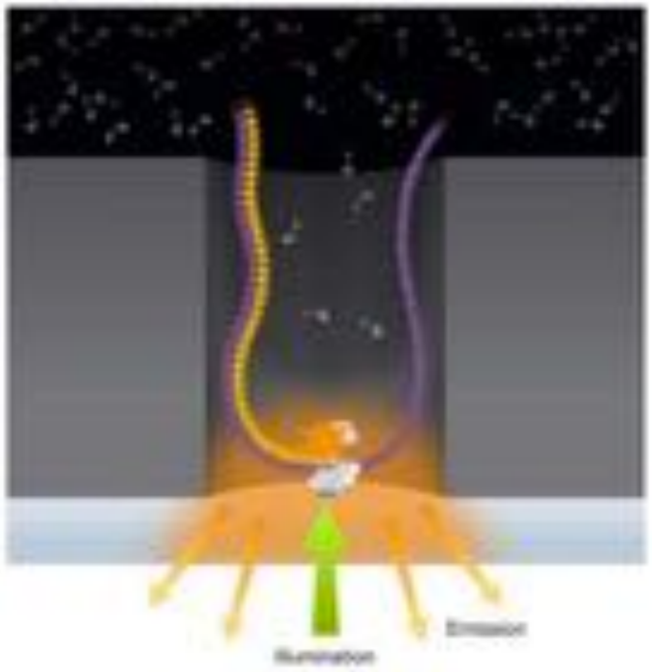


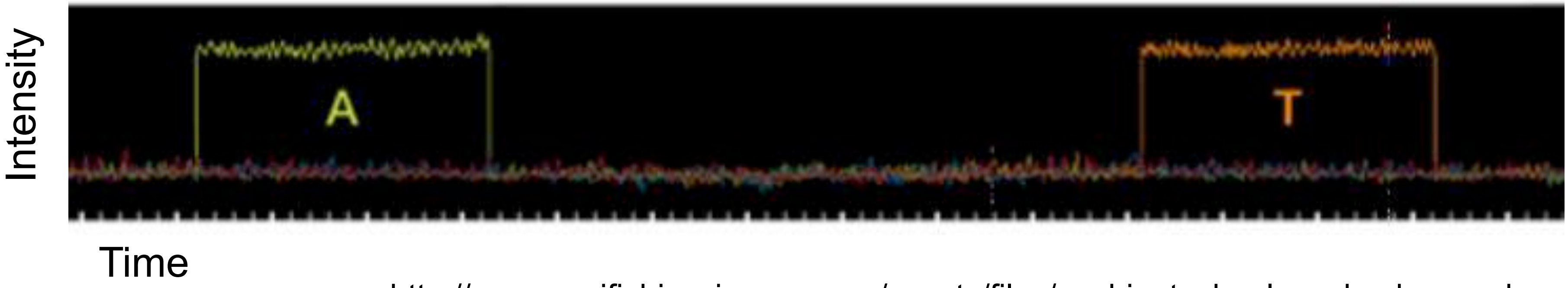
Overall sequence content
In each sector, the top number is the total number of base pairs, the middle number is the number of exonic bases, and the bottom is the gene count. If a gene is partially shared, it is assigned to the sector with the most exonic bases.

PacBio SMRT Sequencing

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

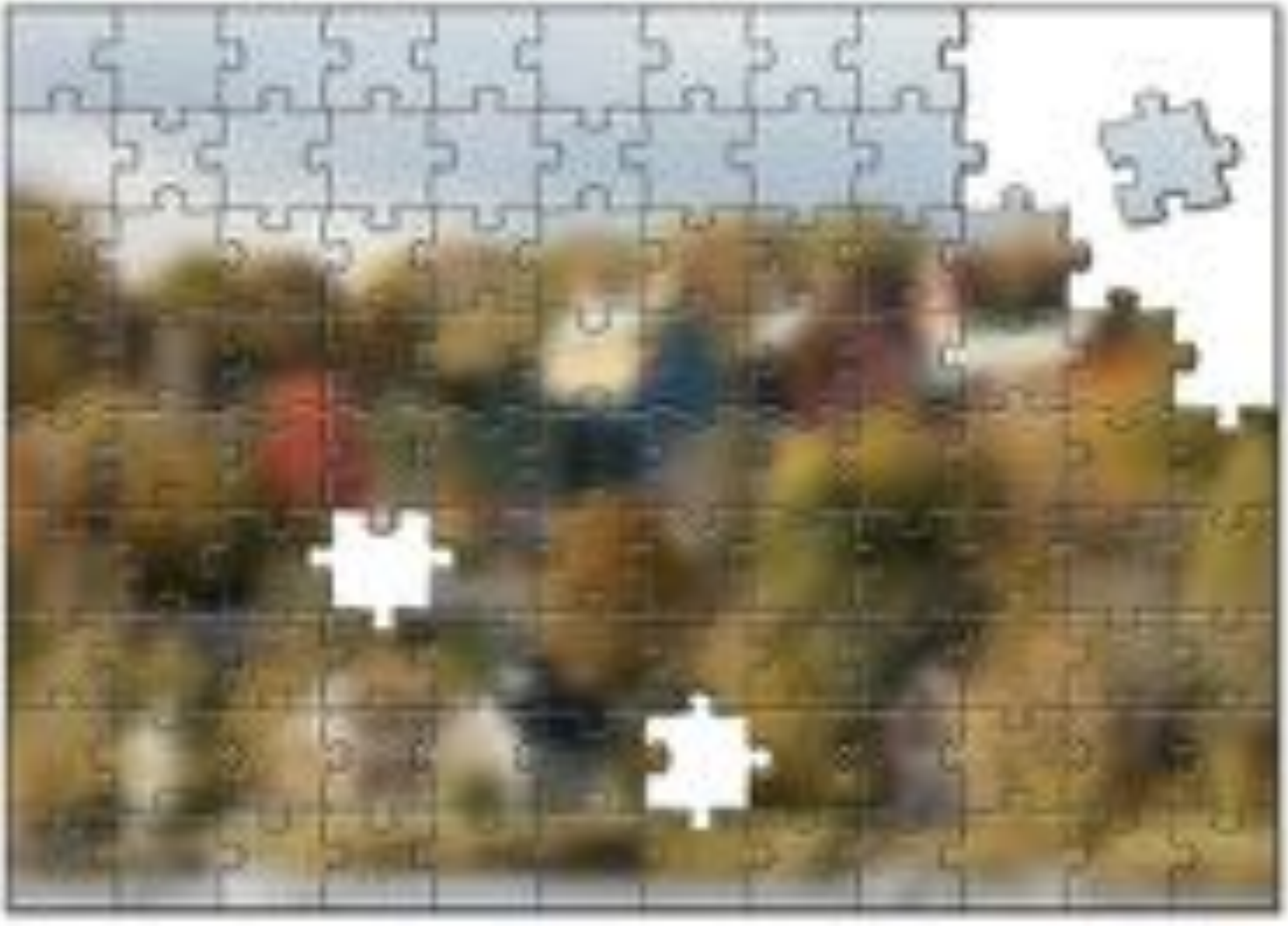




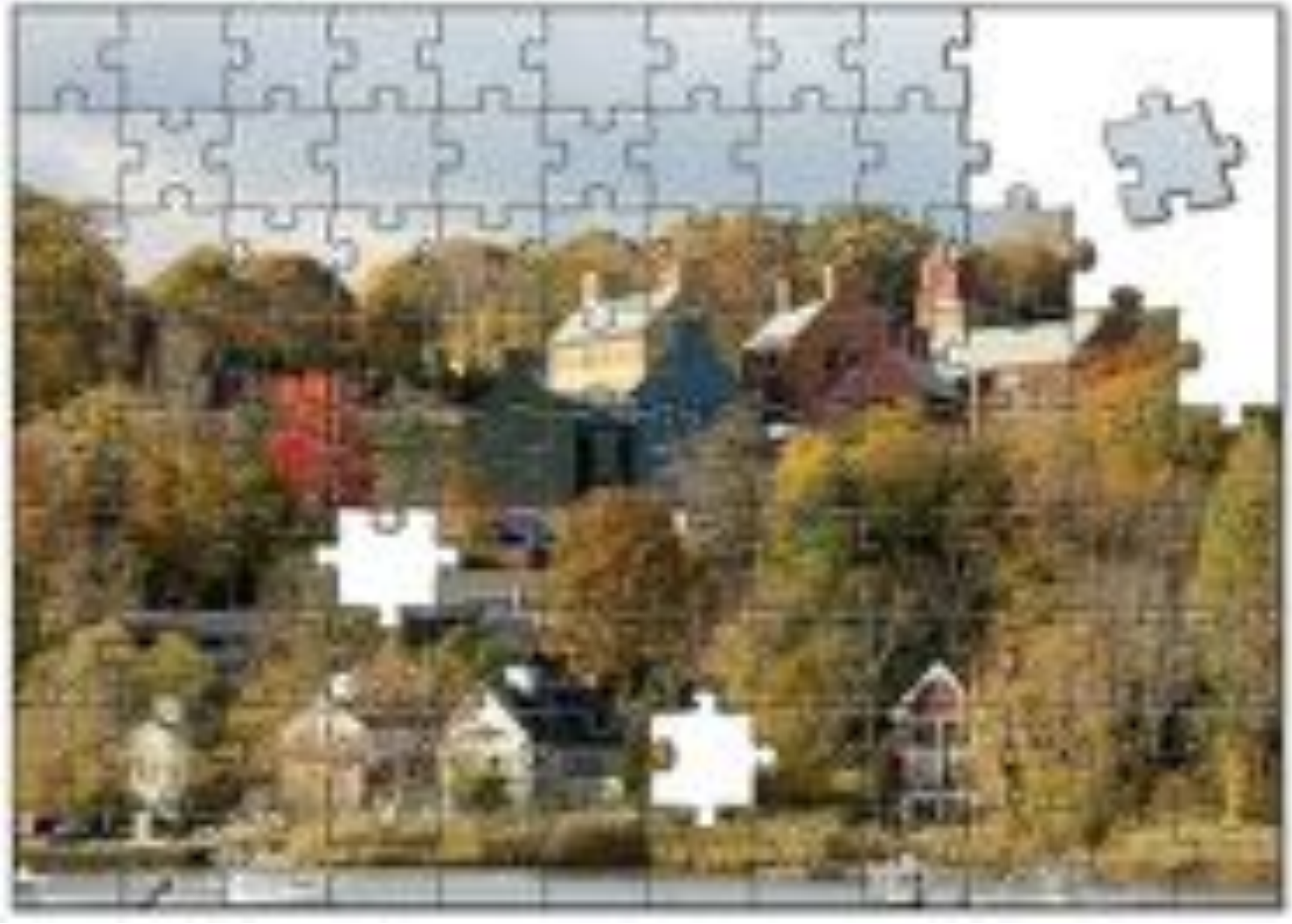


http://www.pacificbiosciences.com/assets/files/pacbio_technology_backgrounder.pdf

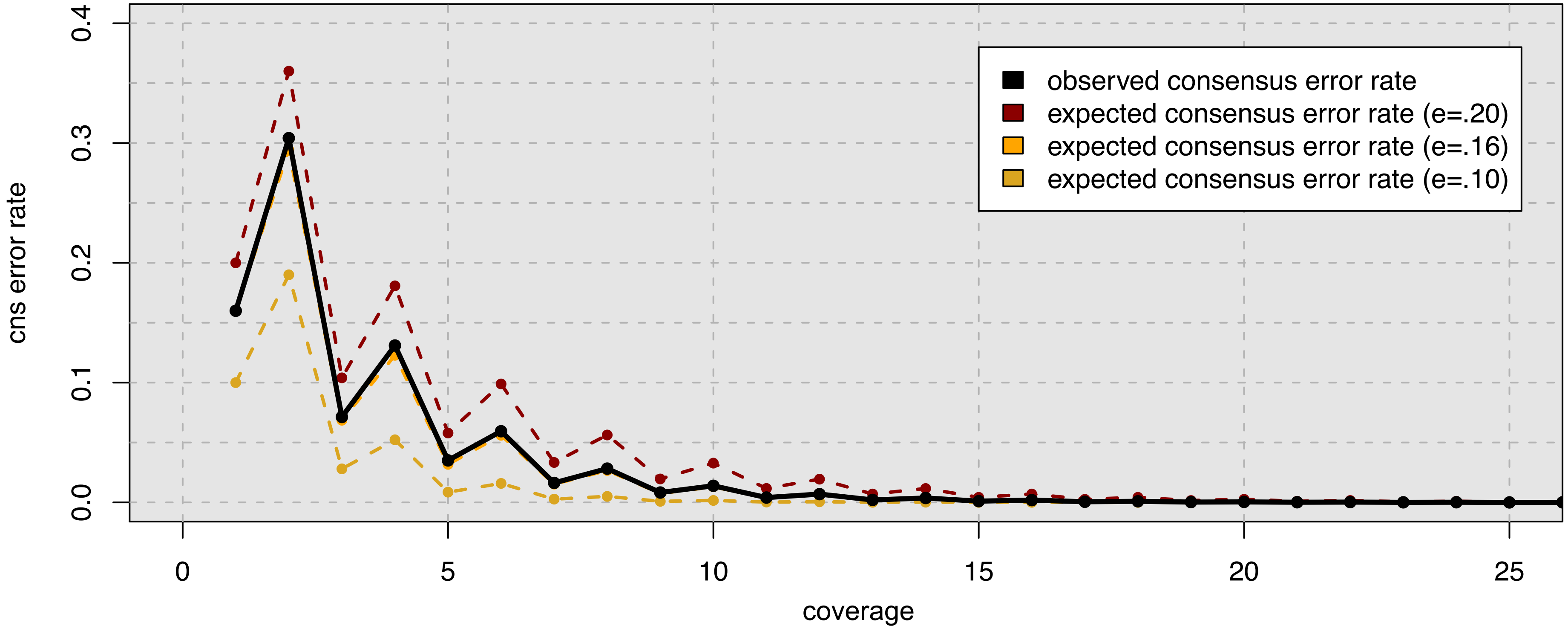
Single Molecule Sequences



“Corrective Lens” for Sequencing



Consensus Accuracy and Coverage



Coverage can overcome random errors

- Dashed: error model from binomial sampling
- Solid: observed accuracy

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

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

PacBio Assembly Algorithms

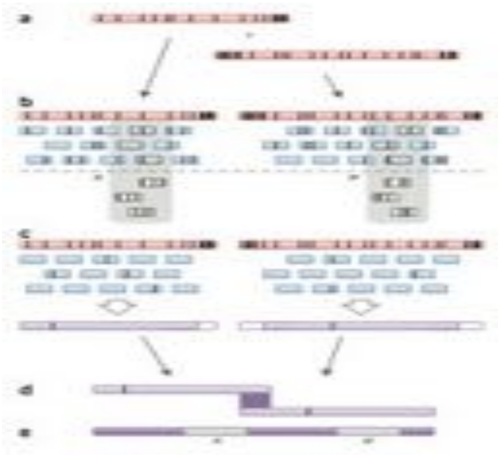
PBJelly

A) Pipeline
B) Support
Scaffold
C) Assembly

Gap Filling and Assembly Upgrade

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

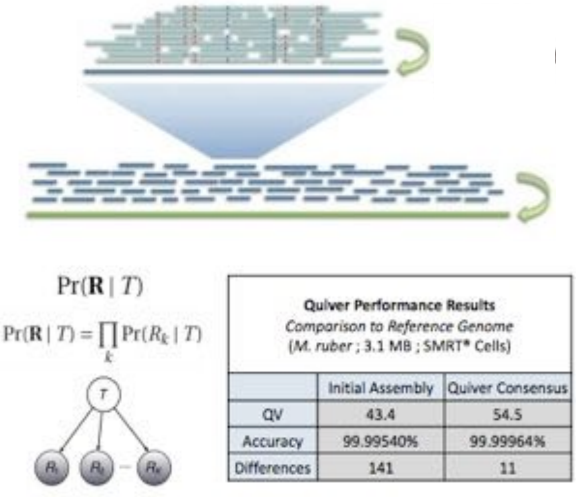
PacBioToCA & ECTools



Hybrid/PB-only 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

PB-only Correction & Polishing

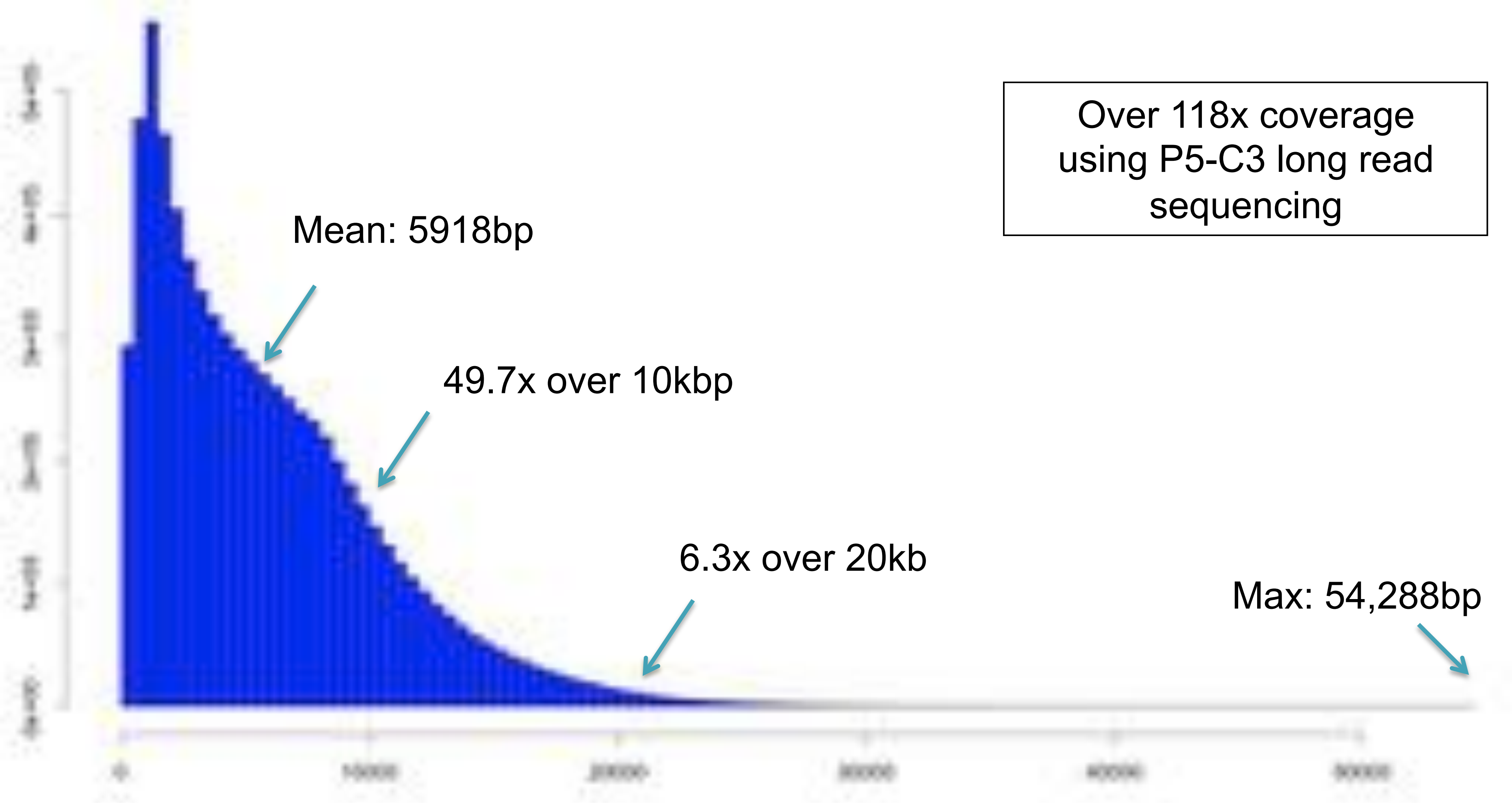
Chin *et al* (2013)
Nature Methods. 10:563–569

< 5x PacBio Coverage > 50x

O. sativa pv Indica (IR64)

PacBio RS II sequencing at PacBio

- Size selection using an 10 Kb elution window on a BluePippin™ device from Sage Science



Mean: 5918bp

49.7x over 10kbp

6.3x over 20kb

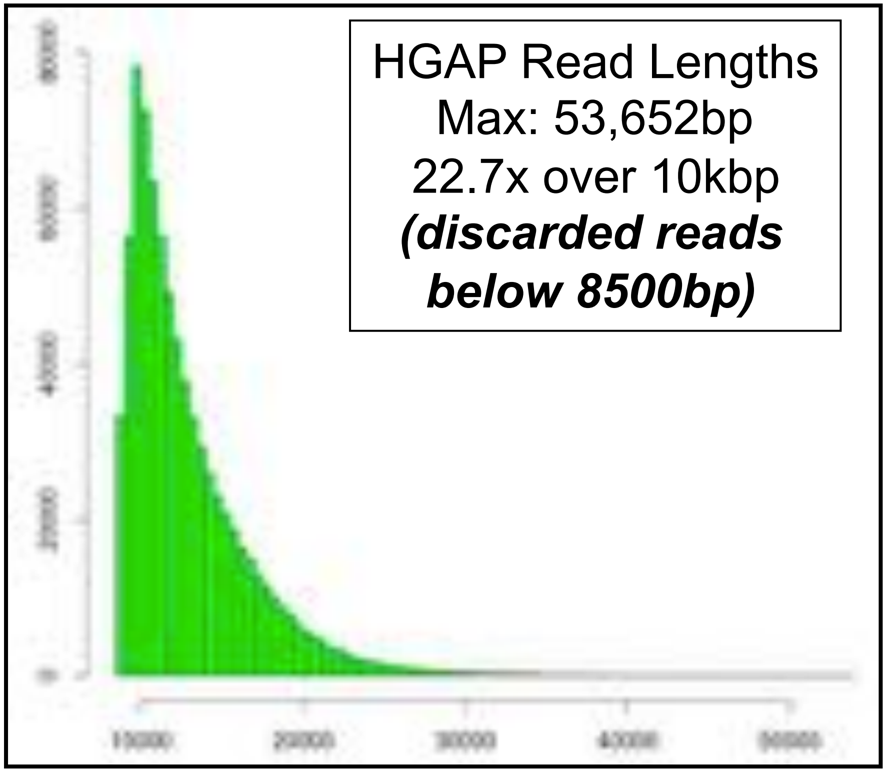
Max: 54,288bp

Over 118x coverage using P5-C3 long read sequencing

O. sativa pv Indica (IR64)

Genome size: ~370 Mb
Chromosome N50: ~29.7 Mbp

Assembly	Contig NG50
MiSeq Fragments 25x 456bp (3 runs 2x300 @ 450 FLASH)	19 kbp
"ALLPATHS-recipe" 50x 2x100bp @ 180 36x 2x50bp @ 2100 51x 2x50bp @ 4800	18 kbp
HGAP + CA 22.7x @ 10kbp	4.0 Mbp
Nipponbare BAC-by-BAC Assembly	5.1 Mbp



HGAP Read Lengths
Max: 53,652bp
22.7x over 10kbp
(discarded reads below 8500bp)

S5 Hybrid Sterility Locus

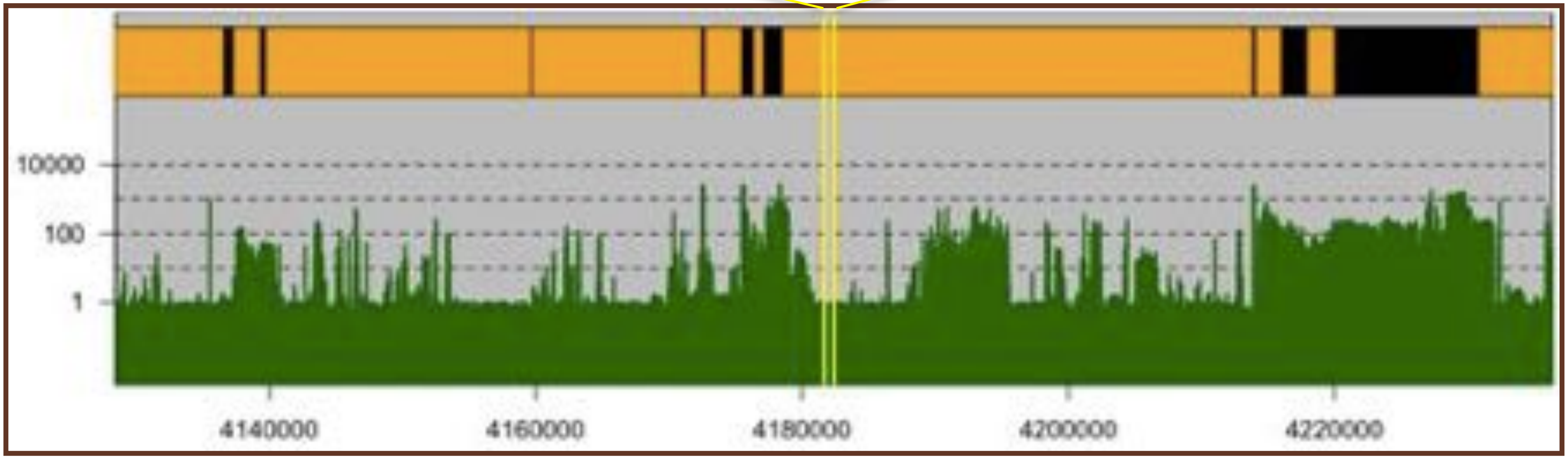
```
Sanger      …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
Illumina    …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
PacBio      …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
```

S5 is a major locus for hybrid sterility in rice that affects embryo sac fertility.

- Genetic analysis of the S5 locus documented three alleles: an indica (S5-i), a japonica (S5-j), and a neutral allele (S5-n)
- Hybrids of genotype S5-i/S5-j are mostly sterile, whereas hybrids of genotypes consisting of S5-n with either S5-i or S5-j are mostly fertile.
- Contains three tightly linked genes that work together in a 'killer-protector'-type system: ORF3, ORF4, ORF5
- The ORF5 indica (ORF5+) and japonica (ORF5-) alleles differ by only **two nucleotides**

S5 Hybrid Sterility Locus

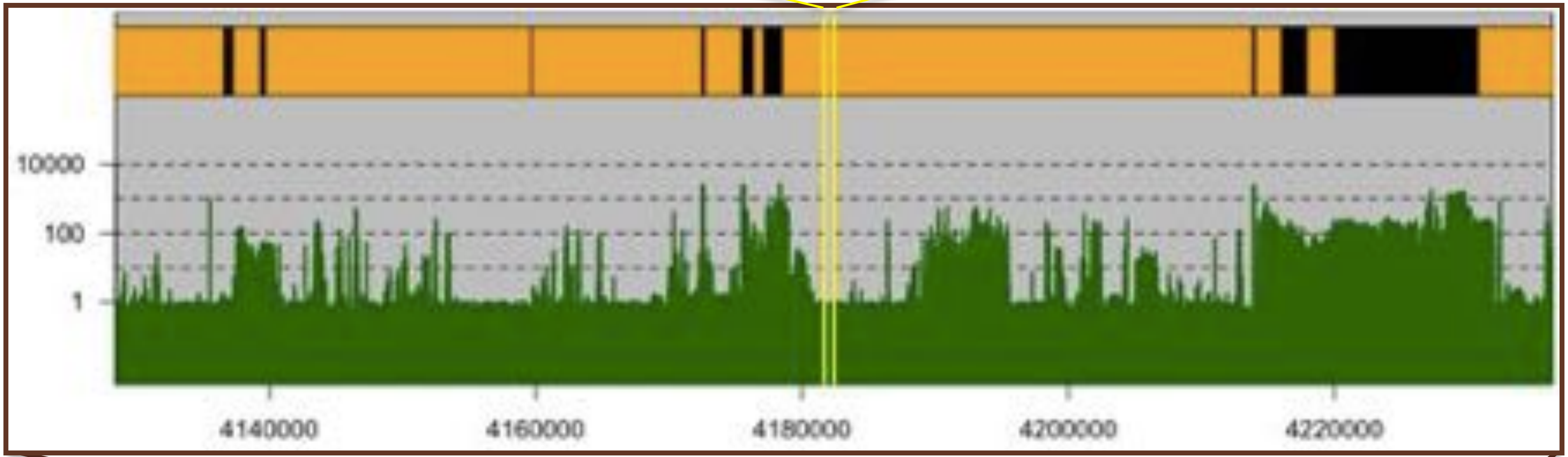
```
Sanger      …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
Illumina    …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
PacBio      …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
```

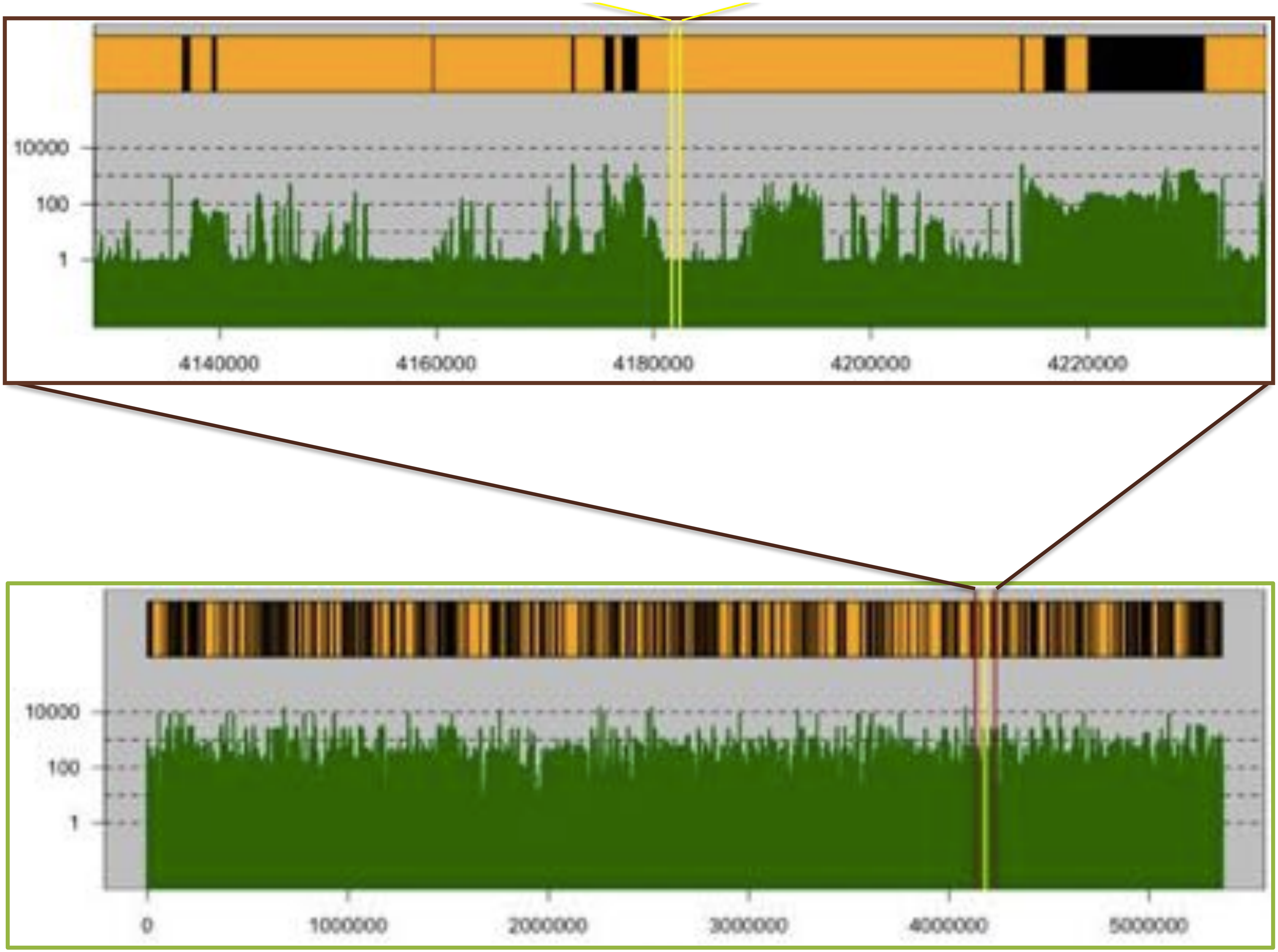


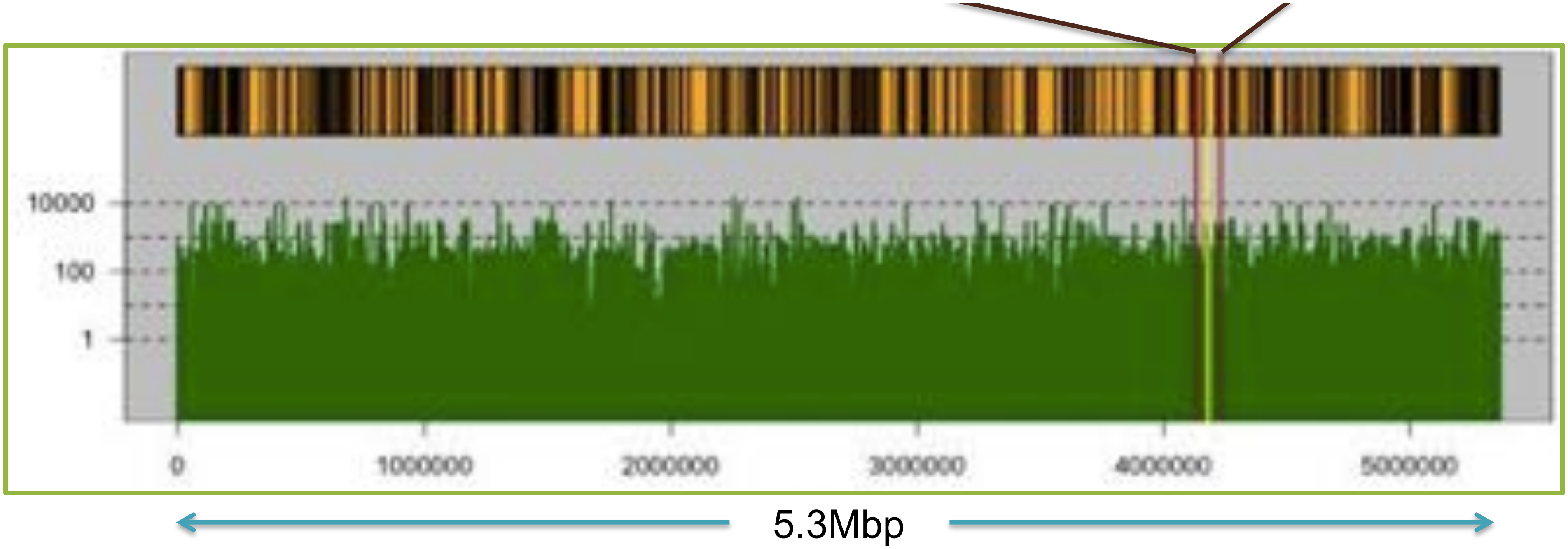
100kbp

S5 Hybrid Sterility Locus

```
Sanger      …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
Illumina    …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
PacBio      …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
```




10000
100
1
4140000
4160000
4180000
4200000
4220000
10000
100
1
0
1000000
2000000
3000000
4000000
5000000

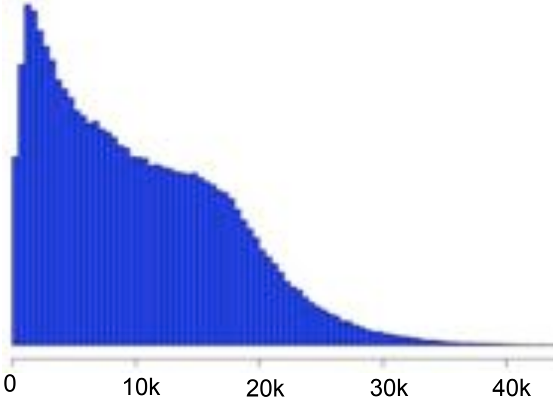


Improvements from 20kbp to 4Mbp contig N50:

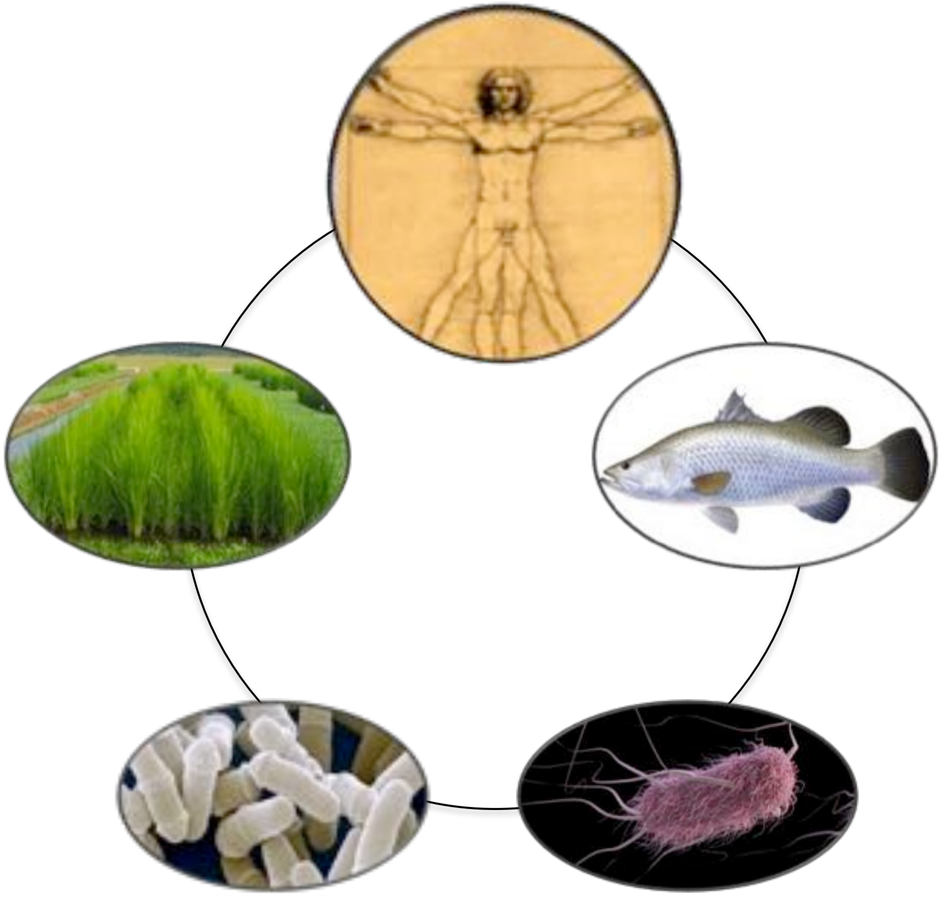
- Over 20 Megabases of additional sequence
  - Extremely high sequence identity (>99.9%)
  - Thousands of gaps filled, hundreds of mis-assemblies corrected
- Complete gene models, promoter regions for nearly every gene
  - True representation of transposons and other complex features
- Opportunities for studying large scale chromosome evolution
  - Largest contigs approach complete chromosome arms

Current Collaborations

PacBio RS II

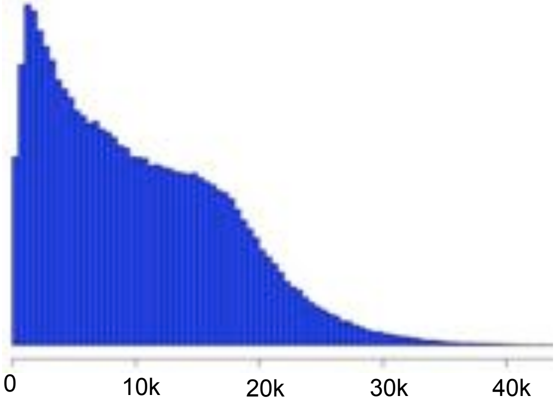


CSHL/PacBio

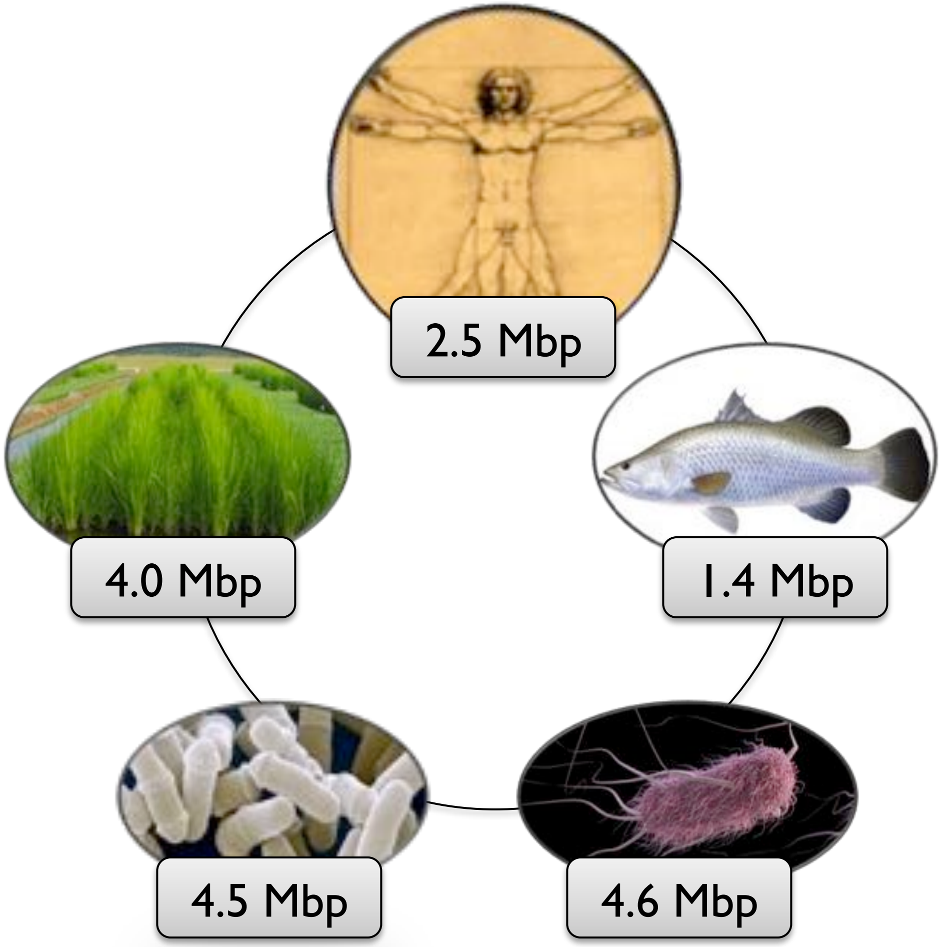


Current Collaborations

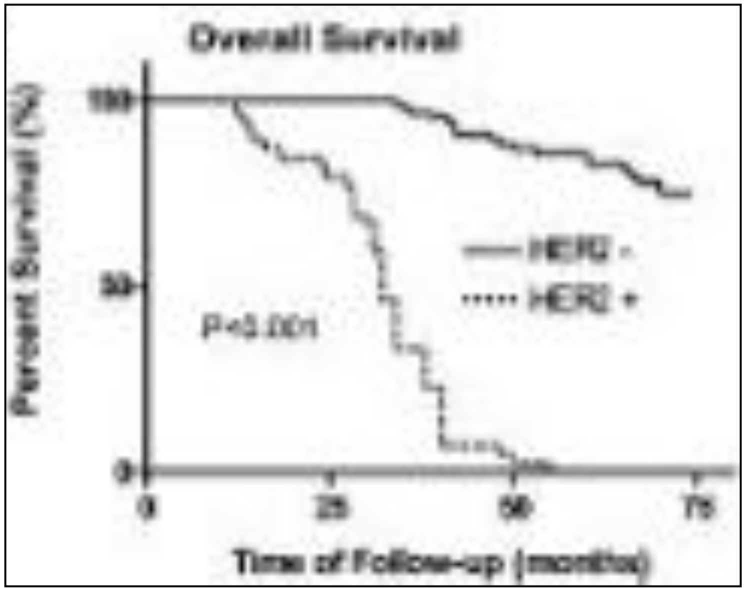
PacBio RS II



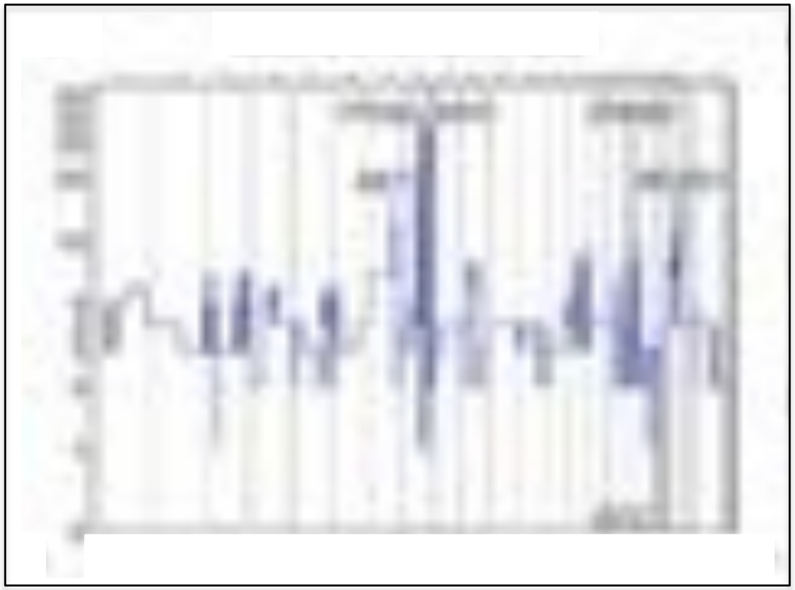
CSHL/PacBio



Long Read Sequencing of SK-BR-3



(Wen-Sheng et al, 2009)



(Navin et al, 2011)

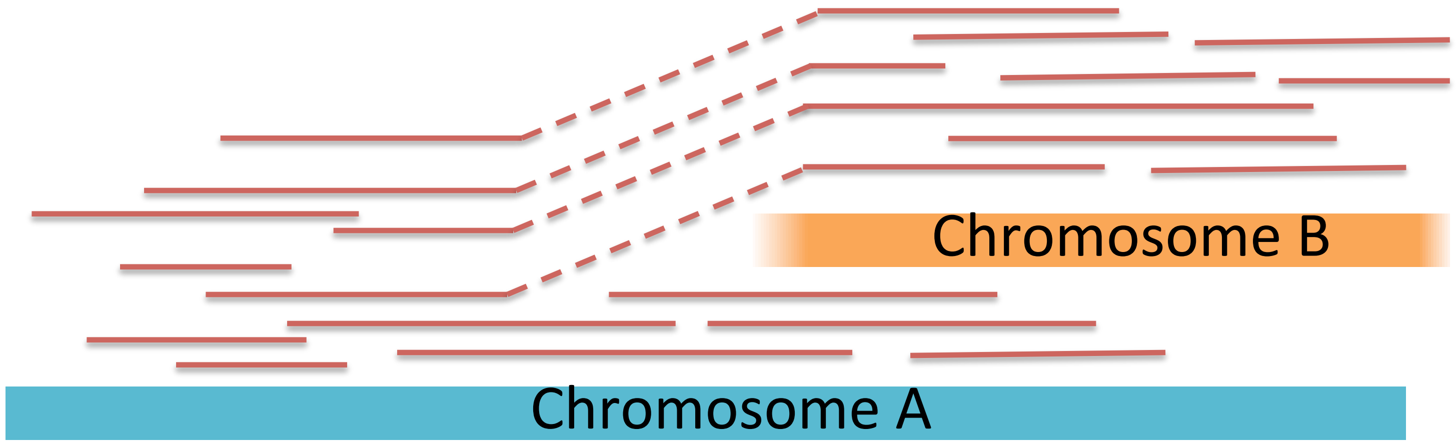
Long read PacBio sequencing of SK-BR-3 breast cancer cell line

- Her2+ breast cancer is one of the most deadly forms of the disease
- SK-BR-3 is one of the most important models, known to have widespread CNVs

- Currently have 72x coverage with long read PacBio sequencing (mean: ~10kbp)
- Analyzing breakpoints in an attempt to infer the mutation history, especially around HER2

In collaboration with McCombie (CSHL) and McPherson (OICR) labs

Structural variant discovery with long reads



1. Alignment-based split read analysis: Efficient capture of most events

BWA-MEM + Lumpy

2. Local assembly of regions of interest: In-depth analysis with *base-pair precision*

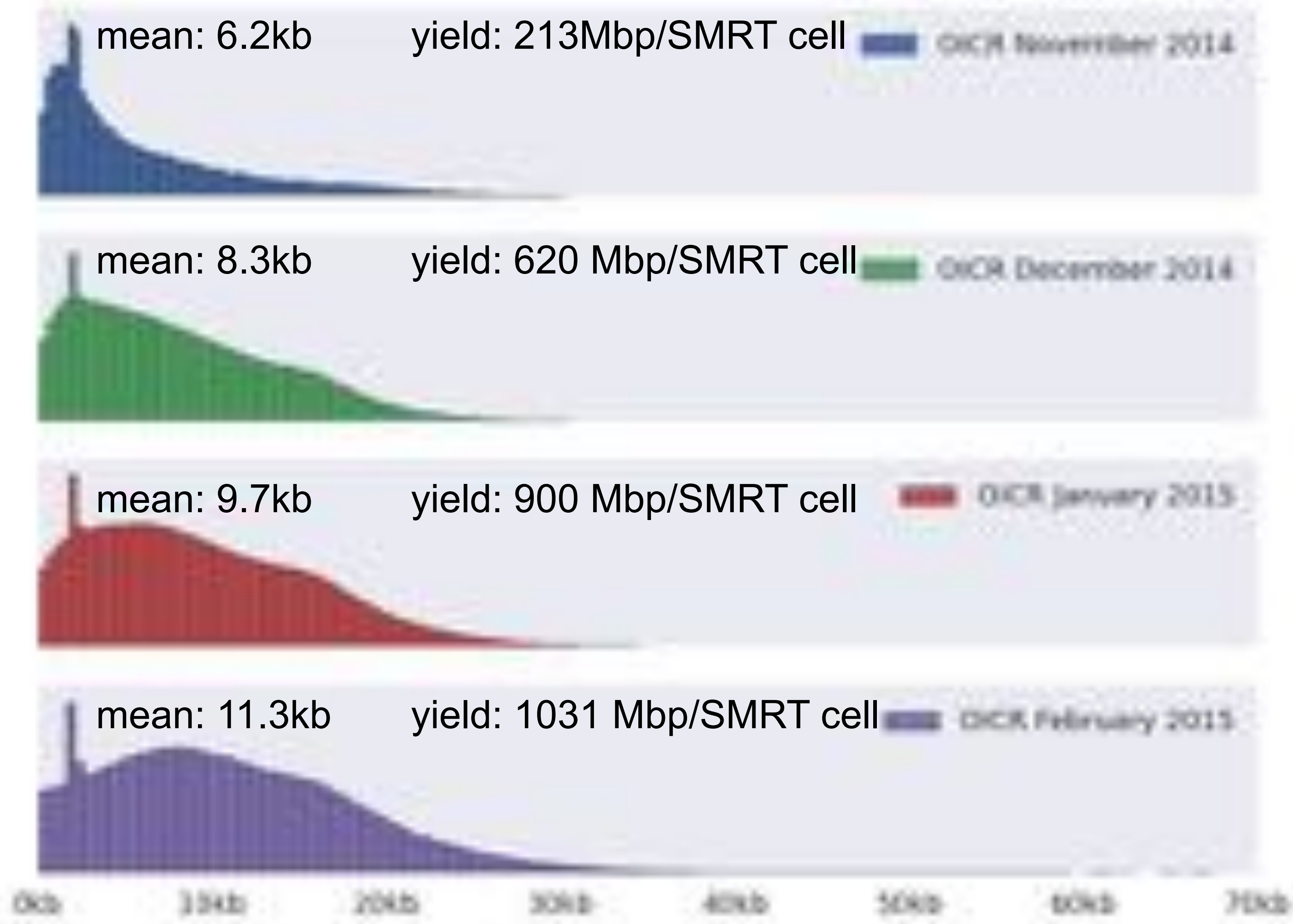
Localized HGAP + Celera Assembler + MUMmer

3. Whole genome assembly: In-depth analysis including *novel sequences*

DNAnexus-enabled version of Falcon

Total Assembly: 2.64Gbp **Contig N50: 2.56 Mbp** **Max Contig: 23.5Mbp**

Improving SMRTcell Performance



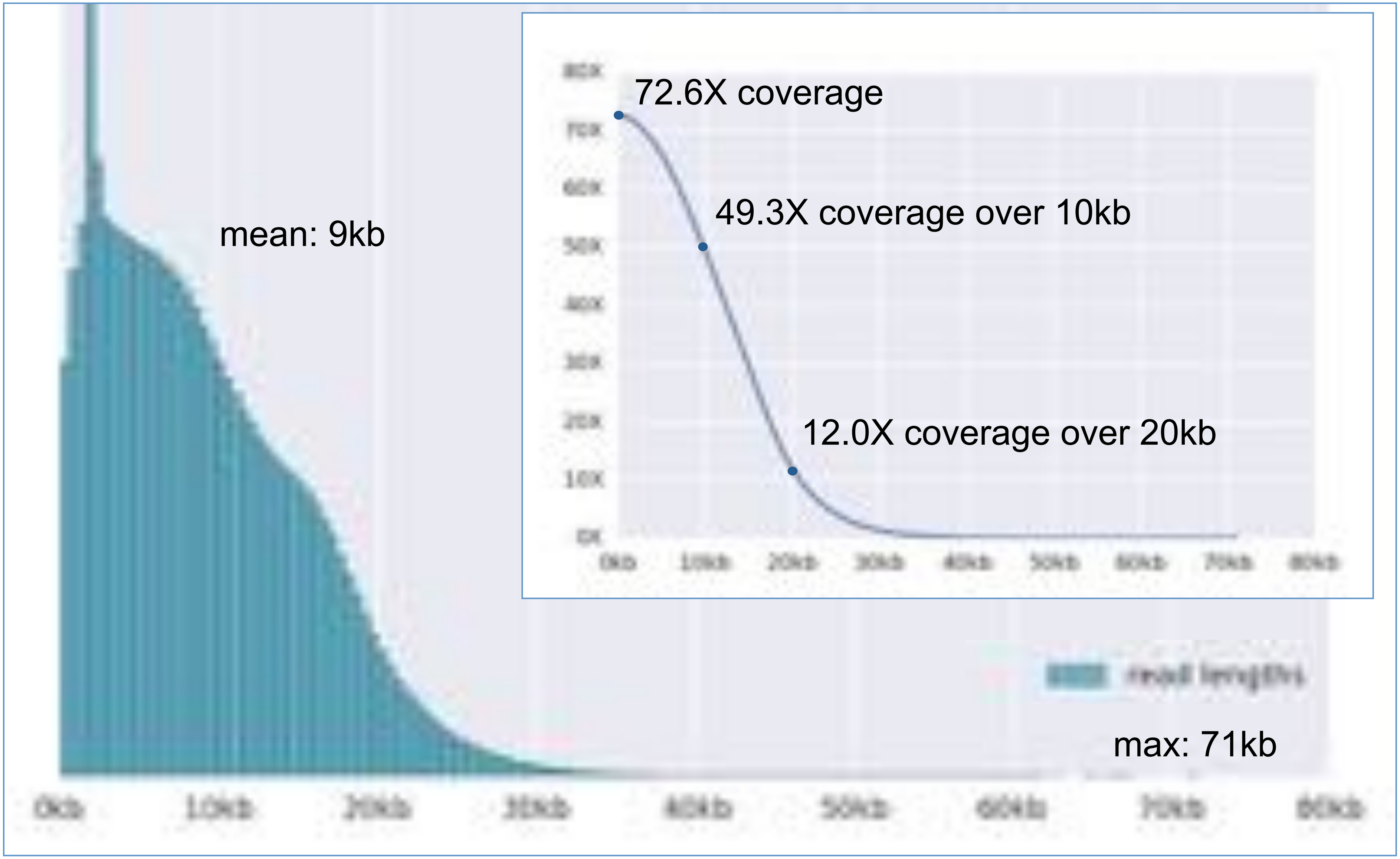
mean: 6.2kb yield: 213Mbp/SMRT cell

mean: 8.3kb yield: 620 Mbp/SMRT cell

mean: 9.7kb yield: 900 Mbp/SMRT cell

mean: 11.3kb yield: 1031 Mbp/SMRT cell

PacBio read length distribution



72.6X coverage

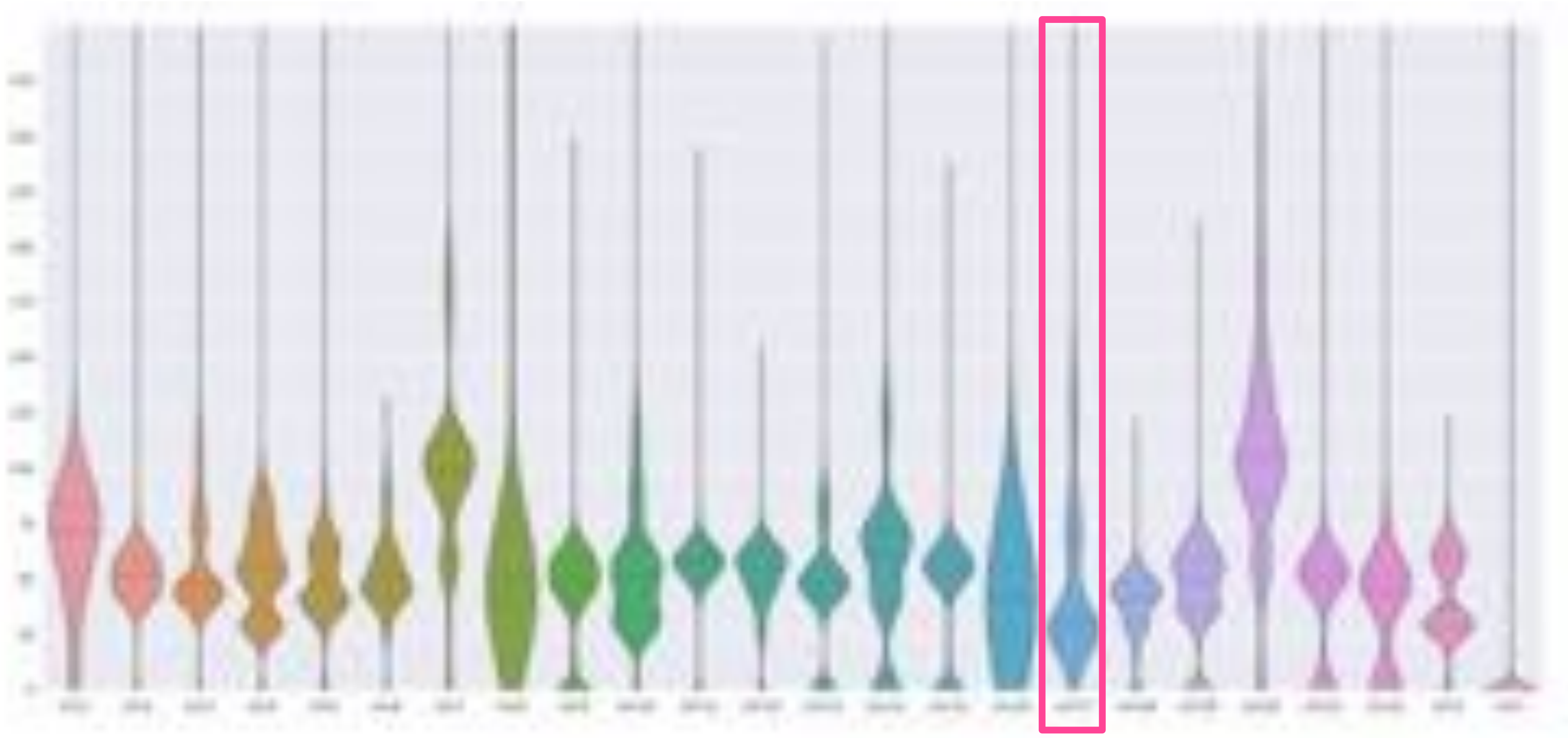
49.3X coverage over 10kb

12.0X coverage over 20kb

mean: 9kb

max: 71kb

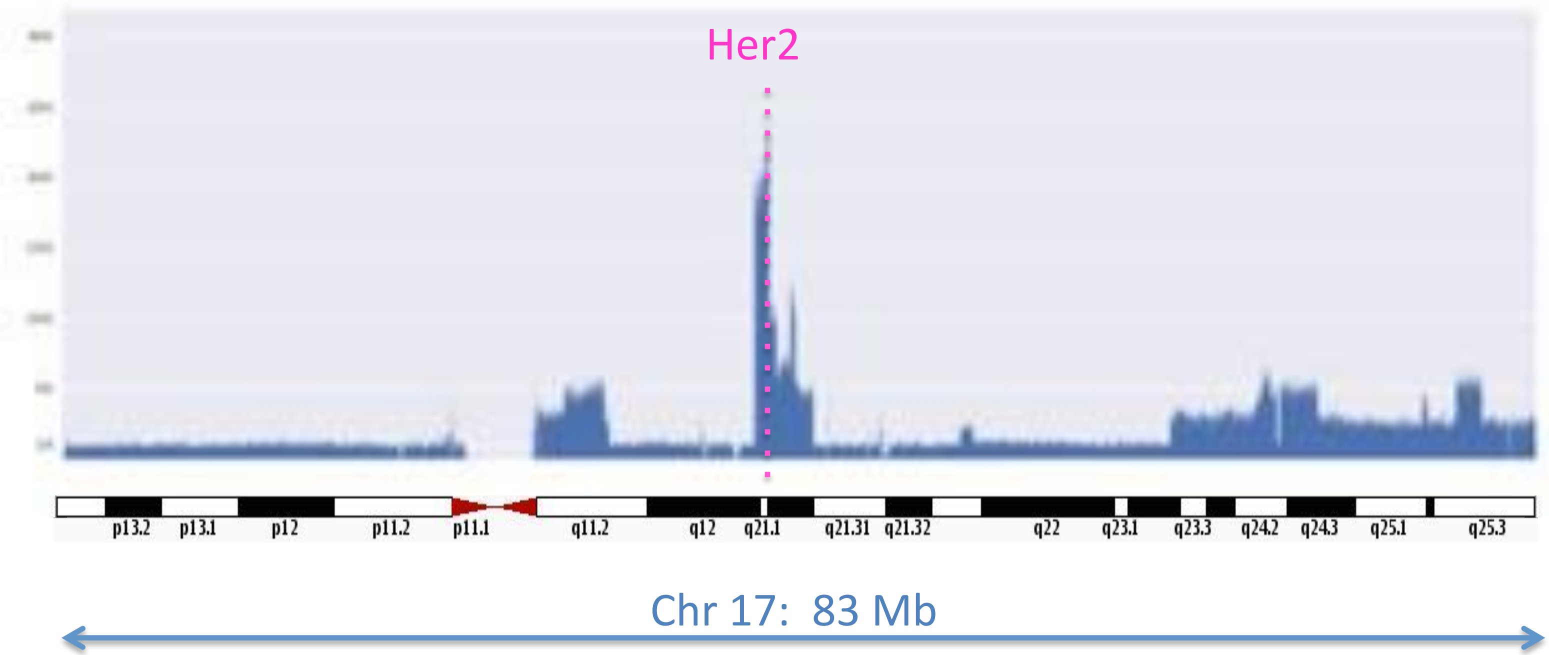
Genome-wide alignment coverage



Genome-wide coverage averages around 54X

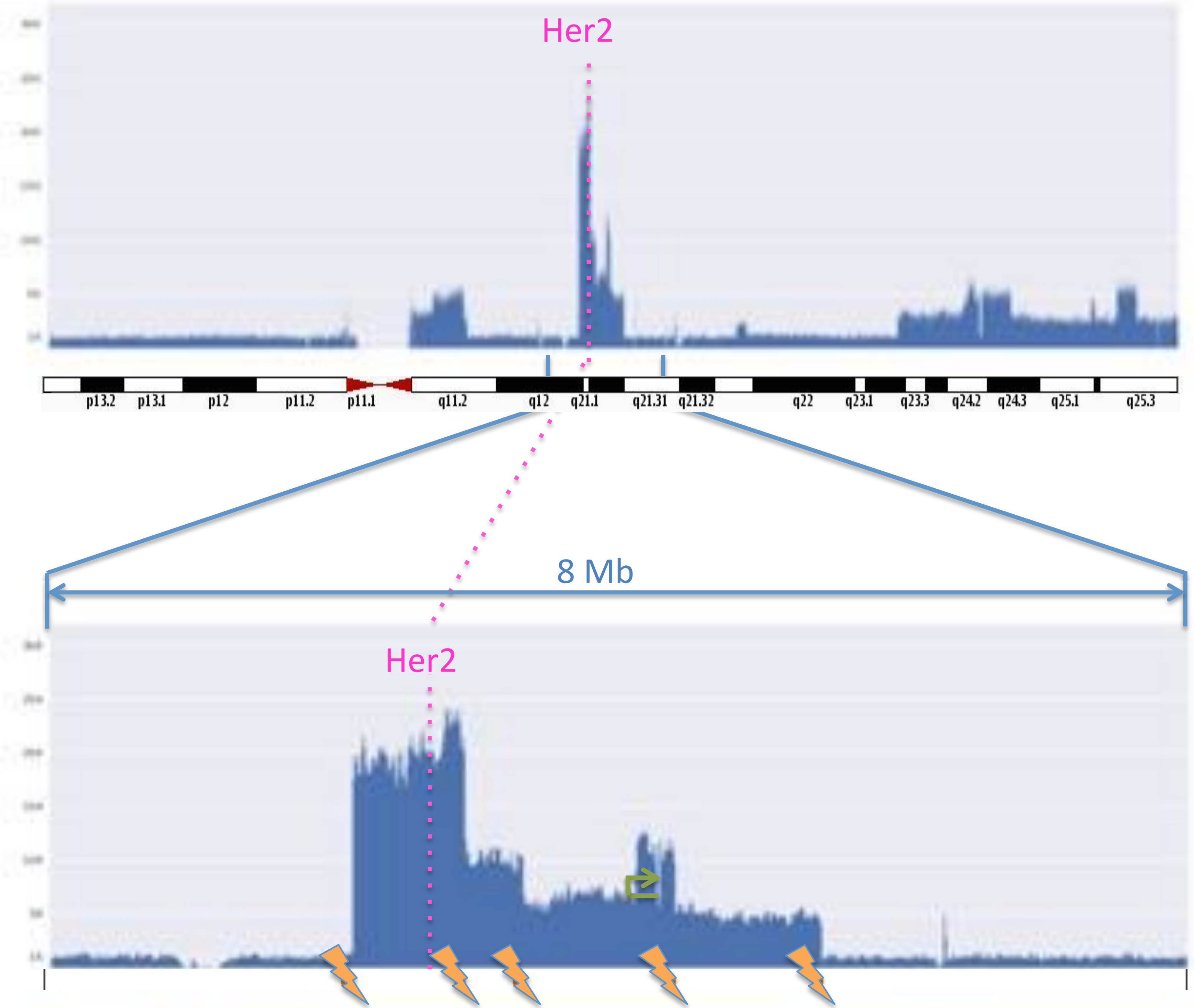
Coverage per chromosome varies greatly as expected from previous karyotyping results

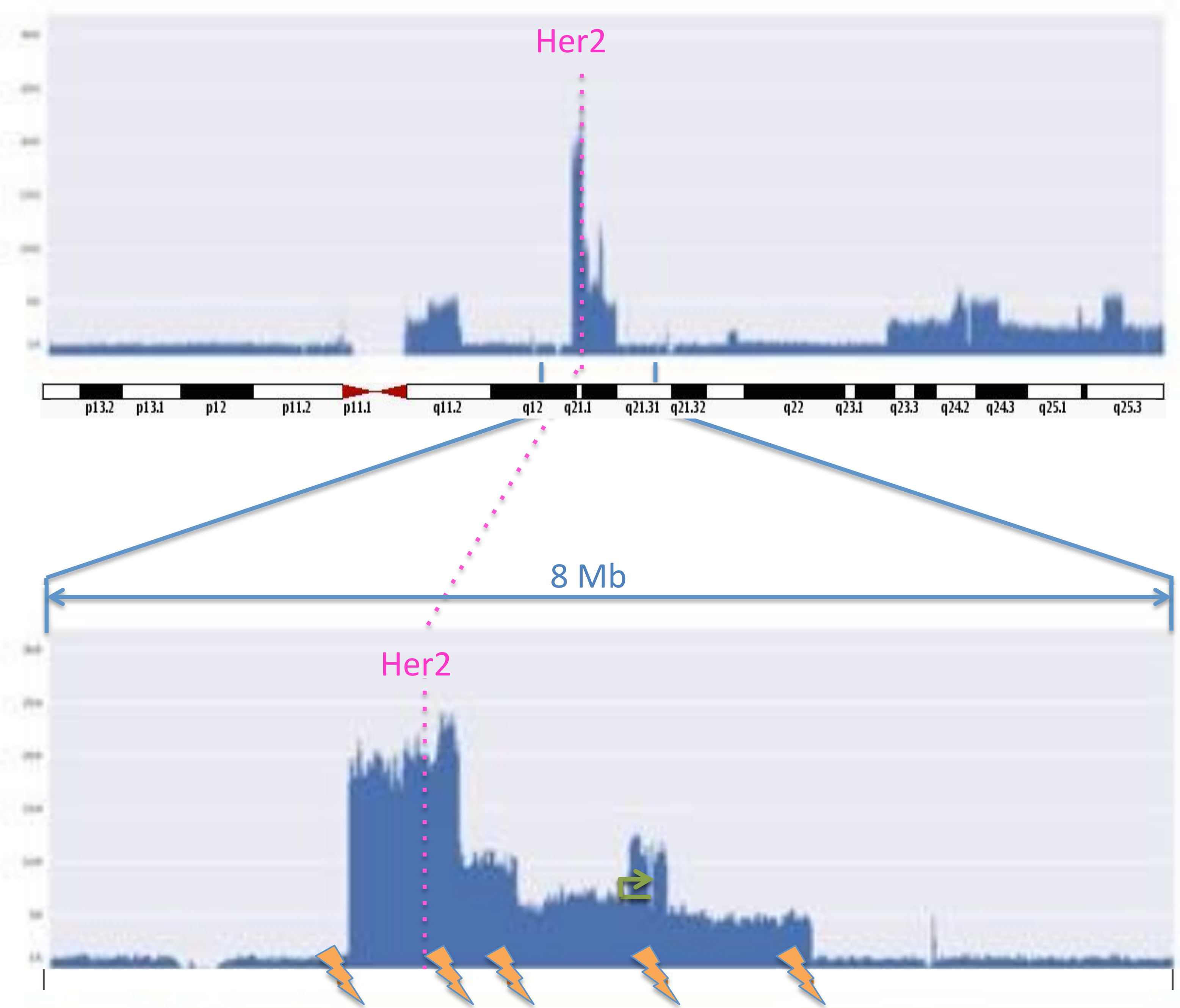
PacBio


Her2
p13.2
p13.1
p12
p11.2
p11.1
q11.2
q12
q21.1
q21.31
q21.32
q22
q23.1
q23.3
q24.2
q24.3
q25.1
q25.3
Chr 17: 83 Mb

PacBio

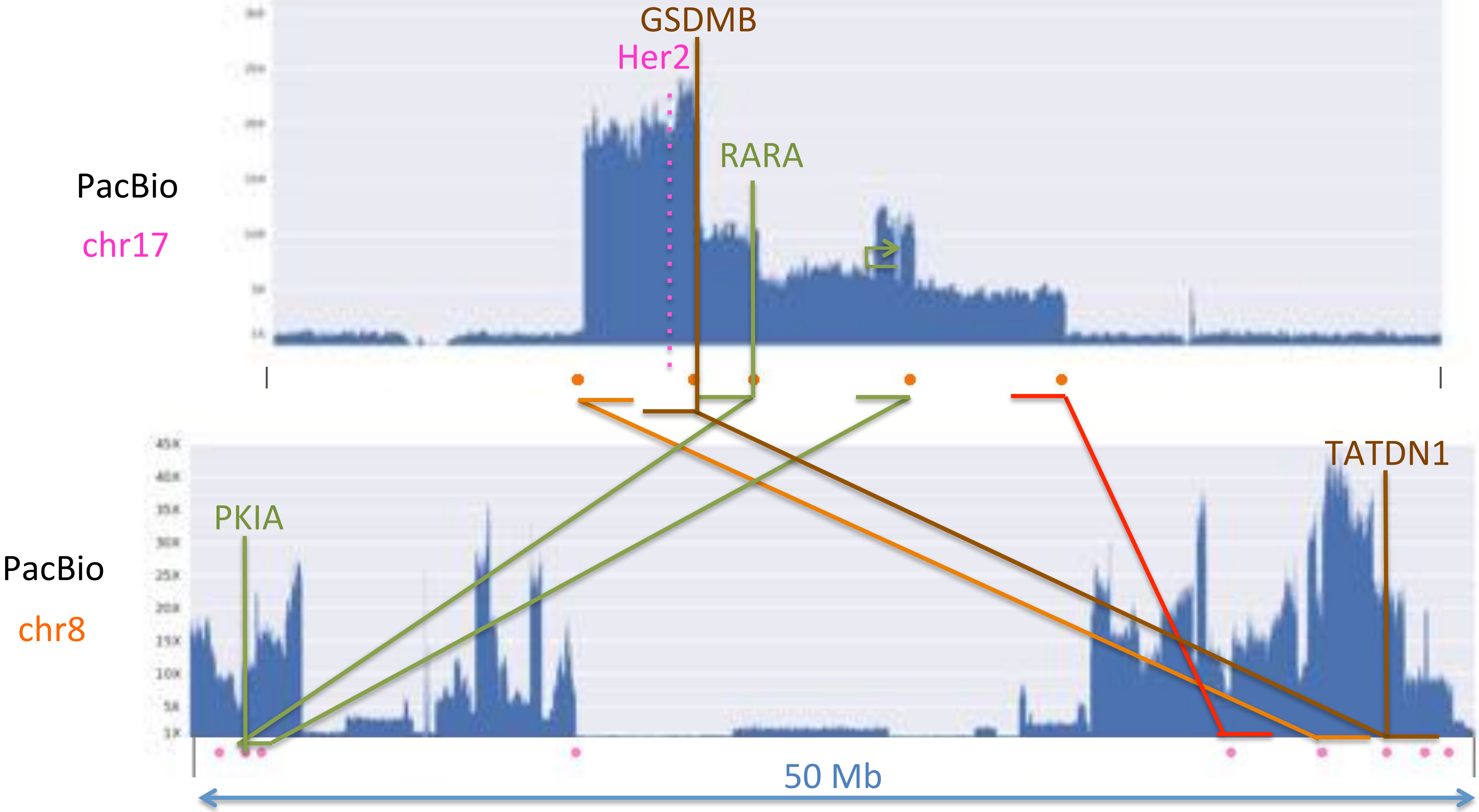
PacBio
chr17



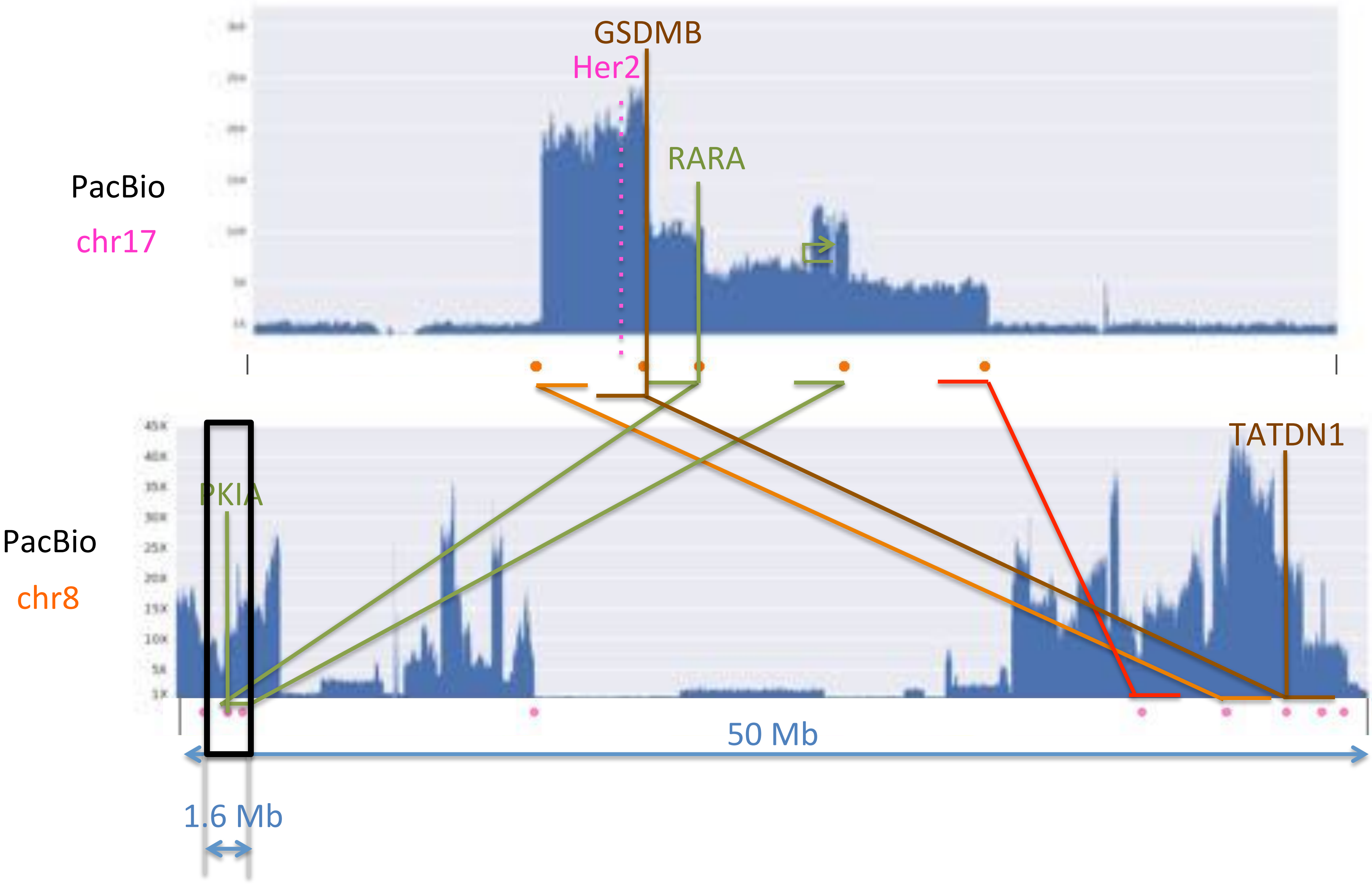


PacBio

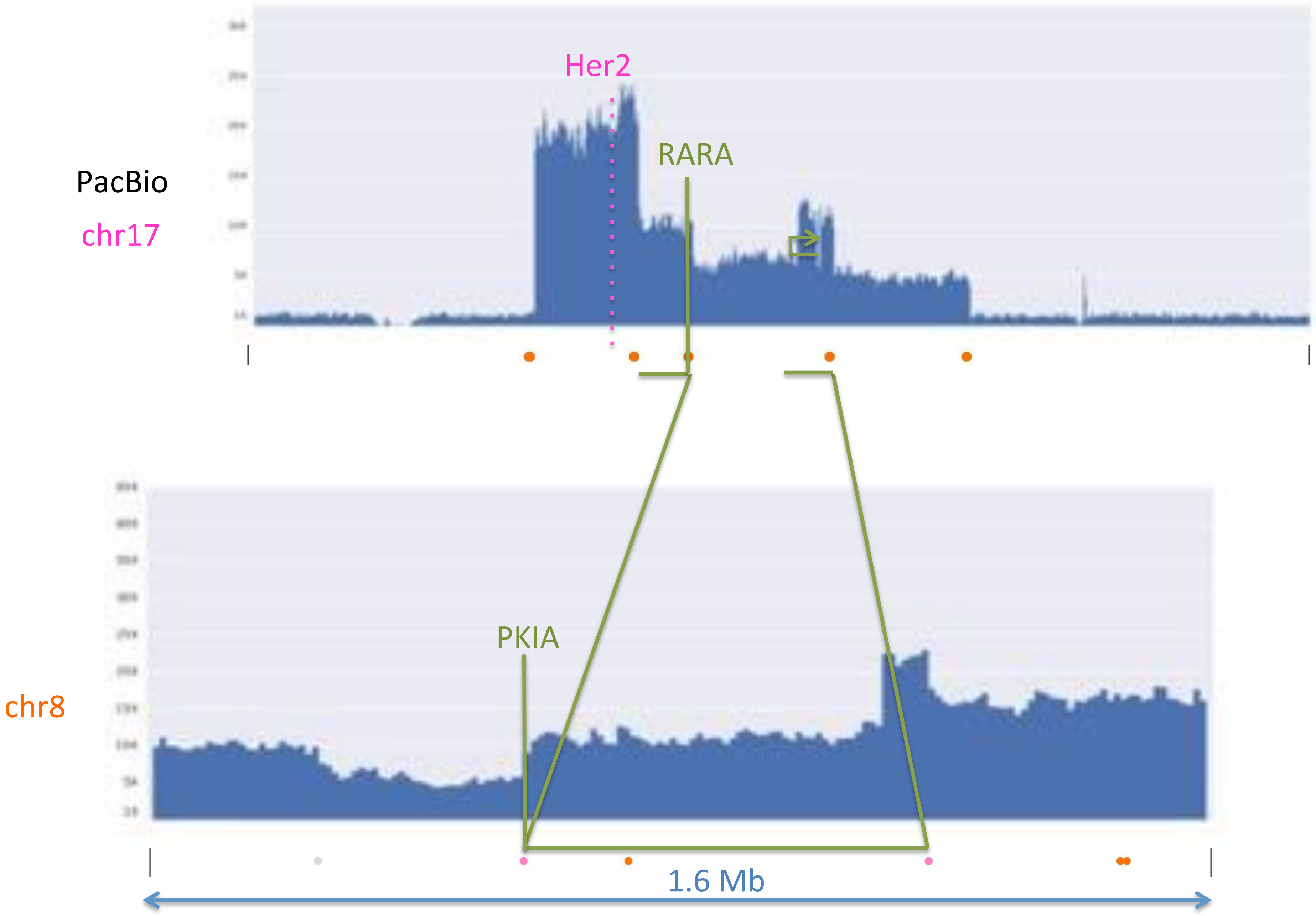
PacBio
chr17



Confirmed both known gene fusions in this region

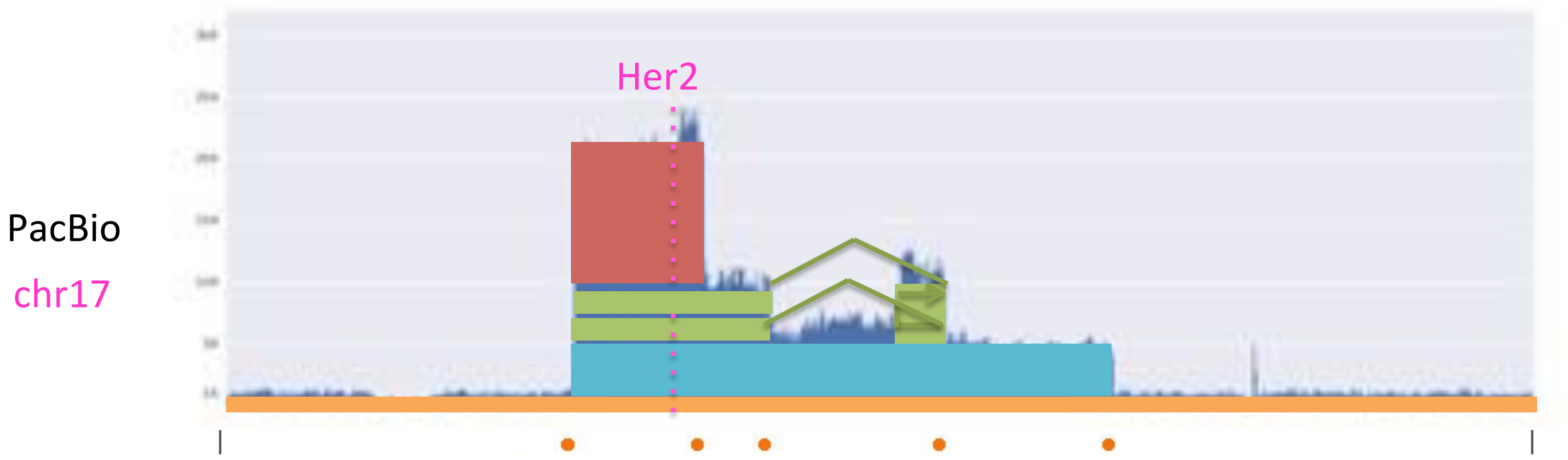


Confirmed both known gene fusions in this region



Joint coverage and breakpoint analysis to discover underlying events

Cancer lesion Reconstruction



By comparing the proportion of reads that are spanning or split at breakpoints we can begin to infer the history of the genetic lesions.

1. Healthy diploid genome

2. Original translocation into chromosome 8

3. Duplication, inversion, and inverted duplication within chromosome 8

4. Final duplication from within chromosome 8

Cancer lesion Reconstruction

Available *today* under the Toronto Agreement:

- Fastq & BAM files of aligned reads
- Interactive Coverage Analysis with BAM.IOBIO
- Whole genome assembly

Available soon

- Whole genome methylation analysis
- Full length cDNA transciptome analysis
- Comparison to single cell analysis of >100 individual cells

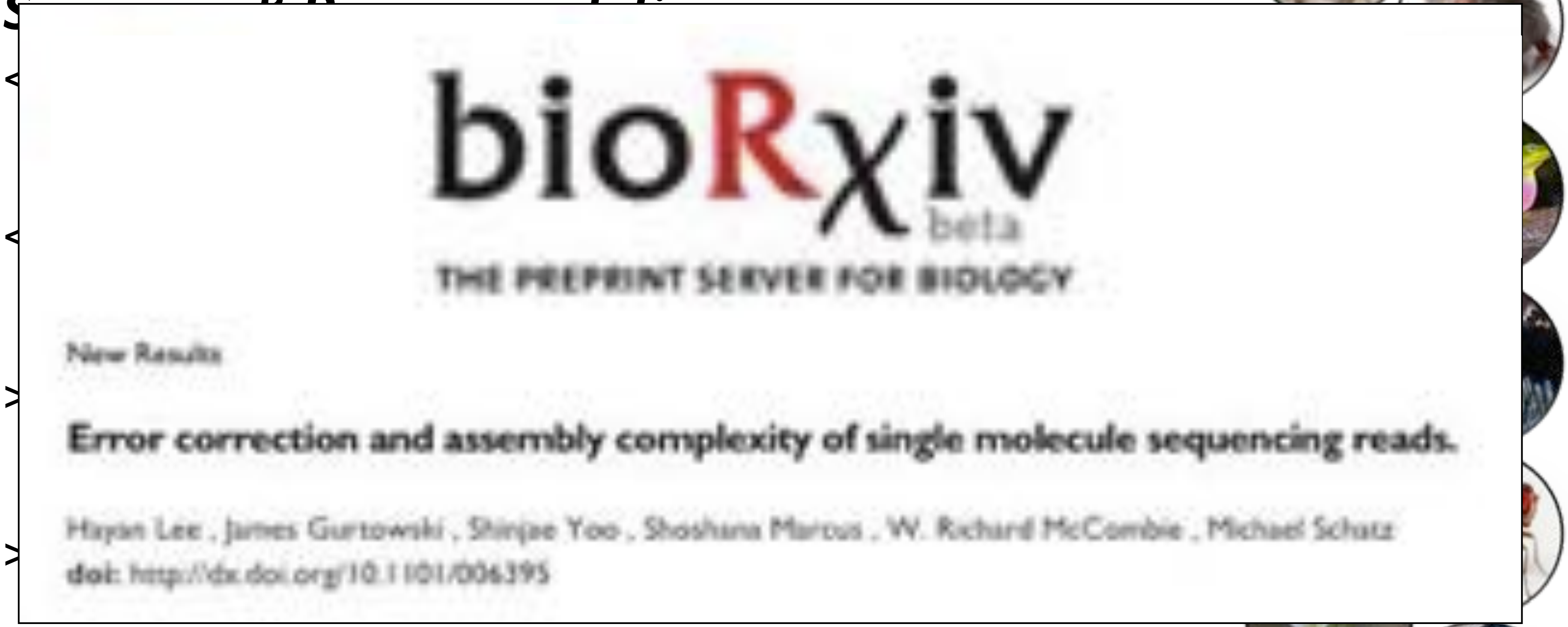
http://schatzlab.cshl.edu/skbr3

3. Duplication, inversion, and inverted duplication within chromosome 8

4. Final duplication from within chromosome 8

What should we expect from an assembly?

The resurgence of reference quality genomes



bioRxiv beta

THE PREPRINT SERVER FOR BIOLOGY

New Results

Error correction and assembly complexity of single molecule sequencing reads.

Hayan Lee , James Gurtowski , Shinjae Yoo , Shoshana Marcus , W. Richard McCombie , Michael Schatz

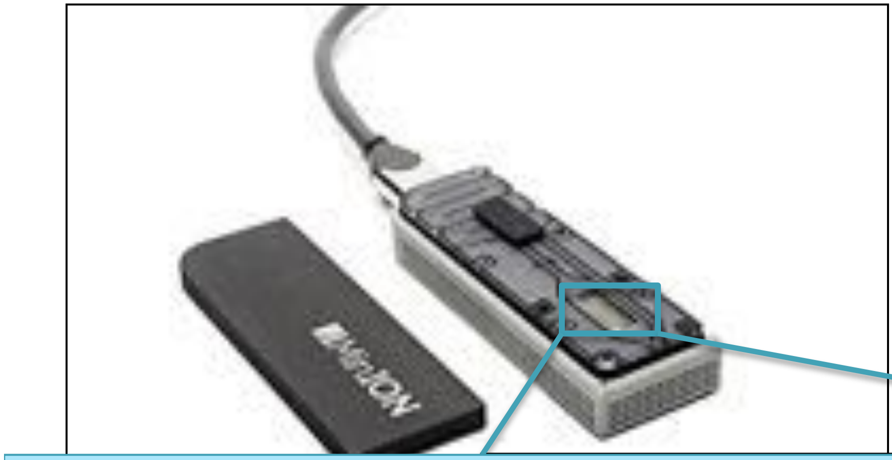
doi: http://dx.doi.org/10.1101/006395

Caveats

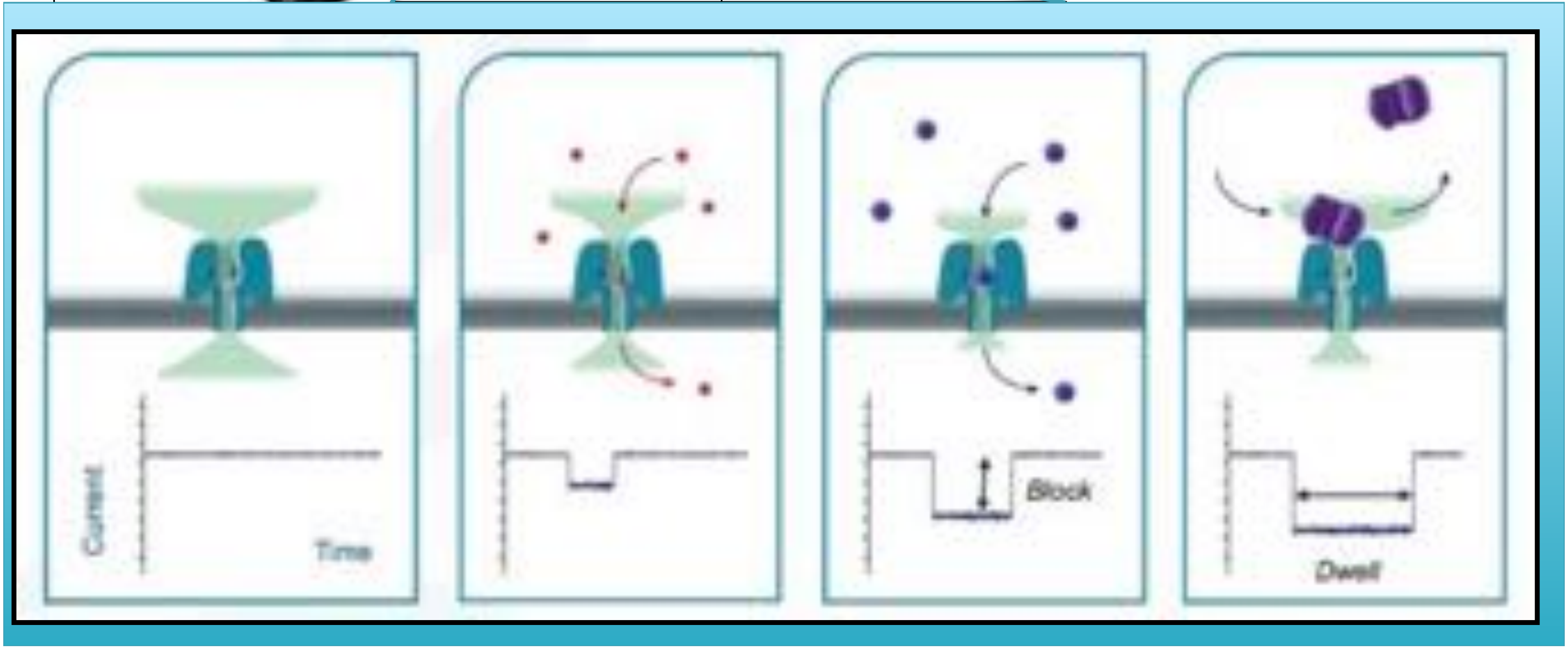
Model only as good as the available references (esp. haploid sequences)

Technologies are quickly improving, exciting new scaffolding technologies

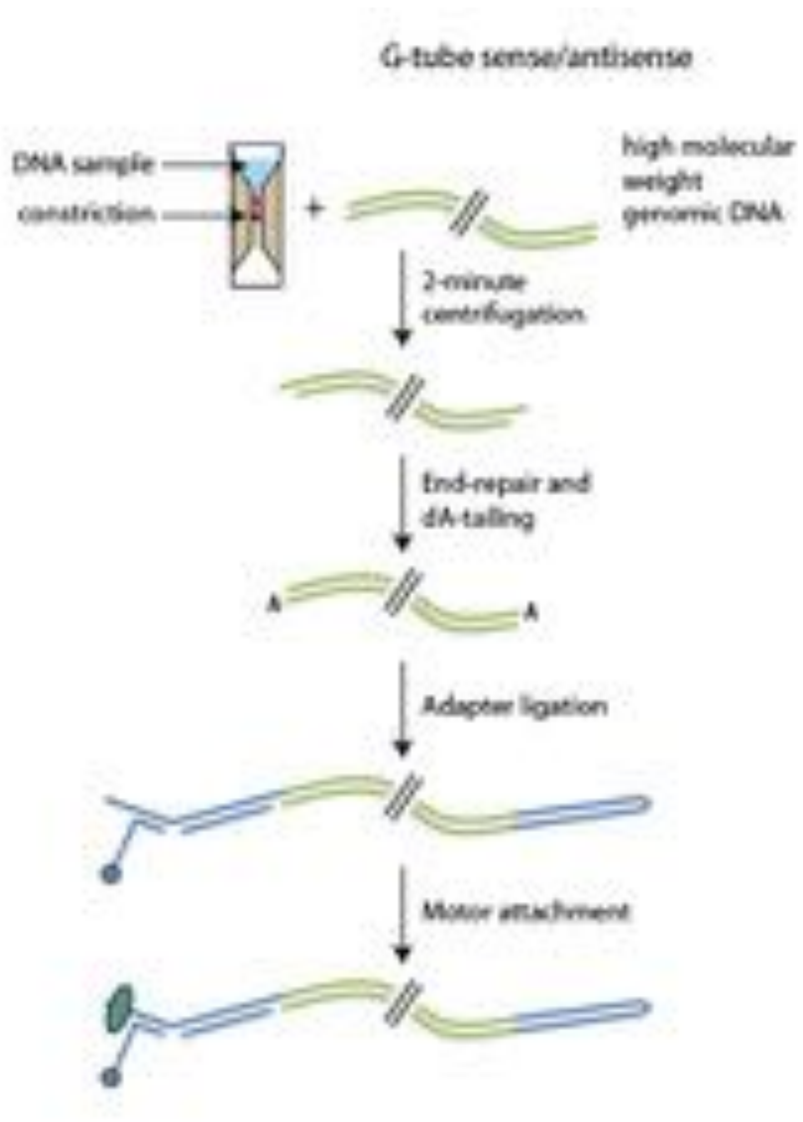
Oxford Nanopore MinION

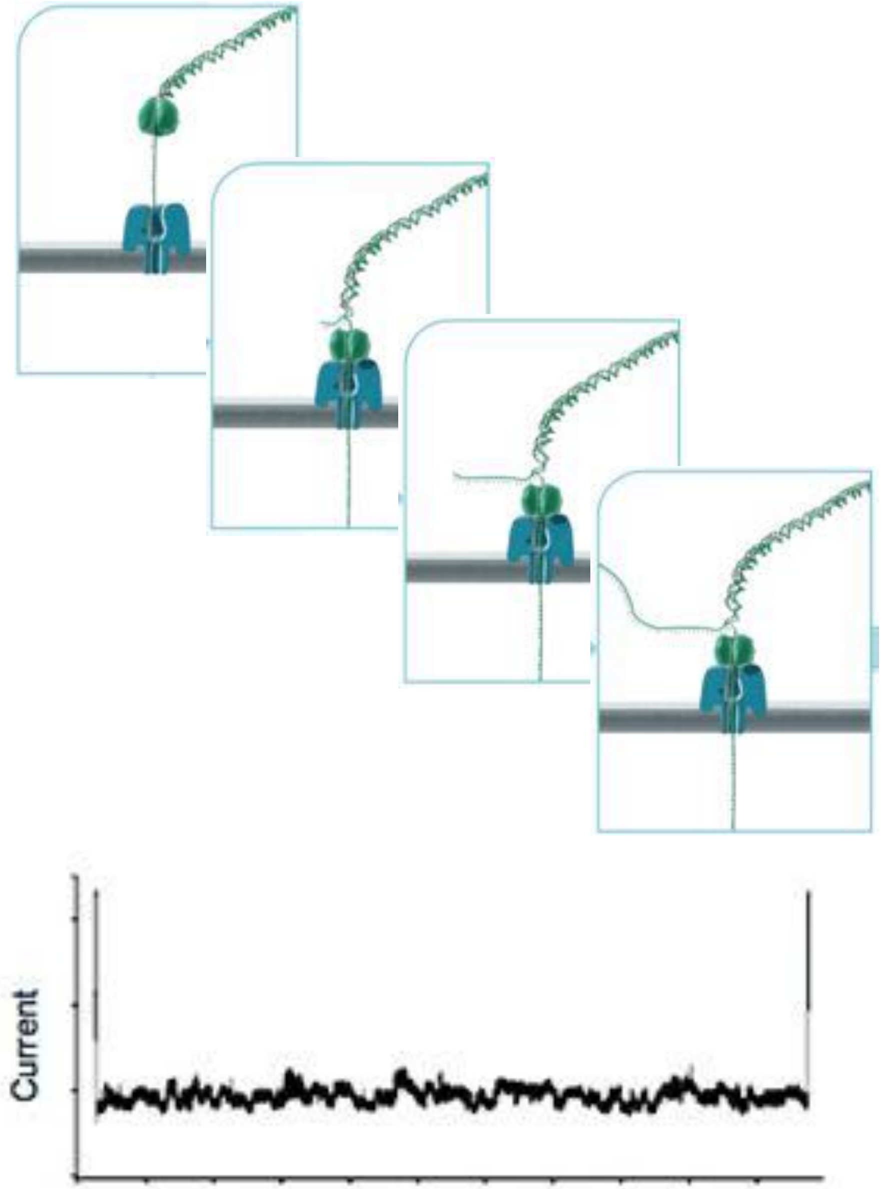


- Thumb drive sized sequencer powered over USB
- Capacity for 512 reads at once
- Senses DNA by measuring changes to ion flow

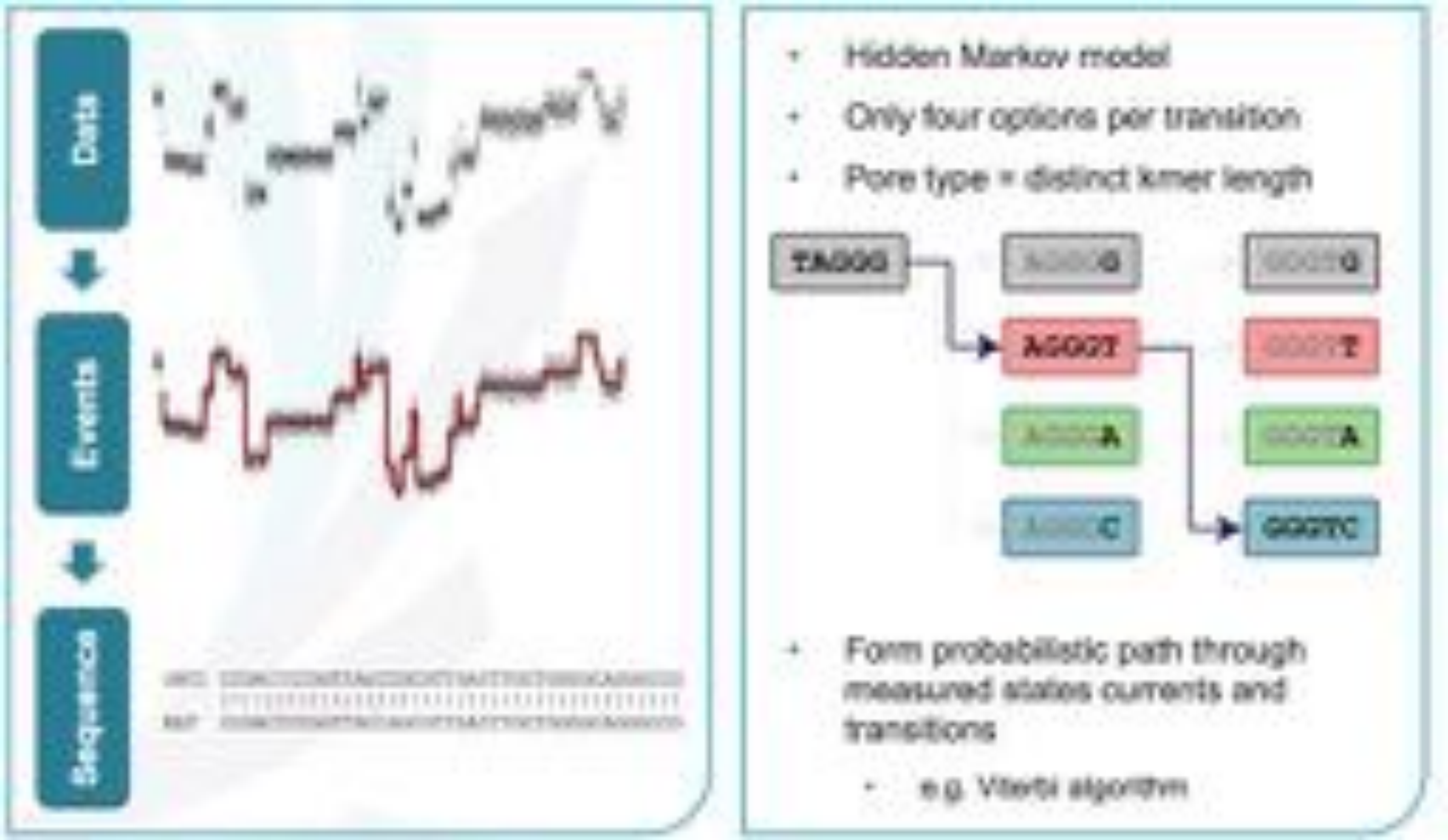


Nanopore Sequencing





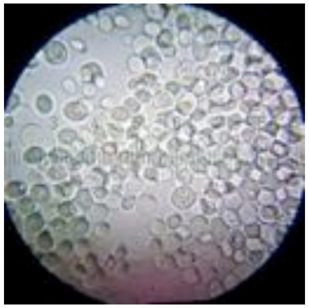
Nanopore Sequencing

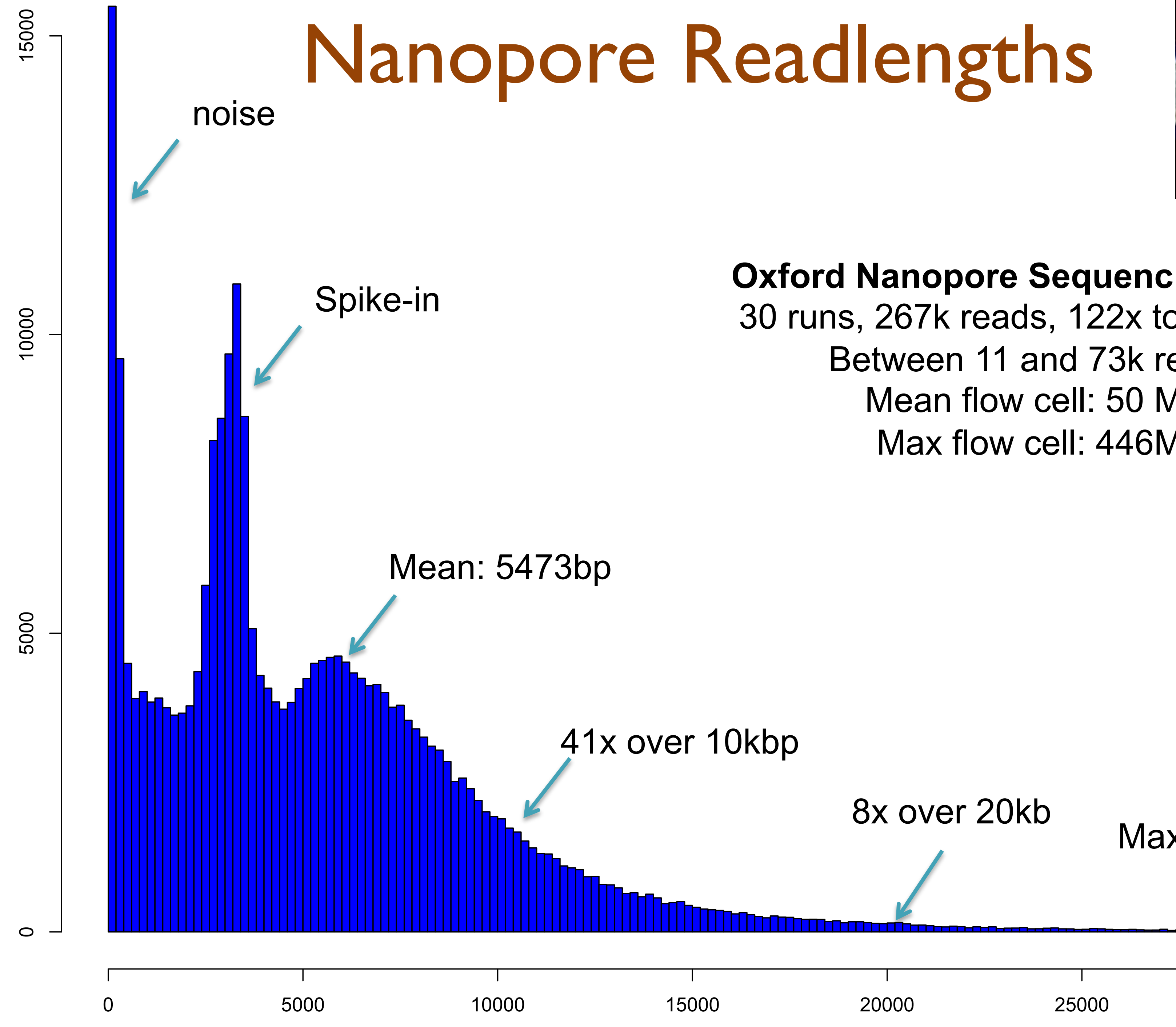


- Hidden Markov model
- Only four options per transition
- Pore type = distinct kmer length
- Form probabilistic path through measured states currents and transitions
  - e.g. Viterbi algorithm

Basecalling currently performed at Amazon with frequent updates to algorithm

Nanopore Readlengths





noise

Spike-in

Mean: 5473bp

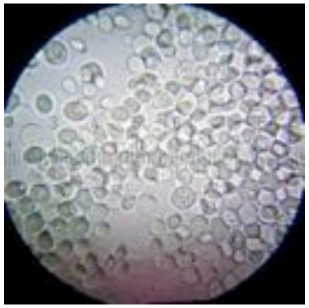
41x over 10kbp

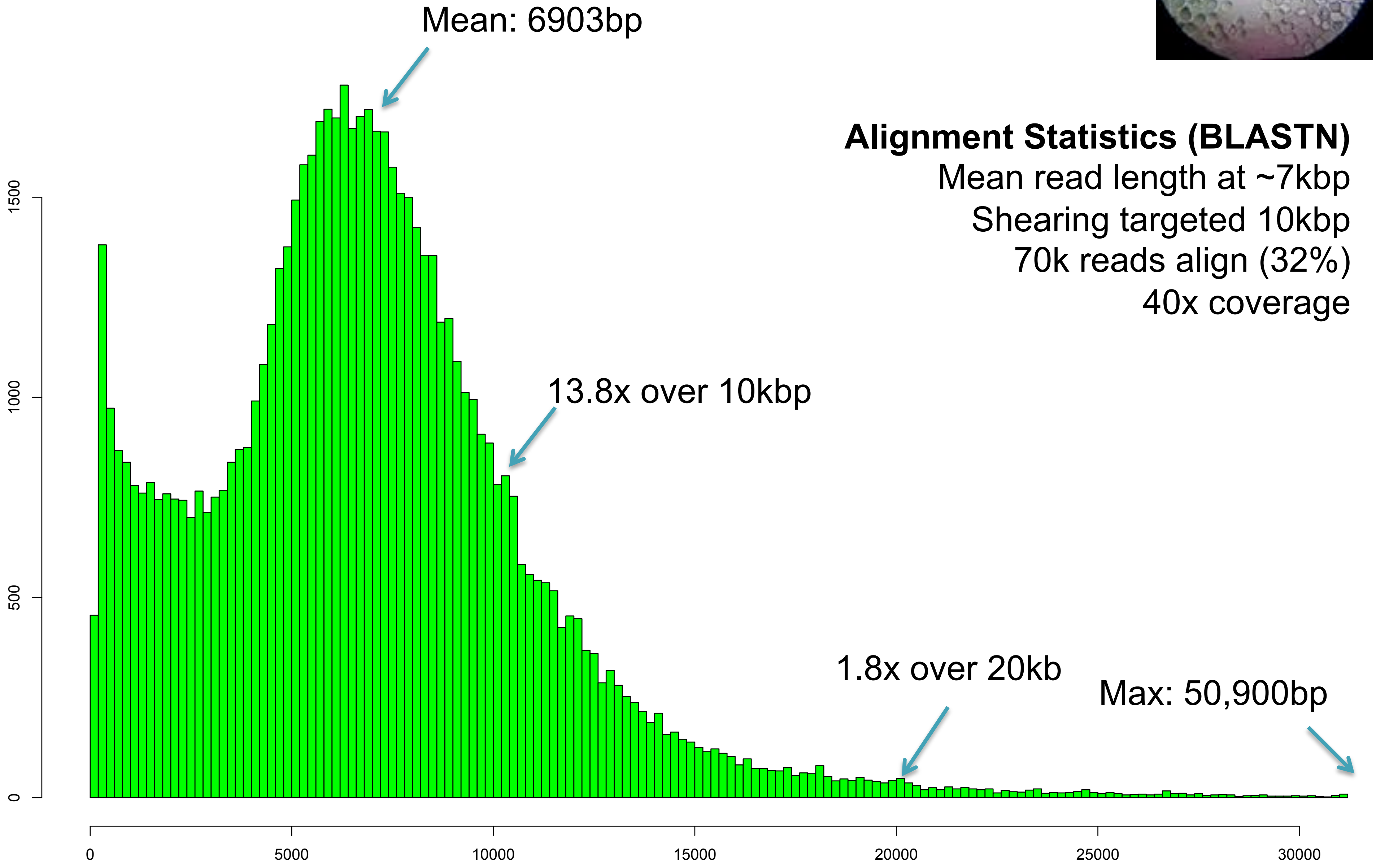
8x over 20kb

Max: 146,992bp

Oxford Nanopore Sequencing at CSHL
30 runs, 267k reads, 122x total coverage
Between 11 and 73k reads per run!
Mean flow cell: 50 Mbp in 2 days
Max flow cell: 446Mbp in 2 days

Nanopore Alignments





Mean: 6903bp

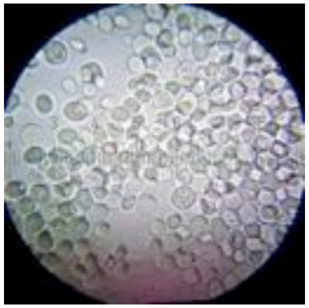
13.8x over 10kbp

1.8x over 20kb

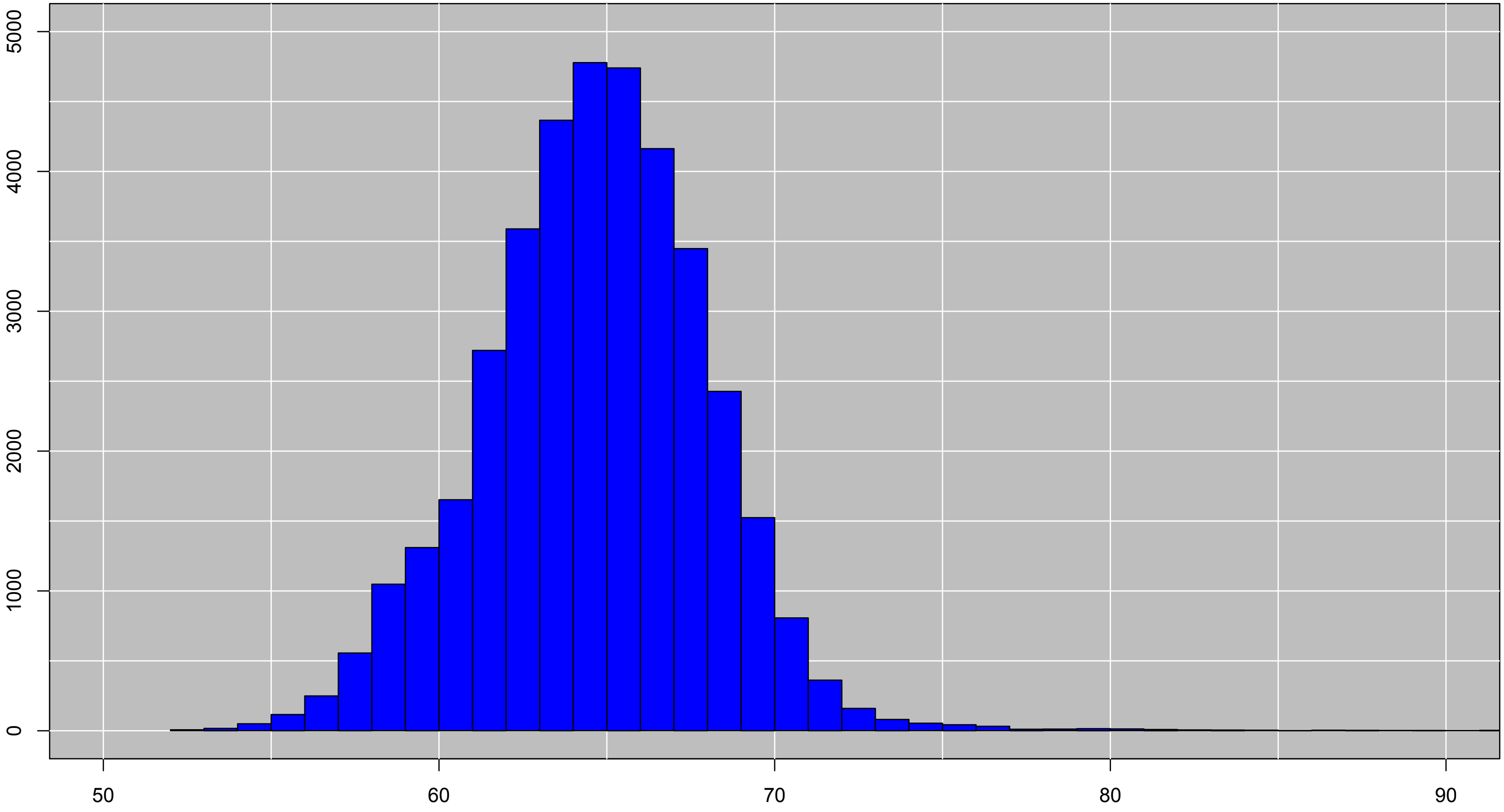
Max: 50,900bp

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

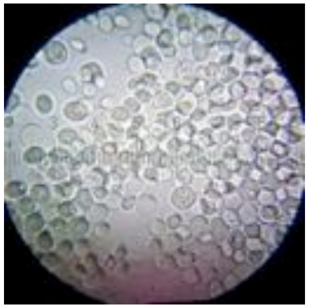
Nanopore Accuracy



Alignment Quality (BLASTN)
Of reads that align, average ~64% identity



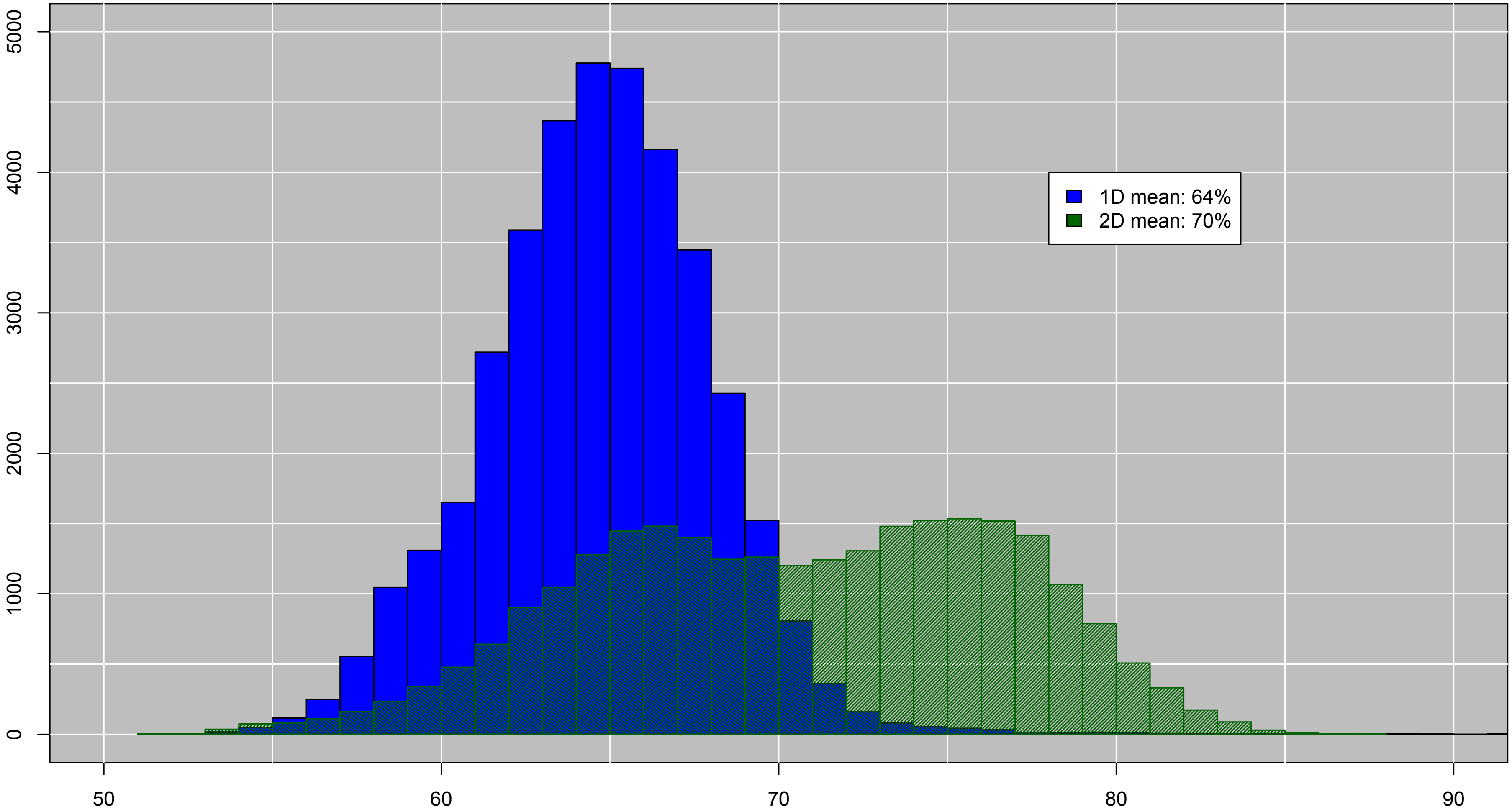
Nanopore Accuracy



Alignment Quality (BLASTN)

Of reads that align, average ~64% identity

“2D base-calling” improves to ~70% identity



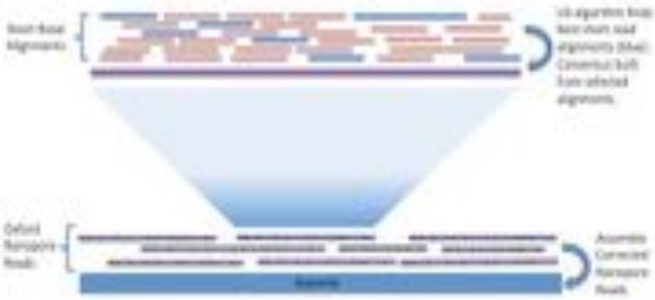
NanoCorr: Nanopore-Illumina Hybrid Error Correction

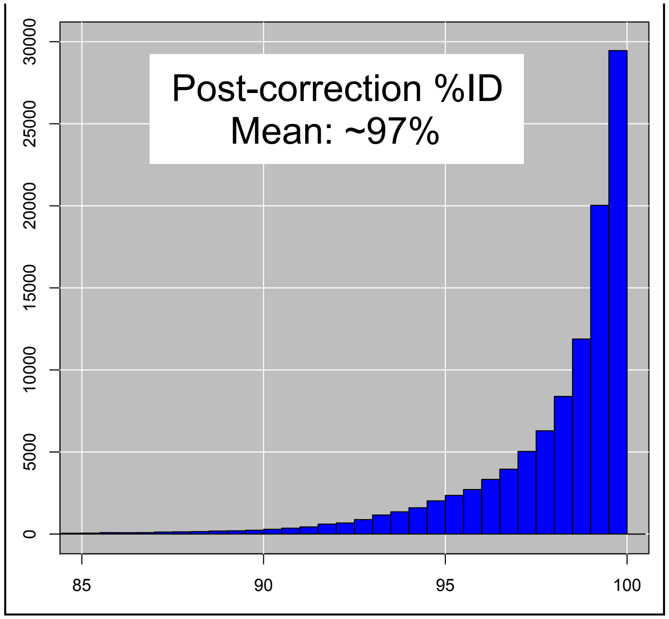
https://github.com/jgurtowski/nanocorr

1. BLAST Miseq reads to all raw Oxford Nanopore reads

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

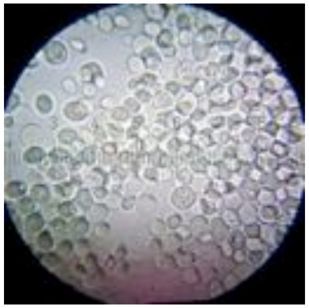
3. Compute consensus of each Oxford Nanopore read
   - ○ State machine of most commonly observed base at each position in read


Short Read Alignments
LIS algorithm finds best short read alignments (blue). Consensus built from selected alignments
Oxford Nanopore Reads
Assemble Corrected Nanopore Reads
Genome



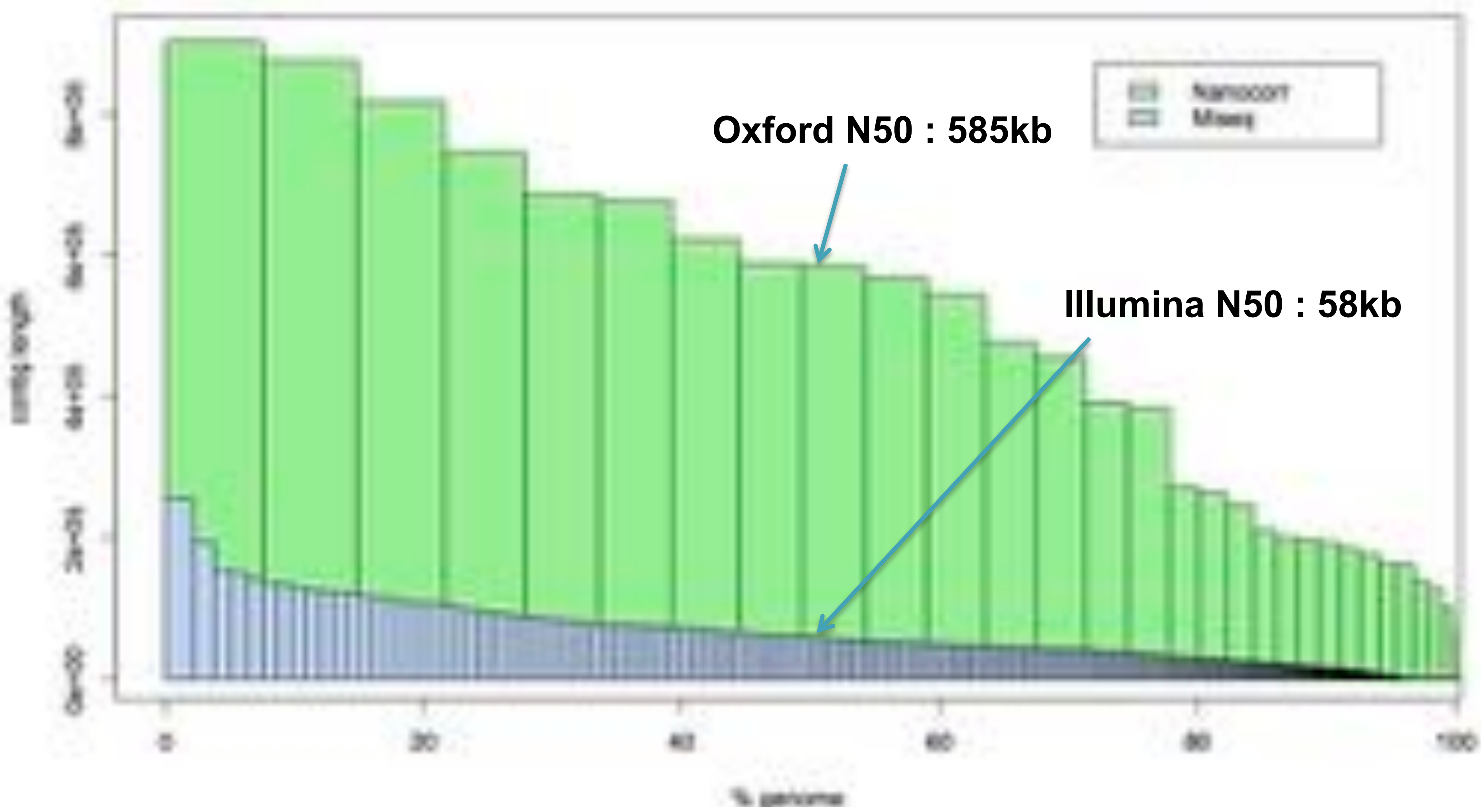
Oxford Nanopore Sequencing and de novo Assembly of a Eukaryotic Genome
Goodwin, S, Gurtowski, J *et al.* (2015) bioRxiv doi: http://dx.doi.org/10.1101/013490

NanoCorr Yeast Assembly



S288C Reference sequence

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



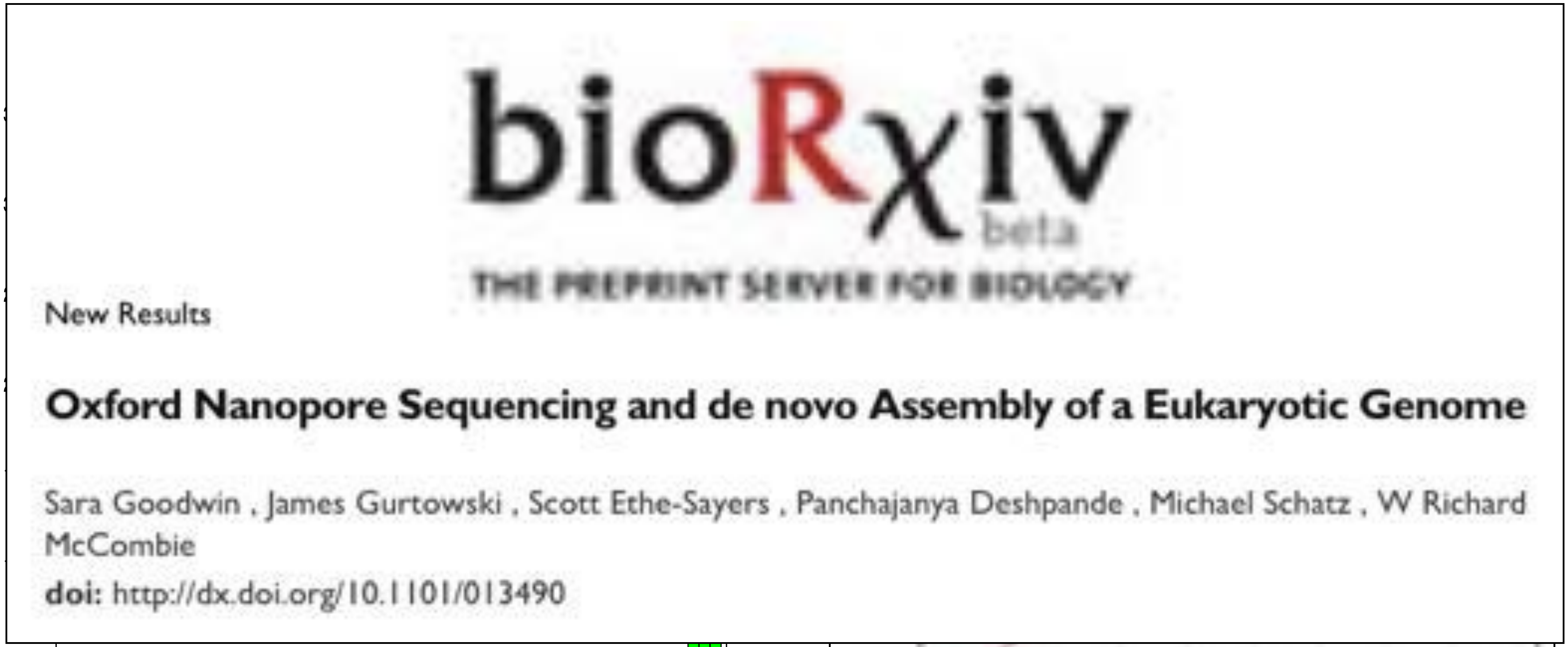
Oxford N50 : 585kb

Illumina N50 : 58kb

NanoCorr E. coli K12 Assembly

Nanocor Correction Results
145x Oxford Nanopore X 35x MiSeq

Single Contig Assembly
99.99% Identity (Pilon polishing)



bioRxiv beta

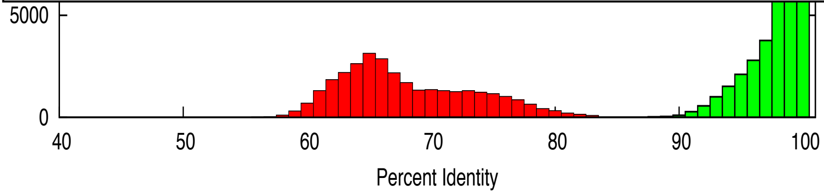
THE PREPRINT SERVER FOR BIOLOGY

New Results

Oxford Nanopore Sequencing and de novo Assembly of a Eukaryotic Genome

Sara Goodwin , James Gurtowski , Scott Ethe-Sayers , Panchajanya Deshpande , Michael Schatz , W Richard McCombie

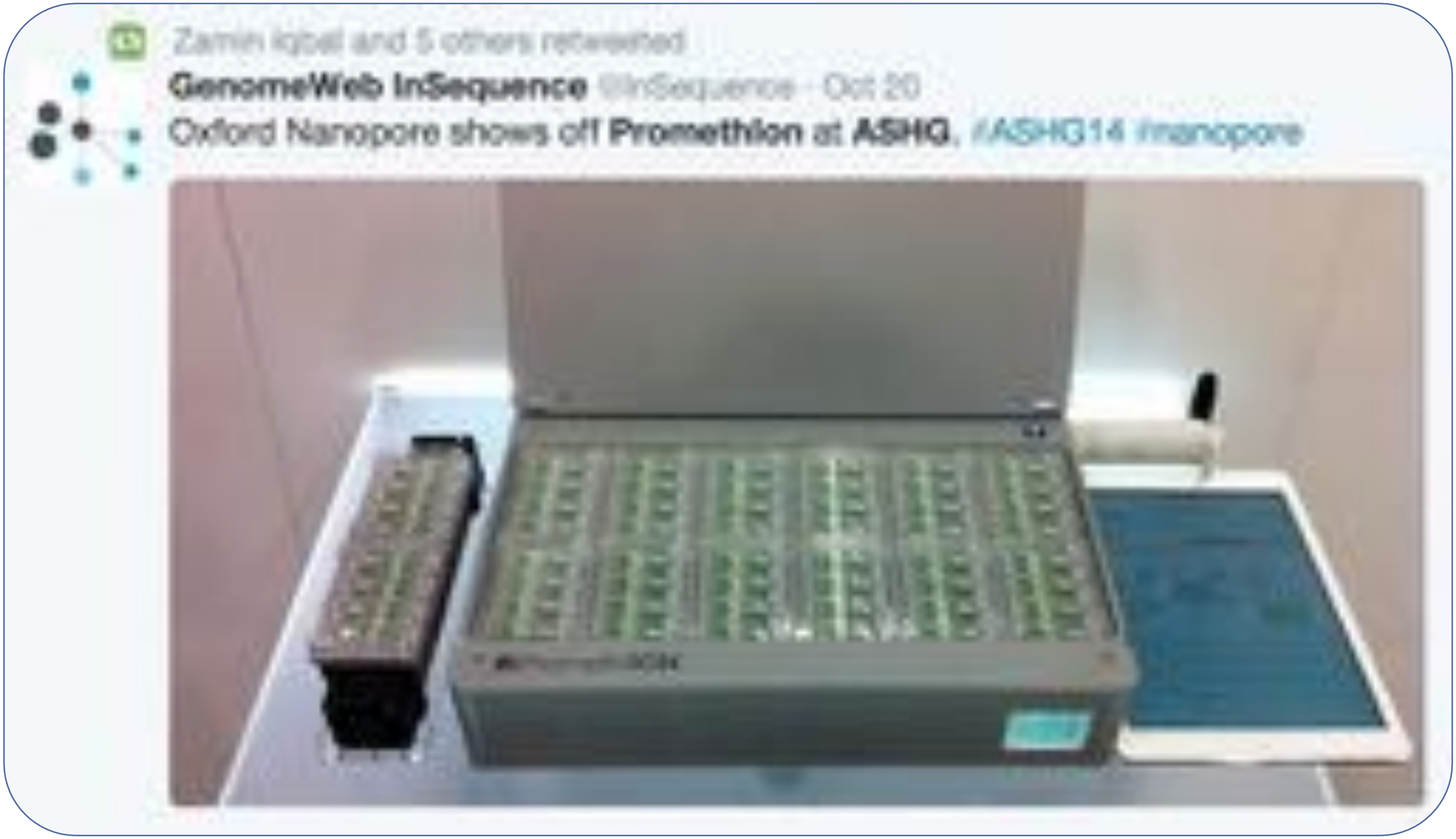
doi: http://dx.doi.org/10.1101/013490



Sequencing Data From:

A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer
Joshua Quick, Aaron R Quinlan and Nicholas J Loman

Genomic Futures?



Zamin Iqbal and 5 others retweeted

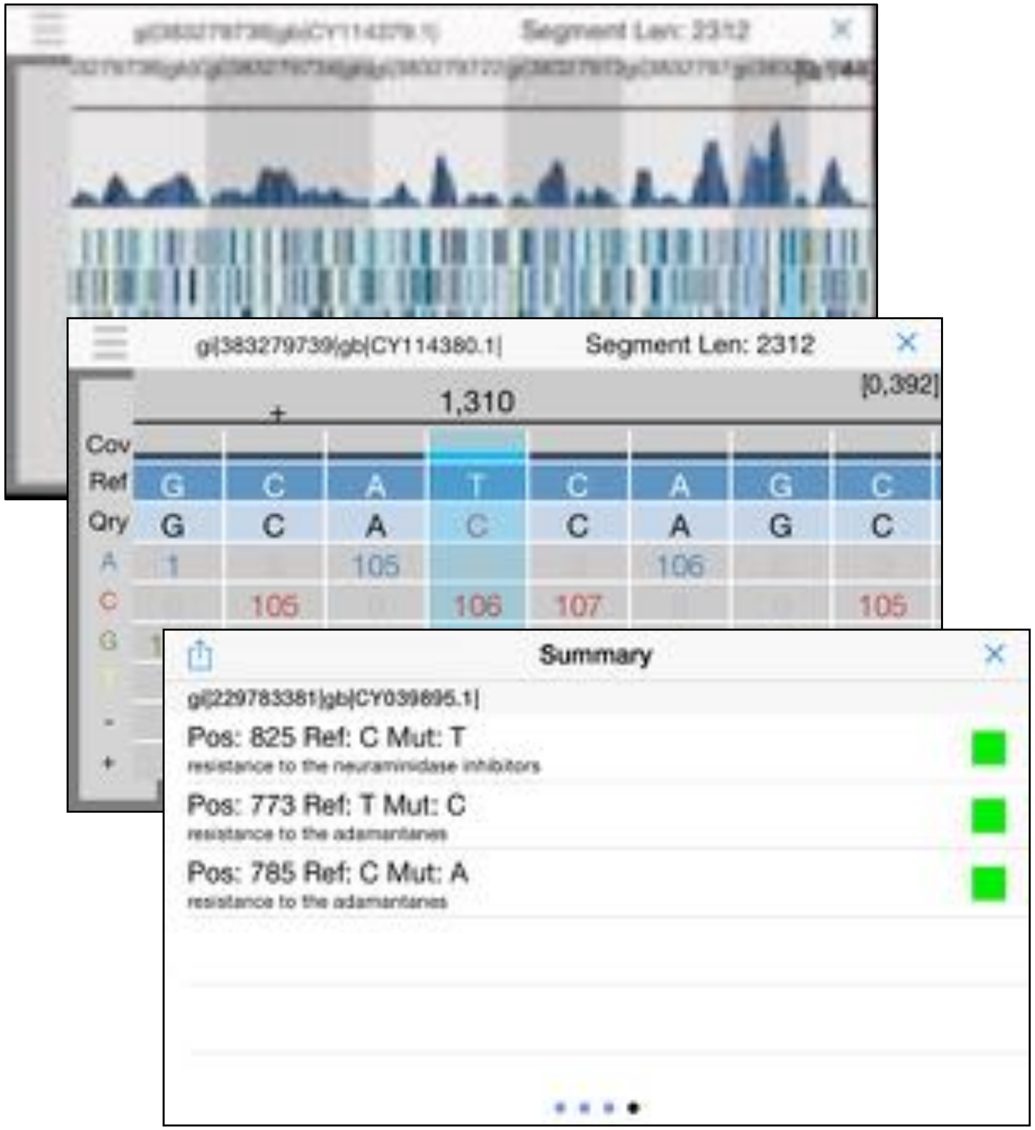
GenomeWeb InSequence @InSequence · Oct 20

Oxford Nanopore shows off **Promethion** at **ASHG**. #ASHG14 #nanopore

Genomic Futures?

iGenomics: Mobile Sequence Analysis

Aspyn Palatnick, Elodie Ghedin, Michael Schatz



gi|383279736|gb|CY114379.1| Segment Len: 2312

gi|383279739|gb|CY114380.1| Segment Len: 2312

1,310 [0,392]

Cov								
Ref	G	C	A	T	C	A	G	C
Qry	G	C	A	C	C	A	G	C
A	1		105			106		
C		105		106	107			105
G	1							

Summary

gi|229783381|gb|CY039895.1|

Pos: 825 Ref: C Mut: T 🟩
resistance to the neuraminidase inhibitors

Pos: 773 Ref: T Mut: C 🟩
resistance to the adamantanes

Pos: 785 Ref: C Mut: A 🟩
resistance to the adamantanes

The worlds first genomics analysis app for iOS devices

BWT + Dynamic Programming + UI

First application:

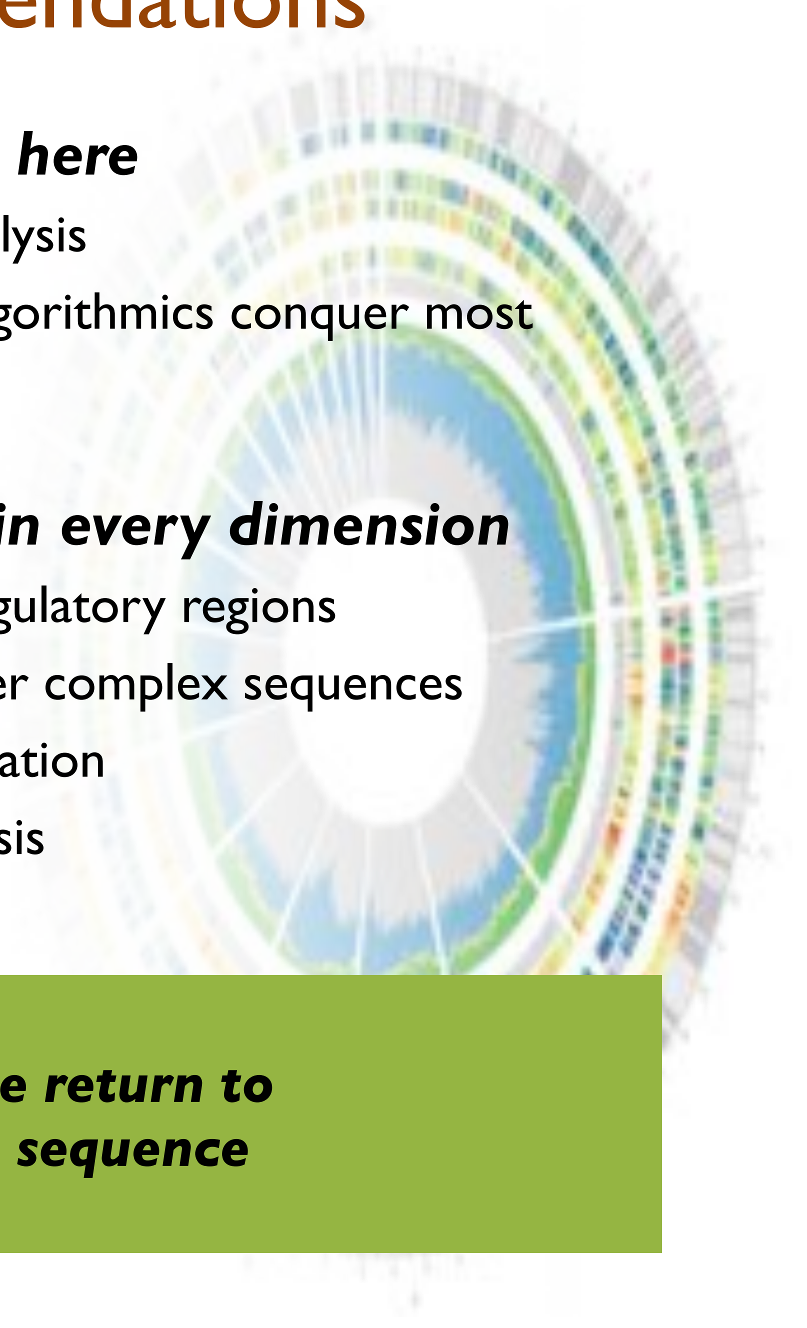
- Handheld diagnostics and therapeutic recommendations for influenza infections

- In the iOS AppStore now!

Future applications

- Pathogen detection
- Food safety
- Biomarkers
- etc..

Summary & Recommendations



Reference quality genome assembly is here

- Use the longest possible reads for the analysis
- Don't fear the error rate, coverage and algorithmics conquer most problems

Megabase N50 improves the analysis 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

The year 2015 will mark the return to reference quality genome sequence

Acknowledgements

Schatz Lab

Rahul Amin
Eric Biggers
Han Fang
Tyler Gavin
James Gurtowski
Ke Jiang
Hayan Lee
Zak Lemmon
Shoshana Marcus
Giuseppe Narzisi
Maria Nattestad
Aspyn Palatnick
Srividya Ramakrishnan
Fritz Sedlazeck
Rachel Sherman
Greg Vurture
Alejandro Wences

CSHL

Hannon Lab
Gingeras Lab
Jackson Lab
Hicks Lab
Iossifov Lab
Levy Lab
Lippman Lab
Lyon Lab
Martienssen Lab
McCombie Lab
Tuveson Lab
Ware Lab
Wigler Lab

Cornell

Susan McCouch
Lyza Maron
Mark Wright

OICR

John McPherson
Karen Ng
Timothy Beck
Yogi Sundaravadanam

NBACC

Adam Phillippy
Serge Koren

Oxford NANOPORE Technologies

PACIFIC BIOSCIENCES®

NSF

National Human Genome Research Institute

U.S. DEPARTMENT OF ENERGY

SFARI
SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

ALFRED P. SLOAN FOUNDATION

Thank you

http://schatzlab.cshl.edu

@mike_schatz