A near perfect de novo assembly of a eukaryotic genome using sequence reads of greater than 10 kilobases generated by the Pacific Biosciences RS II

W. Richard McCombie

Disclosures

Orion Genomics – Founder and Shareholder
Cancer epigenetics and plant genomics

Previously Compensated Speaker for Illumina, Inc.

Previously Compensated Speaker for Pacific Biosciences, Inc.

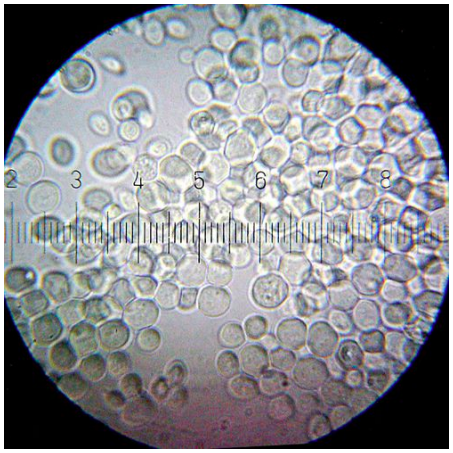
Introduction to the challenge

- Short read NGS has revolutionized resequencing
- *De novo* assembly is possible but not optimal with short reads
- Long reads improve the ability do *de novo* assembly dramatically
- Even in organisms with a good reference, such as humans, resequencing misses some structural differences relative to the reference
- Plant genomes are very large in general
- There are significant structural differences between different strains of the same plant such as rice
- These structural differences contribute to salient biological differences

Potential uses for rice genomes

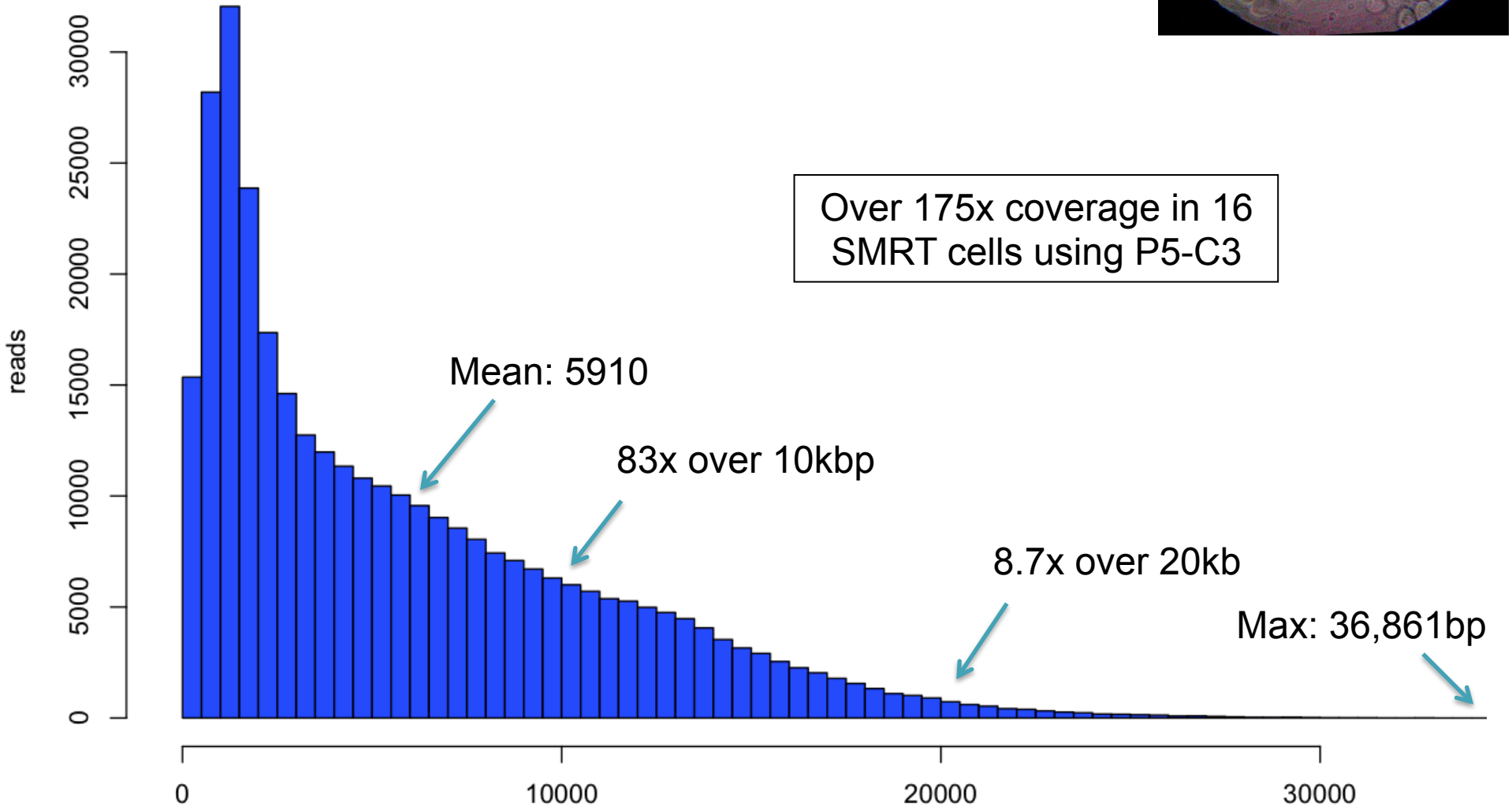
- Understand basis of differences among subpopulations and varieties (duplications, CNVs, etc.)that lead to important phenotypic differences - this requires de novo assemblies - not simple resequencing
- Many of these differences relate to ability to grow in less than optimum conditions
- Low phosphorus
- Submergence
- Drought
- Disease exposure

A test genome – yeast: S. cerevisiae W303

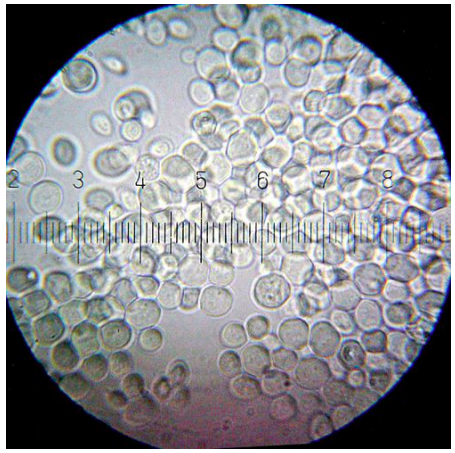


PacBio RS II sequencing at CSHL

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



S. cerevisiae W303

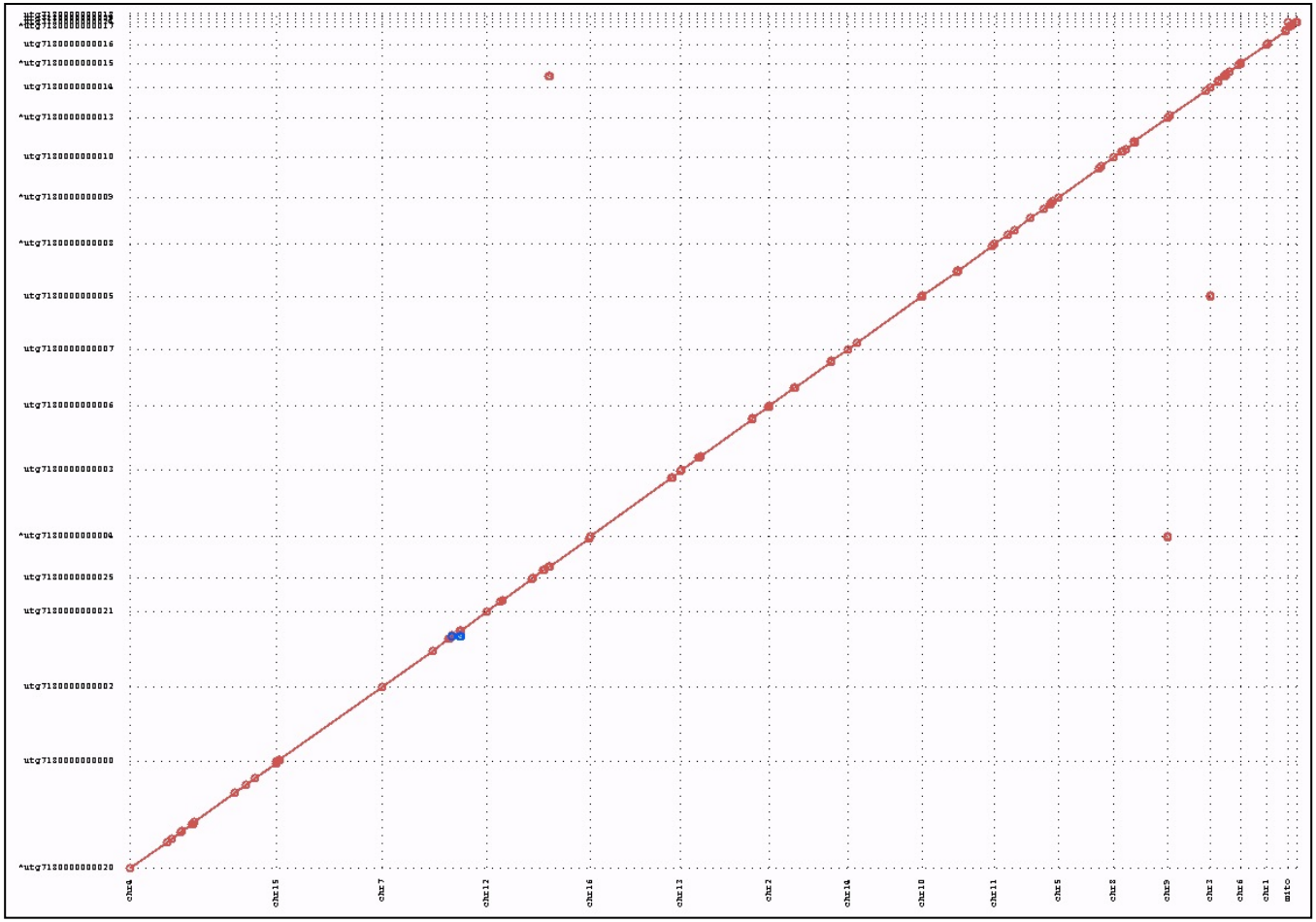


S288C Reference sequence

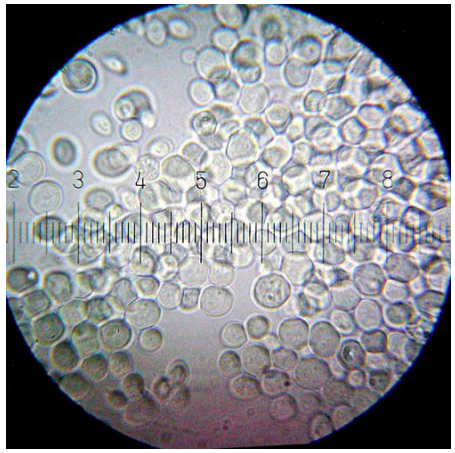
- 12.1Mbp; 16 chromo + mitochondria; N50: 924kbp

PacBio assembly using HGAP + Celera Assembler

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



S. cerevisiae W303

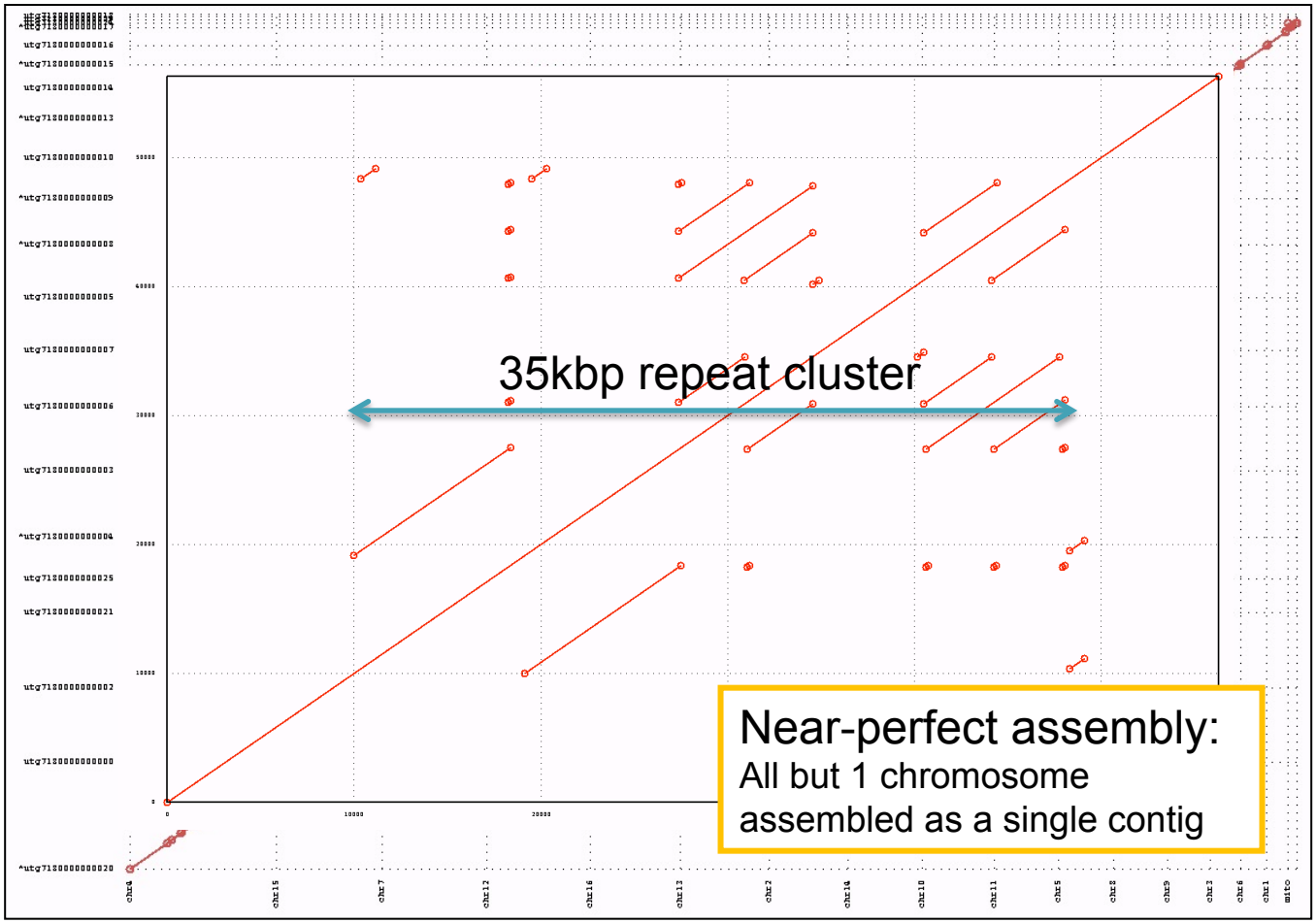


S288C Reference sequence

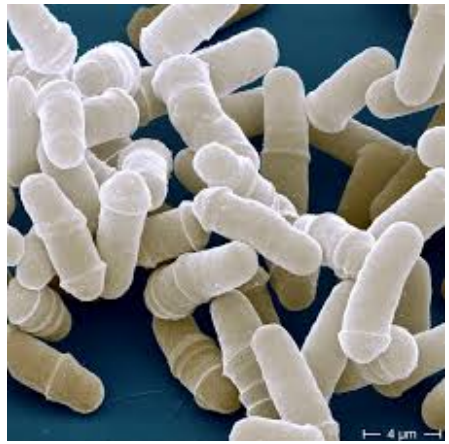
- 12.1Mbp; 16 chromo + mitochondria; N50: 924kbp

PacBio assembly using HGAP + Celera Assembler

- 12.4Mbp; 21 non-redundant contigs; N50: 811kbp; >99.9% id

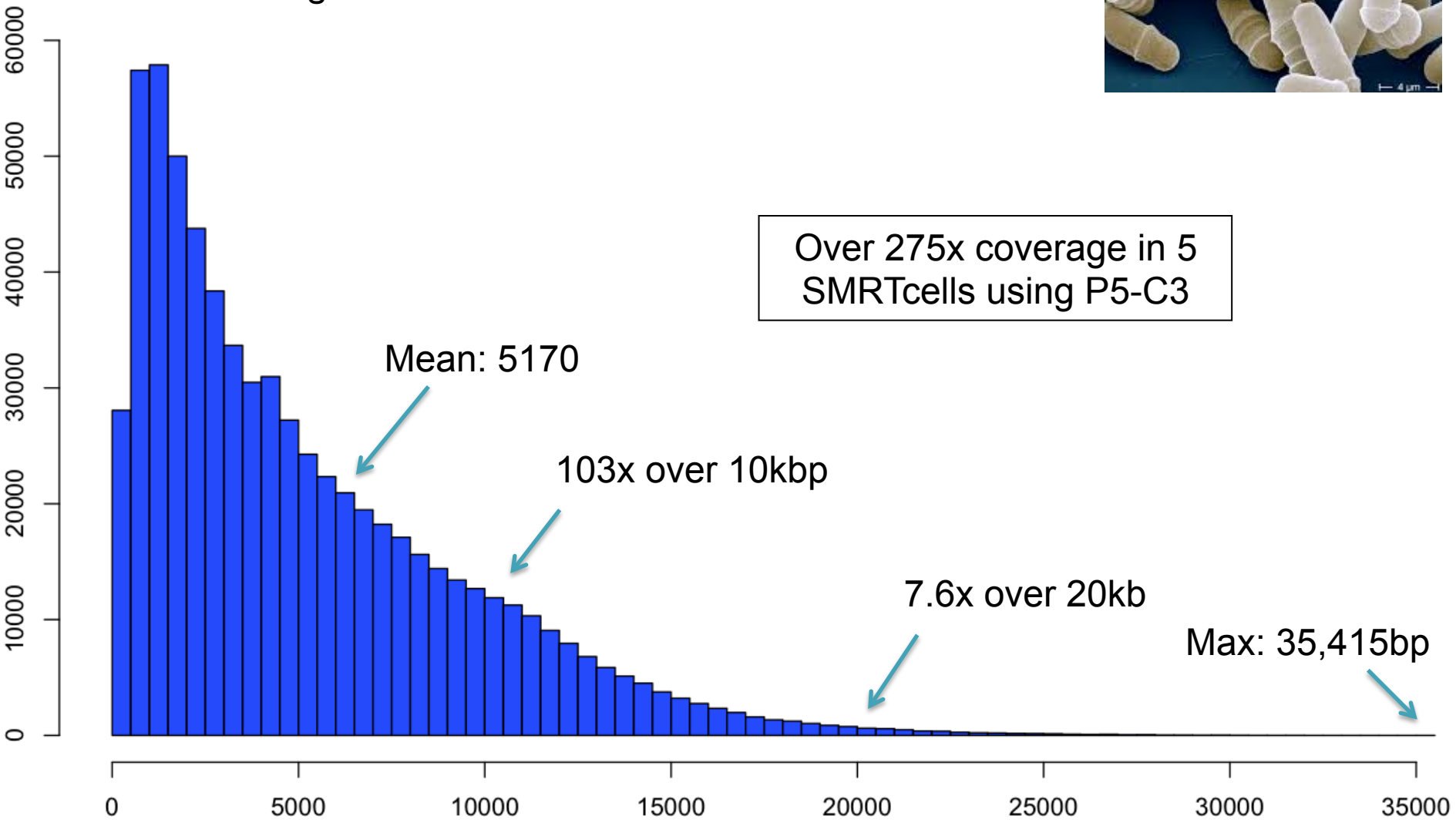


S. pombe dg21

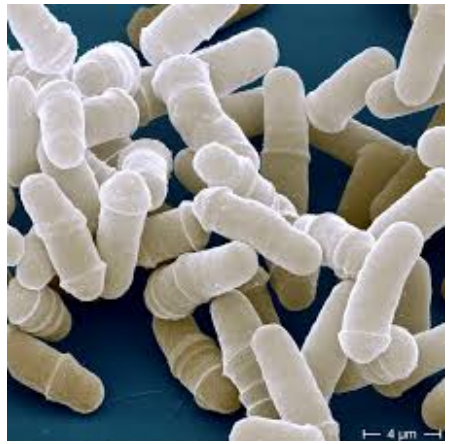


PacBio RS II sequencing at CSHL

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



S. pombe dg21

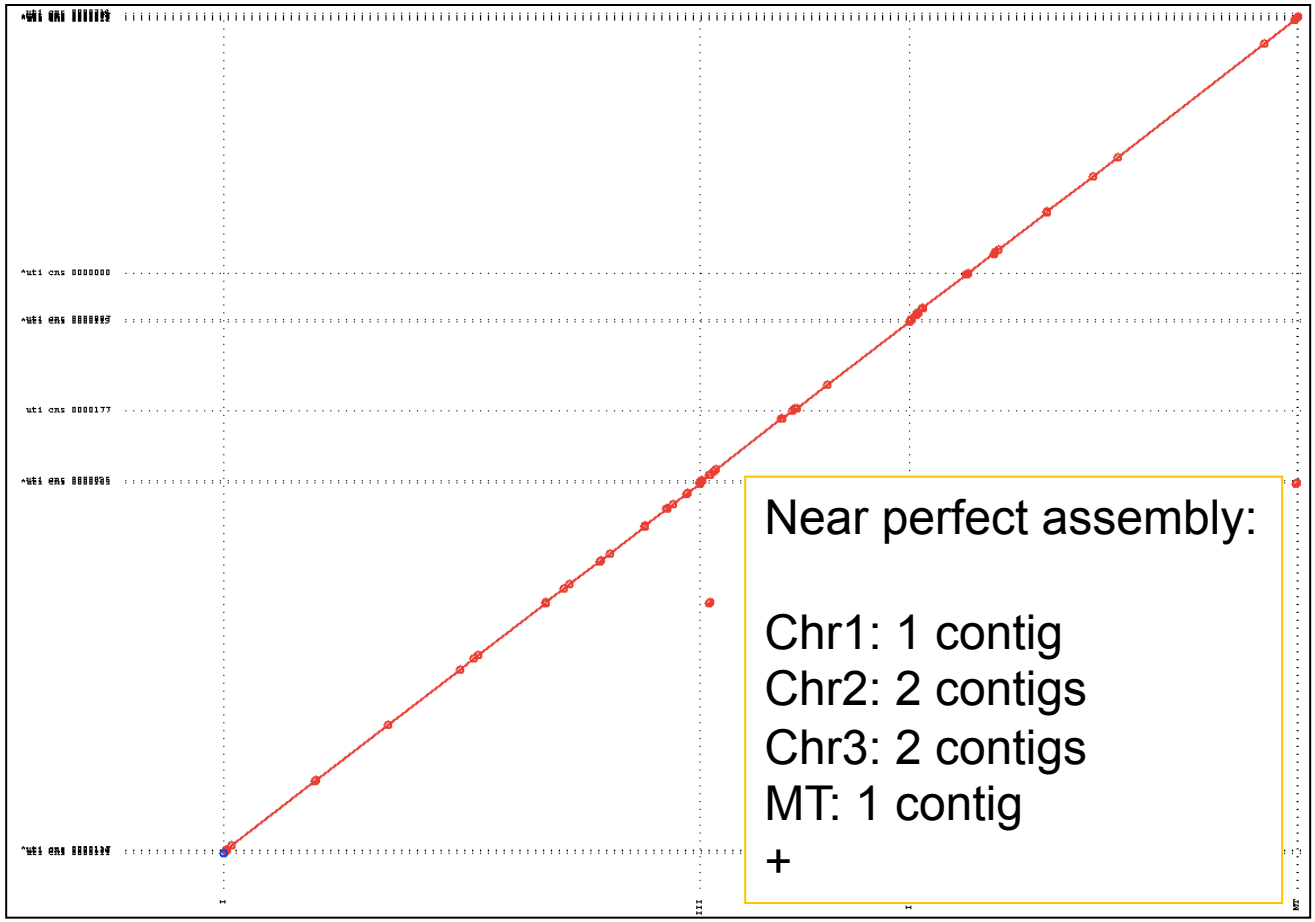


ASM294 Reference sequence

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

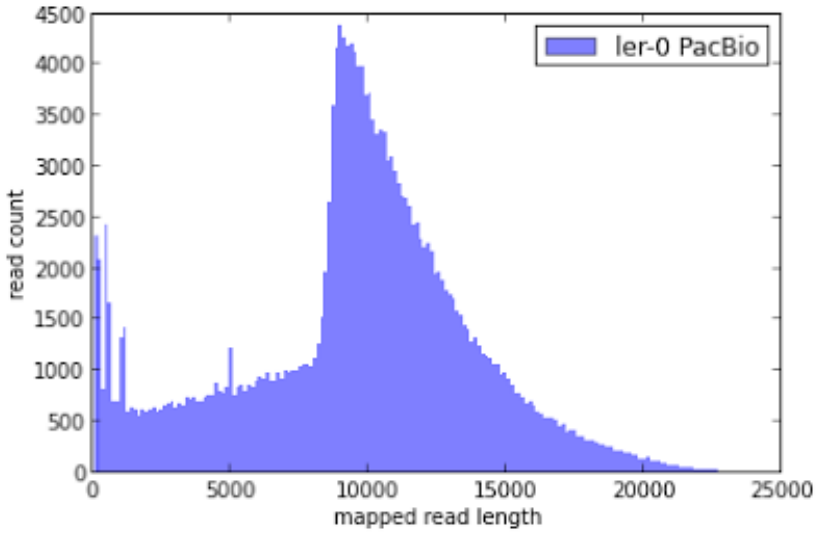
PacBio assembly using HGAP + Celera Assembler

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



A. thaliana Ler-0

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



A. thaliana Ler-0 sequenced at PacBio

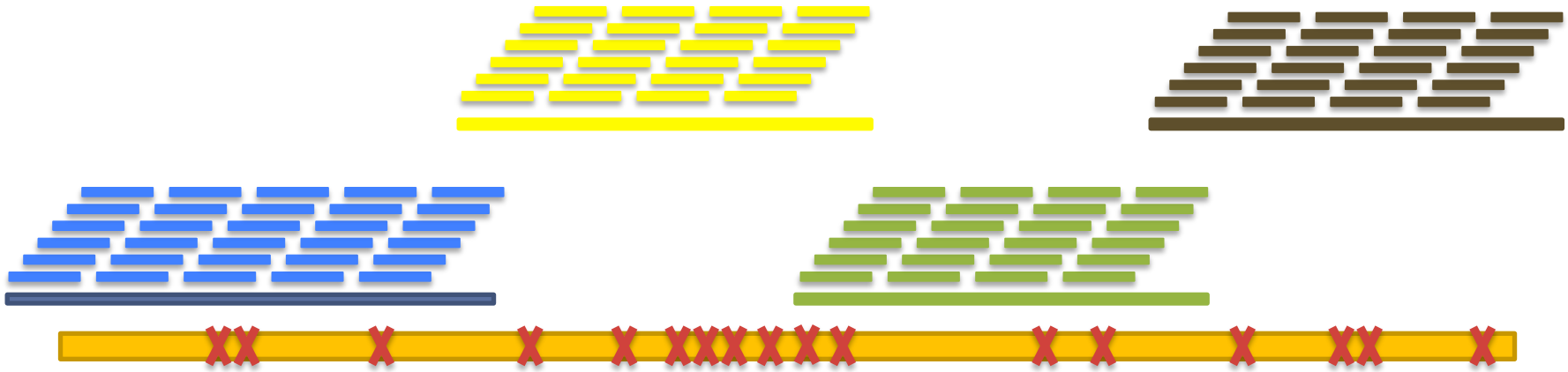
- Sequenced using the previous P4 enzyme and C2 chemistry
- Size selection using an 8 Kb to 50 Kb elution window on a BluePippin™ device from Sage Science
- Total coverage >119x

Genome size:	124.6 Mbp	Sum of Contig Lengths:	149.5Mb
Chromosome N50:	23.0 Mbp	N50 Contig Length:	8.4 Mb
Corrected coverage:	20x over 10kb	Number of Contigs:	1788

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

ECTools: Error Correction with pre-assembled reads

https://github.com/jgurtowski/ectools



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

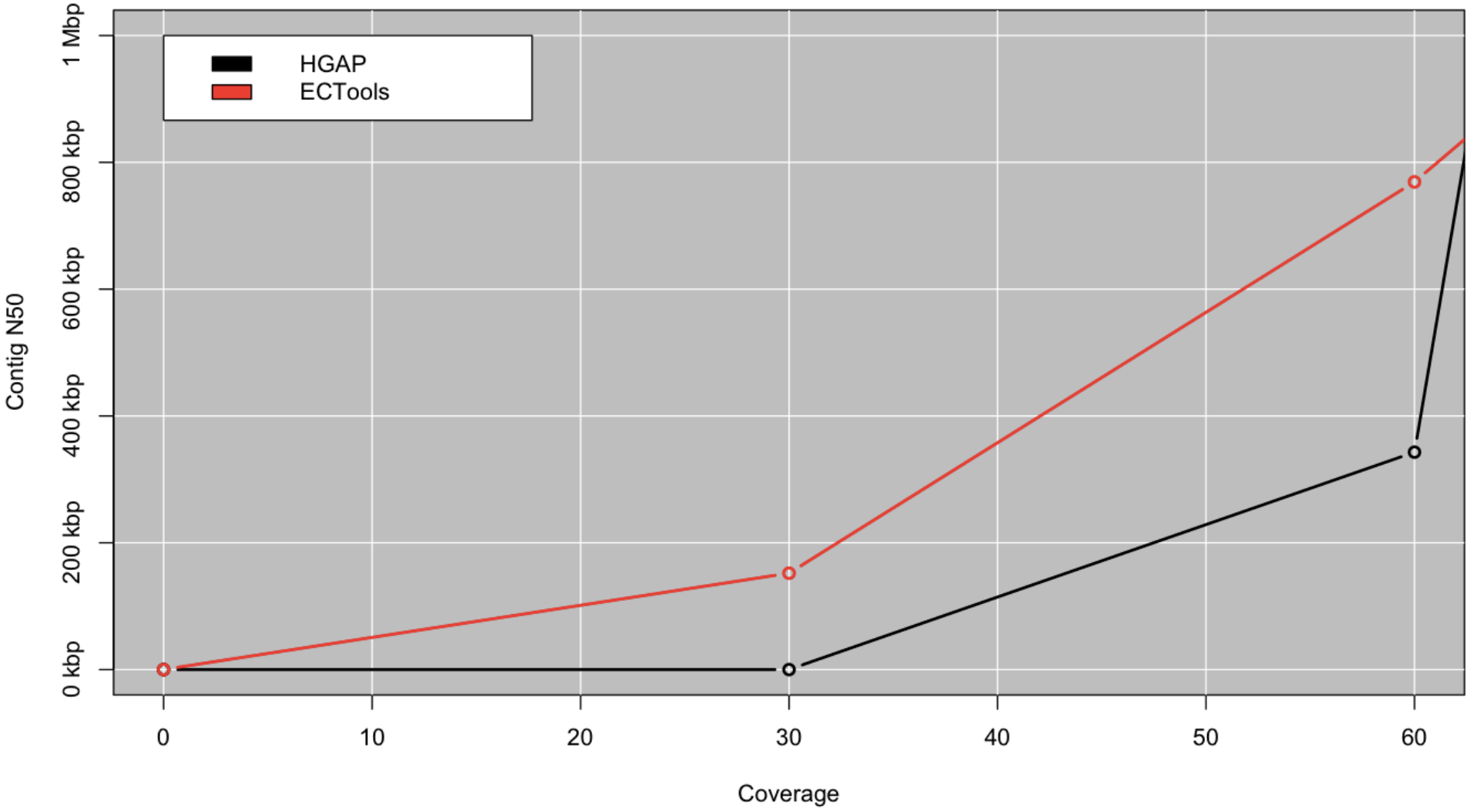
Can Help us overcome:

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

A. thaliana Ler-0

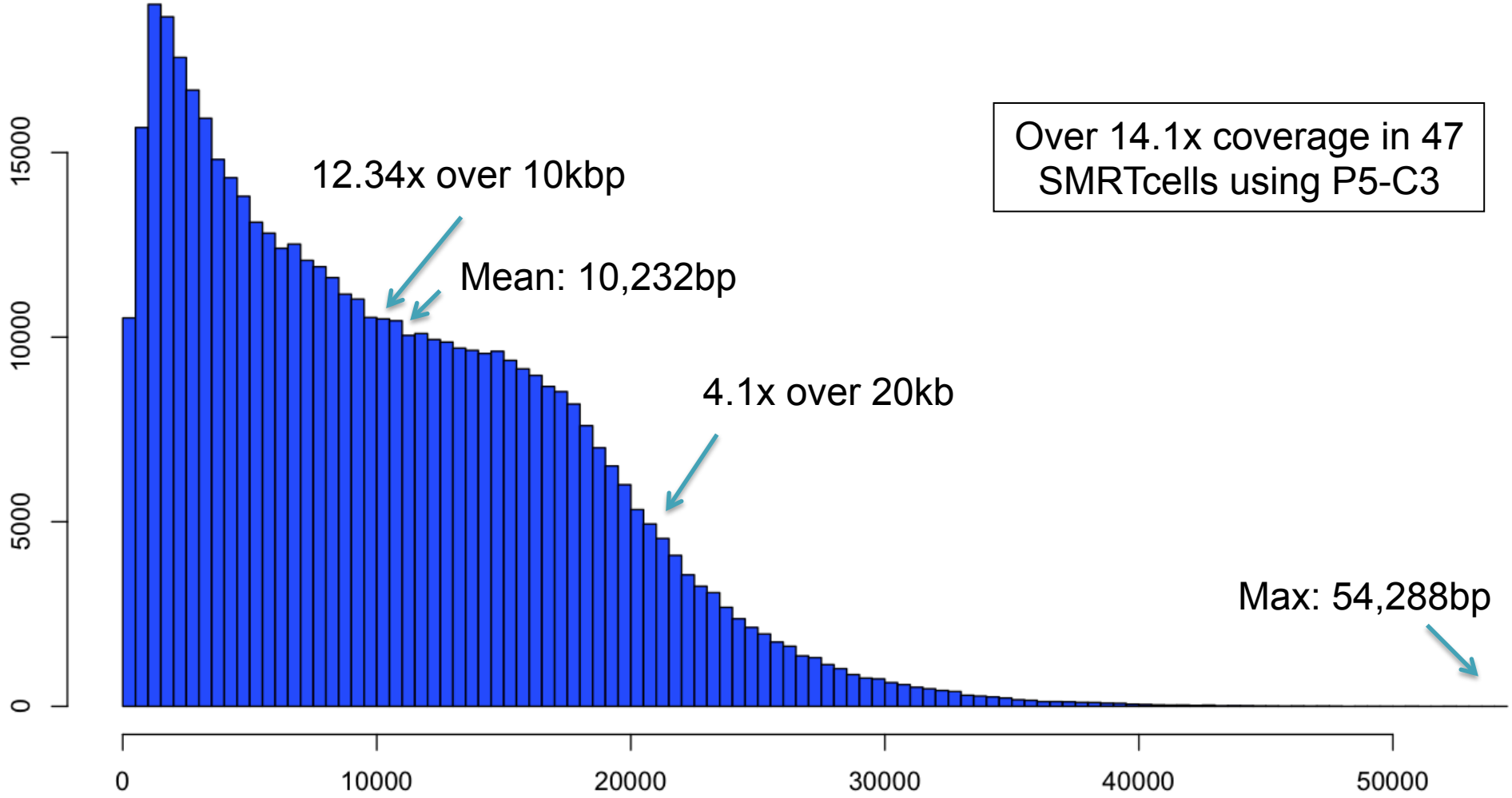
http://blog.pacificbiosciences.com/2013/08/new-data-release-arabidopsis-assembly.html



O. sativa pv Indica (IR64)

PacBio RS II sequencing at PacBio

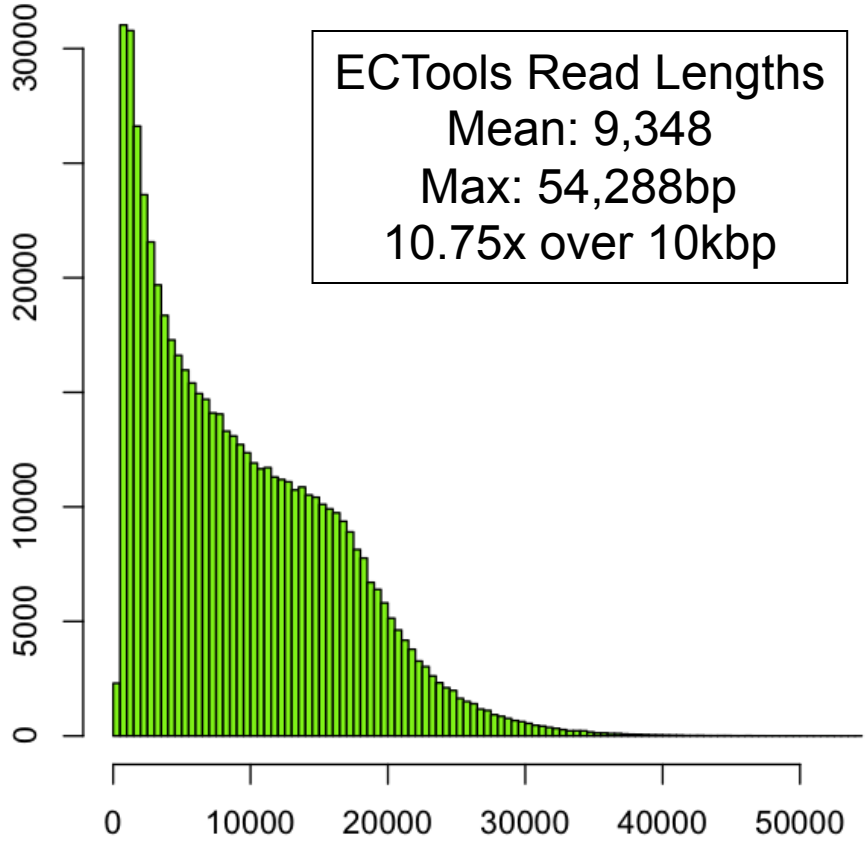
- Size selection using an 10 Kb elution window on a BluePippin™ device from Sage Science



O. sativa pv Indica (IR64)

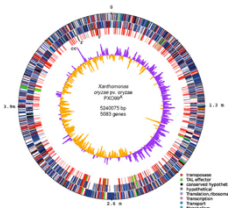
Genome size: ~370 Mb
Chromosome N50: ~29.7 Mbp

Assembly	Contig NG50
MiSeq Fragments 25x 456bp (3 runs 2x300 @ 450 FLASH)	19,078
"ALLPATHS-recipe" 50x 2x100bp @ 180 36x 2x50bp @ 2100 51x 2x50bp @ 4800	18,450
ECTools 10.7x @ 10kbp	271,885



Next steps

- Optimization of large fragment isolation and purification
- Optimization of loading SMRT cells – efficiency and consistency
- Target other yeast genomes – fermentation strains and various mutation containing strains
- Complete coverage of rice IR64
- Complete coverage of rice DJ123 (aus group)



Summary

- **Long read sequencing of eukaryotic genomes is here**
- **Technologies are quickly improving, exciting new scaffolding technologies**

- **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
  - ➢ Poster Schatz Poster #221 @ 5:00pm

Acknowledgements

McCombie Lab

Panchajanya Deshpande

Senem Mavruk Eskipehlivan

Melissa Kramer

Scott Ethe Sayers

Sara Goodwin

Eric Antoniou

Schatz Lab

James Gurtowski

Hayan Lee

Shoshana Marcus

PacBio

Cheryl Heiner

Greg Khitrov

Cornell University

Susan McCouch

Lyza Maron

Jim Hicks – CSHL

Rob Martienssen - CSHL

Pacific Biosciences

National Human Genome Research Institute

NSF