

The Resurgence of Reference Quality Genomes using 3rd Gen Sequencing

Michael Schatz

Jan 6, 2015
Penn State



Schatzlab Overview



Informatics

Ultra-large scale
biocomputing

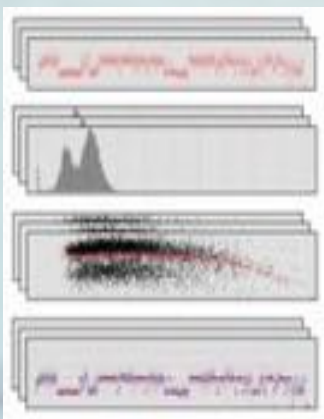
Blood *et al.* (2014)
Schatz *et al.* (2013)



Human Genetics

Autism, Cancer,
Psychiatric Disorders

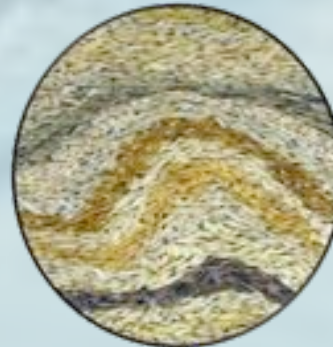
Narzisi *et al.* (2014)
Iossifov *et al.* (2014)



Biotechnology

Single Cell &
Single Molecule
Analysis

Garvin *et al.* (2014)
Roberts *et al.* (2013)



Plant Biology

Genomes &
Transcriptomes

Schatz *et al.* (2014)
Ming *et al.* (2013)

Outline

1. Assembly theory

1. Assembly by analogy
2. De Bruijn and Overlap graph
3. Coverage, read length, errors, and repeats

2. Sequencing and Assembly options

1. Illumina/ALLPATHS-LG
2. Pacific Biosciences
3. Oxford Nanopore

3. Summary & Recommendations



Shredded Book Reconstruction

- Dickens accidentally shreds the first printing of A Tale of Two Cities
 - Text printed on 5 long spools

It was	the	best	of	times,	it	was	the	worst	of	times,	it	was	the	age	of	wisdom,	it	was	the	age	of	foolishness, ...	
It was	the	best	of	times,	it	was	the	worst	of	times,	it	was	the	age	of	wisdom,	it	was	the	age	of	foolishness, ...	
It was	the	best	of	times,	it	was	the	worst	of	times,	it	was	the	age	of	wisdom,	it	was	the	age	of	foolishness, ...	
It was	the	best	of	times,	it	was	the	worst	of	times,	it	was	the	age	of	wisdom,	it	was	the	age	of	foolishness, ...	
It	was	the	best	of	times,	it	was	the	worst	of	times,	it	was	the	age	of	wisdom,	it	was	the	age	of	foolishness, ...

- How can he reconstruct the text?
 - 5 copies x 138,656 words / 5 words per fragment = 138k fragments
 - The short fragments from every copy are mixed together
 - Some fragments are identical

Greedy Reconstruction

It was the best of
age of wisdom, it was
best of times, it was
it was the age of
it was the age of
it was the worst of
of times, it was the
of times, it was the
of wisdom, it was the
the age of wisdom, it
the best of times, it
the worst of times, it
times, it was the age
times, it was the worst
was the age of wisdom,
was the age of foolishness,
was the best of times,
was the worst of times,
wisdom, it was the age
worst of times, it was

It was the best of
was the best of times,
the best of times, it
best of times, it was
of times, it was the
of times, it was the
times, it was the worst
times, it was the age

The repeated sequence make the correct reconstruction ambiguous

- It was the best of times, it was the [worst/age]

Model the assembly problem as a graph problem

de Bruijn Graph Construction

- $D_k = (V, E)$
 - $V =$ All length- k subfragments ($k < l$)
 - $E =$ Directed edges between consecutive subfragments
 - Nodes overlap by $k-1$ words

Original Fragment

It was the best of

Directed Edge

It was the best → was the best of

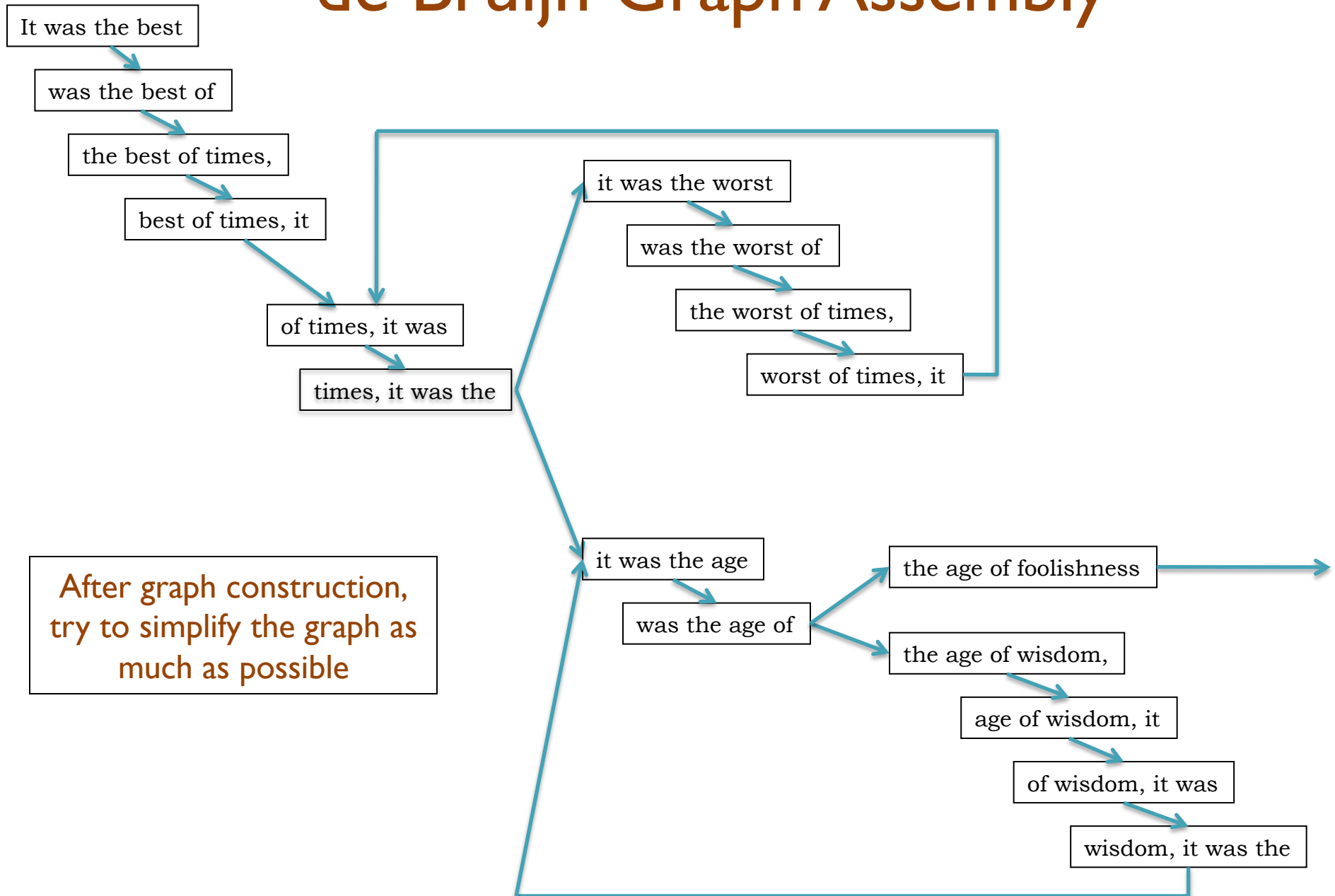
- Locally constructed graph reveals the global sequence structure
 - Overlaps between sequences implicitly computed

de Bruijn, 1946

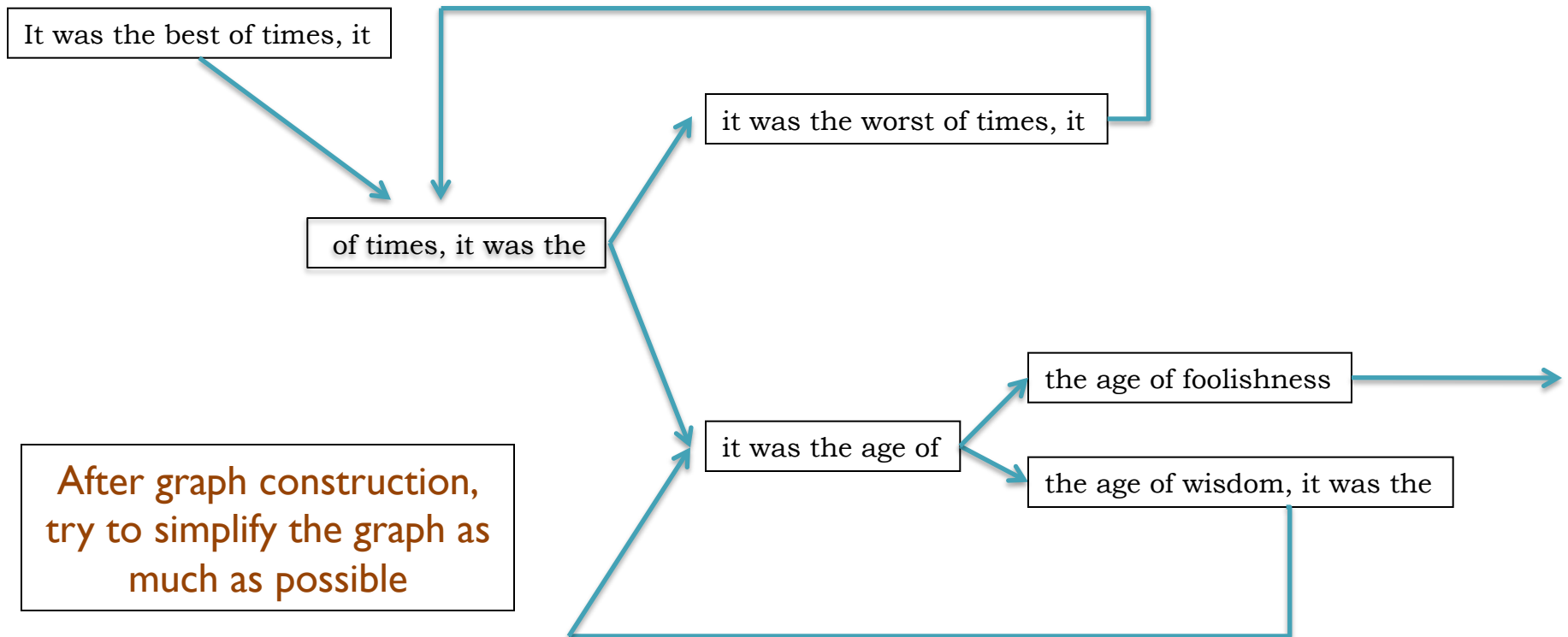
Idury and Waterman, 1995

Pevzner, Tang, Waterman, 2001

de Bruijn Graph Assembly

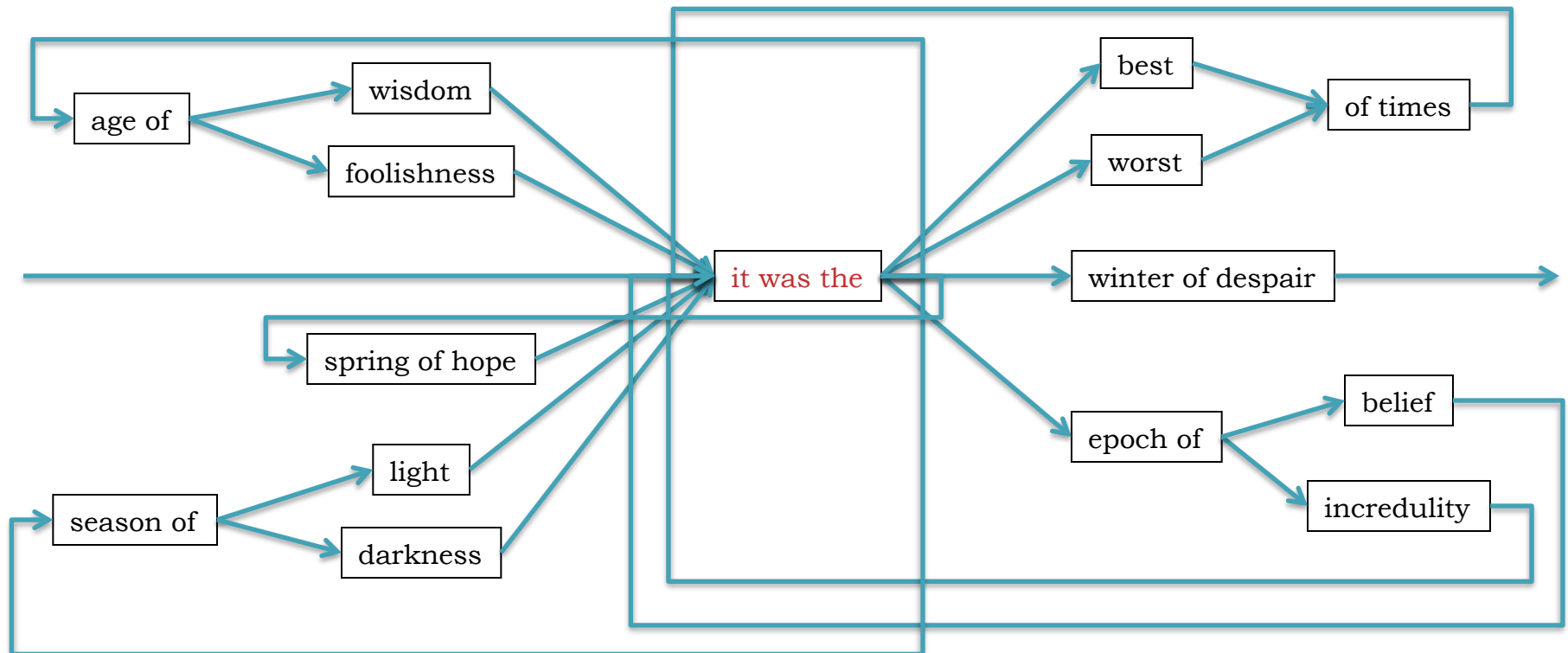


de Bruijn Graph Assembly



The full tale

... it was the best of times it was the worst of times ...
... it was the age of wisdom it was the age of foolishness ...
... it was the epoch of belief it was the epoch of incredulity ...
... it was the season of light it was the season of darkness ...
... it was the spring of hope it was the winter of despair ...



N50 size

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

Example: 1 Mbp genome

50%



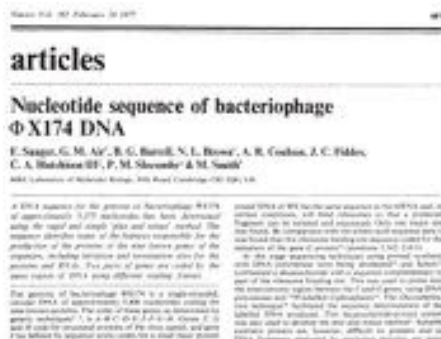
N50 size = 30 kbp

(300k+100k+45k+45k+30k = 520k \geq 500kbp)

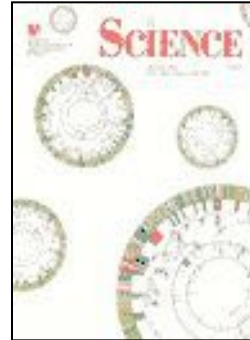
A greater N50 is indicative of improvement in every dimension:

- 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

Milestones in Genome Assembly



1977. Sanger *et al.*
1st Complete Organism
5375 bp



1995. Fleischmann *et al.*
1st Free Living Organism
TIGR Assembler. 1.8Mbp



1998. C.elegans SC
1st Multicellular Organism
BAC-by-BAC Phrap. 97Mbp



2000. Myers *et al.*
1st Large WGS Assembly.
Celera Assembler. 116 Mbp



2001. Venter *et al.*, IHGSC
Human Genome
Celera Assembler/GigaAssembler. 2.9 Gbp



2010. Li *et al.*
1st Large SGS Assembly.
SOAPdenovo 2.2 Gbp

Like Dickens, we must computationally reconstruct a genome from short fragments

Assembly Applications

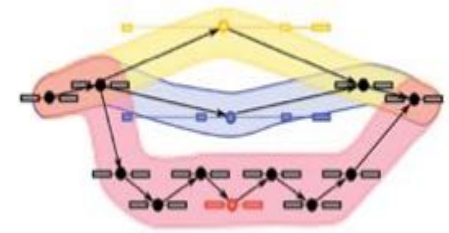
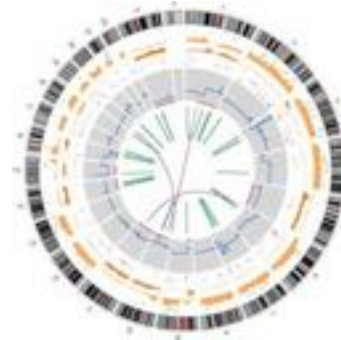
- Novel genomes



- Metagenomes

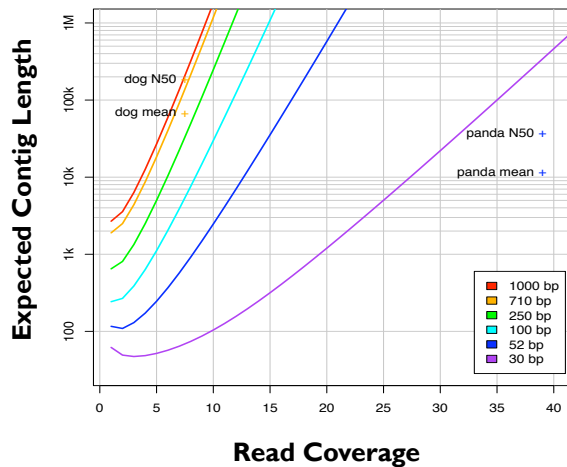


- Sequencing assays
 - Structural variations
 - Transcript assembly
 - ...



Ingredients for a good assembly

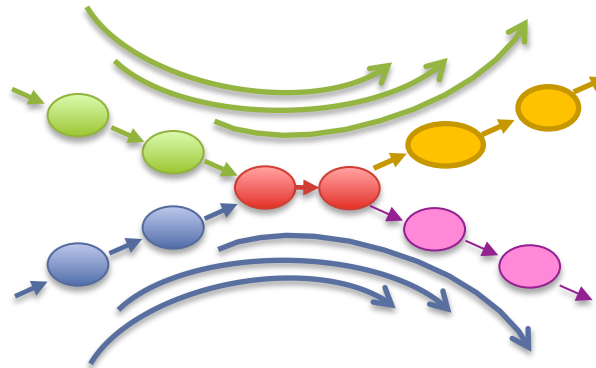
Coverage



High coverage is required

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

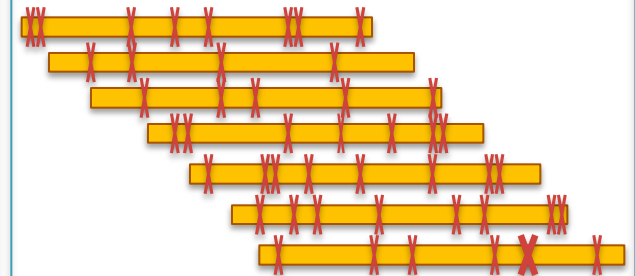
Read Length



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



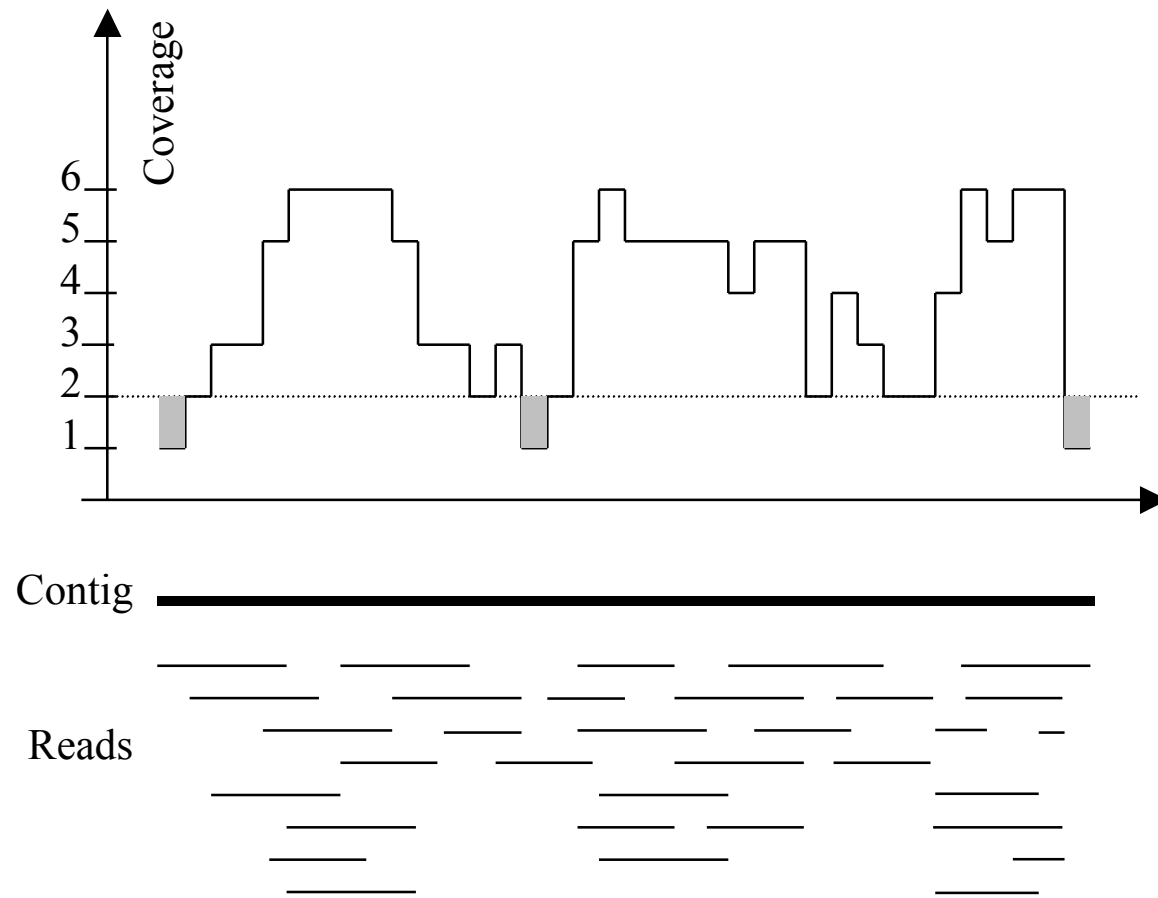
Errors obscure overlaps

- 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, VWR (2012) *Genome Biology*. 12:243

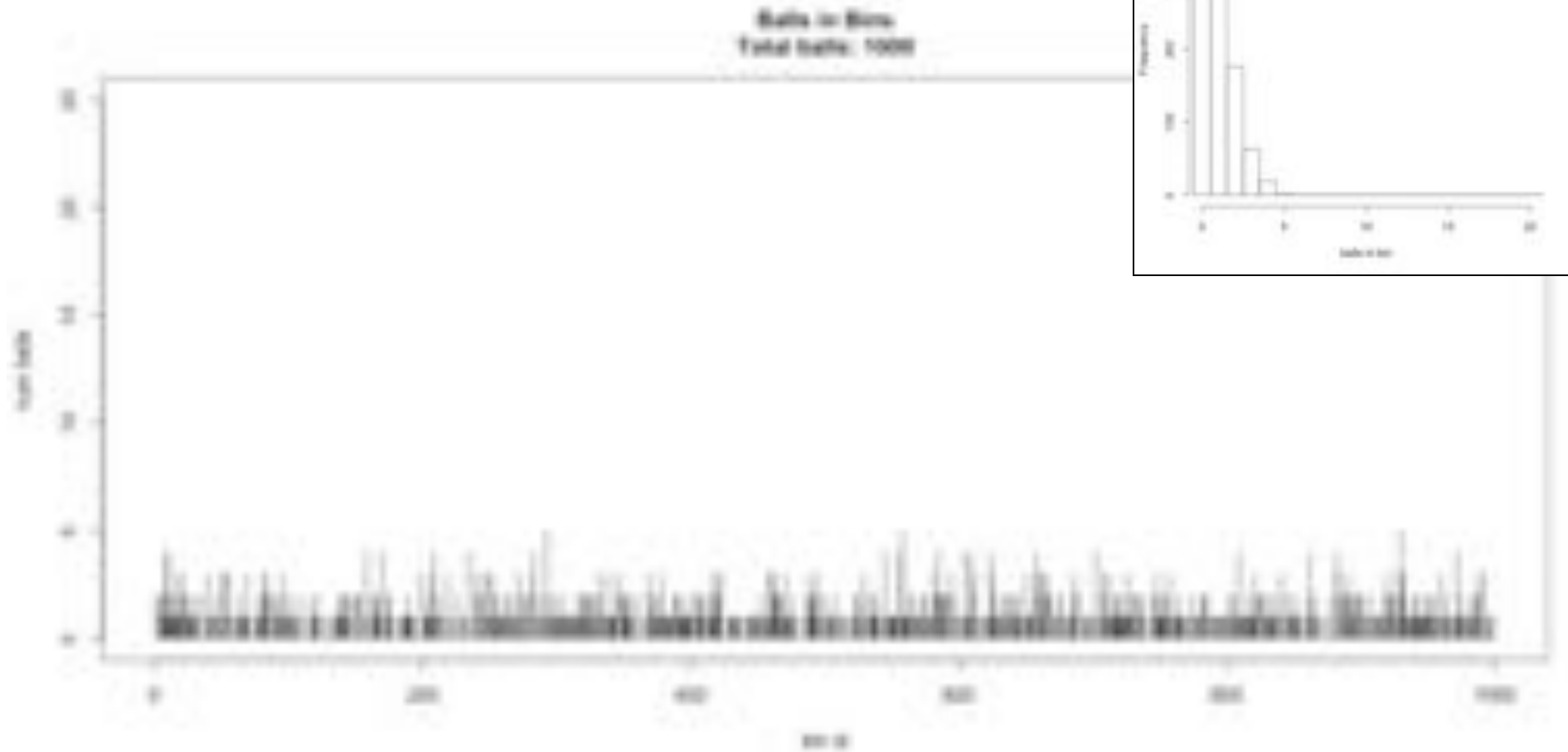
Typical sequencing coverage



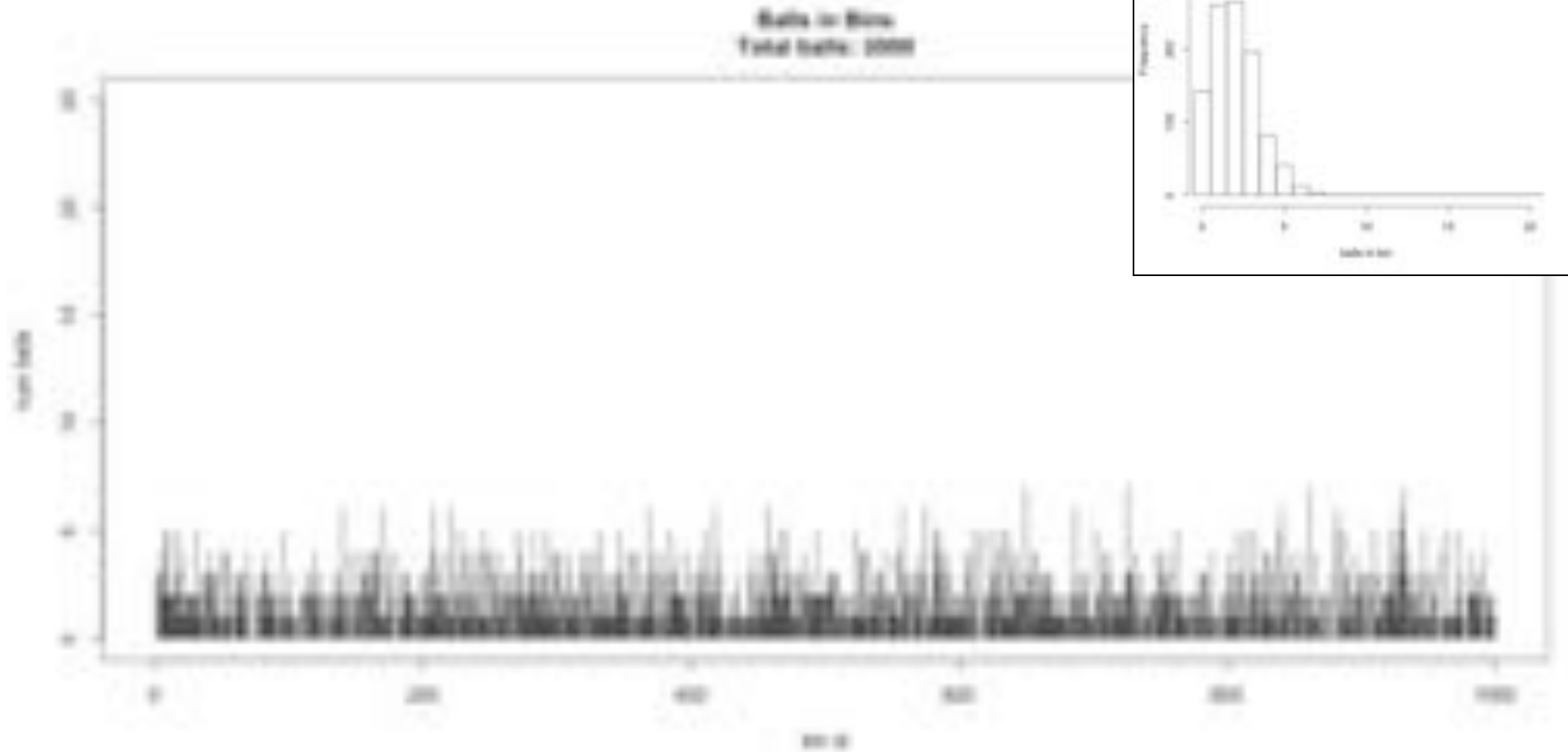
Imagine raindrops on a sidewalk

We want to cover the entire sidewalk but each drop costs \$1

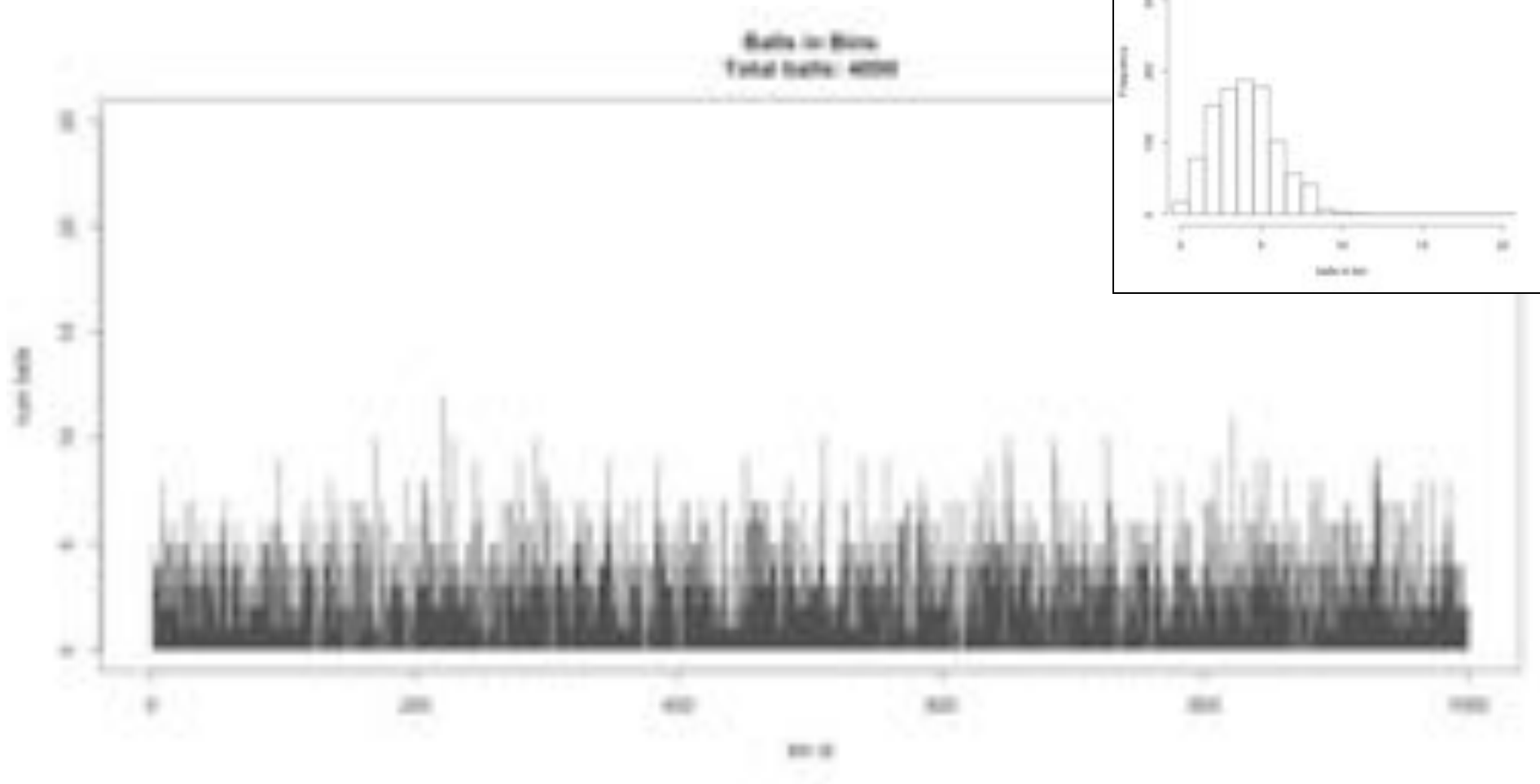
Ix sequencing



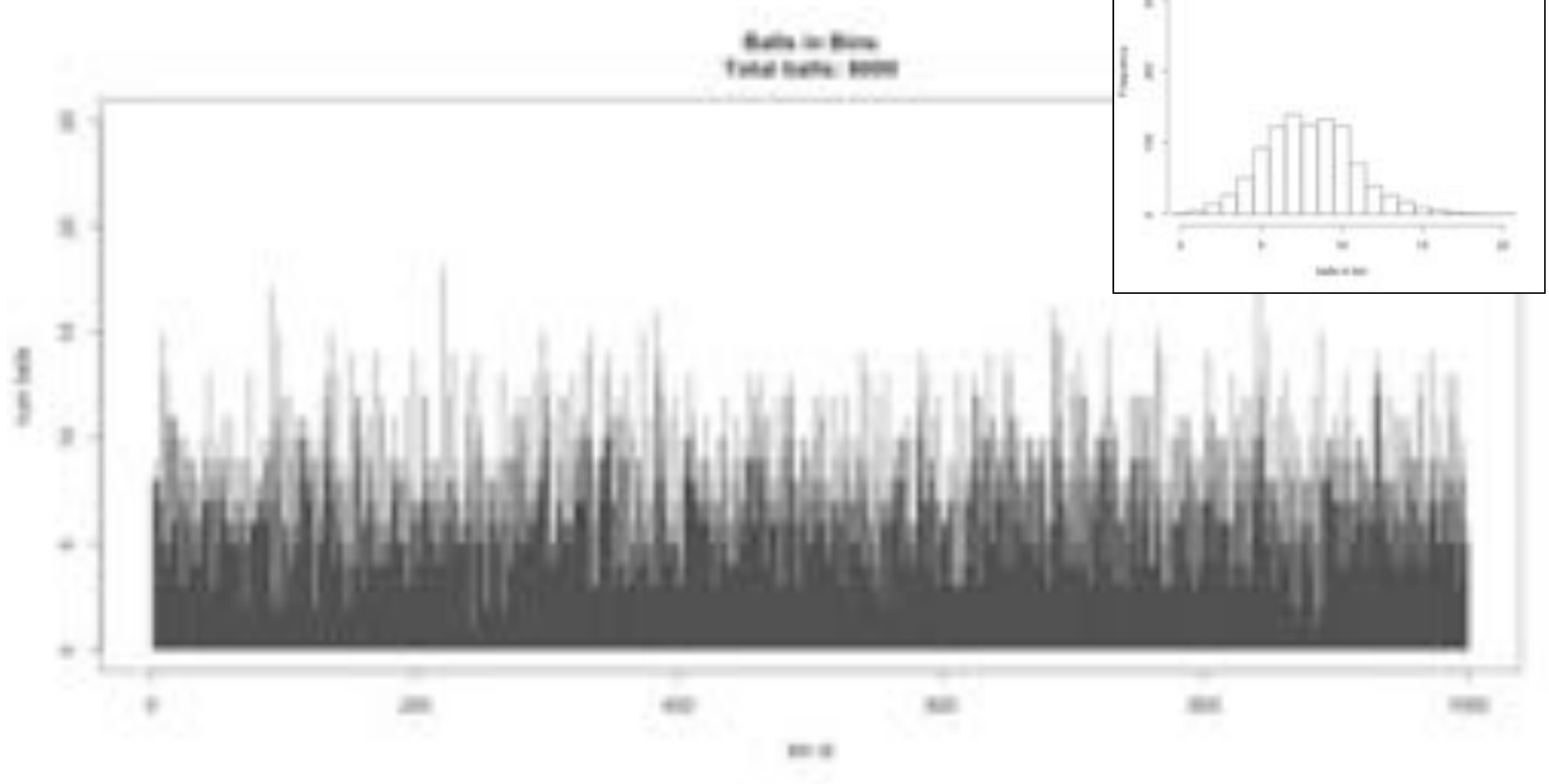
2x sequencing



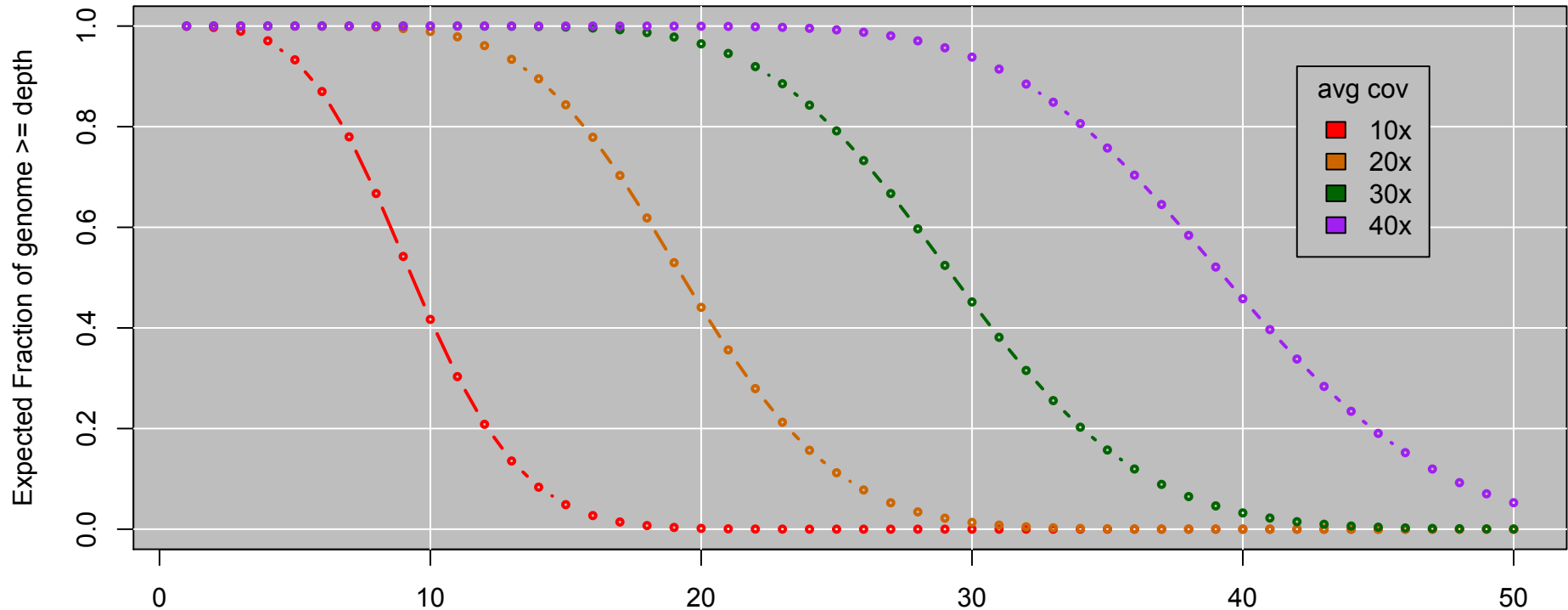
4x sequencing



8x sequencing



Genome Coverage Distribution



Expect Poisson distribution on depth

- Standard Deviation = $\sqrt{\text{cov}}$

This is the mathematical model => reality may be much worse

- Double your coverage for diploid genomes
- Can use somewhat lower coverage in a population to find common variants

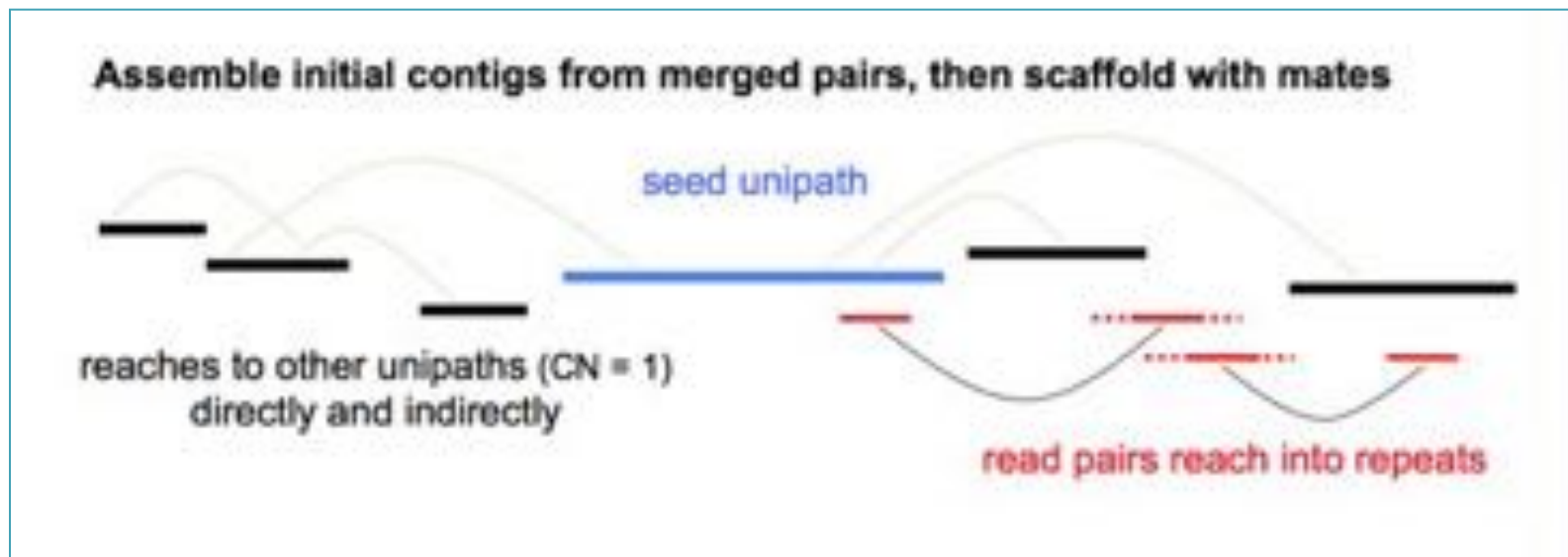
Initial Assembly Attempts with early Illumina sequencers circa 2007-2008

(older Illumina PE70 library with small insert size ~150bp)

Assembler	Contig set	N50 contig size	Max contig size	Total assembly size
Velvet	25X Agropurum	1049bp	21830bp	305.8 Mbp
Velvet	50X Agropurum	4716bp	23094bp	421.6 Mbp
Abyss	25X Agropurum	1853bp	12684bp	286.4 Mbp
Abyss	50X Agropurum	2847bp	34800bp	317.4 Mbp
Abyss	30X peach	2123bp	27079bp	187.2 Mbp

Short Read Assembly with ALLPATHS

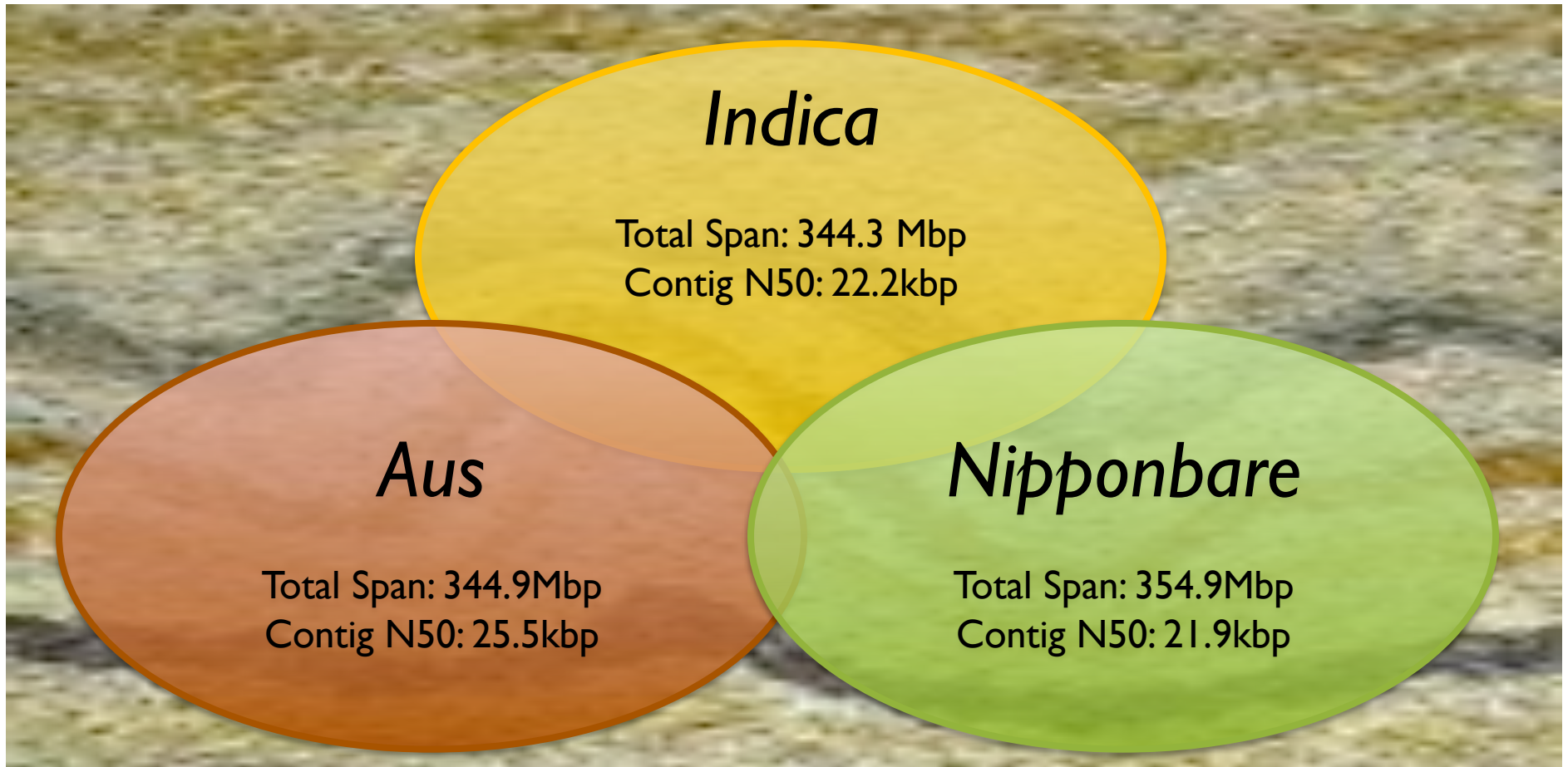
Libraries (insert types)	Fragment size (bp)	Read length (bases)	Sequence coverage (x)	Required
Fragment	180*	≥ 100	45	yes
Short jump	3,000	≥ 100 preferable	45	yes
Long jump	6,000	≥ 100 preferable	5	no**
Fosmid jump	40,000	≥ 26	1	no**



High-quality draft assemblies of mammalian genomes from massively parallel sequence data

Gnerre et al (2010) *PNAS*. doi: 10.1073/pnas.1017351108

Population structure of *Oryza sativa*

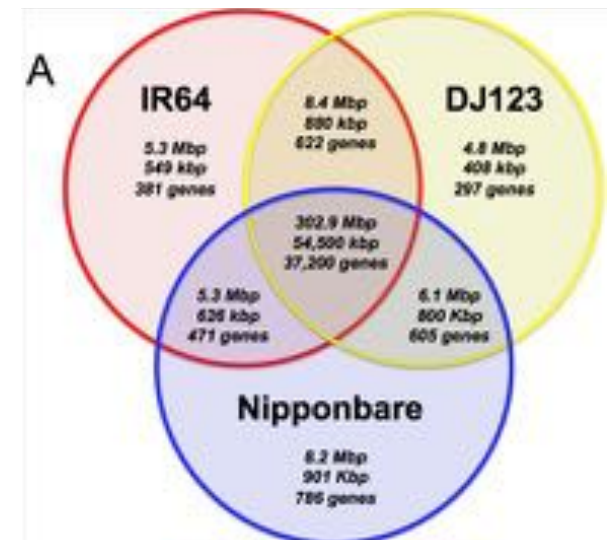


Whole genome de novo assemblies of three divergent strains of rice (*O. sativa*) documents novel gene space of *aus* and *indica*

Schatz, Maron, Stein et al (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

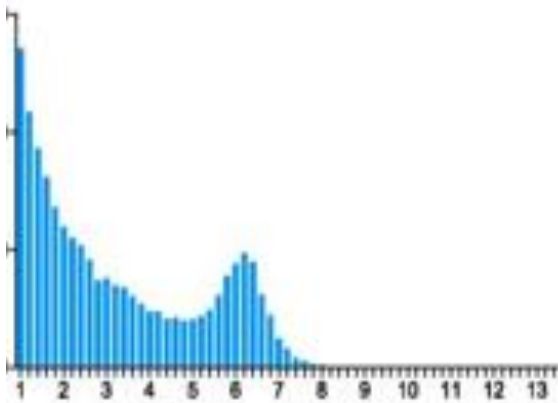


Overall sequence content

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.

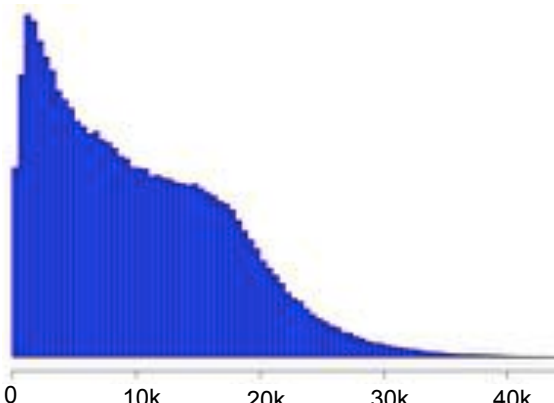
Long Read Sequencing Technology

Moleculo



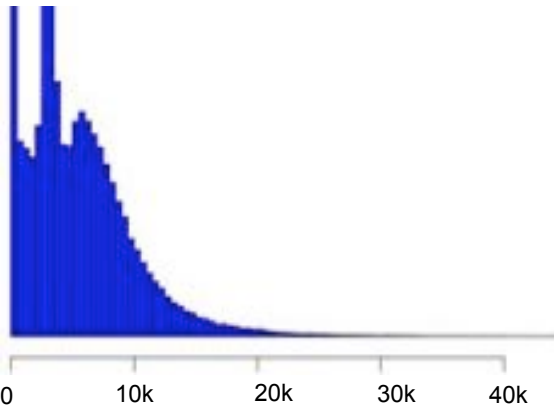
(Voskoboynik et al. 2013)

PacBio RS II



CSHL/PacBio

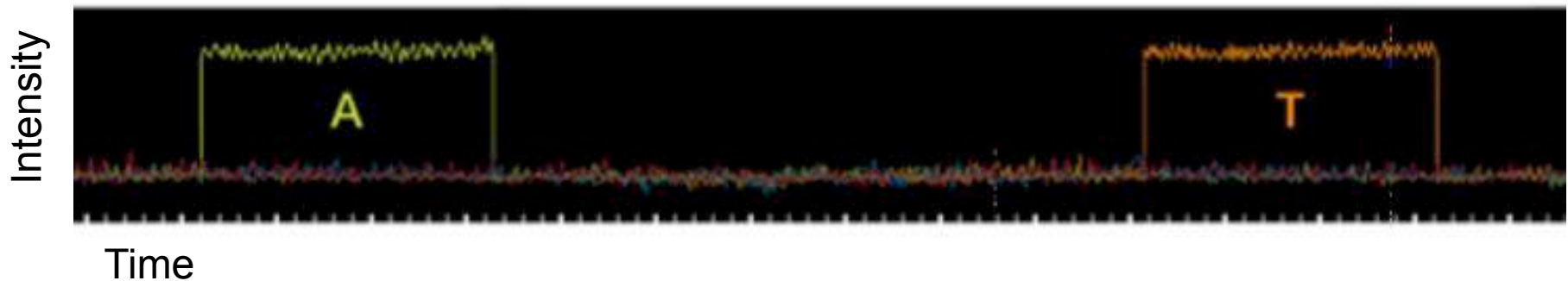
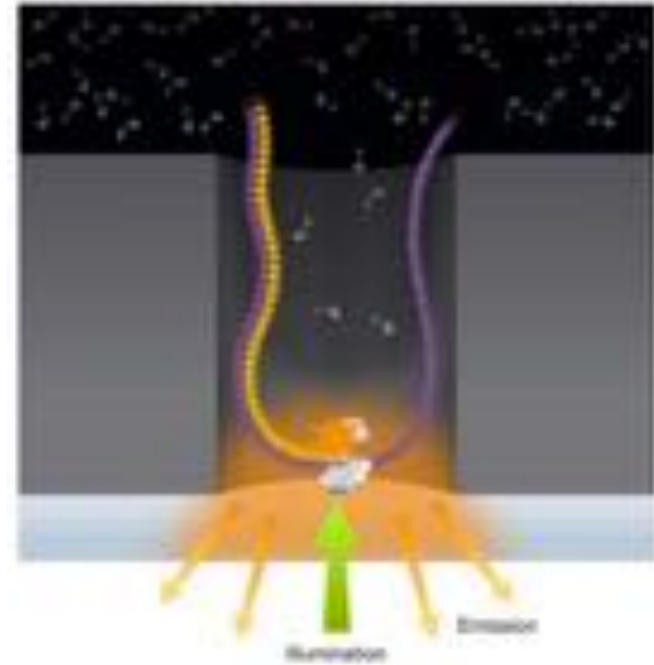
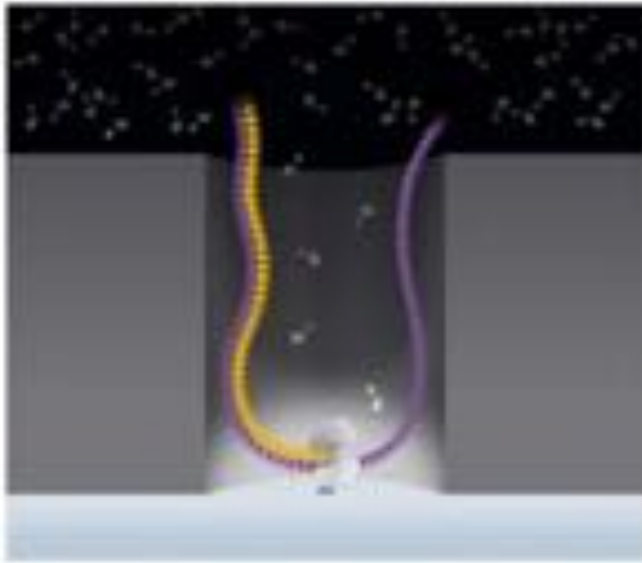
Oxford Nanopore



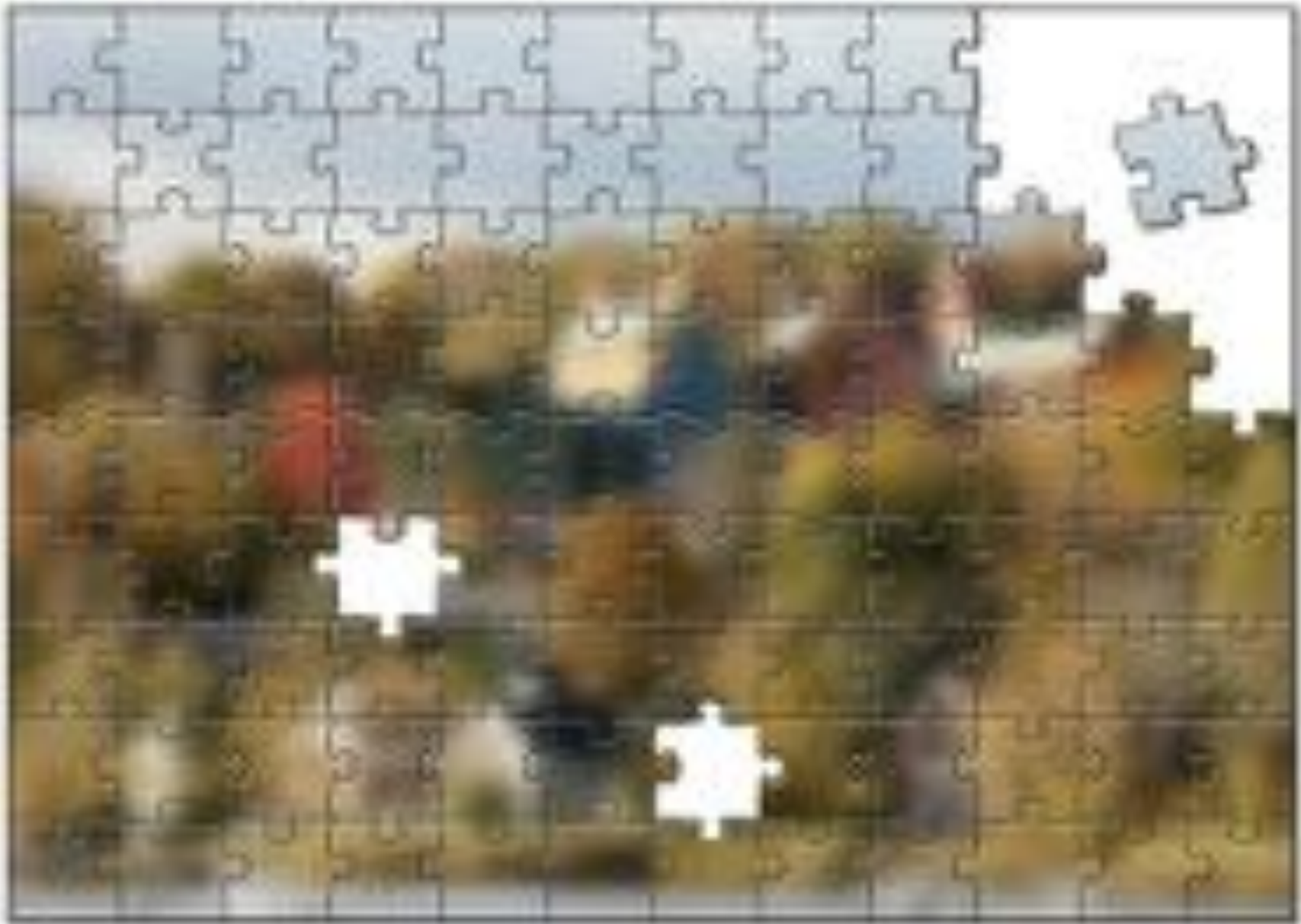
CSHL/ONT

PacBio SMRT Sequencing

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



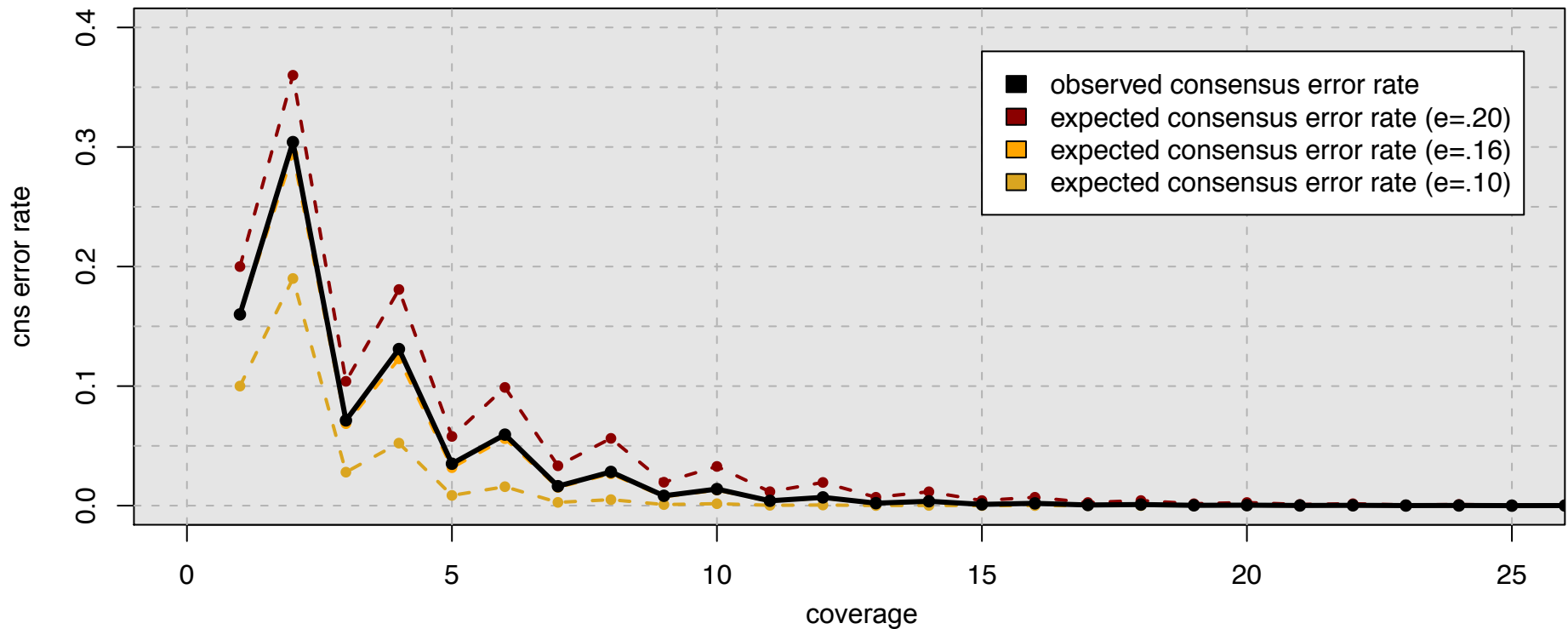
Single Molecule Sequencing



“Corrective Lens” for Sequencing



Consensus Accuracy and Coverage



Coverage can overcome random errors

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

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

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

PacBio Assembly Algorithms

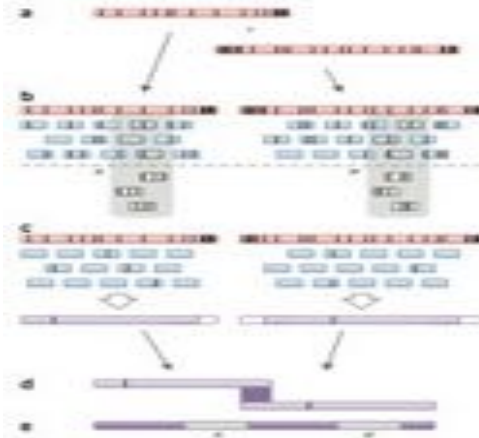
PBJelly



**Gap Filling
and Assembly Upgrade**

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

PacBioToCA & ECTools



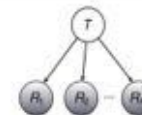
**Hybrid/PB-only Error
Correction**

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

HGAP & Quiver



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



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

**PB-only Correction &
Polishing**

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

< 5x

PacBio Coverage

> 50x

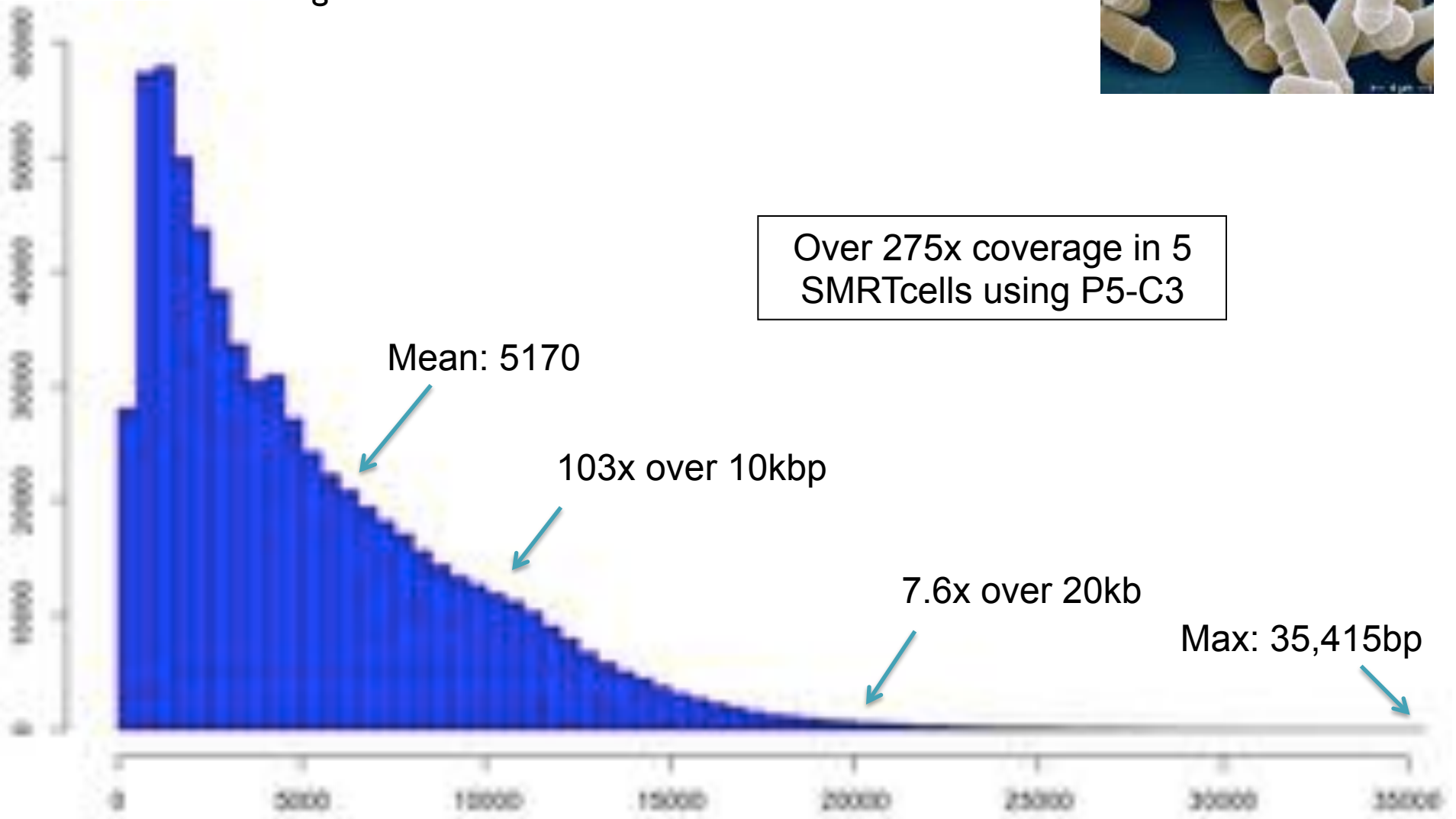
S. pombe dg2 I

PacBio RS II sequencing at CSHL

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



Over 275x coverage in 5 SMRTcells using P5-C3



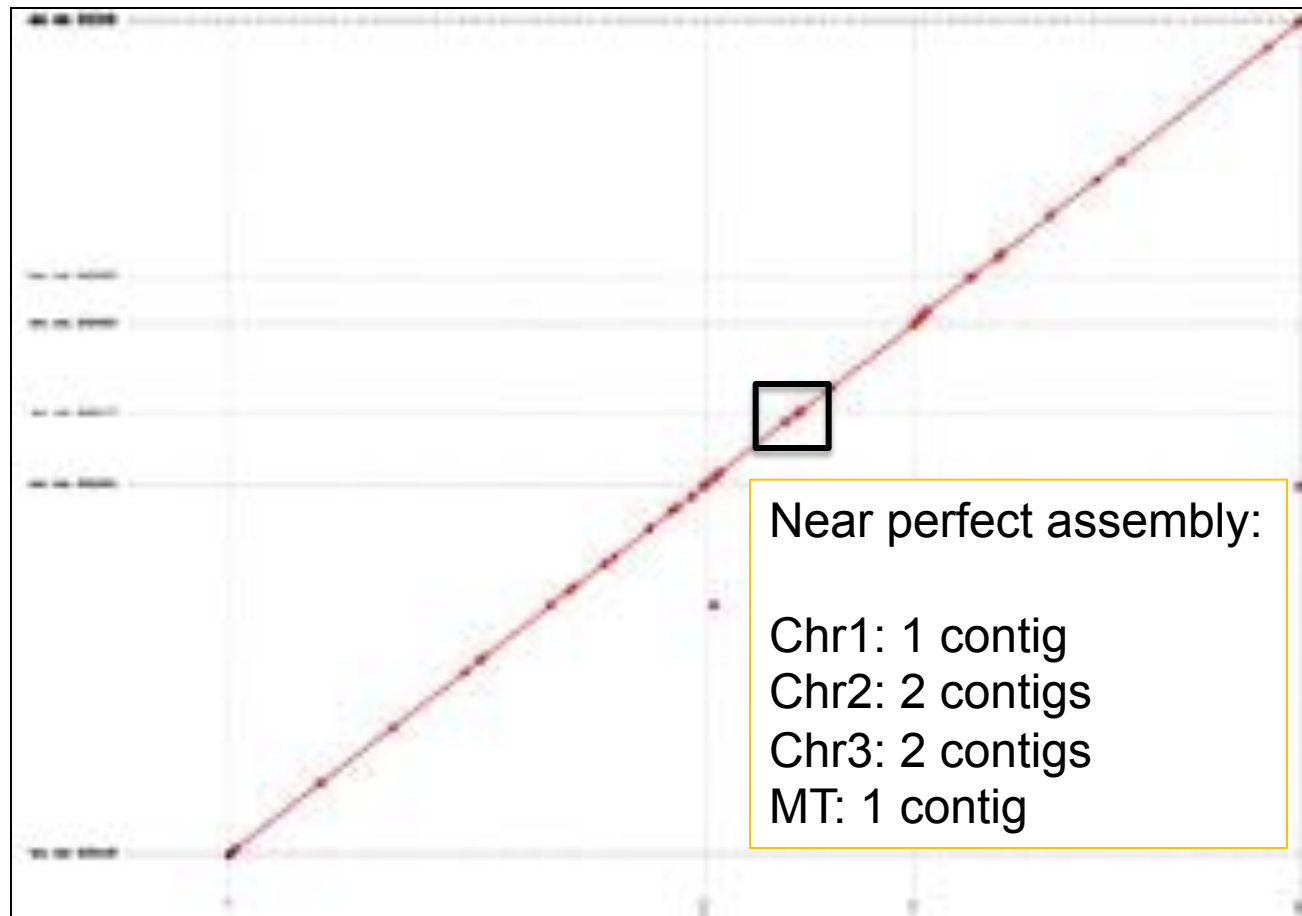
S. pombe dg2 I

ASM294 Reference sequence

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

PacBio assembly using HGAP + Celera Assembler

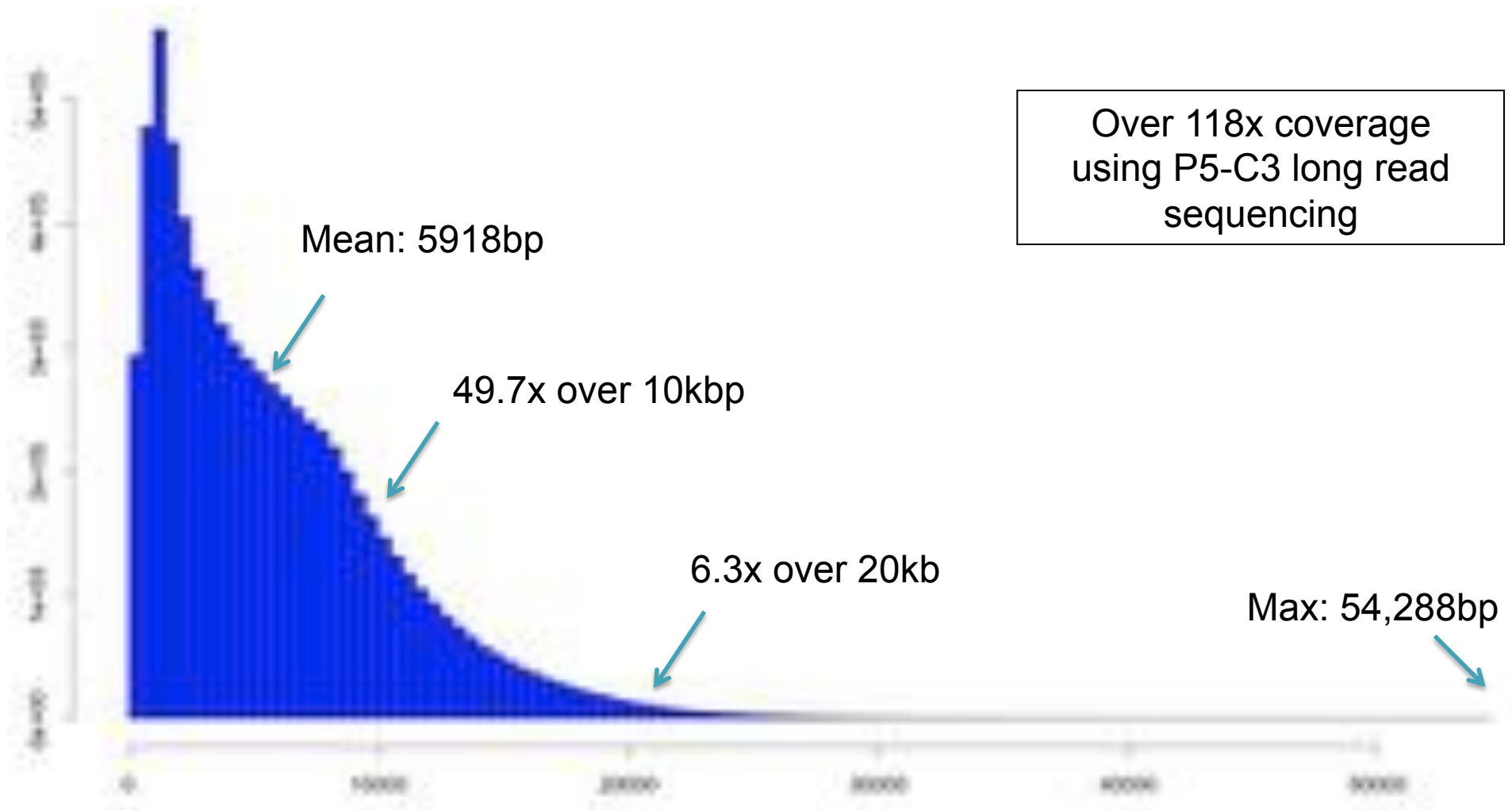
- 12.7Mbp; 13 non-redundant contigs; N50: 3.83Mbp; >99.98% id



O. sativa pv Indica (IR64)

PacBio RS II sequencing at PacBio

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

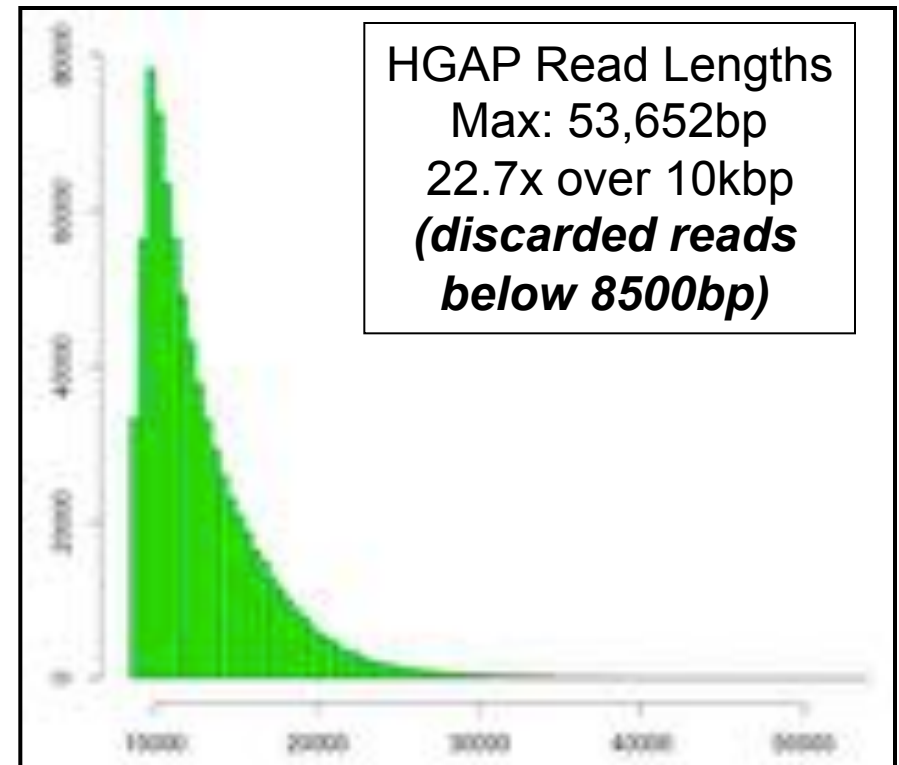


O. sativa pv Indica (IR64)

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



Current Collaborations



Pinapple
UIUC



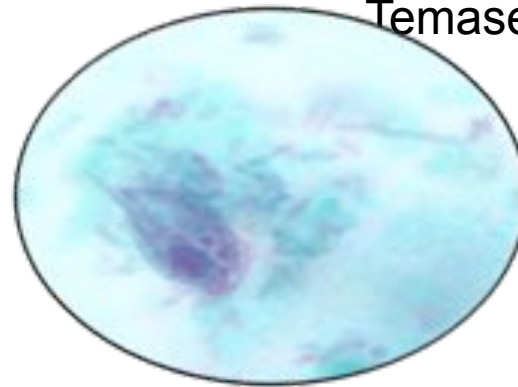
Human
CSHL/OICR/PacBio



Asian Sea Bass
Temasek Life Sciences

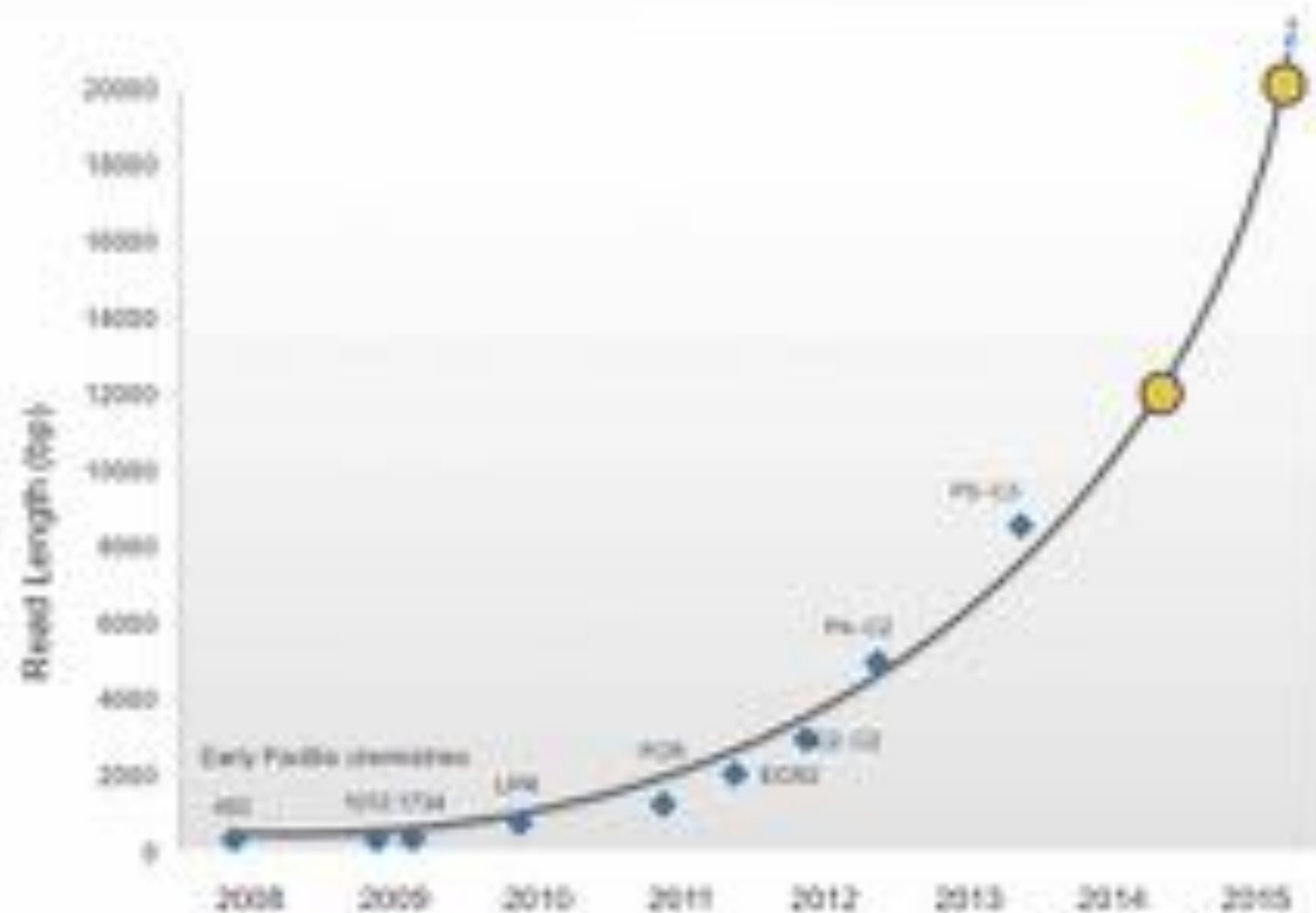


M. ligano
Hannon

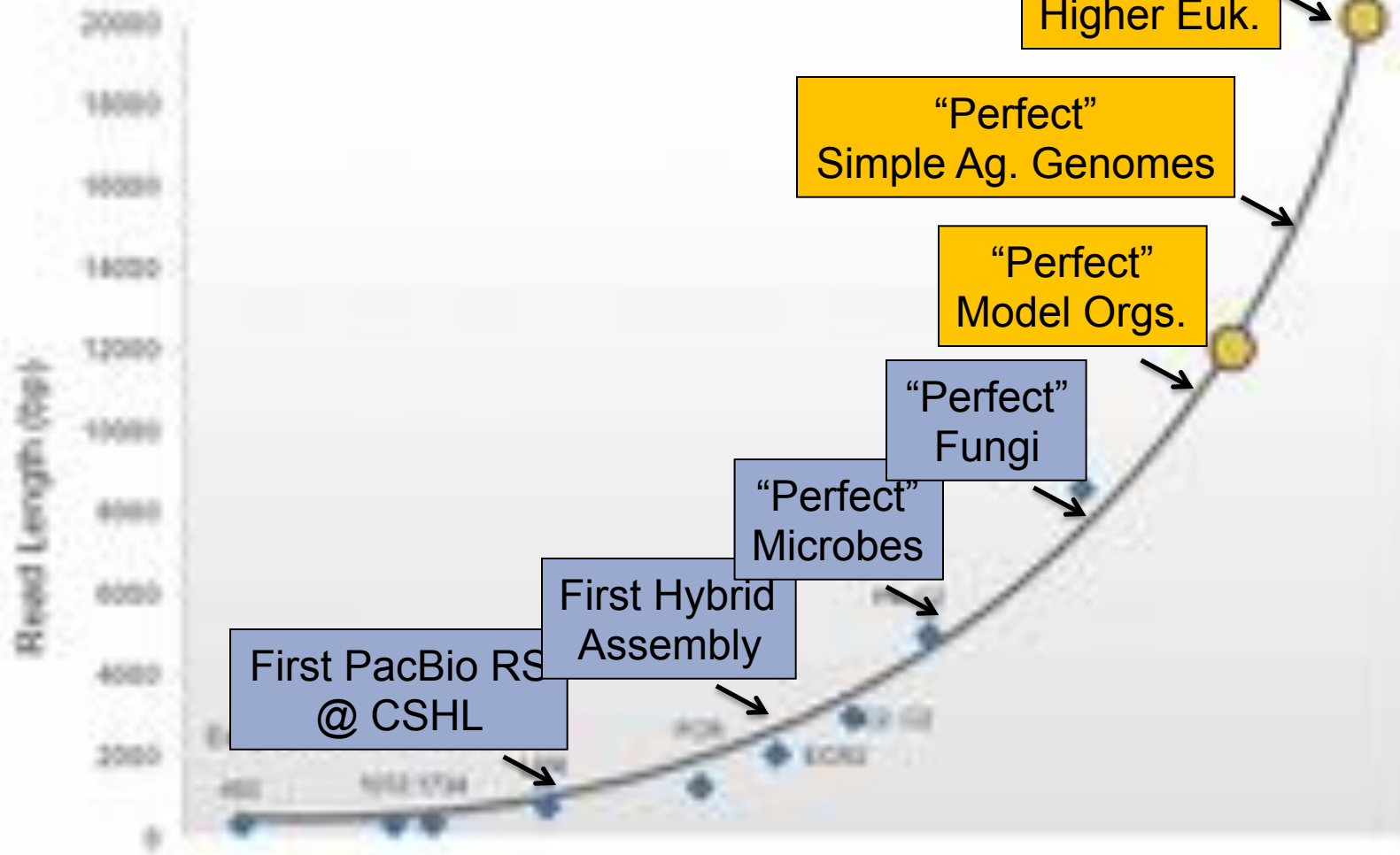


P. hominis
NYU

PacBio® Advances in Read Length



Advances in Assembly



Error correction and assembly complexity of single molecule sequencing reads.

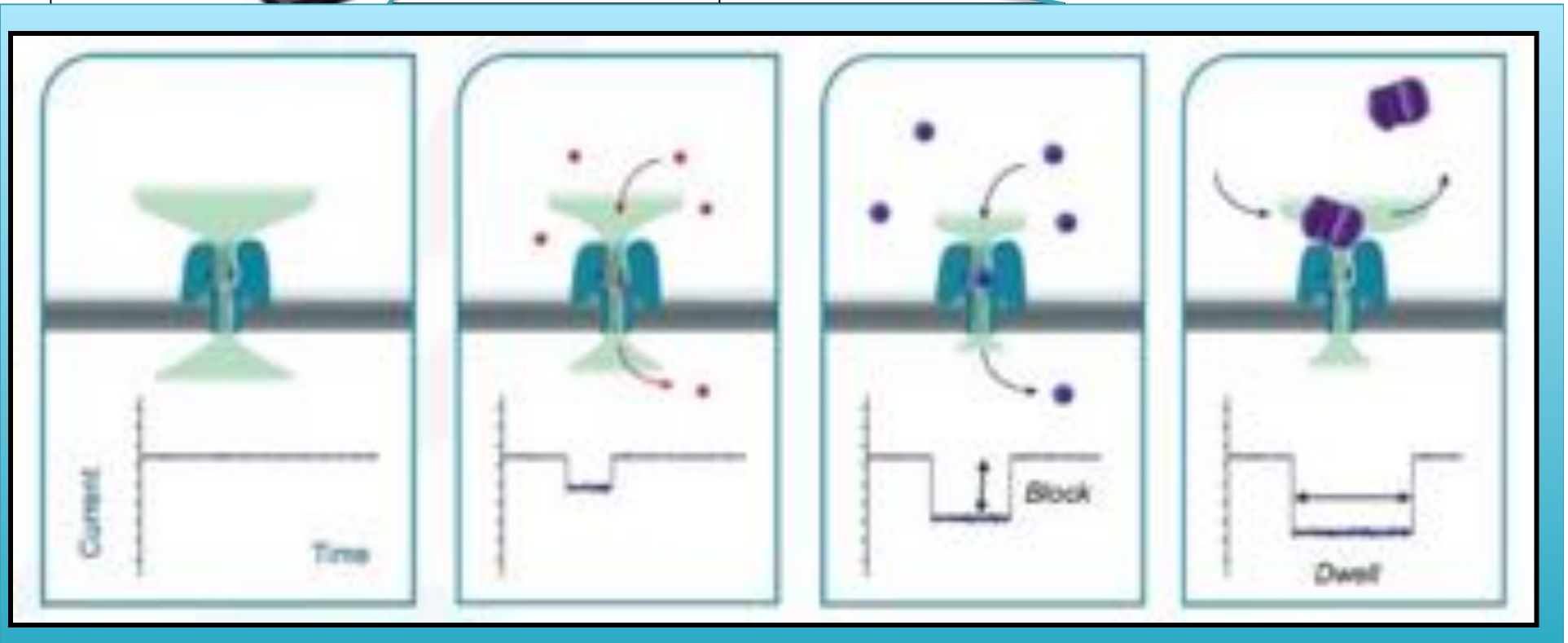
Lee, H*, Gurtowski, J*, Yoo, S, Marcus, S, McCombie, WR, Schatz, MC

<http://www.biorxiv.org/content/early/2014/06/18/006395>

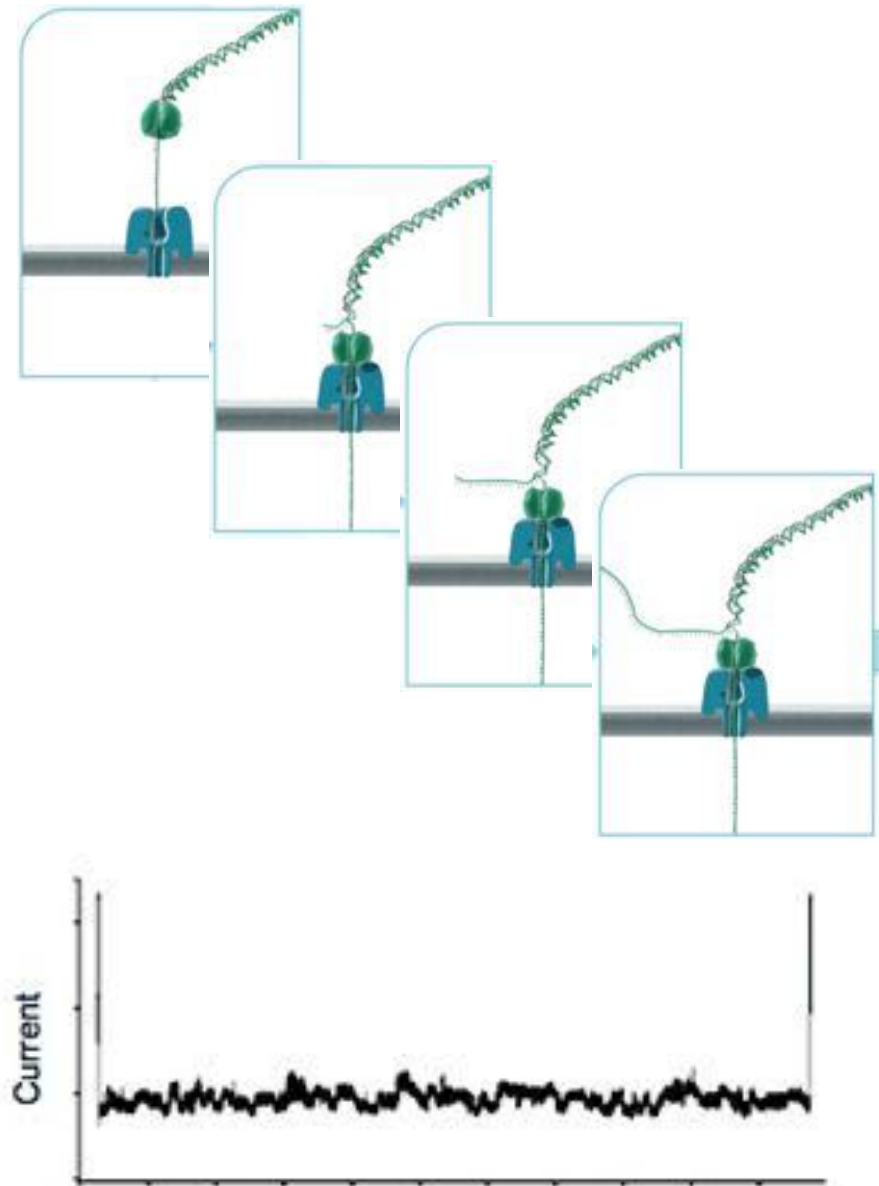
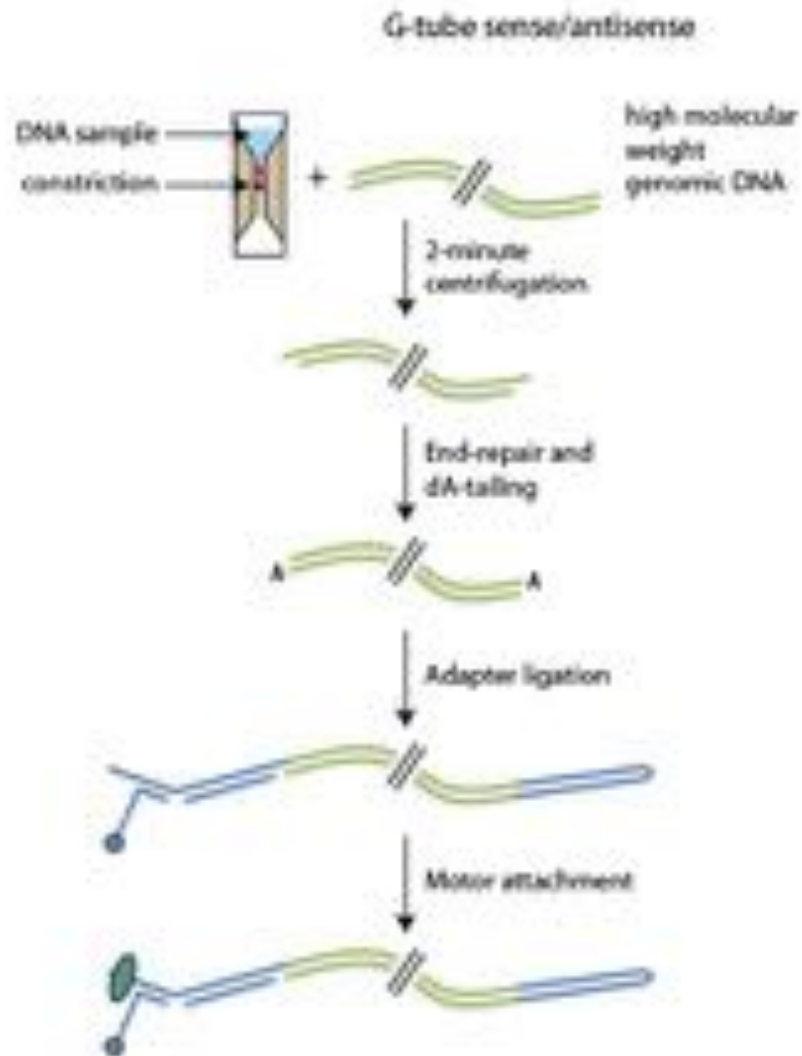
Oxford Nanopore MinION



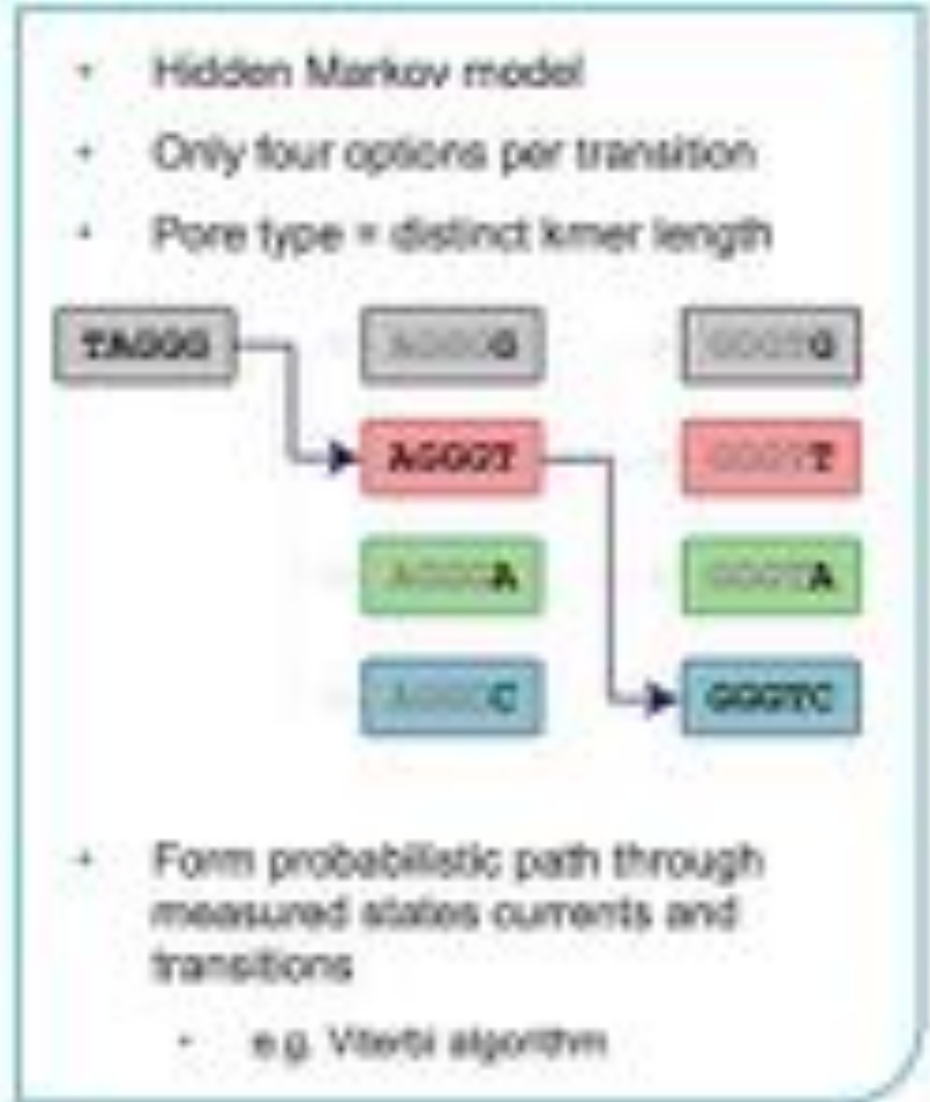
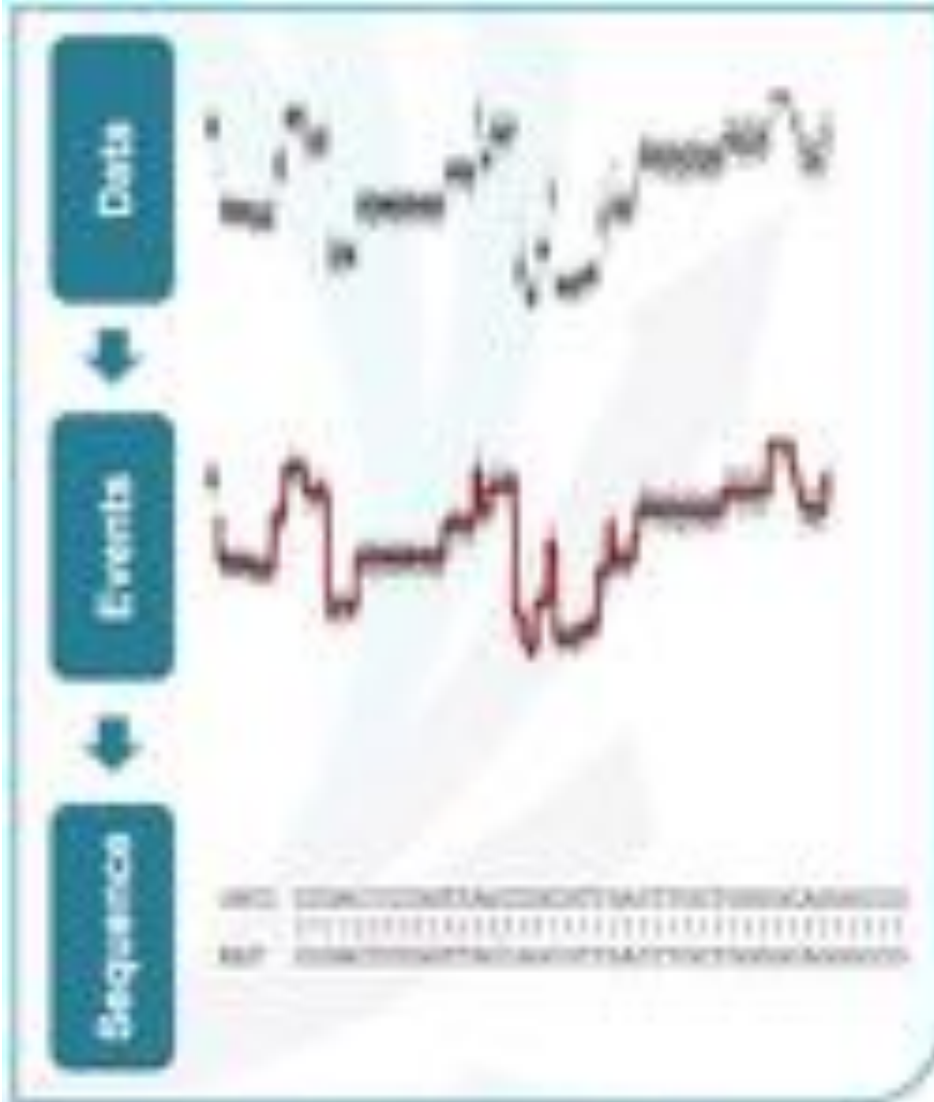
- Thumb drive sized sequencer powered over USB
- Capacity for 512 reads at once
- Senses DNA by measuring changes to ion flow



Nanopore Sequencing

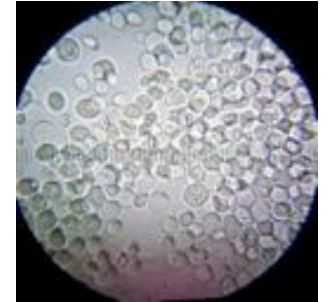


Nanopore Basecalling

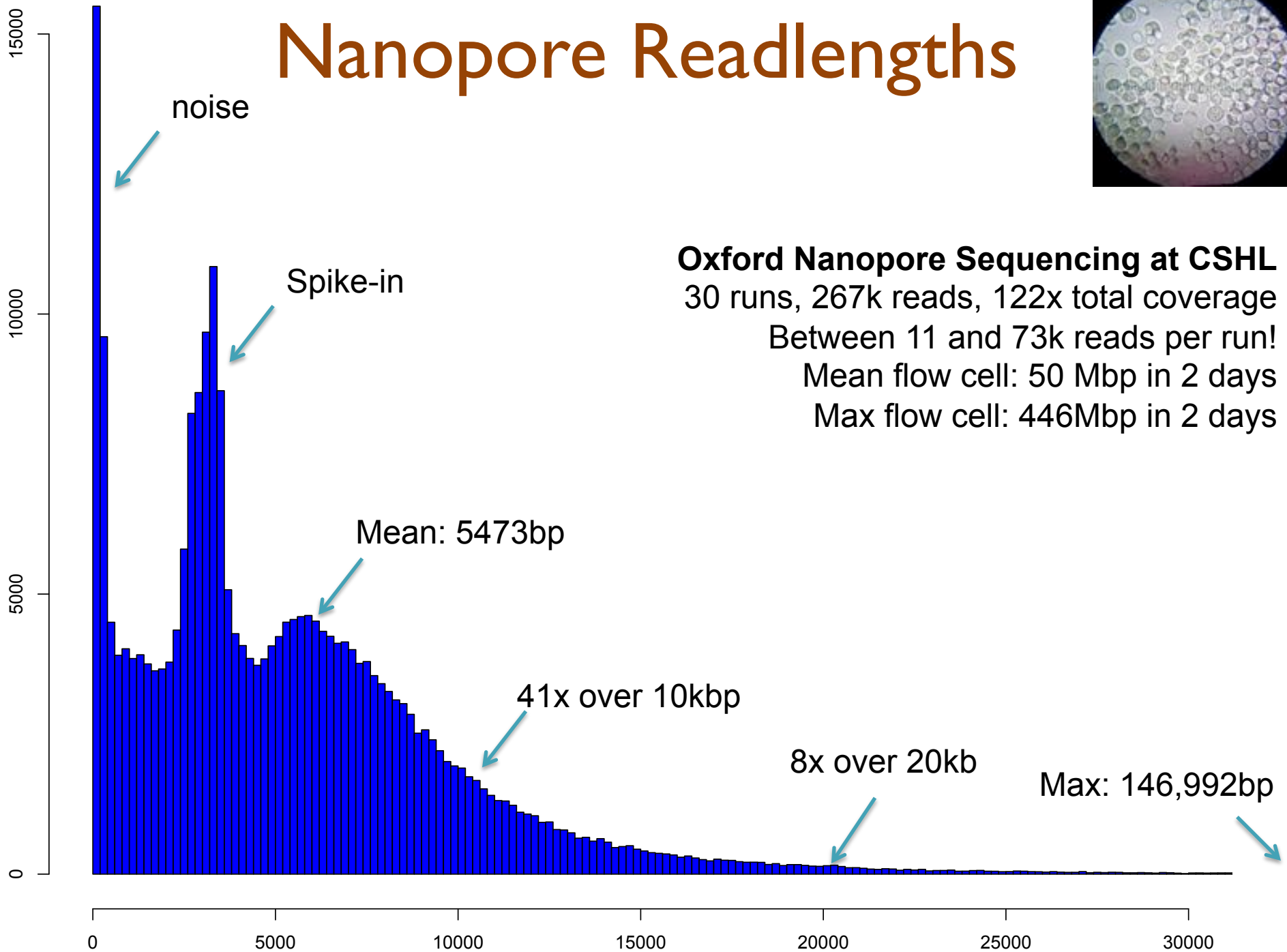


Basecalling currently performed at Amazon with frequent updates to algorithm

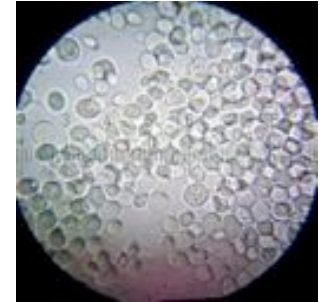
Nanopore Readlengths



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



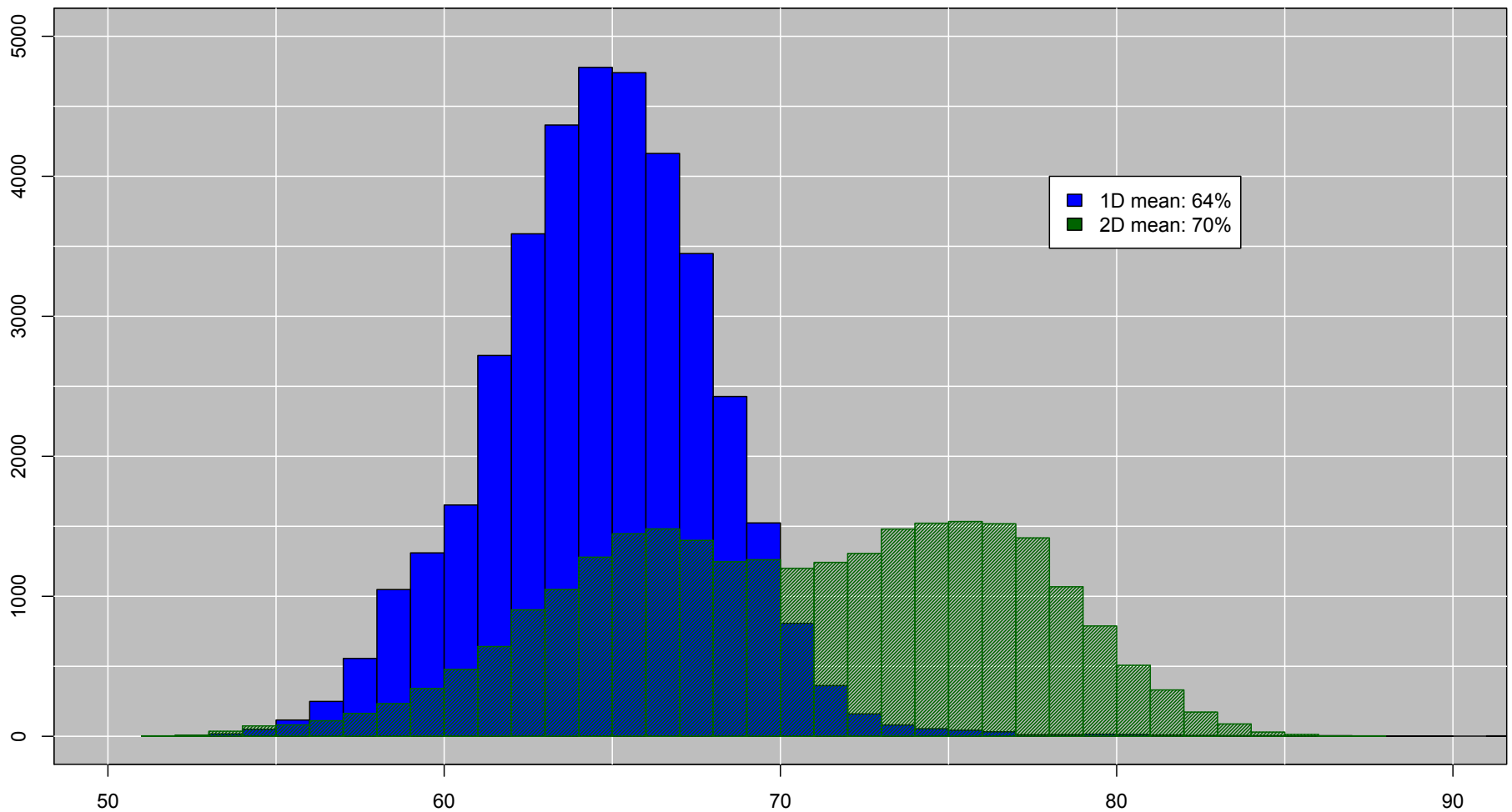
Nanopore Accuracy



Alignment Quality (BLASTN)

Of reads that align, average ~64% identity

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

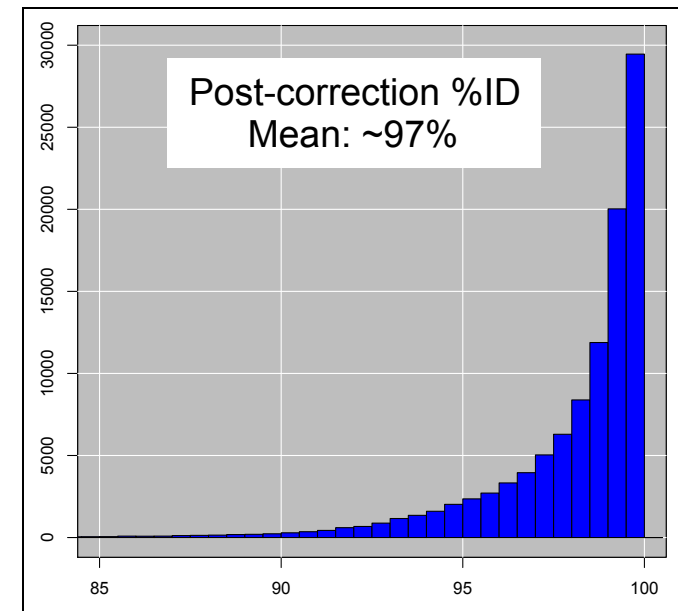


NanoCorr: Nanopore-Illumina Hybrid Error Correction

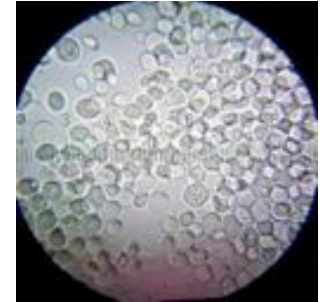


<https://github.com/jgurtowski/nanocorr>

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
 - Currently using Pacbio’s pbdagcon



Long Read Assembly



S288C Reference sequence

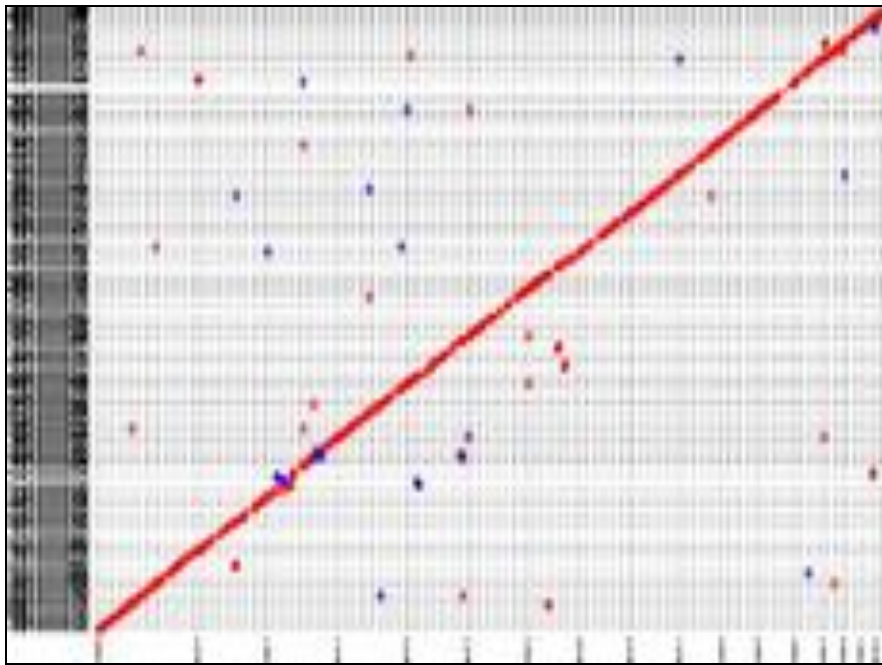
- 12.1Mbp; 16 chromo + mitochondria; N50: 924kbp

Illumina MiSeq



30x, 300bp PE (Flashed)

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

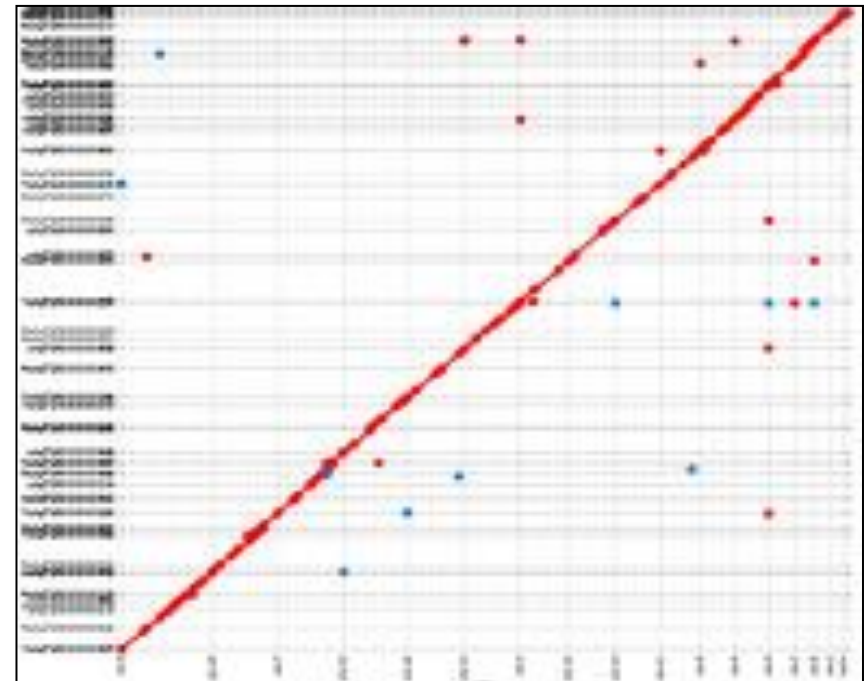


Oxford Nanopore



NanoCorr + Celera Assembler

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



Genomic Futures?



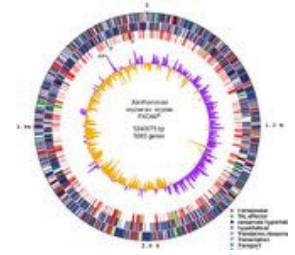
Zamin Iqbal and 5 others retweeted

GenomeWeb InSequence @InSequence · Oct 20

Oxford Nanopore shows off Promethion at ASHG, #ASHG14 #nanopore



Assembly Summary



Assembly quality depends on

1. **Coverage**: low coverage is mathematically hopeless
 2. **Repeat composition**: high repeat content is challenging
 3. **Read length**: longer reads help resolve repeats
 4. **Error rate**: errors reduce coverage, obscure true overlaps
- Assembly is a hierarchical, starting from individual reads, build high confidence contigs/unitigs, incorporate the mates to build scaffolds
 - Extensive error correction is the key to getting the best assembly possible from a given data set
 - Watch out for collapsed repeats & other misassemblies
 - Globally/Locally reassemble data from scratch with better parameters & stitch the 2 assemblies together

What should we expect from an assembly?

Analysis of dozens of genomes from across the tree of life with real and simulated data

Summary & Recommendations

- < 100 Mbp: HGAP/PacBio2CA @ 100x PB C3-P5
expect near perfect chromosome arms
- < 1GB: HGAP/PacBio2CA @ 100x PB C3-P5
high quality assembly: contig N50 over 1Mbp
- > 1GB: hybrid/gap filling
expect contig N50 to be 100kbp – 1Mbp
- > 5GB: Email mschatz@cshl.edu



Error correction and assembly complexity of single molecule sequencing reads.

Lee, H*, Gurtowski, J*, Yoo, S, Marcus, S, McCombie, WR, Schatz, MC

<http://www.biorxiv.org/content/early/2014/06/18/006395>

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

Rachel Sherman

Greg Vulture

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

IT & Meetings Depts.

Pacific Biosciences

Oxford Nanopore



National Human
Genome Research
Institute



U.S. DEPARTMENT OF
ENERGY

SFARI

SIMONS FOUNDATION
AUTISM RESEARCH INITIATIVE



Thank you

<http://schatzlab.cshl.edu>

@mike_schatz