The Resurgence of Reference Quality Genome Sequence

Michael Schatz

Jan 13, 2015
PAG XXIII

@mike_schatz / #PAGXXIII

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 fo
389 Mb genome, inclu
transposable-element-
***Arabidopsis*. In a recipr**
proteome. Twenty-nin
classes of transposabl
maize and sorghum ge
nuclear chromosomes.
traits. The additional s
accelerate improveme

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

Chr	Sequenced bases (bp)	Gaps on arm regions No.	Gaps on arm regions Length (Mb)	Telomeric gaps* (Mb)	Centromeric gap† (Mb)	rDNA‡ (Mb)	Total (Mb)	Coverage§ (%)
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	12688bp	288.4 Mbp
Abyss	50X Nipponbare	2847bp	34893bp	317.4 Mbp

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

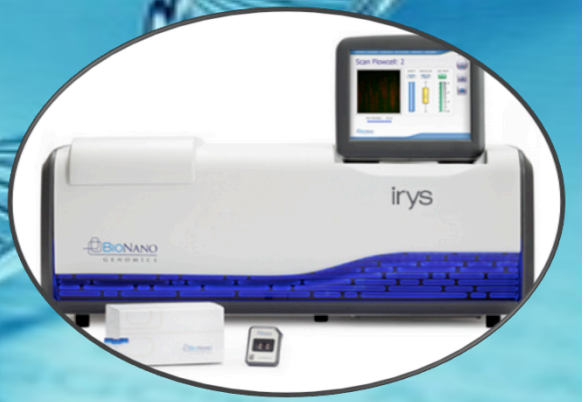
W.R. McCombie

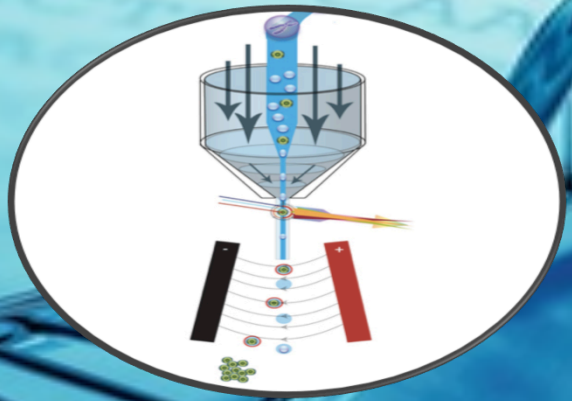
Genomics Arsenal in the year 2015

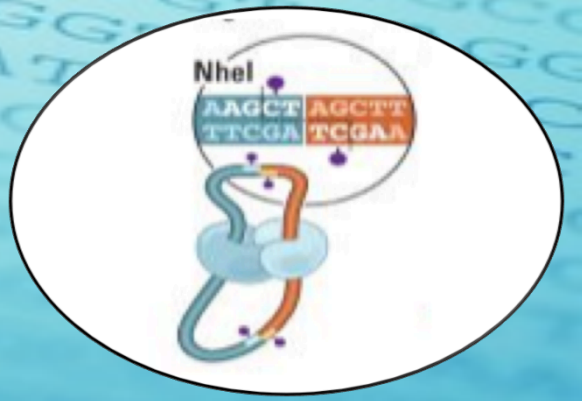
Sample Preparation

Sequencing

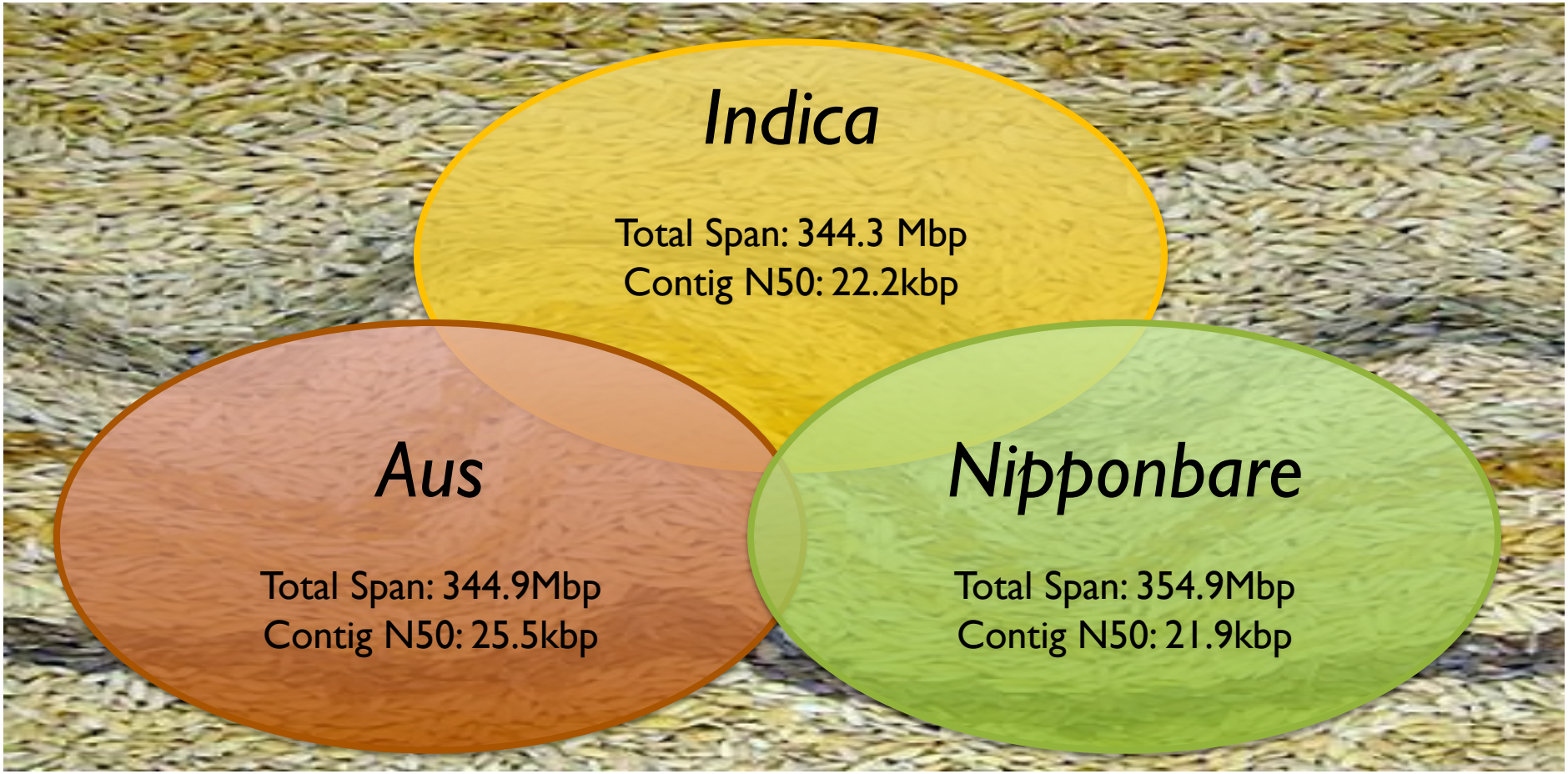
Chromosome Mapping







Population structure of *Oryza sativa*

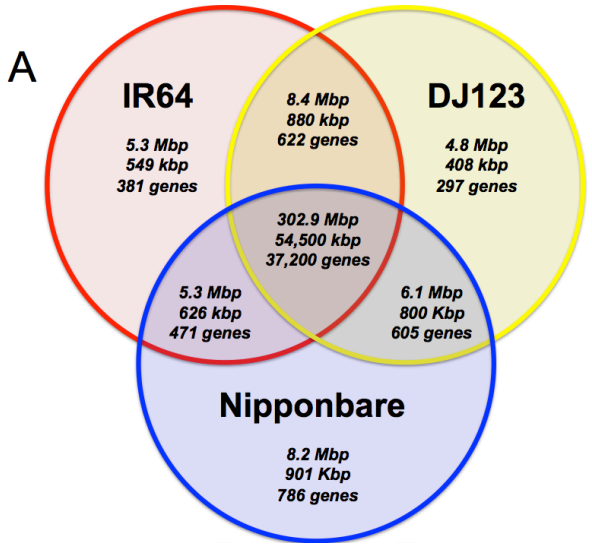


Whole genome de novo assemblies of three divergent strains of rice (*O. sativa*) documents novel gene space of *aus* and *indica*

Schatz, Maron, Stein et al (2014) *Genome Biology.* 15:506 doi:10.1186/s13059-014-0506-z

Oryza sativa Gene Diversity

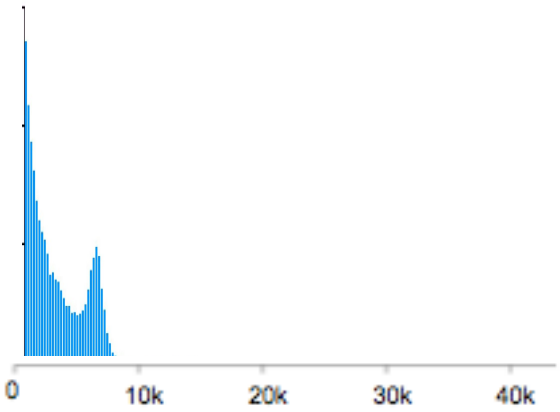
- Very high quality representation of the "gene-space"
  - Overall identity ~99.9%
  - Less than 1% of exonic bases missing
- Genome-specific genes enriched for disease resistance
  - Reflects their geographic and environmental diversity
- Assemblies fragmented at (high copy) repeats
  - Difficult to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

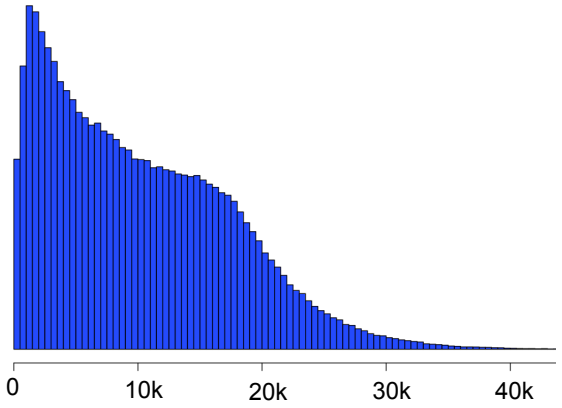
Long Read Sequencing Technology

Moleculo



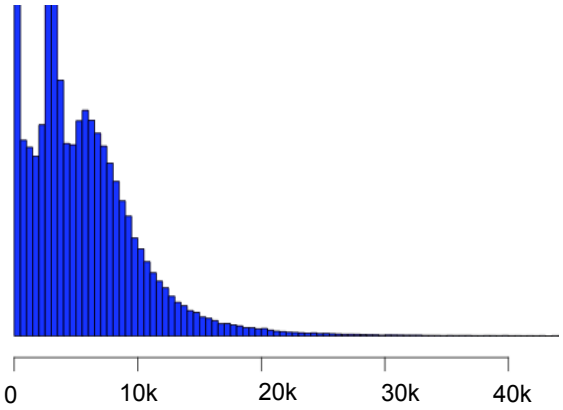
(Voskoboynik et al. 2013)

PacBio RS II



CSHL/PacBio

Oxford Nanopore

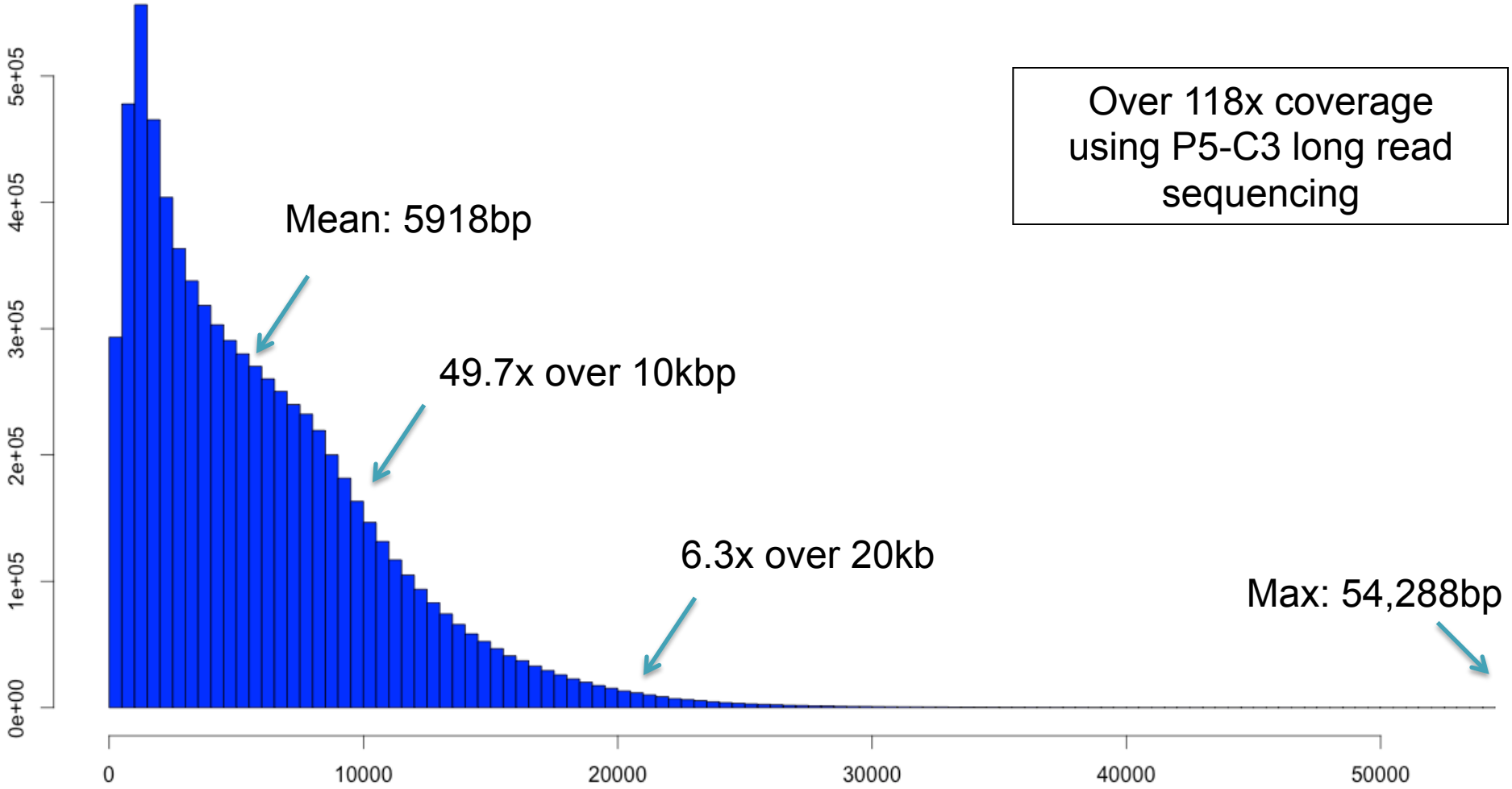


CSHL/ONT

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

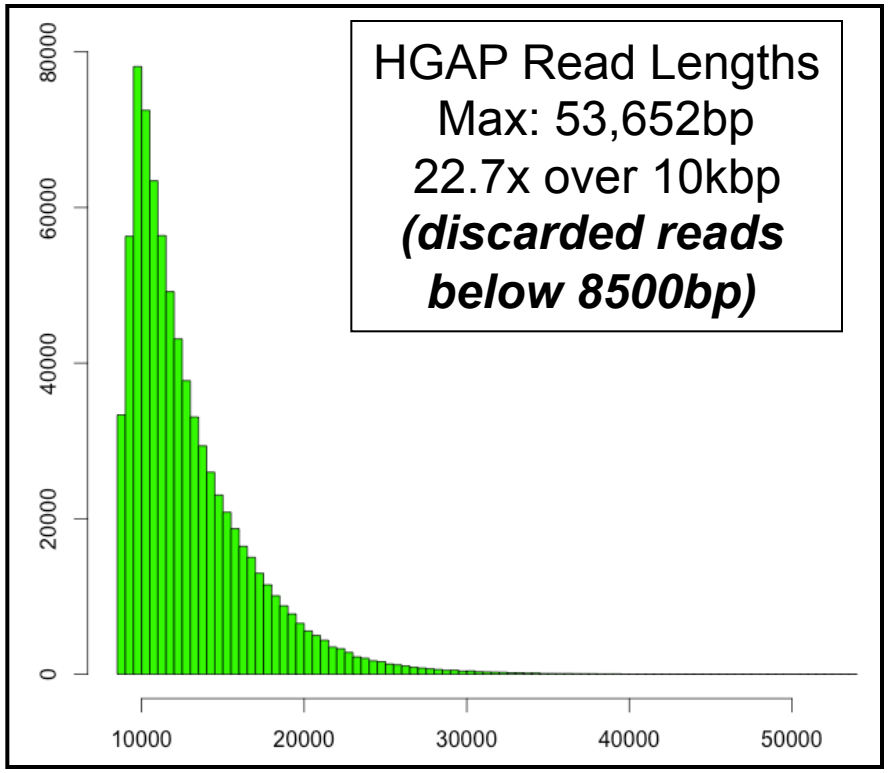


Over 118x coverage using P5-C3 long read sequencing

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



S5 Hybrid Sterility Locus

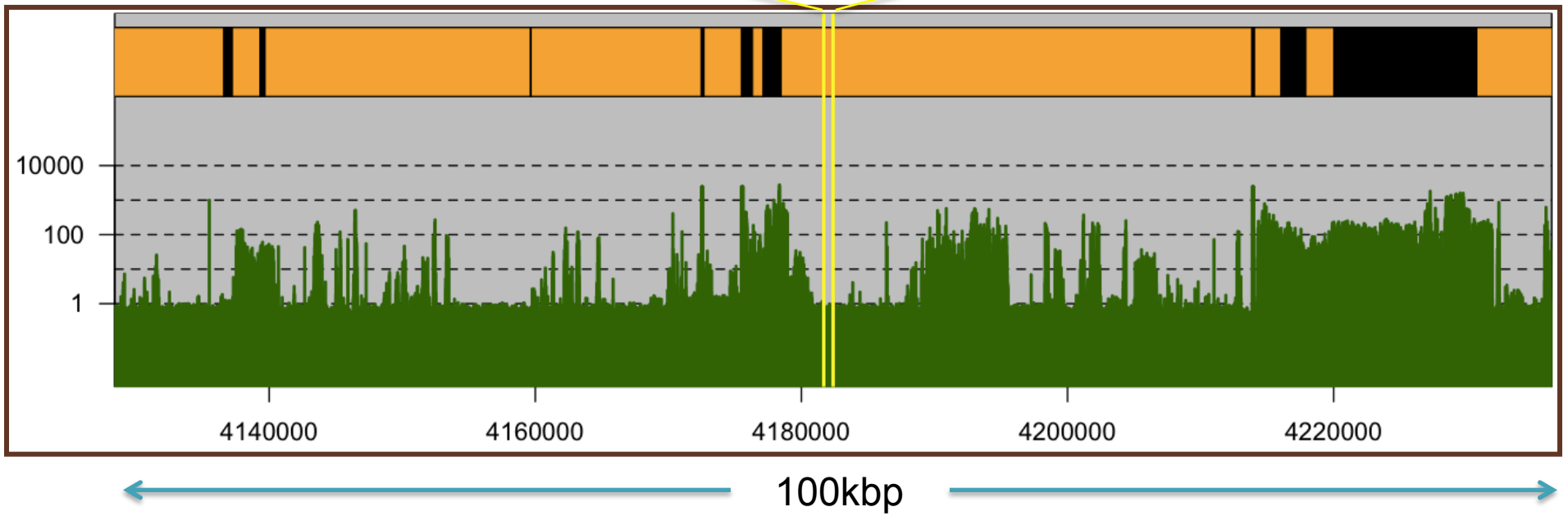
```
Sanger      …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
Illumina    …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
PacBio      …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
```

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**

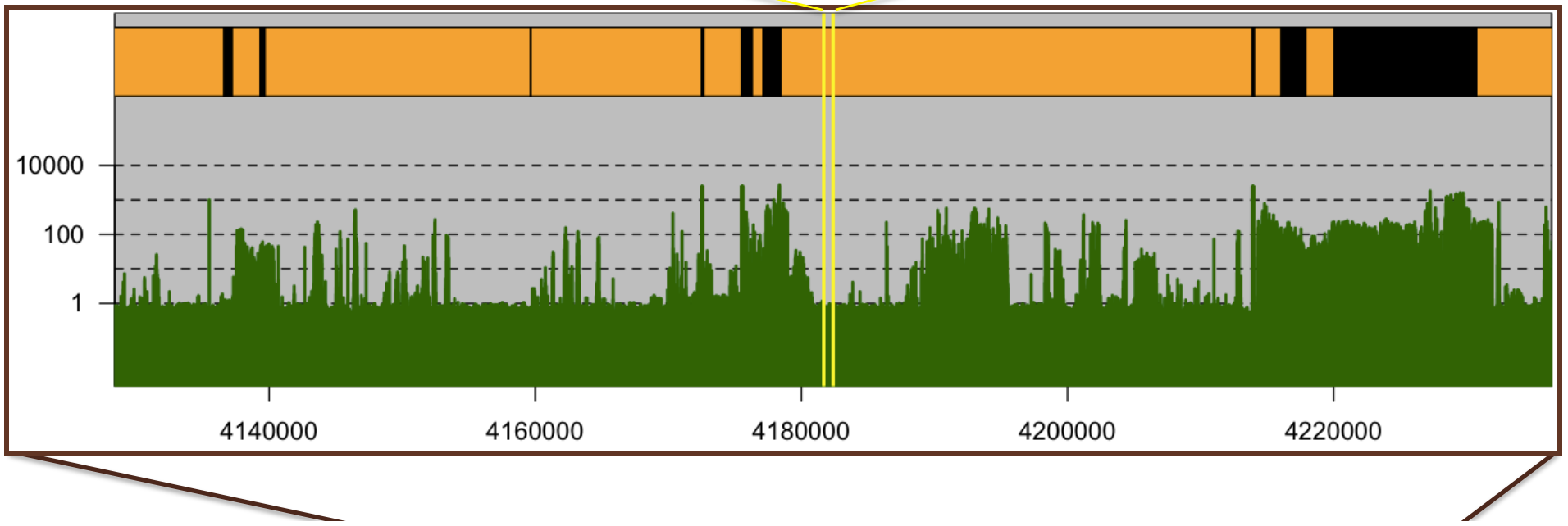
S5 Hybrid Sterility Locus

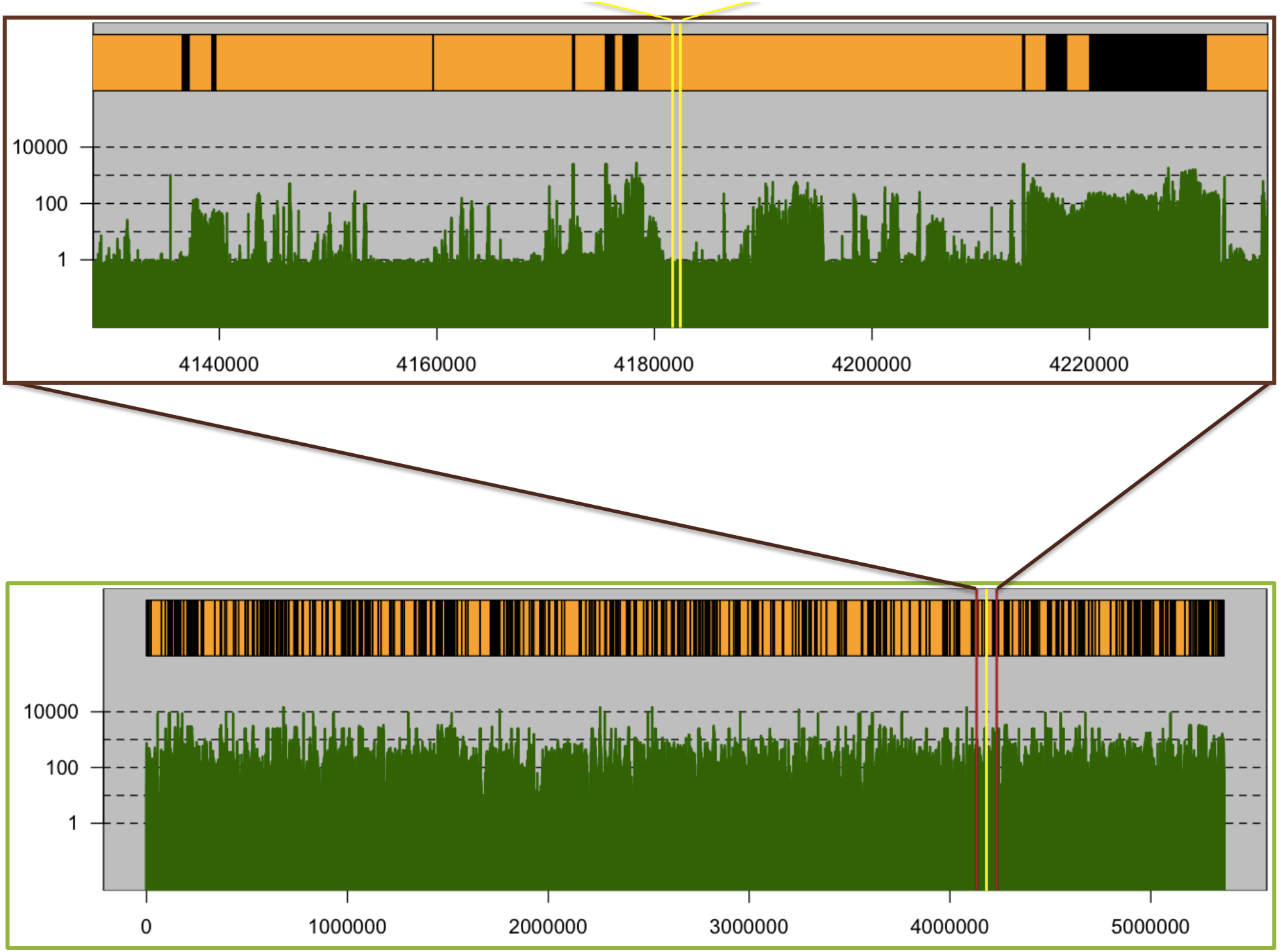
```
Sanger      …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
Illumina    …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
PacBio      …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
```

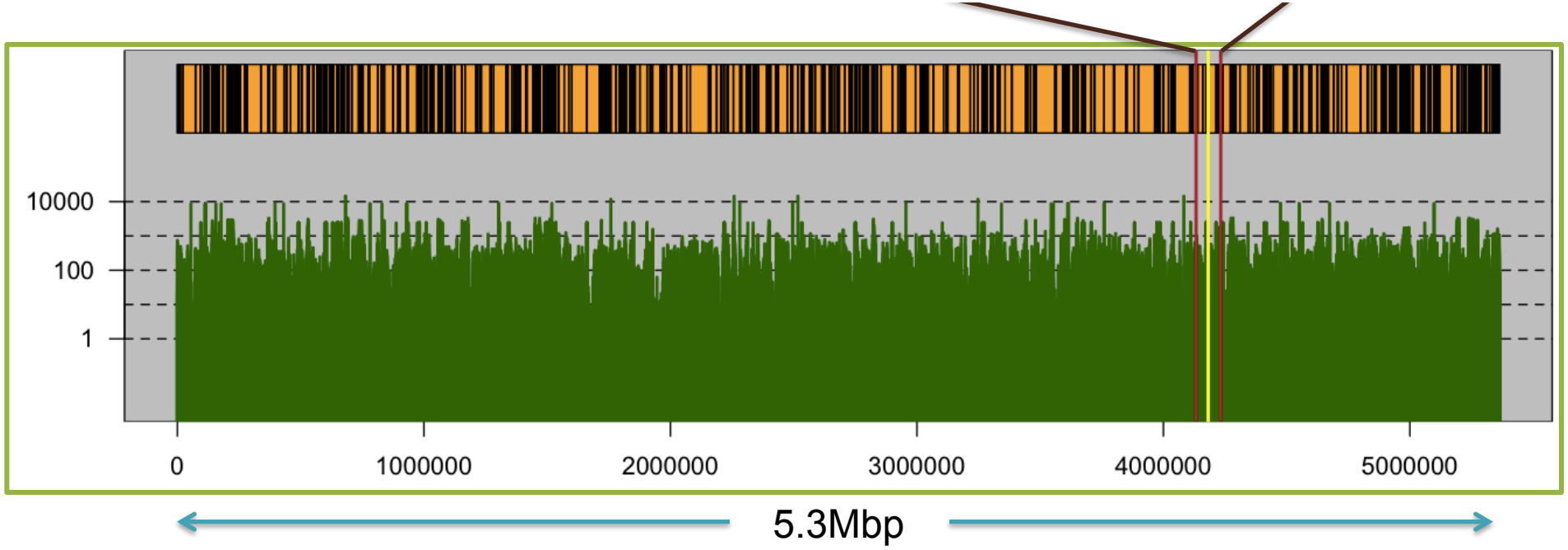


S5 Hybrid Sterility Locus

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




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



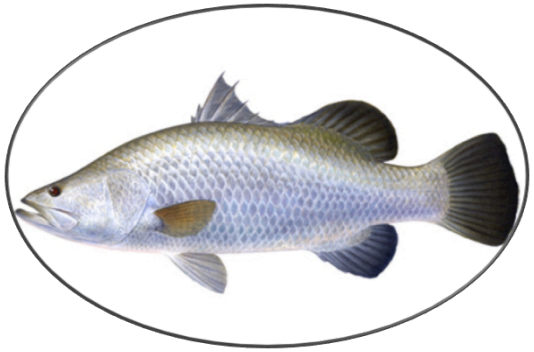
Improvements from 20kbp to 4Mbp contig N50:

- 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

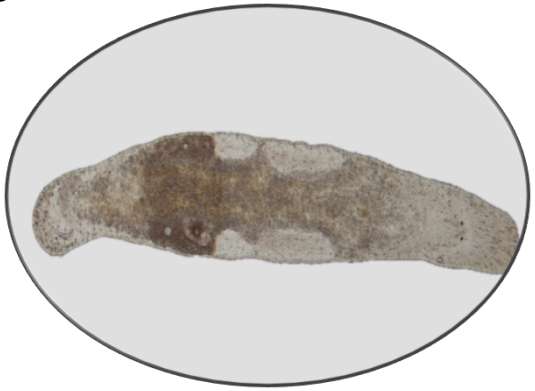
Current Collaborations

Pineapple
UIUC

Human
CSHL/OICR



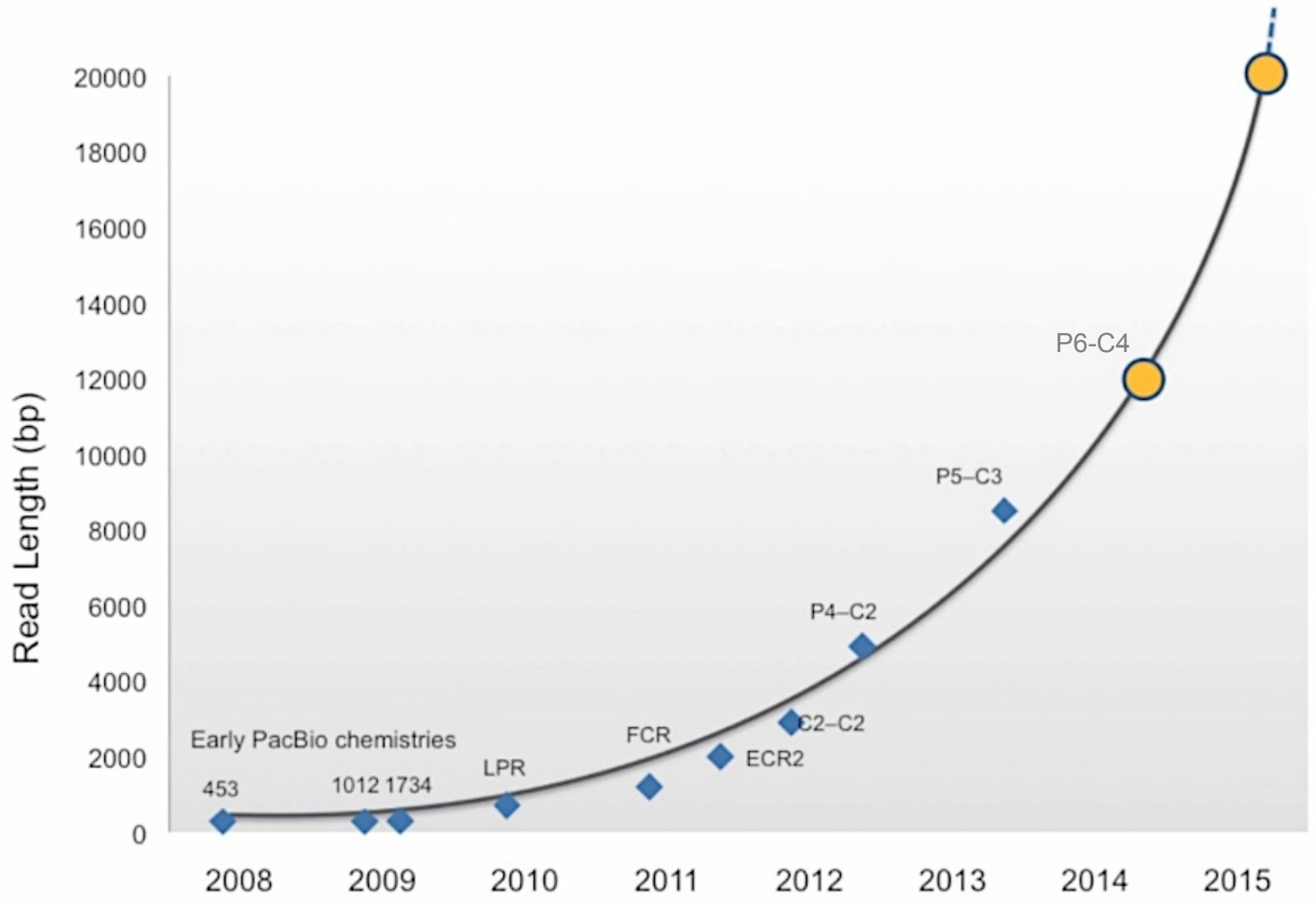
Asian Sea Bass
Temasek Life Sciences



M. ligano
Hannon

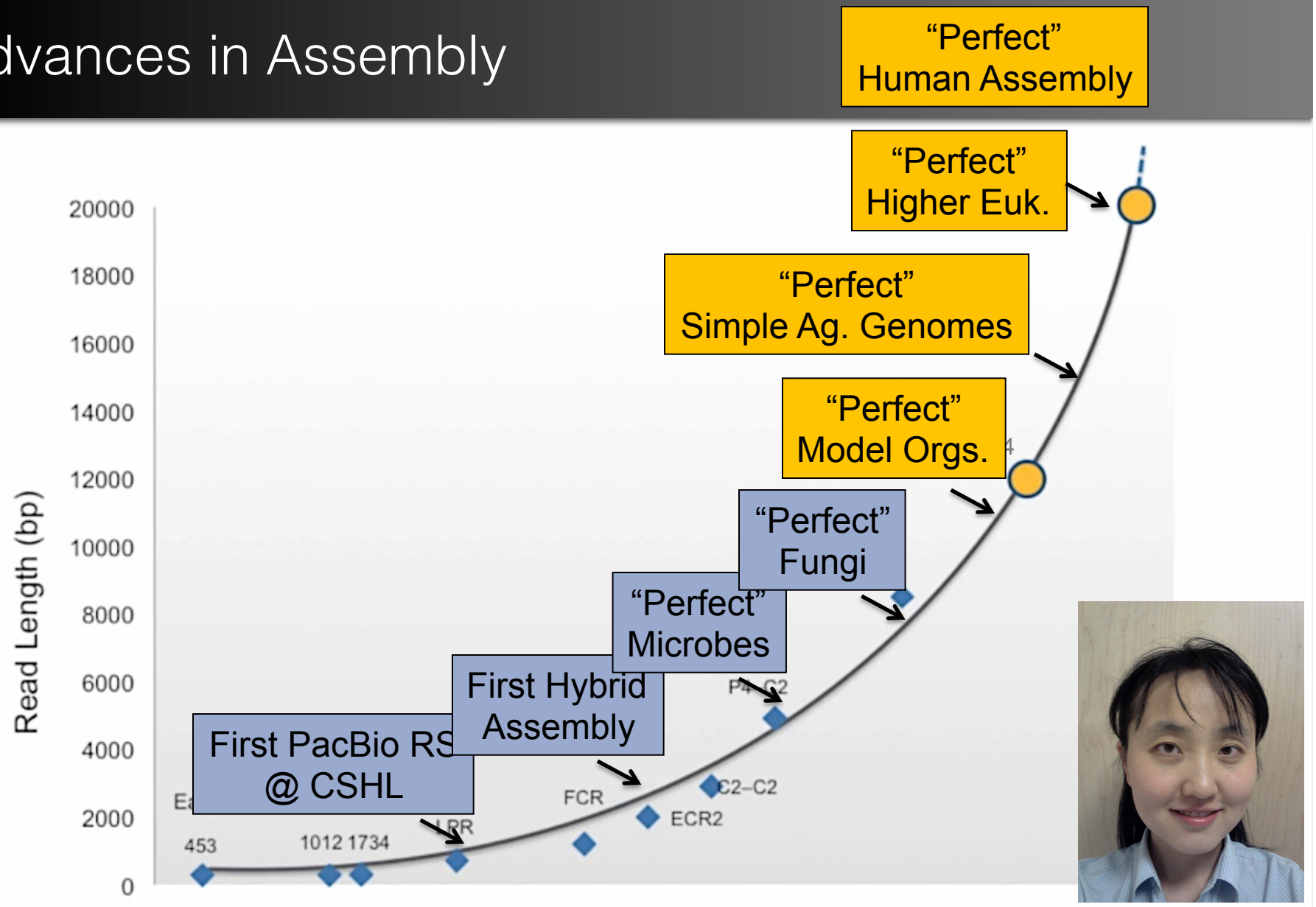
T. vaginalis
NYU

PacBio® Advances in Read Length



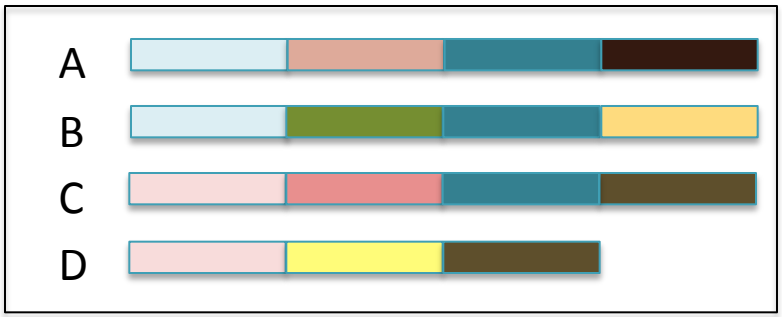
PACIFIC BIOSCIENCES®

Advances in Assembly



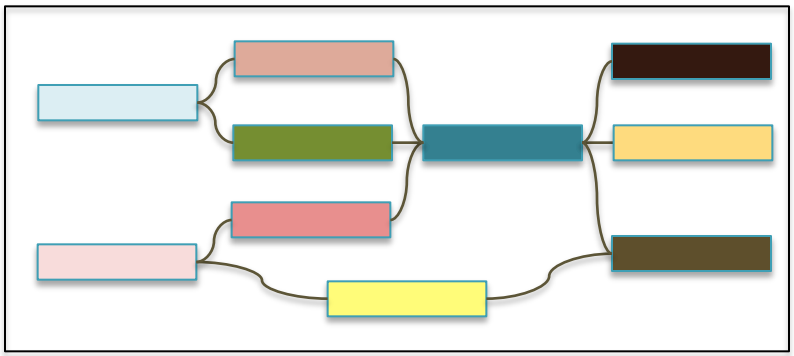
Error correction and assembly complexity of single molecule sequencing reads.
Lee, H*, Gurtowski, J*, Yoo, S, Marcus, S, McCombie, WR, Schatz, MC
http://www.biorxiv.org/content/early/2014/06/18/006395

Pan-Genome Alignment & Assembly



Time to start considering problems for which N complete genomes is the input to study the "pan-genome"

- Available today for many microbial species, near future for higher eukaryotes



Pan-genome colored de Bruijn graph

- Encodes all the sequence relationships between the genomes
- How well conserved is a given sequence?
- What are the pan-genome network properties?

SplitMEM: A graphical algorithm for pan-genome analysis with suffix skips
Marcus, S, Lee, H, Schatz MC (2014) *Bioinformatics*. doi: 10.1093/bioinformatics/btu756

Extending reference assembly models
Church, D. *et al.* (2015) *Genome Biology*. In Press.

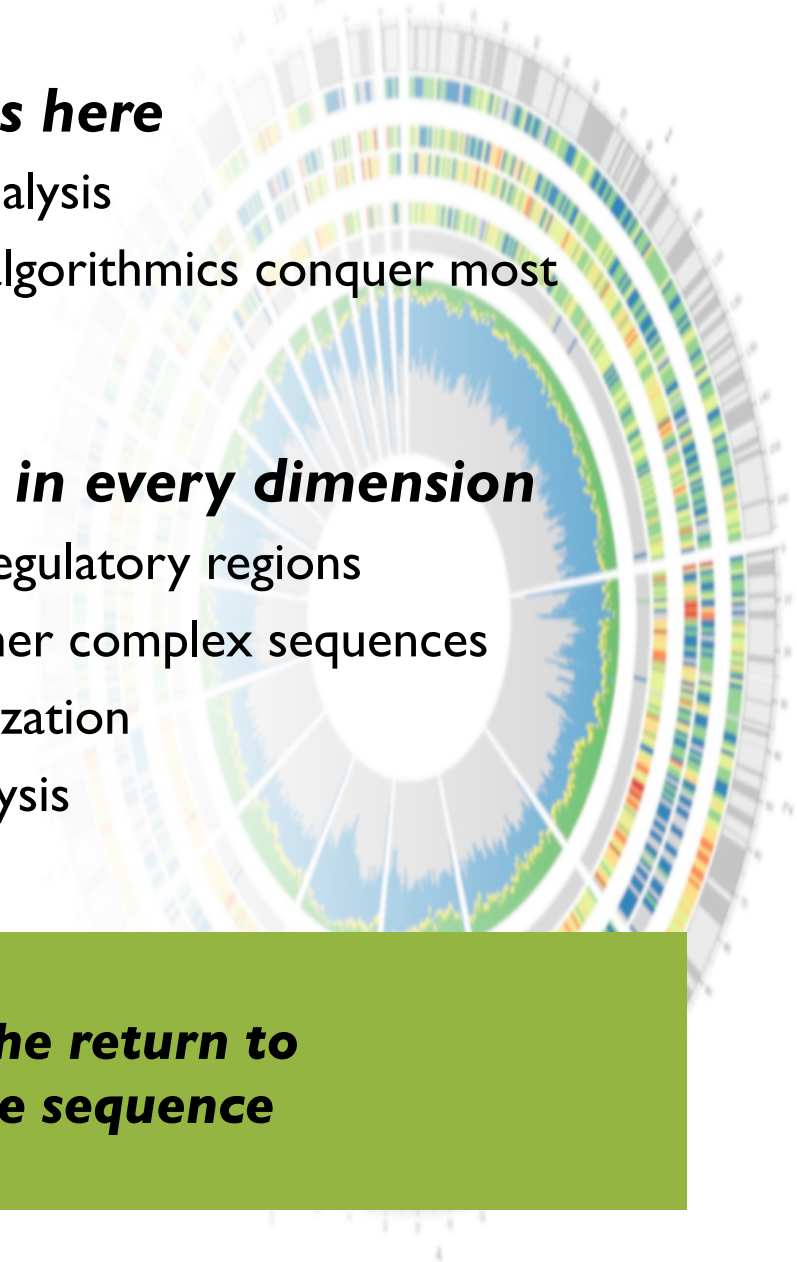
Summary & Recommendations

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
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

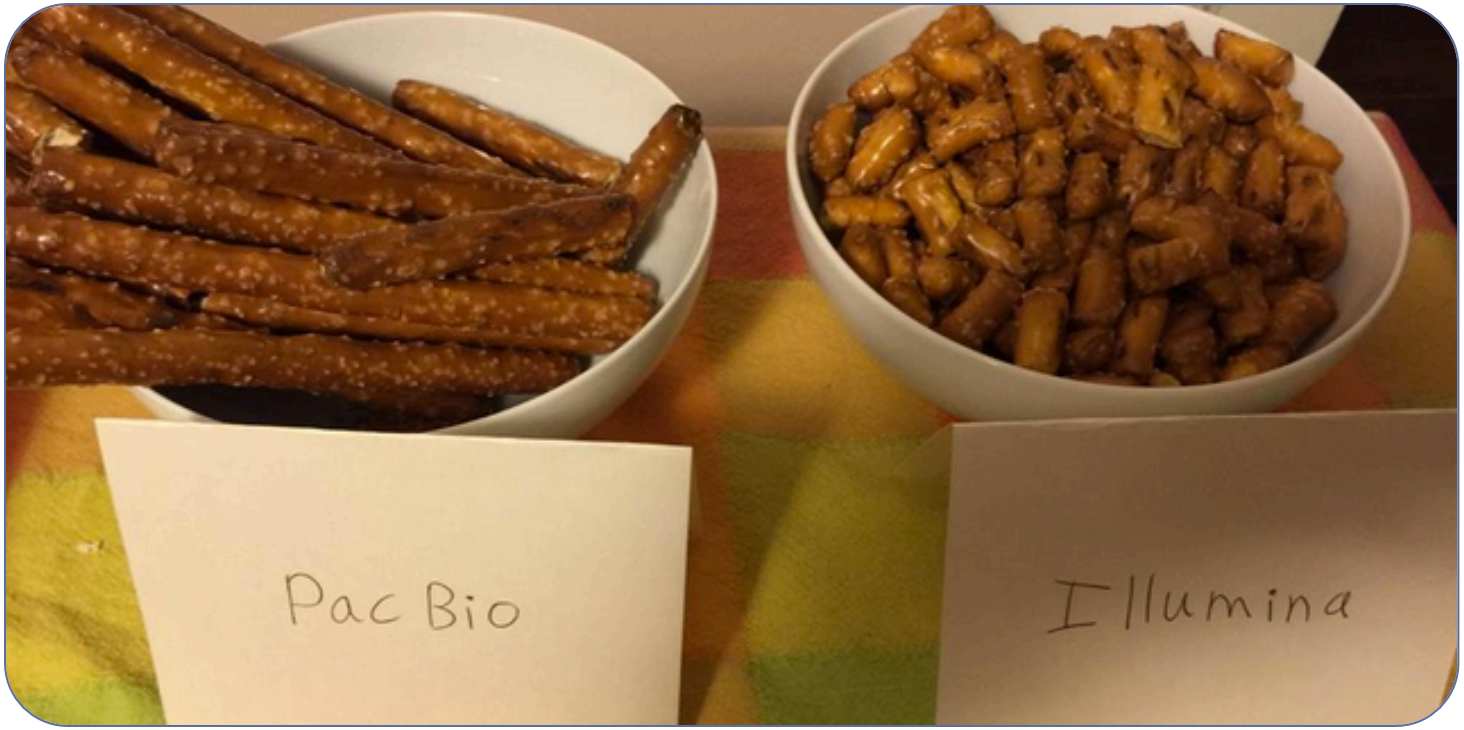
IT & Meetings Depts.
Pacific Biosciences
Oxford Nanopore

NSF

National Human Genome Research Institute

U.S. DEPARTMENT OF ENERGY

SFARI
SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE



Thank you

http://schatzlab.cshl.edu

@mike_schatz / PAGXXIII

O. sativa pv Indica (IR64)
S5 Hybrid Sterility Locus

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