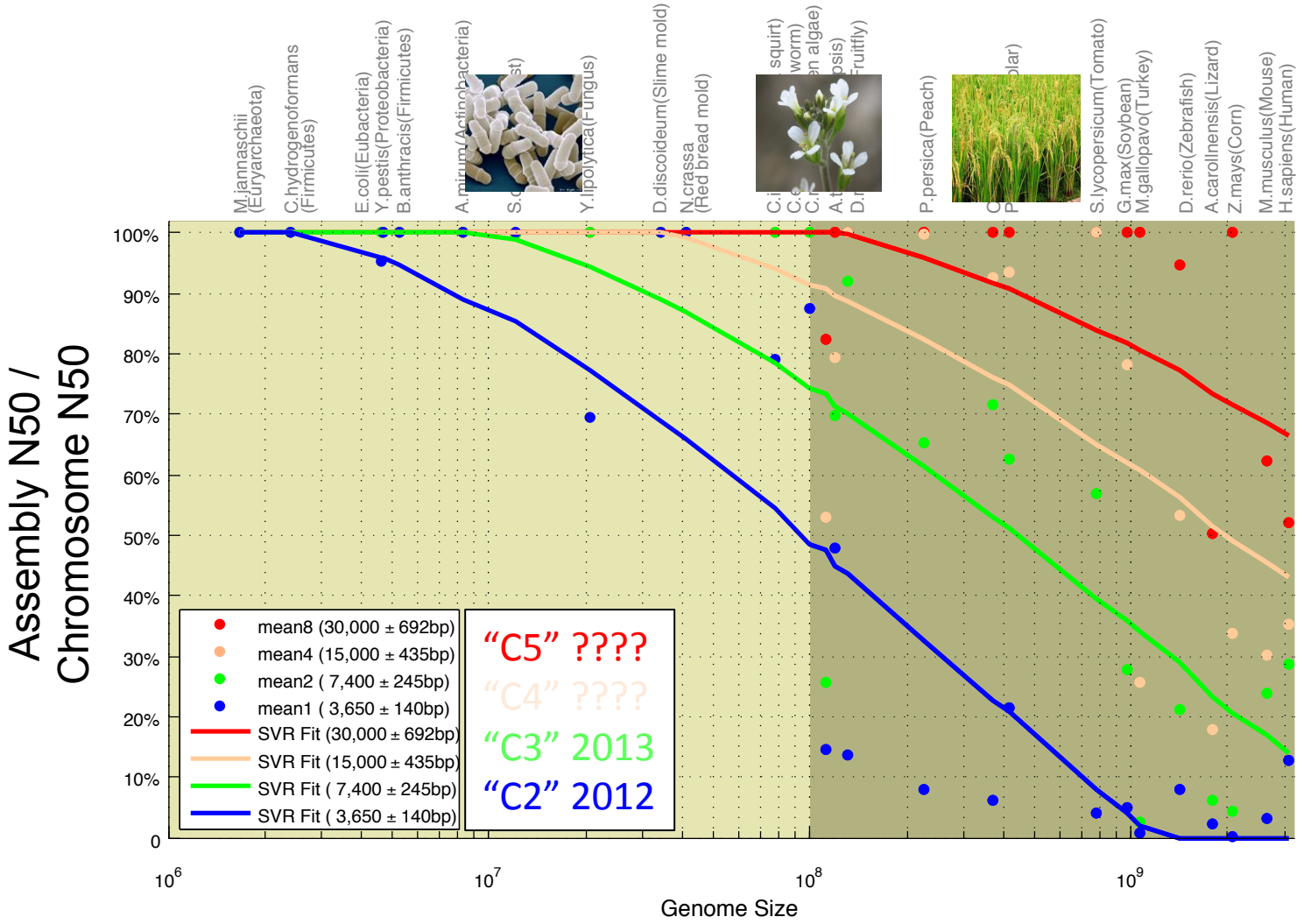
Near perfect de novo assemblies of eukaryotic genomes using PacBio long read sequencing

James Gurtowski

Schatz Lab

5/29/2014

Assembly Complexity of Long Reads



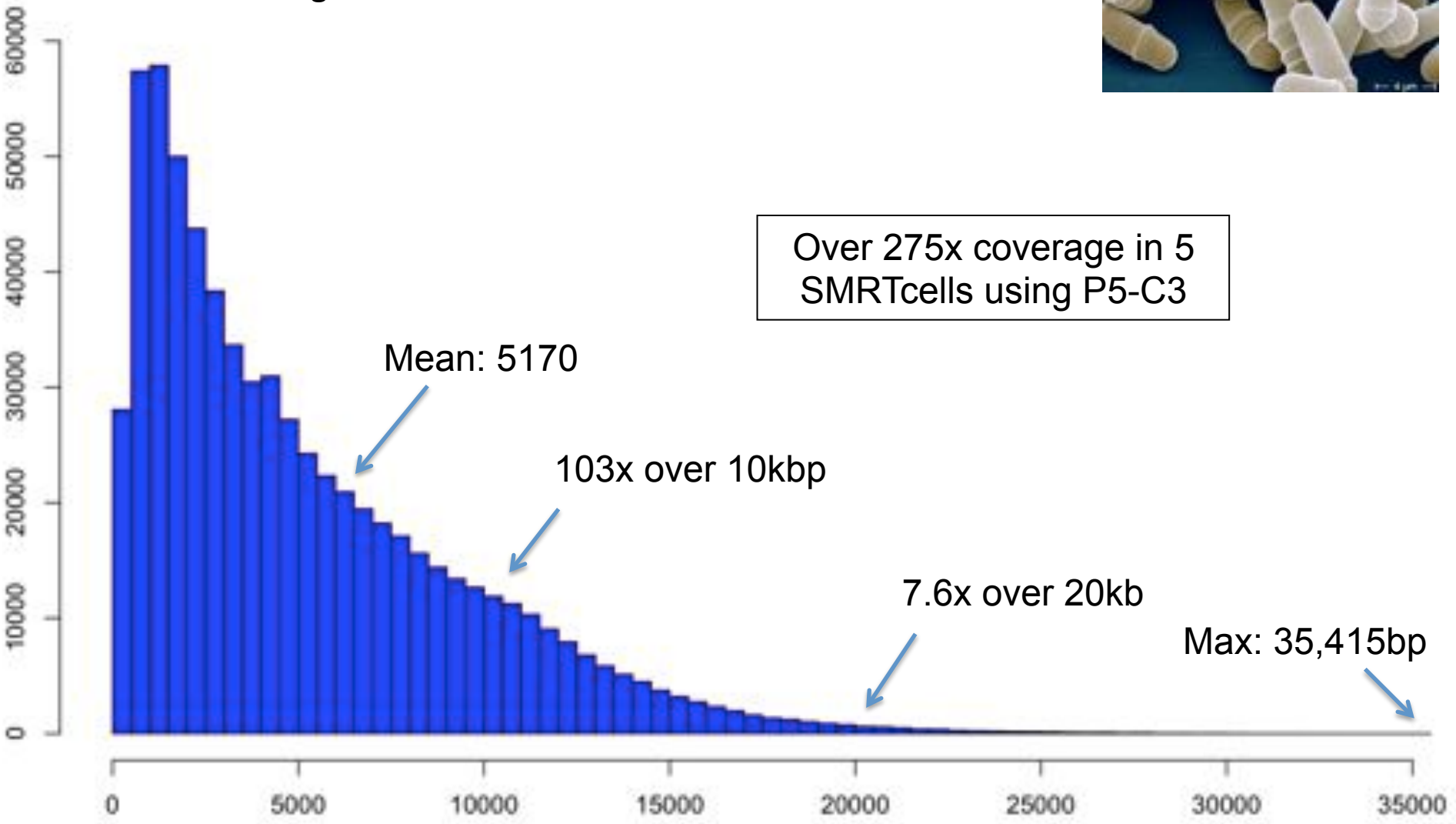
Assembly complexity of long read sequencing

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

S. pombe dg21

PacBio RS II *sequencing* at CSHL

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



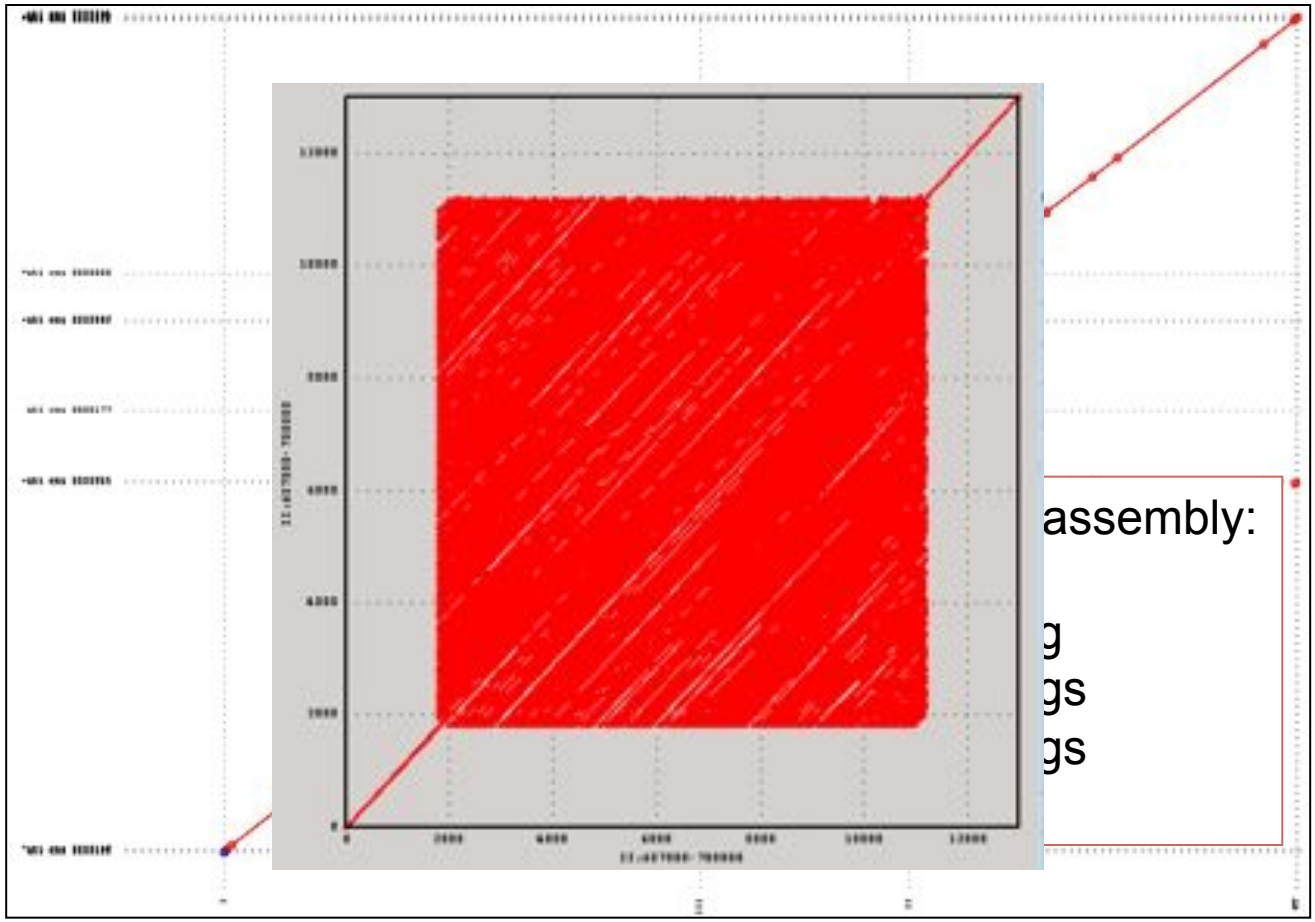
S. pombe dg21

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



Spanning vs Standard Coverage

Standard Coverage
(SpanLength = 1bp)

$$\frac{\sum_{reads} Length(read)}{GenomeSize}$$

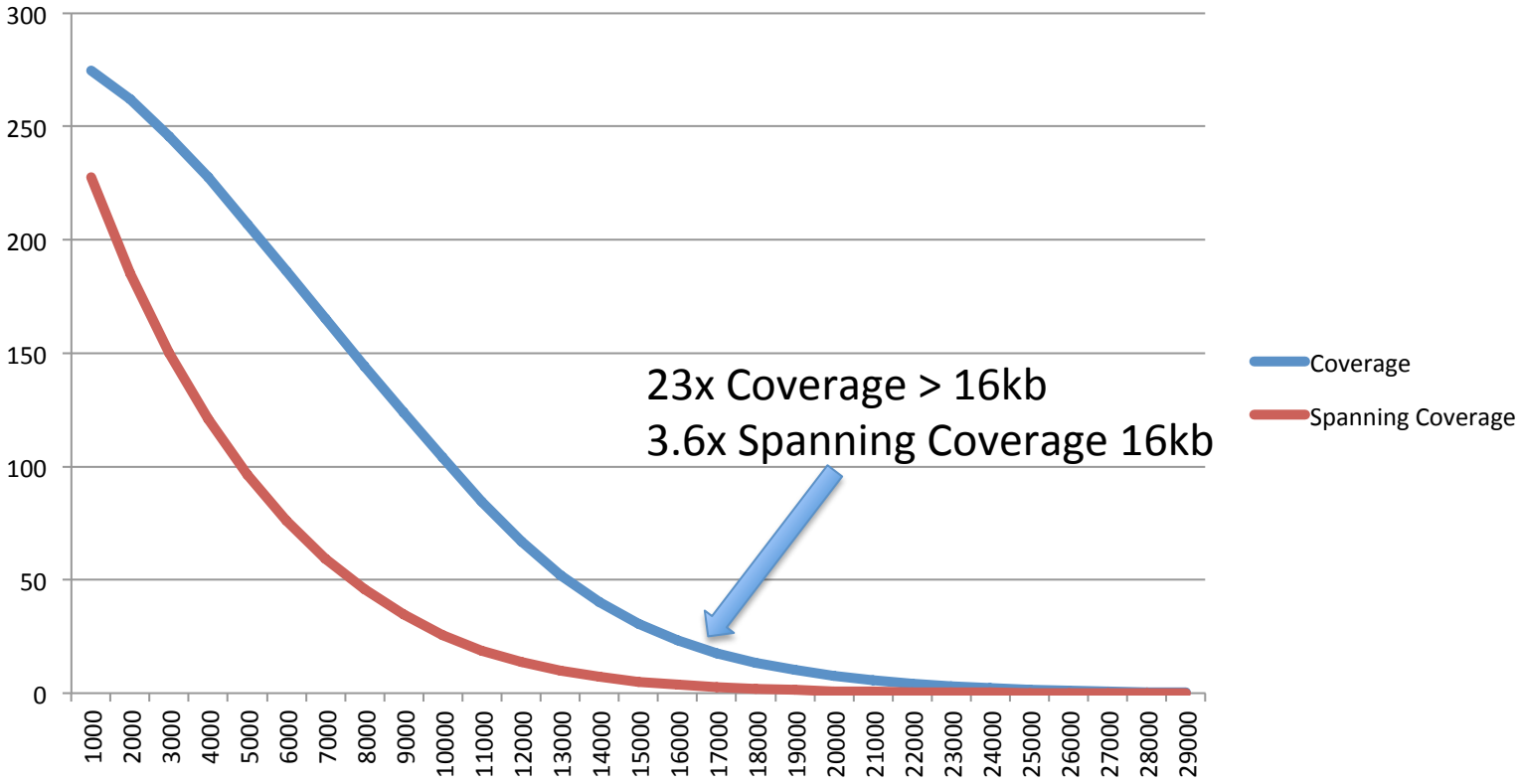
Spanning Coverage
(SpanLength > 1bp)

$$\frac{\sum_{reads} \max(0, Length(read) - SpanLength)}{GenomeSize}$$

Spanning Coverage (S. pombe)

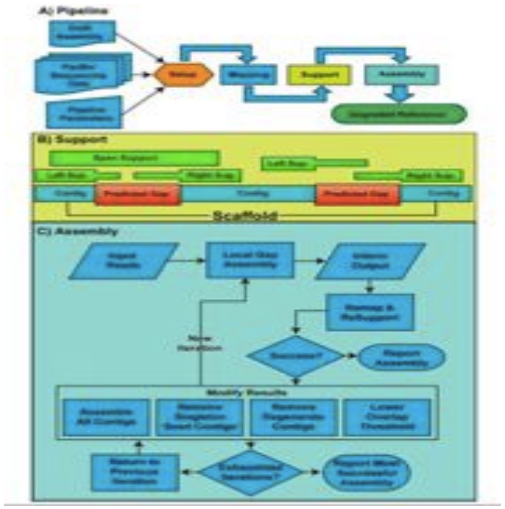
- **How many reads span a particular 16kb region?**

  23x Coverage of reads > 16kb, but only expect **3.6** reads to span a particular 16kb region



PacBio Correction/Assembly Algorithms

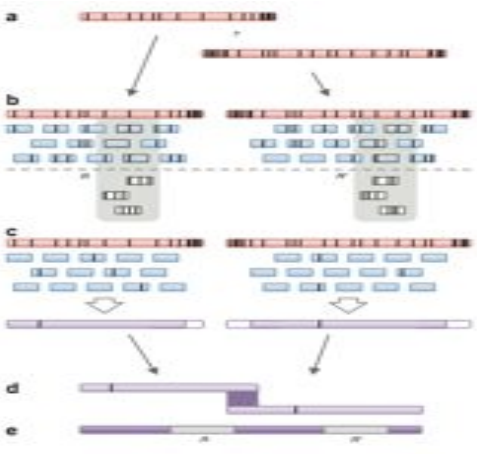
PBJelly



Gap Filling and Assembly Upgrade

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

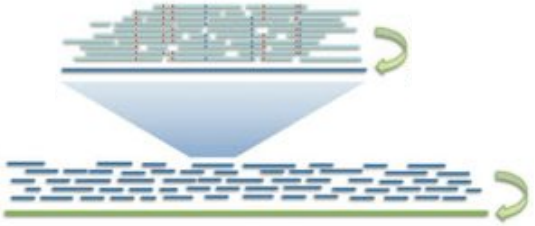
PacBioToCA & ECTools



Hybrid/PB-only Error Correction

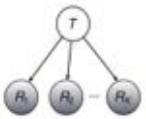
Koren, Schatz, *et al* (2012)
Nature Biotechnology. 30:693–700

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

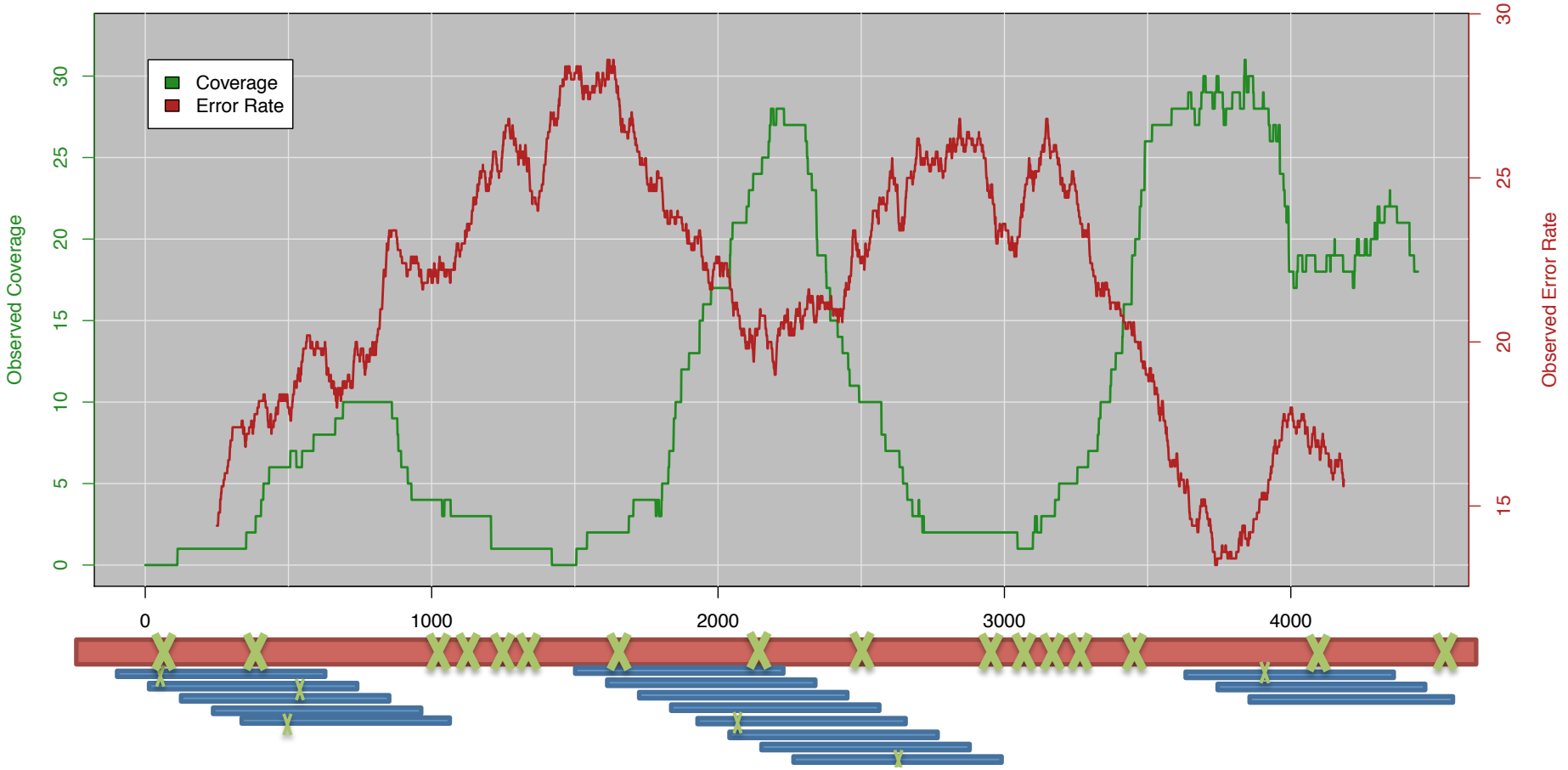
Chin *et al* (2013)
Nature Methods. 10:563–569

< 5x PacBio Coverage > 50x

Hybrid Approaches for Larger Genomes

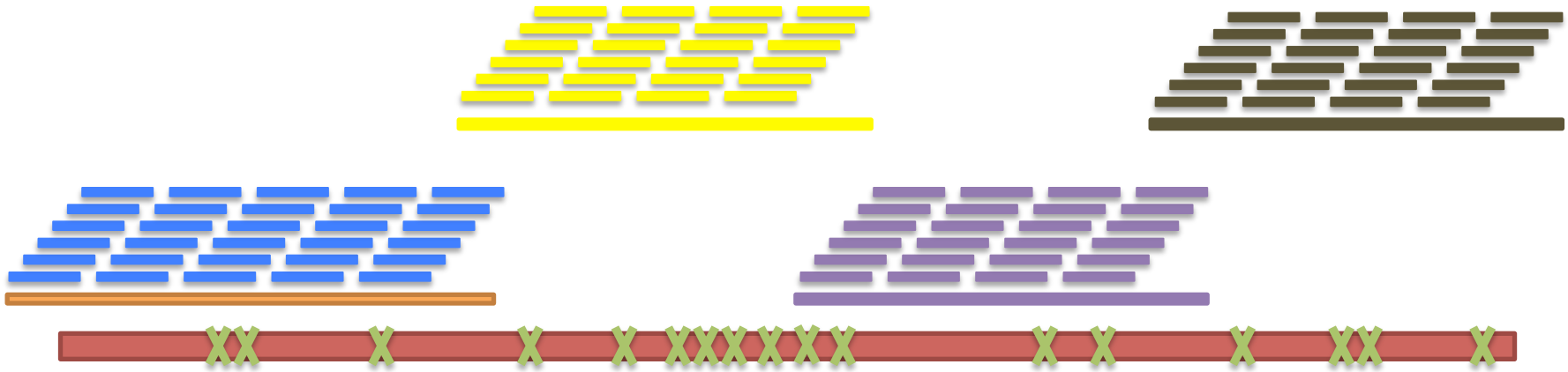
PacBioToCA fails in complex regions

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



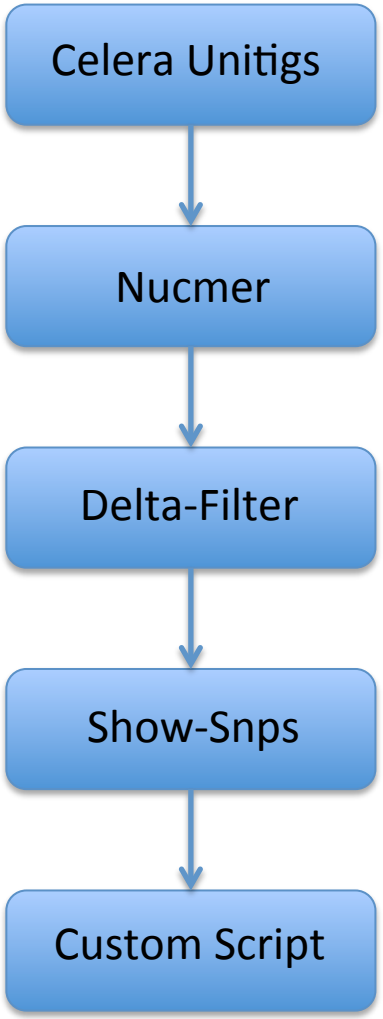
Short Reads -> Assemble Unitigs -> Align & Select - > Error Correct

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

ECTools Pipeline



Generate Unitigs with Celera Using Illumina or another high identity sequencing technology

Align Unitigs to Pacbio Reads With Nucmer

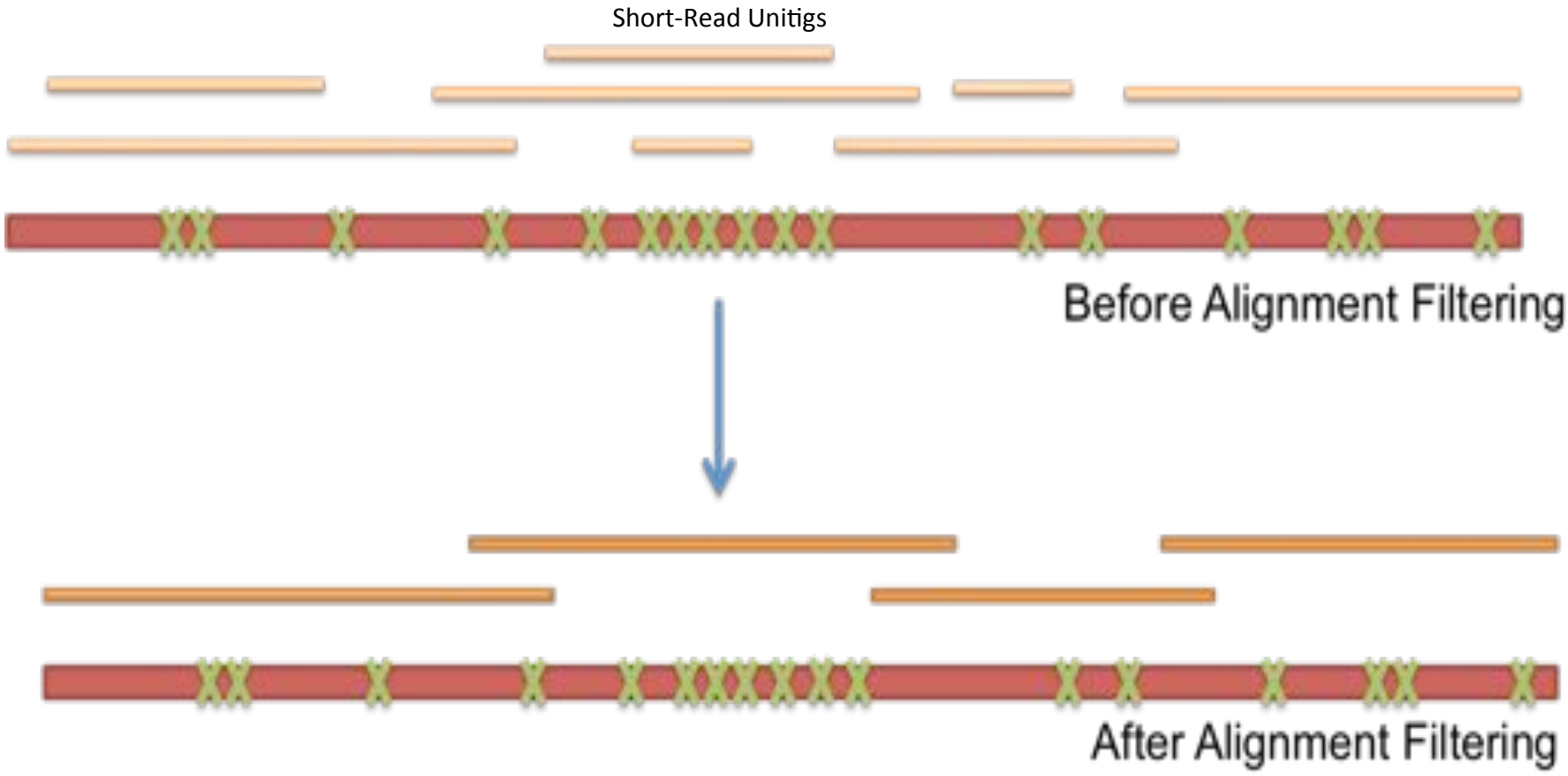
Use Delta-Filter to Generate Unitig Layout

Show-Snps shows differences between trusted Illumina Unitig Sequence and Pacbio Read

Script to “Correct” Pacbio Read

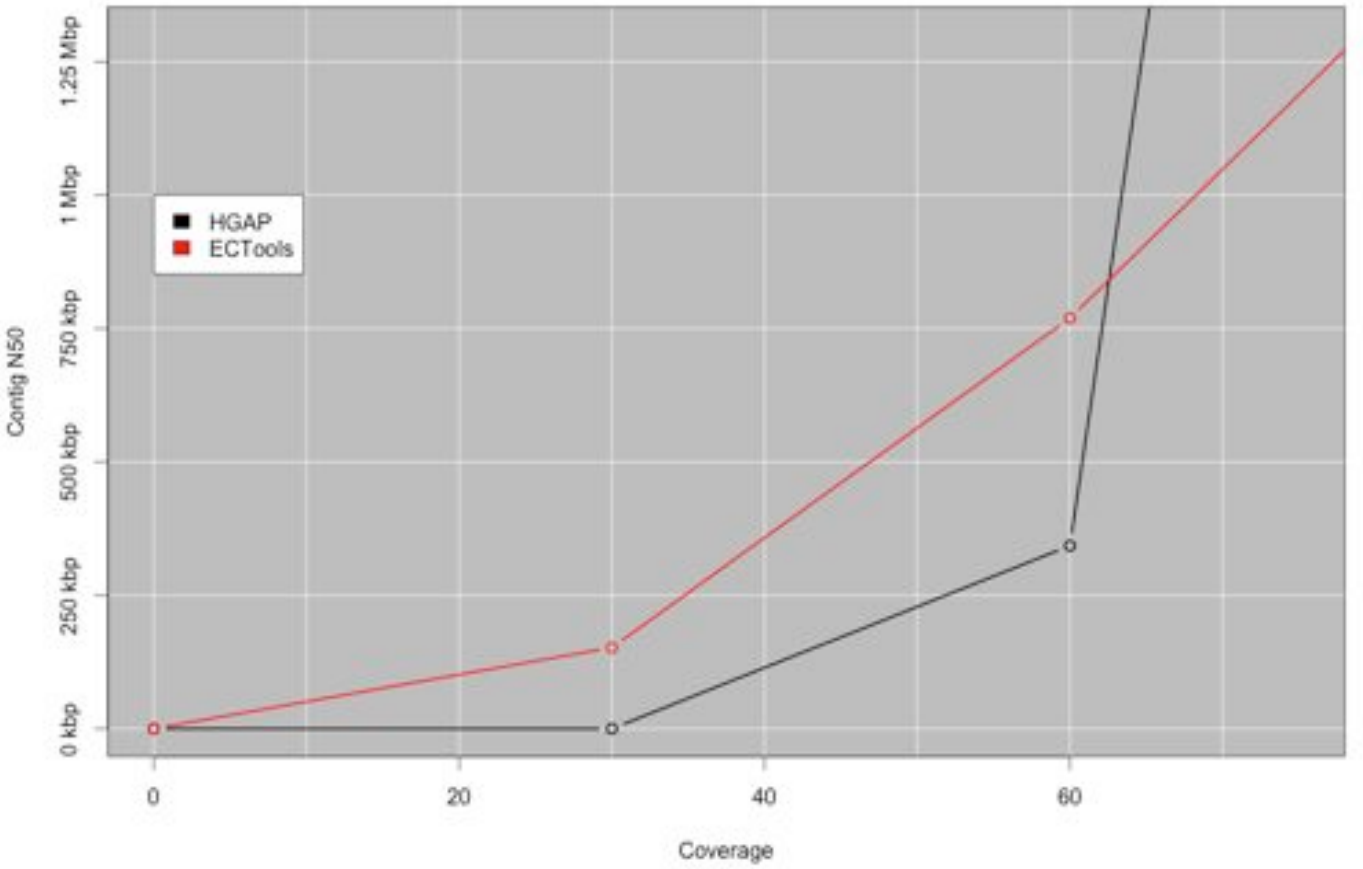
Note: Reads are never split or trimmed

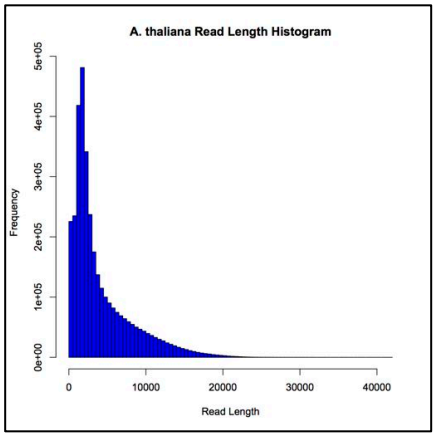
Delta-Filter Alignment filtering



A. thaliana Ler-0

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





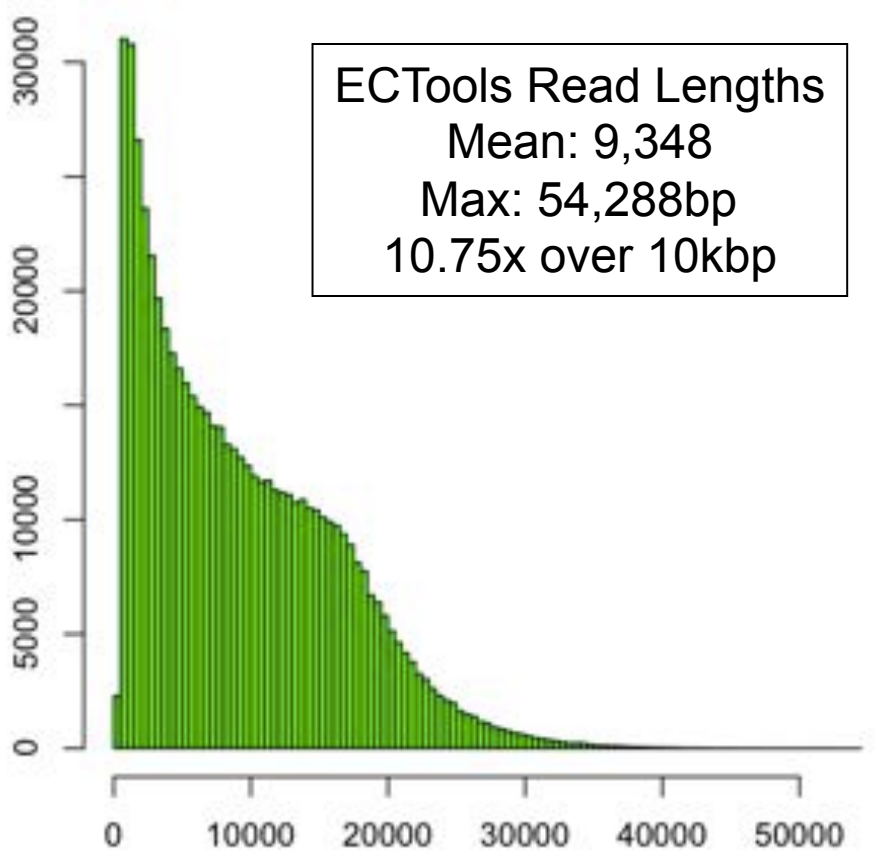
Mean: 4,137bp
Max: 41,753bp
Cov: 118x

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

O. sativa pv Indica (IR64)

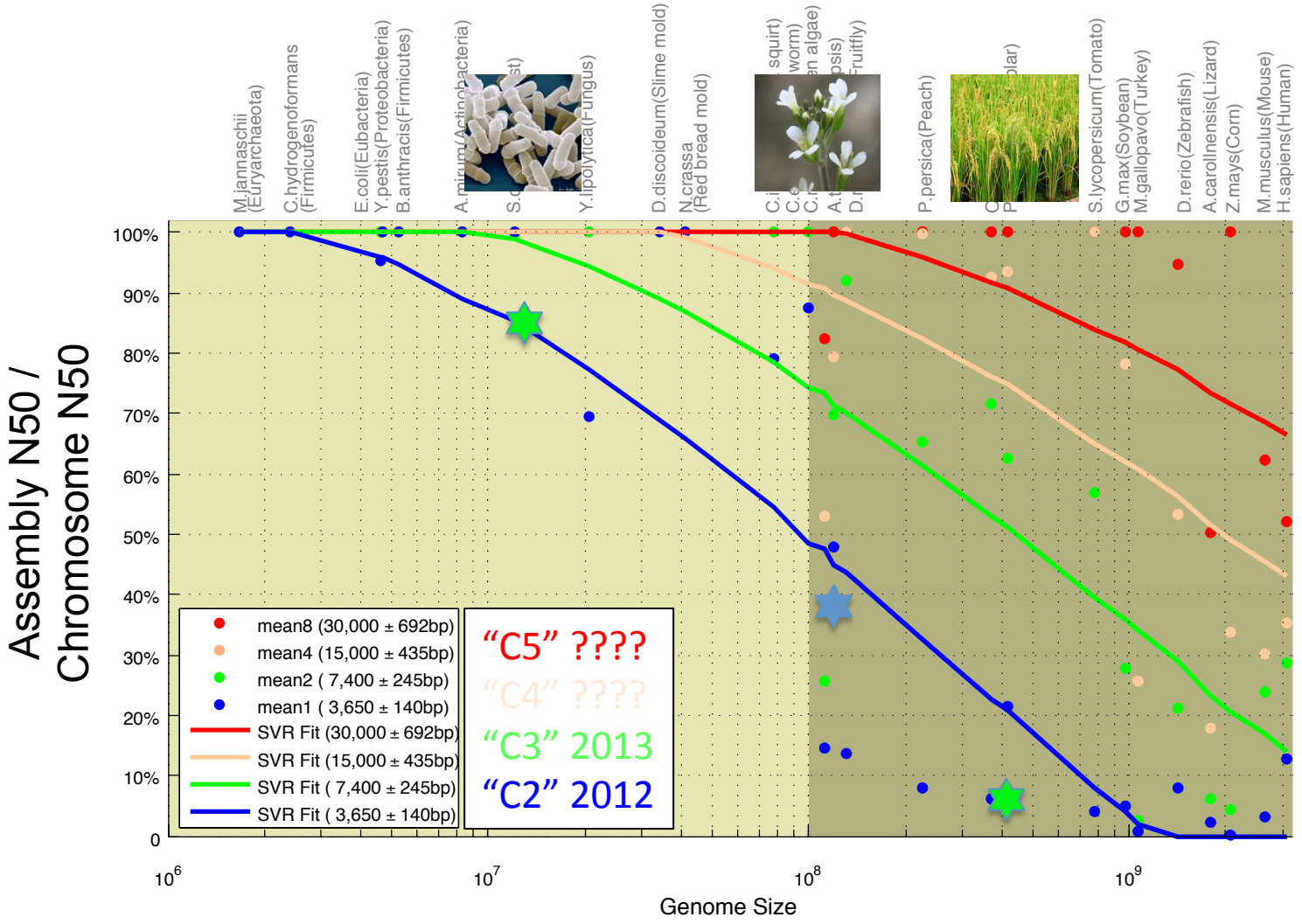
Genome size: ~370 Mb
Chromosome N50: ~29.7 Mbp

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	272,137
NEW HGAP – 114 SMRTCells 29.2x @ 10kbp	600,021



ECTools Read Lengths
Mean: 9,348
Max: 54,288bp
10.75x over 10kbp

Real Data Results



Assembly complexity of long read sequencing

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

Acknowledgements

Cold Spring Harbor Laboratory

Schatz Lab
Mike Schatz
Hayan Lee

McCombie Lab
Dick McCombie
Panchajanya Deshpande
Senem Eskipehlivan
Melissa Kramer
Sara Goodwin
Eric Antoniou

Pacbio
Cheryl Heiner
Greg Khitrov

Email:

gurtowsk@cshl.edu

ECTools:

https://github.com/jgurtowski/ectools