De novo assembly of complex genomes using single molecule sequencing

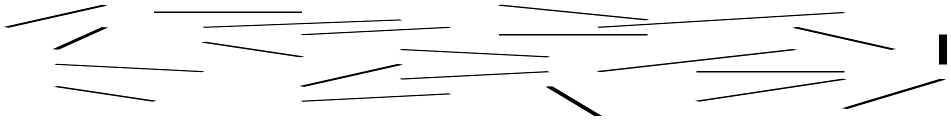
Michael Schatz

Jan 14, 2014
PAG XXII

@mike_schatz / #PAGXXII

Assembling a Genome

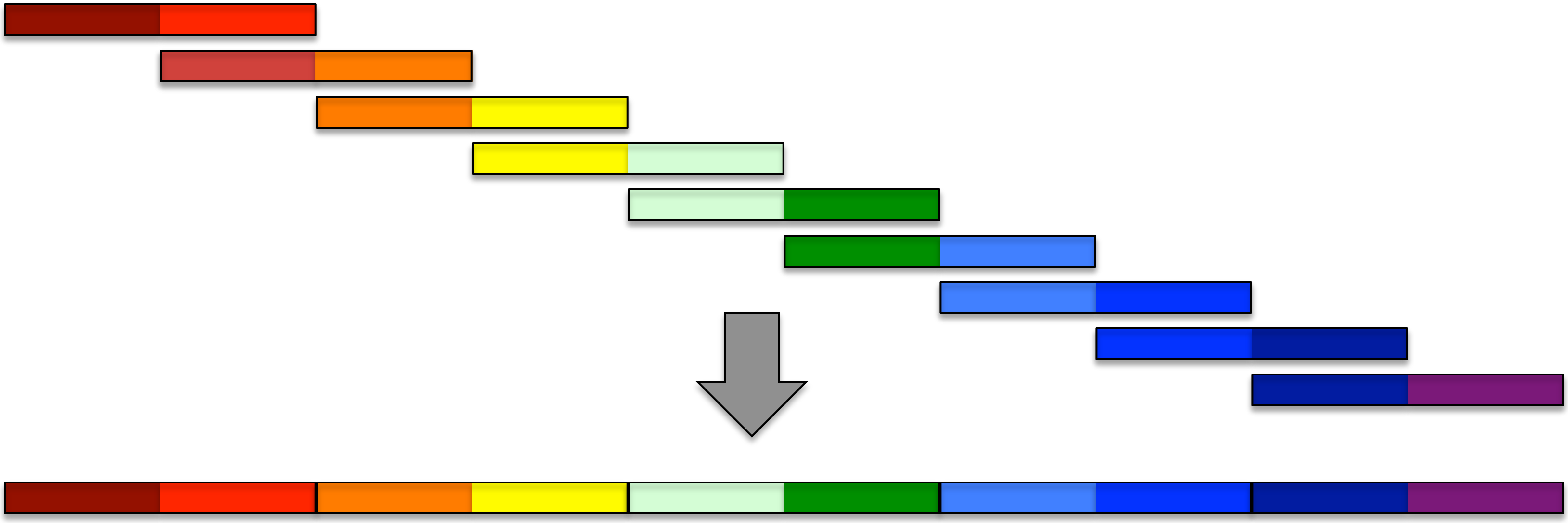
1. Shear & Sequence DNA



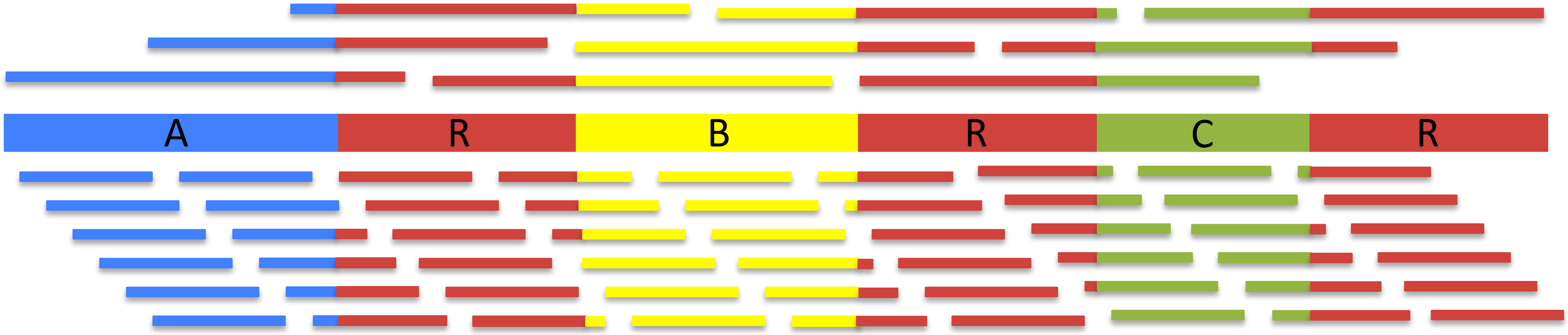
2. Construct assembly graph from overlapping reads

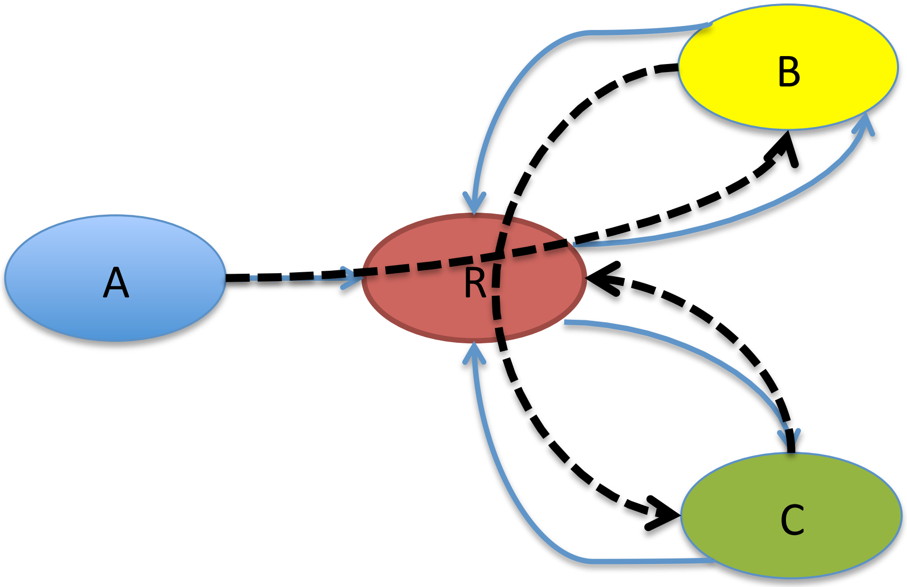
```
...AGCCTAGGGATGCGCGACACGT
          GGATGCGCGACACGTCGCATATCCGGTTTGGTCAACCTCGGACGGAC
                                          CAACCTCGGACGGACCTCAGCGAA...
```

3. Simplify assembly graph

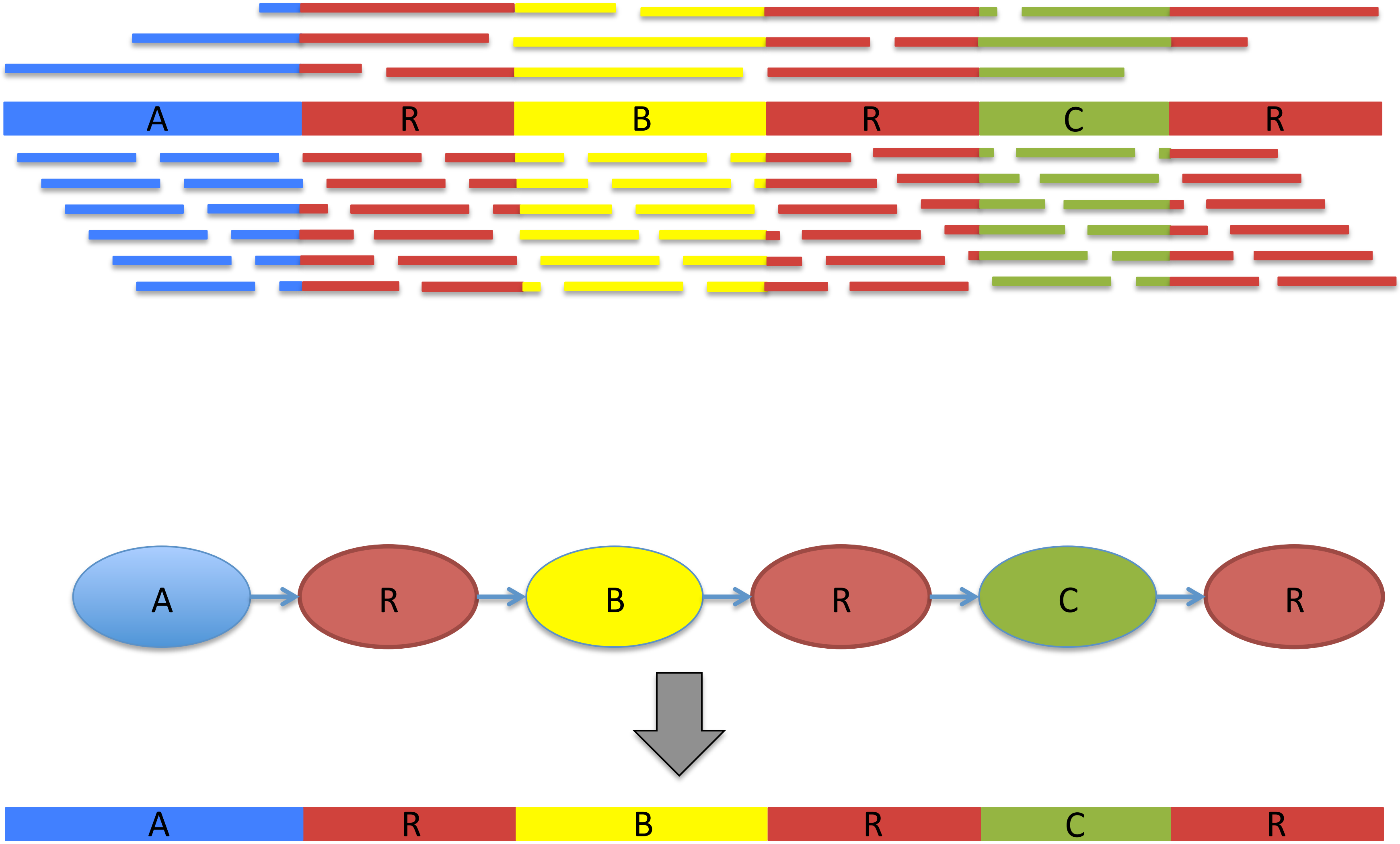


Assembly Complexity



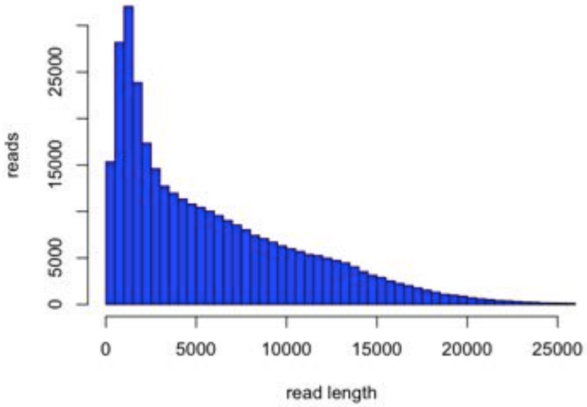


Assembly Complexity

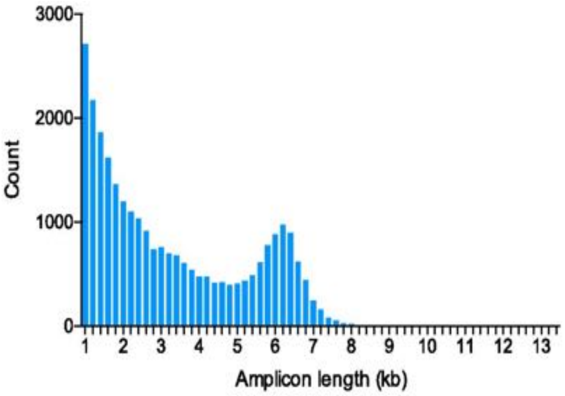


Single Molecule Sequencing Technology

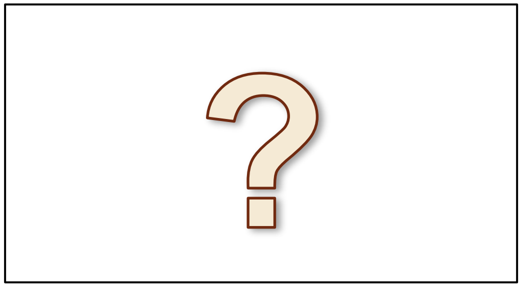
PacBio RS II

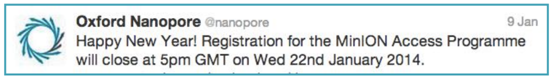


Moleculo



Oxford Nanopore

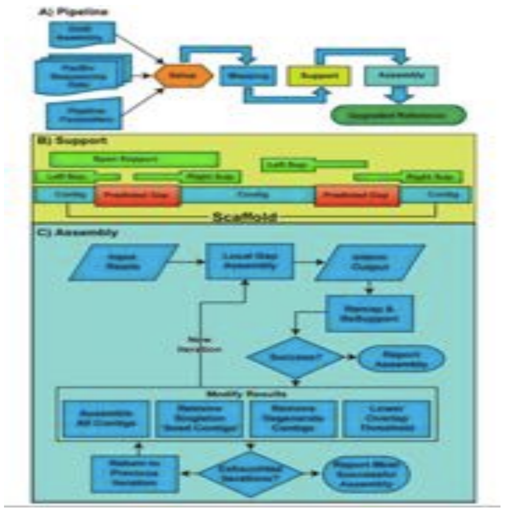




Oxford Nanopore @nanopore 9 Jan
Happy New Year! Registration for the MinION Access Programme will close at 5pm GMT on Wed 22nd January 2014.

PacBio Assembly Algorithms

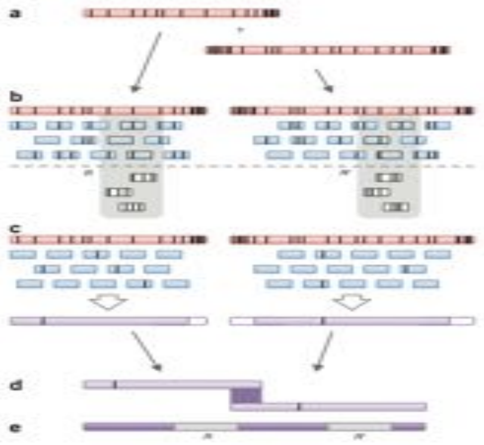
PBJelly



Gap Filling and Assembly Upgrade

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

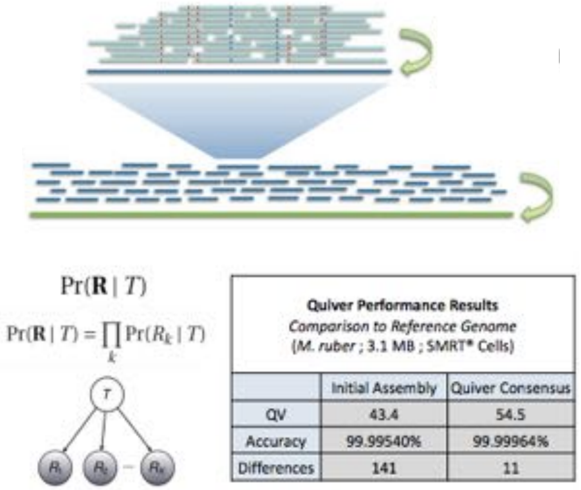
PacBioToCA & ECTools



Hybrid/PB-only Error Correction

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

HGAP & Quiver



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

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

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

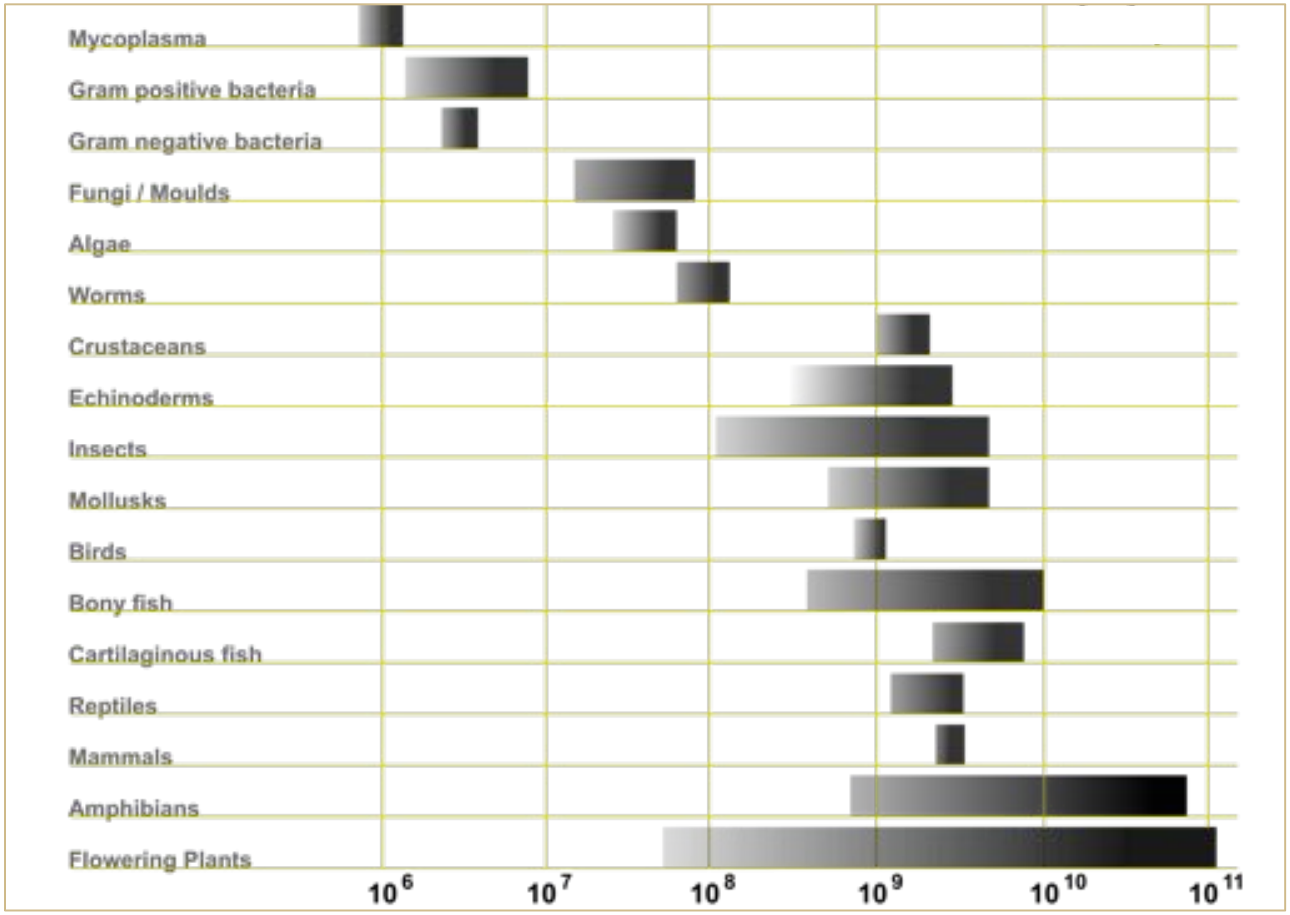
	Initial Assembly	Quiver Consensus
QV	43.4	54.5
Accuracy	99.99540%	99.99964%
Differences	141	11

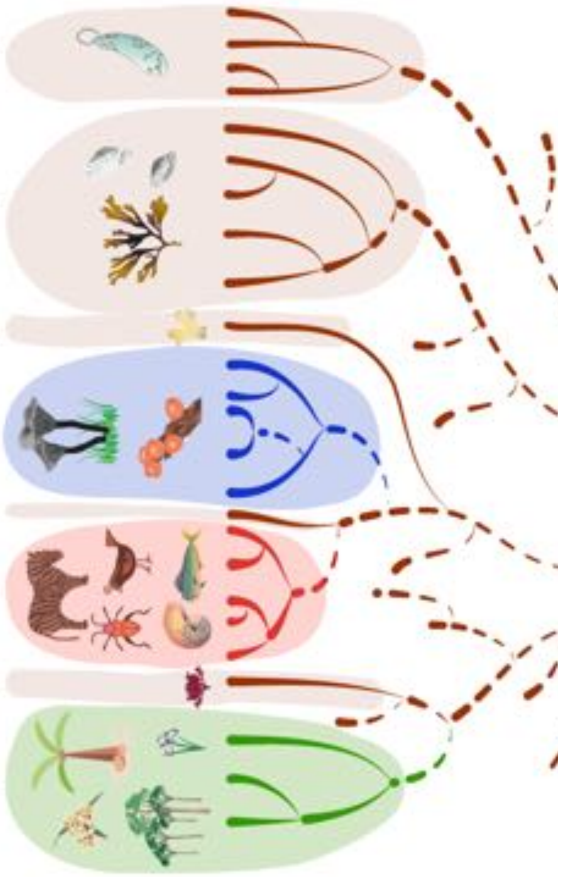
PB-only Correction & Polishing

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

< 5x PacBio Coverage > 50x

What should we expect from an assembly?



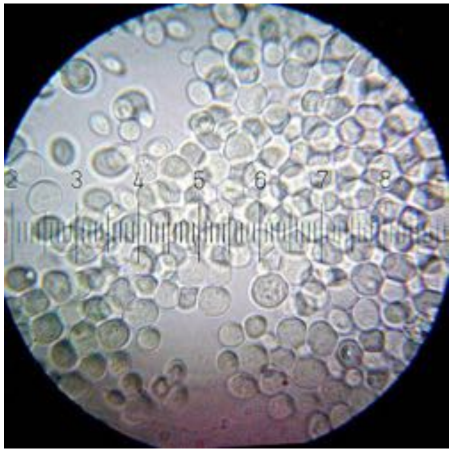


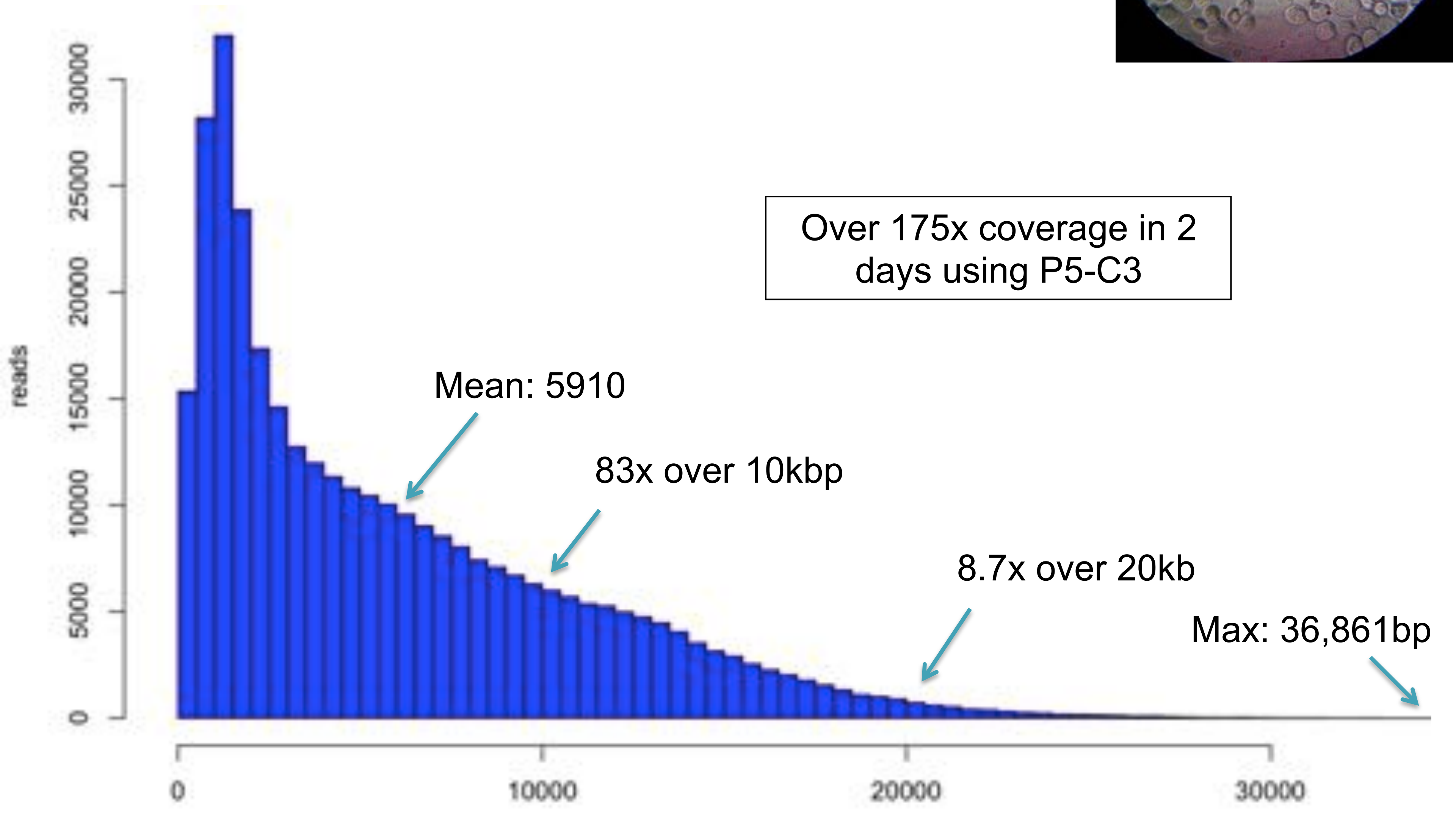
https://en.wikipedia.org/wiki/Genome_size

S. cerevisiae W303

PacBio RS II sequencing at CSHL by Dick McCombie

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





Over 175x coverage in 2 days using P5-C3

Mean: 5910

83x over 10kbp

8.7x over 20kb

Max: 36,861bp

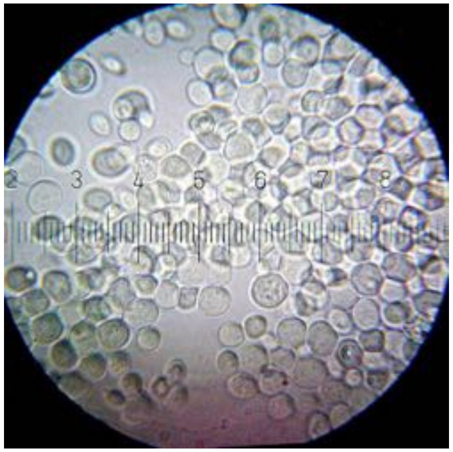
S. cerevisiae W303

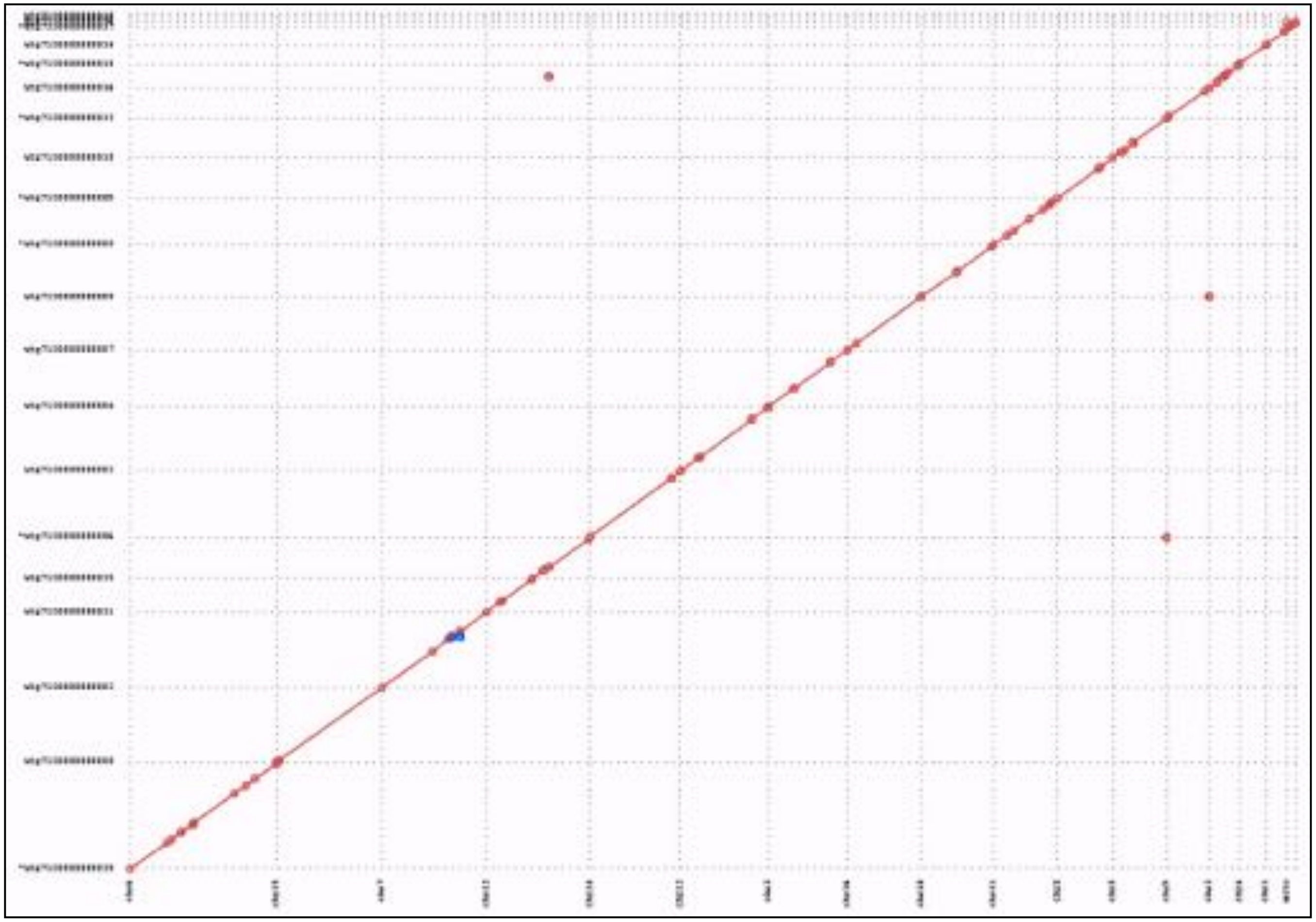
S288C Reference sequence

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

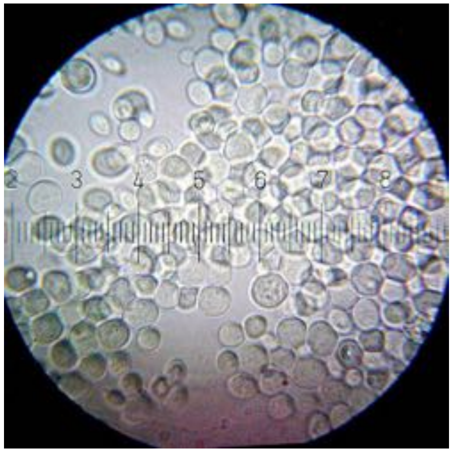
PacBio assembly using HGAP + Celera Assembler

- 12.4Mbp; 21 non-redundant contigs; N50: 811kbp; >99.8% id





S. cerevisiae W303

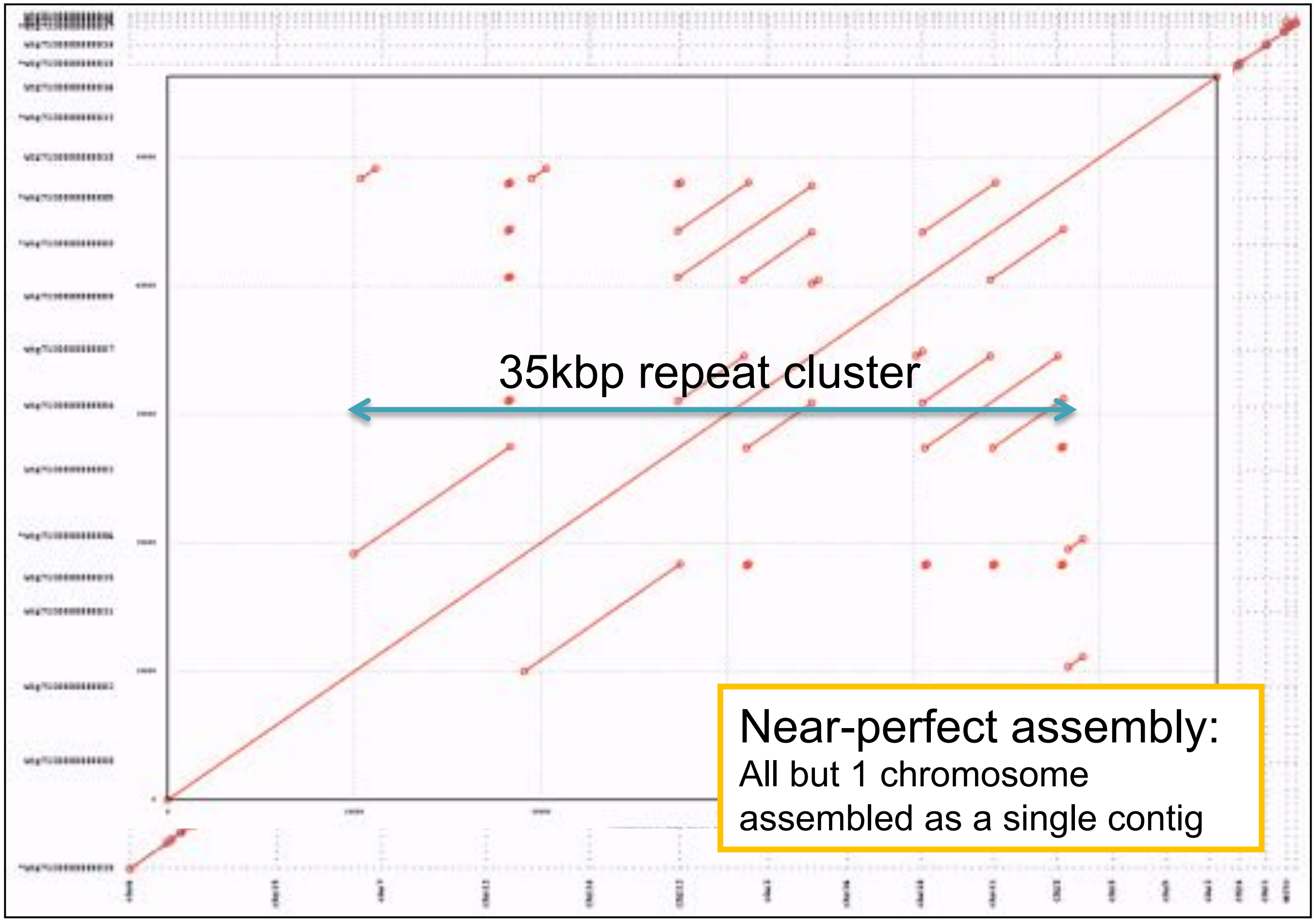


S288C Reference sequence

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

PacBio assembly using HGAP + Celera Assembler

- 12.4Mbp; 21 non-redundant contigs; N50: 811kbp; >99.8% id

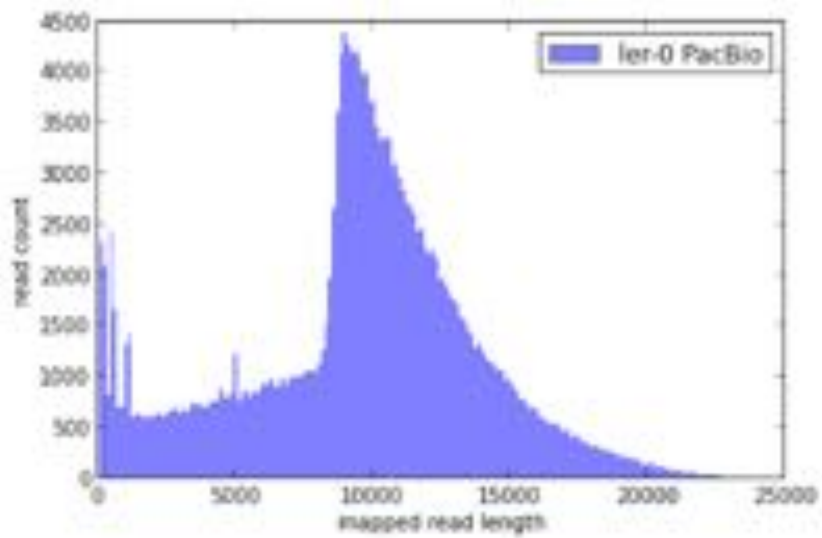


35kbp repeat cluster

Near-perfect assembly:
All but 1 chromosome
assembled as a single contig

A. thaliana Ler-0

http://blog.pacificbiosciences.com/2013/08/new-data-release-arabidopsis-assembly.html



A. thaliana Ler-0 sequenced at PacBio

- Sequenced using the previous P4 enzyme and C2 chemistry
- Size selection using an 8 Kb to 50 Kb elution window on a BluePippin™ device from Sage Science
- Total coverage >119x

Genome size:	124.6 Mbp
Chromosome N50:	23.0 Mbp
Raw data:	11 Gb

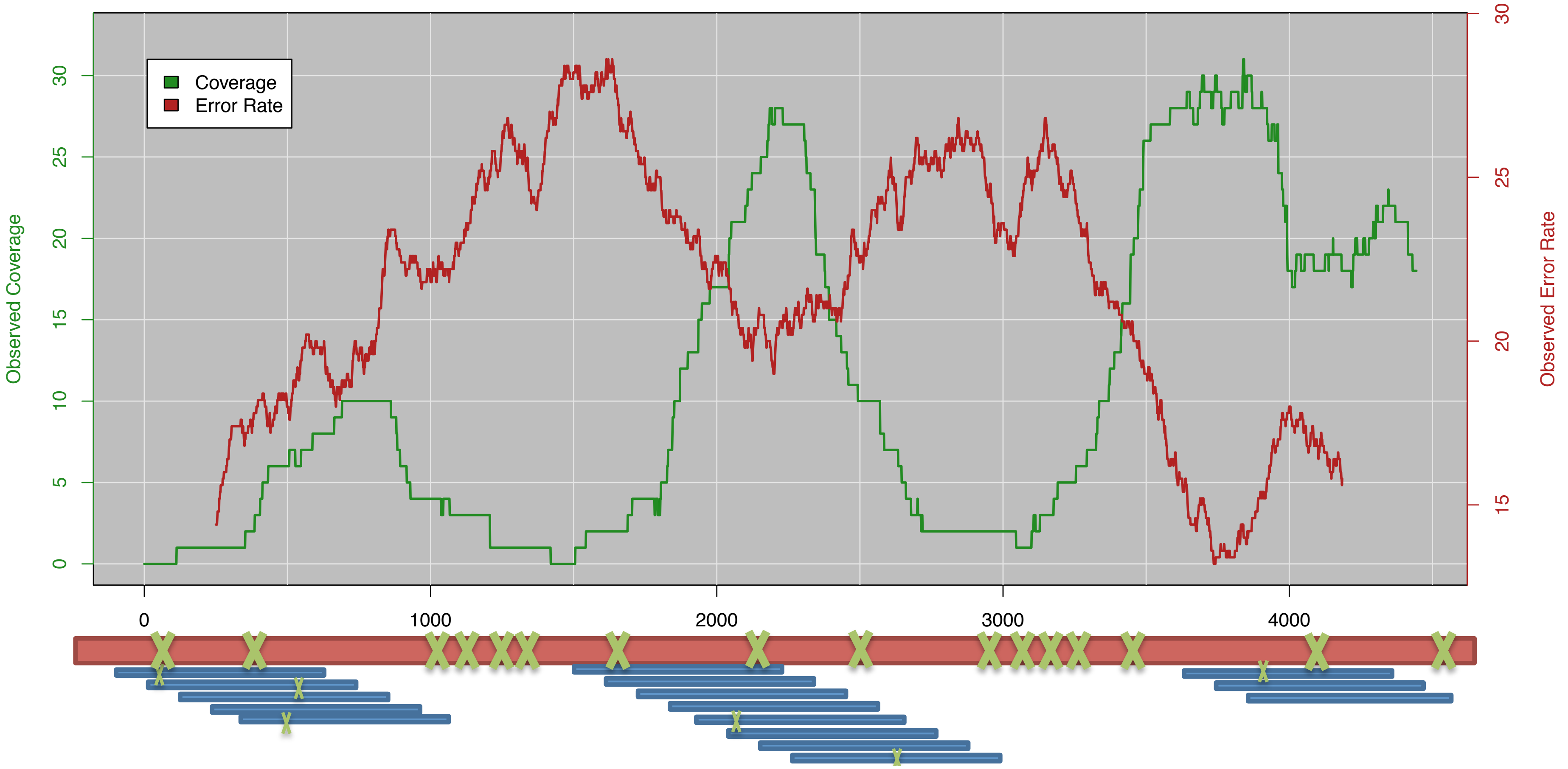
Sum of Contig Lengths:	149.5Mb
N50 Contig Length:	8.4 Mb
Number of Contigs:	1788

High quality assembly of chromosome arms
Assembly Performance: 8.4Mbp/23Mbp = 36%
MiSeq assembly: 63kbp/23Mbp [.2%]

Hybrid Approaches for Larger Genomes

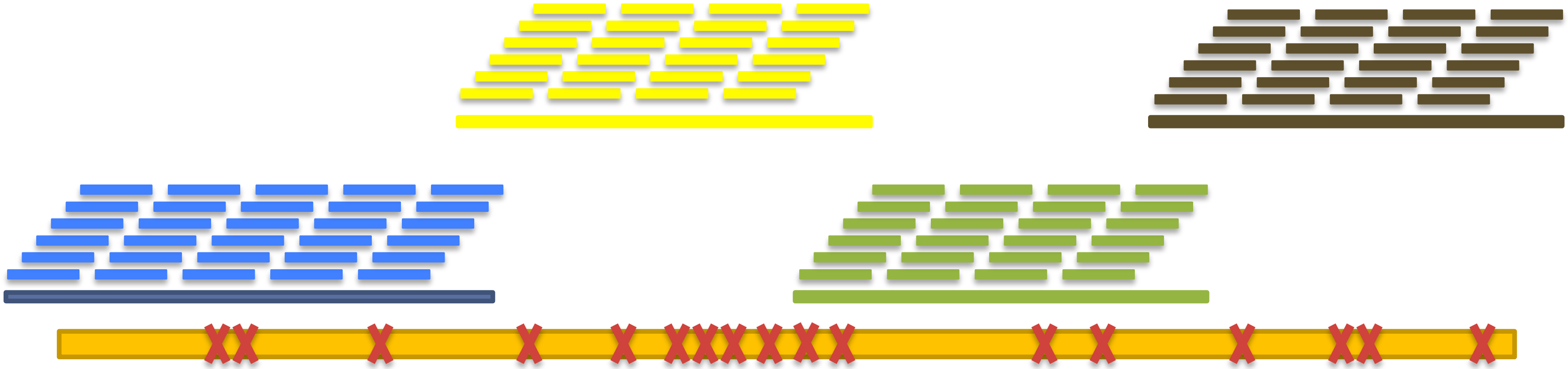
PacBioToCA fails in complex regions

1. Error Dense Regions – Difficult to compute overlaps with many errors
2. Simple Repeats – Kmer Frequency Too High to Seed Overlaps
3. Extreme GC – Lacks Illumina Coverage



ECTools: Error Correction with pre-assembled reads

https://github.com/jgurtowski/ectools



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

Can Help us overcome:

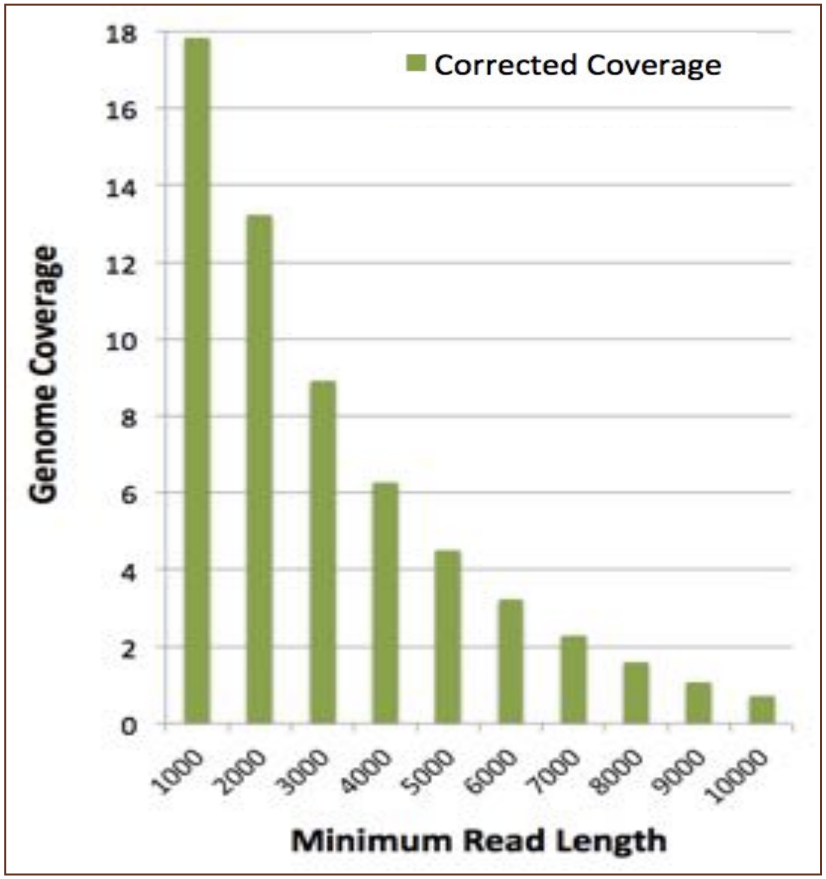
1. Error Dense Regions – Longer sequences have more seeds to match
2. Simple Repeats – Longer sequences easier to resolve

However, cannot overcome Illumina coverage gaps & other biases

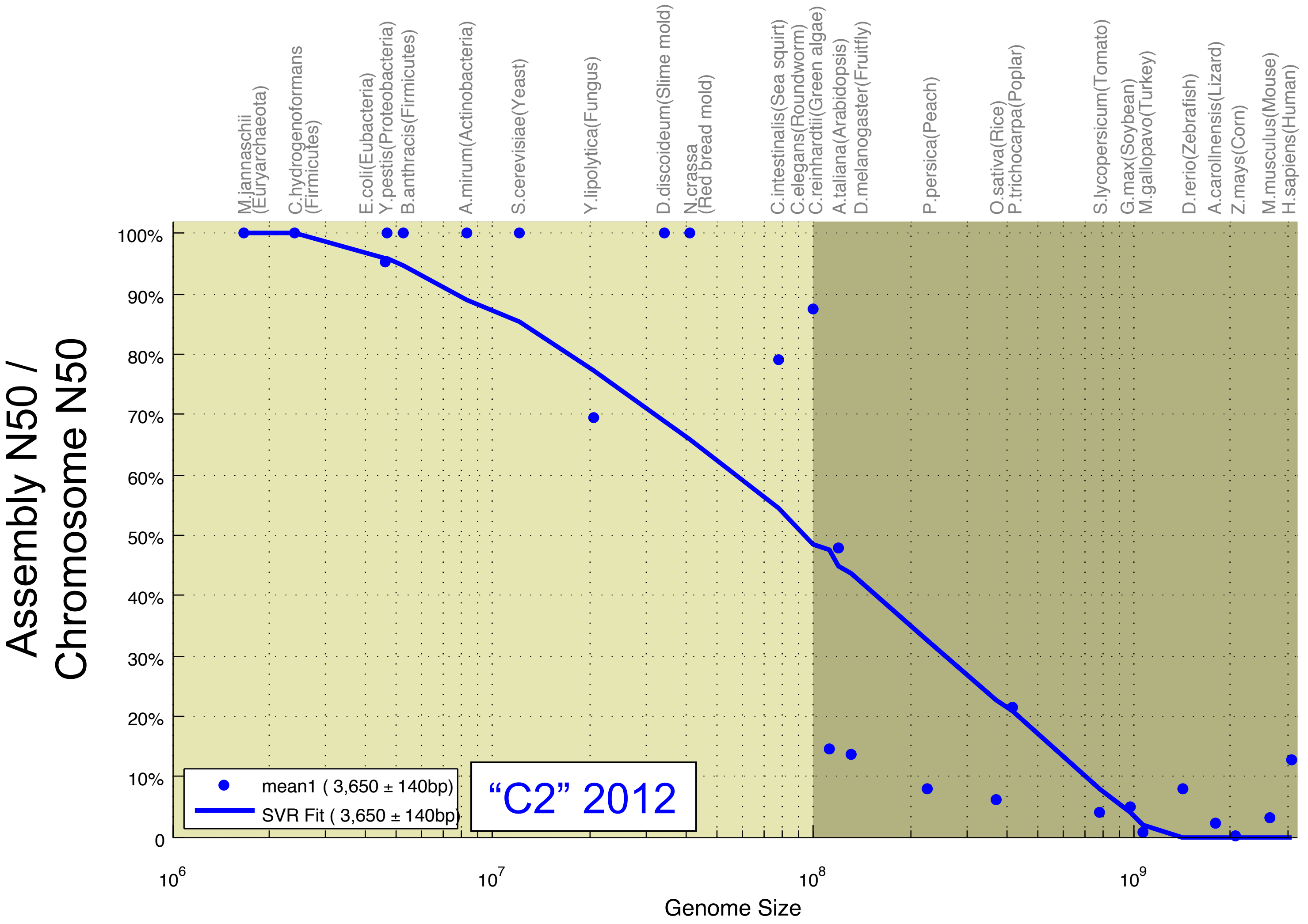
O. sativa pv Nipponbare

Genome size: 370 Mb
Chromosome N50: 29.7 Mbp
19x PacBio C2XL sequencing at CSHL from Summer 2012

Assembly	Contig NG50
MiSeq Fragments 23x 459bp 8x 2x251bp @ 450	6,332
"ALLPATHS-recipe" 50x 2x100bp @ 180 36x 2x50bp @ 2100 51x 2x50bp @ 4800	18,248
PacBioToCA 19x @ 3500 ** MiSeq for correction	50,995
ECTools 19x @ 3500 ** MiSeq for correction	155,695



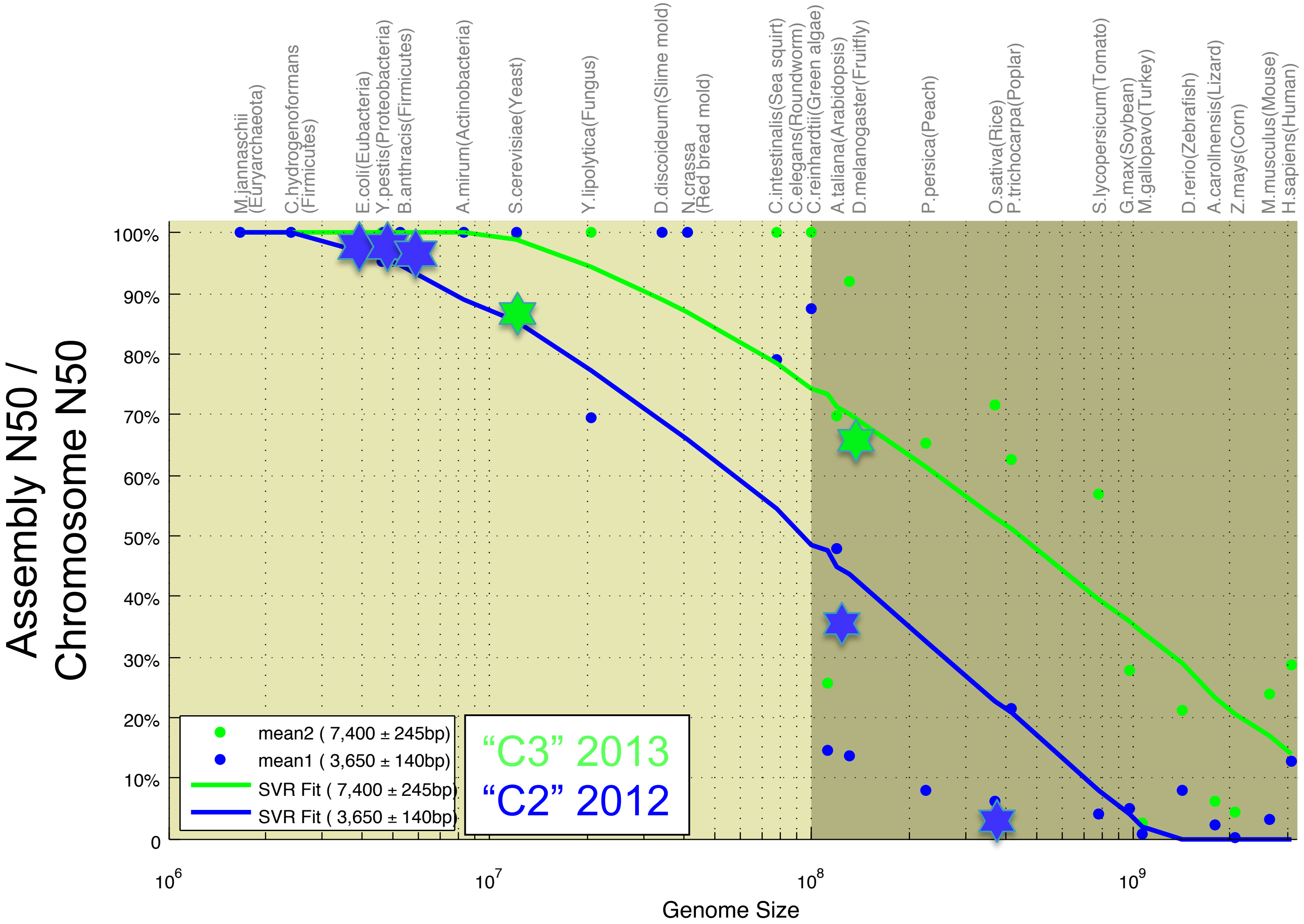
Assembly Complexity of Long Reads



Assembly complexity of long read sequencing

Lee, H*, Gurtowski, J*, Yoo, S, Marcus, S, McCombie, WR, Schatz MC *et al.* (2014) *In preparation*

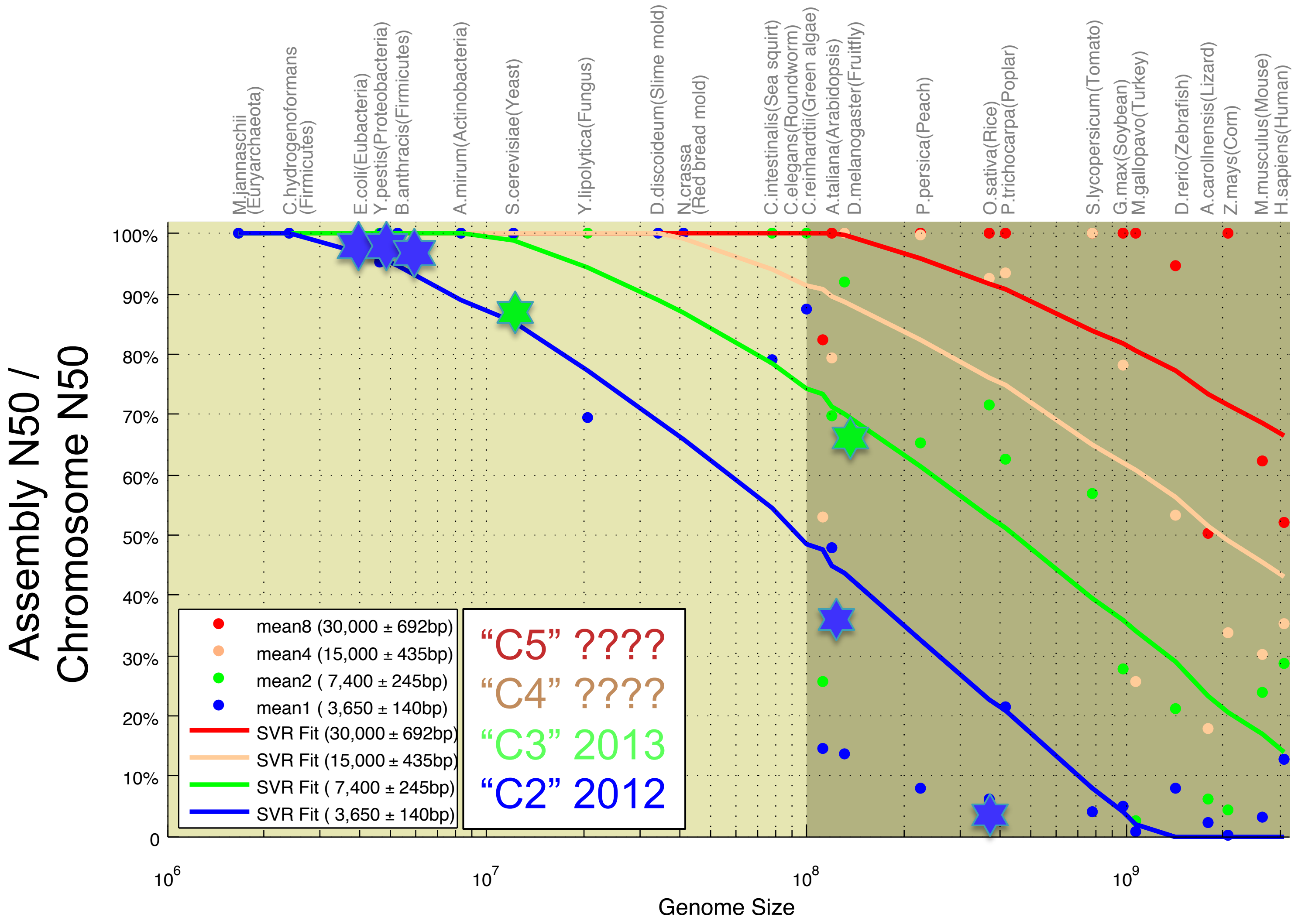
Assembly Complexity of Long Reads



Assembly complexity of long read sequencing

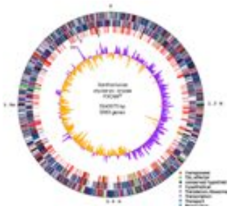
Lee, H*, Gurtowski, J*, Yoo, S, Marcus, S, McCombie, WR, Schatz MC *et al.* (2014) *In preparation*

Assembly Complexity of Long Reads



Assembly complexity of long read sequencing

Lee, H*, Gurtowski, J*, Yoo, S, Marcus, S, McCombie, WR, Schatz MC *et al.* (2014) *In preparation*



Summary

- **Long read sequencing of eukaryotic genomes is here**

- **Recommendations**

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

- **Caveats**
  - Model only as good as the available references (esp. haploid sequences)
  - Technologies are quickly improving, exciting new scaffolding technologies

Acknowledgements

Schatz Lab

James Gurtowski

Hayan Lee

Shoshana Marcus

Alejandro Wences

Giuseppe Narzisi

Srividya Ramakrishnan

Rob Aboukhalil

Mitch Bekritsky

Charles Underwood

Tyler Gavin

Greg Vurture

Eric Biggers

Aspyn Palatnick

CSHL

McCombie Lab

Hannon Lab

Gingeras Lab

Jackson Lab

Iossifov Lab

Levy Lab

Lippman Lab

Lyon Lab

Martienssen Lab

Tuveson Lab

Ware Lab

Wigler Lab

NBACC

Serge Koren

Adam Phillippy

PACIFIC BIOSCIENCES®

National Human Genome Research Institute

NSF

Big Data in Biology

March 23–25, 2014

Fairmont San Francisco
San Francisco, California, USA

Scientific Organizers: Lincoln D. Stein, Doreen Ware and Michael Schatz

KEYSTONE SYMPOSIA™ on Molecular and Cellular Biology

Accelerating Life Science Discovery

Thank You!

http://schatzlab.cshl.edu
@mike_schatz / #PAGXXII

Variant Calling and RNA-seq
@ 4:25 in the KBase Workshop