Algorithms for studying the structure and function of genomes

Michael Schatz

Feb 5, 2015

JHU Dept. of Biology

Schatzlab Overview

Human Genetics

Role of mutations in disease

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

Plant Biology

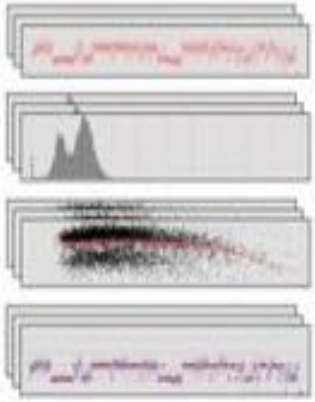
Genomes & Transcriptomes

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

Algorithmics & Systems Research

Ultra-large scale biocomputing

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



Single Cell & Single Molecule

CNVs, SVs, & Cell Phylogenetics

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

Genome Structure & Function

1. **Structure: Sequencing and Assembly**

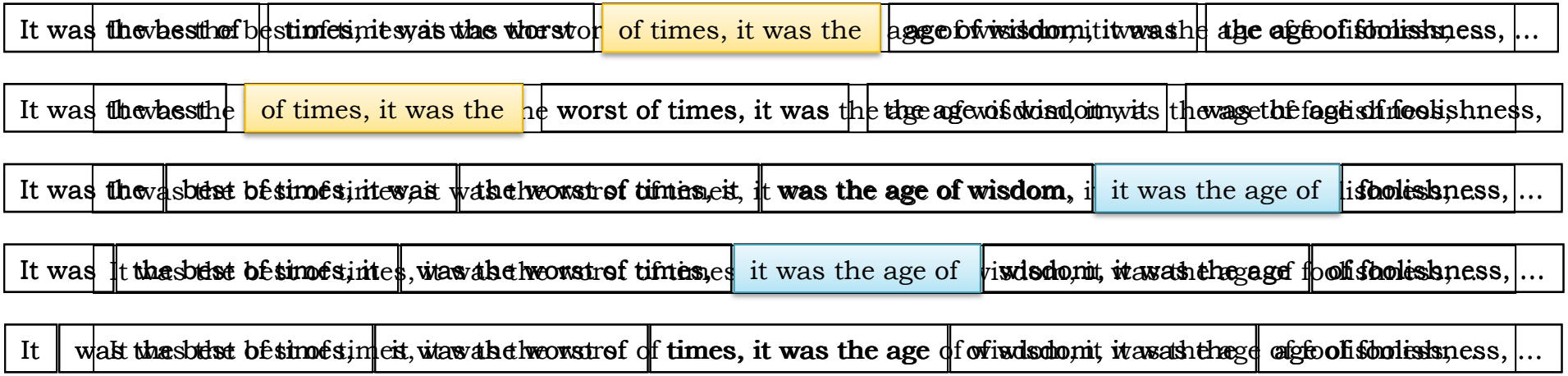
   Long Read Single Molecule Sequencing

2. **Function: Disease Analytics**

   The role of indels in autism spectrum disorders

Shredded Book Reconstruction

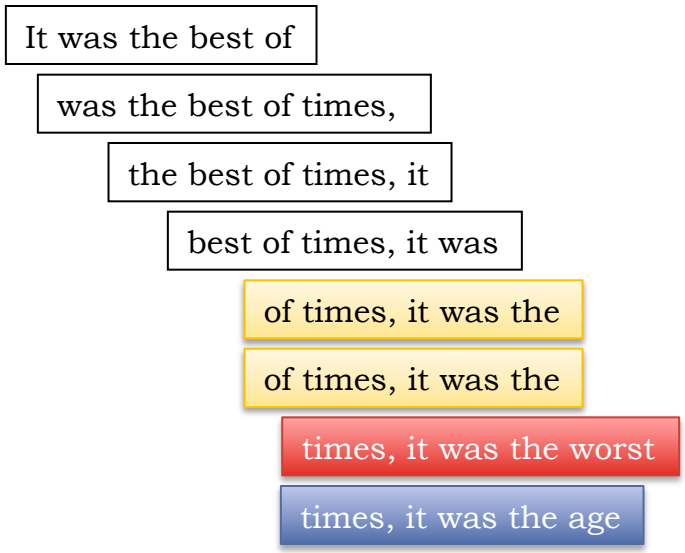
- Dickens accidentally shreds the first printing of A Tale of Two Cities
  - Text printed on 5 long spools



- 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



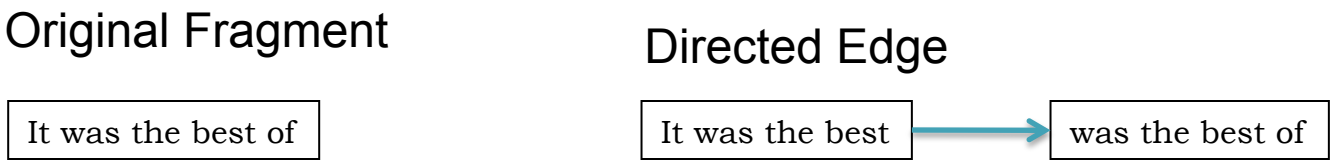
The repeated sequence make the correct reconstruction ambiguous

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

Model the assembly problem as a graph problem

de Bruijn Graph Construction

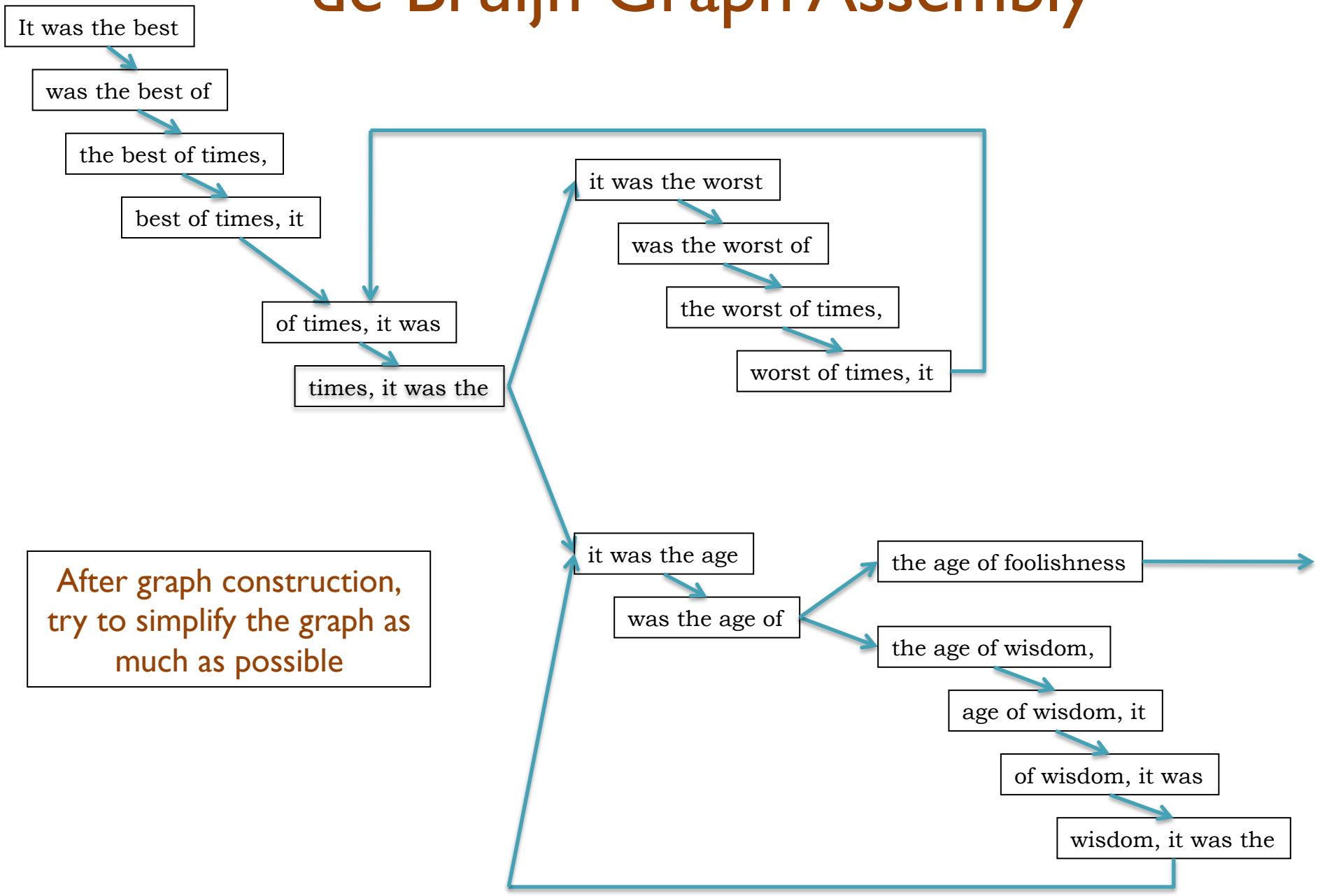
- $D_k = (V,E)$
  - V = All length-k subfragments (k < l)
  - E = Directed edges between consecutive subfragments
    - Nodes overlap by k-1 words



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

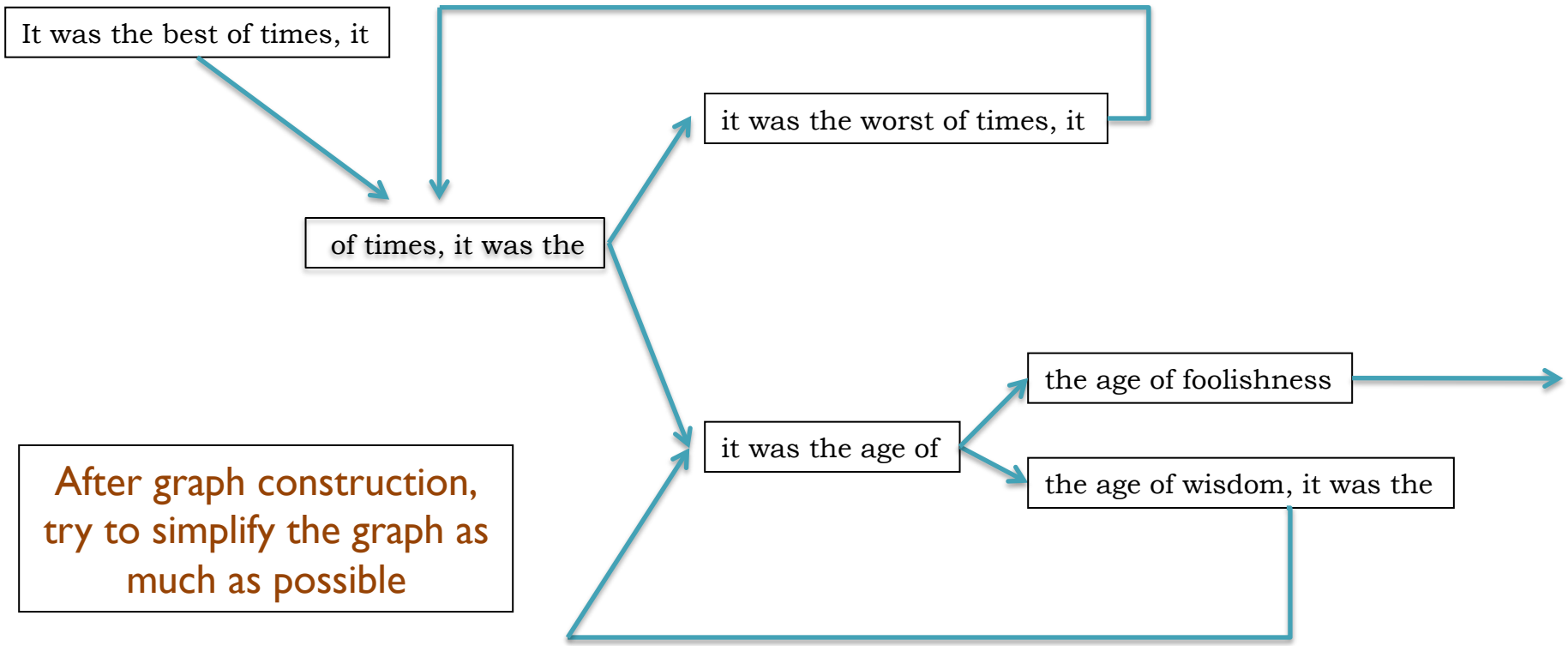
de Bruijn, 1946
Idury and Waterman, 1995
Pevzner, Tang, Waterman, 2001

de Bruijn Graph Assembly



After graph construction, try to simplify the graph as much as possible

de Bruijn Graph Assembly



After graph construction, try to simplify the graph as much as possible

The full tale

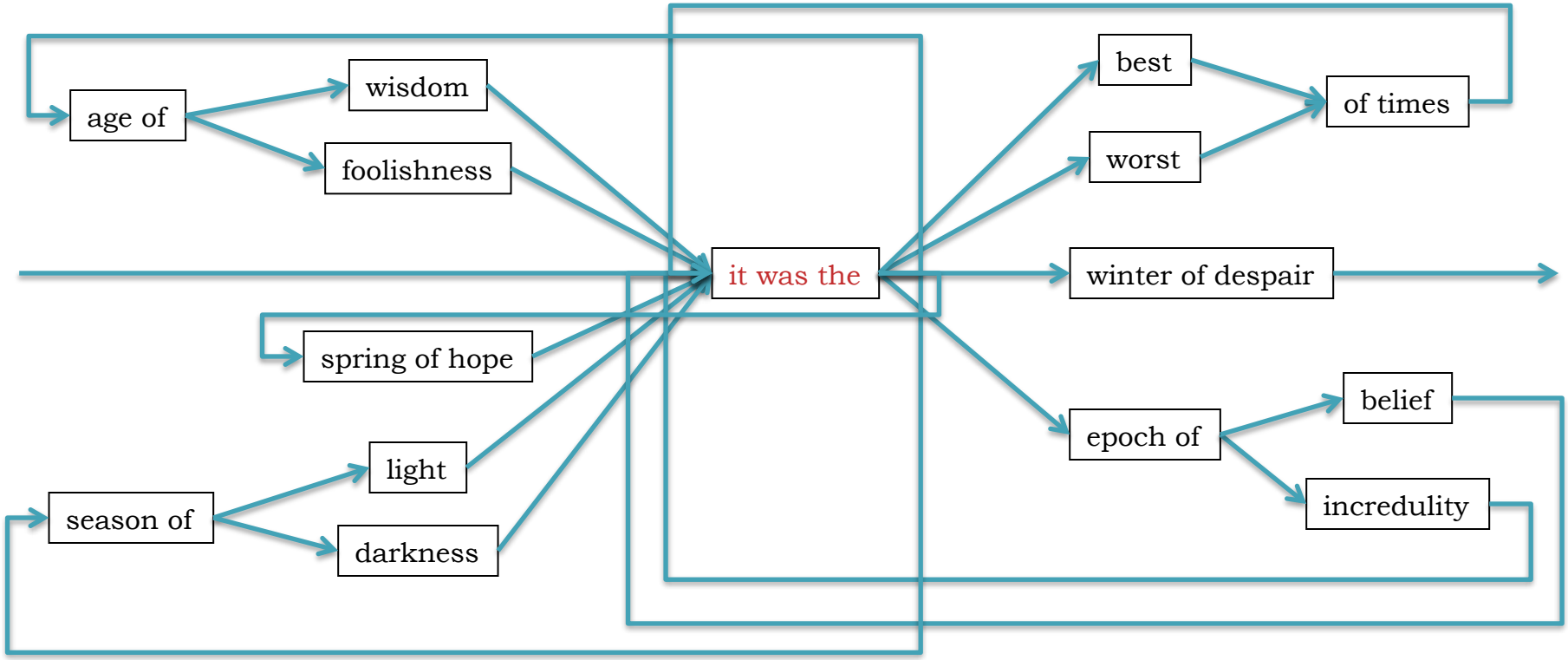
… it was the best of times it was the worst of times …

… 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 …

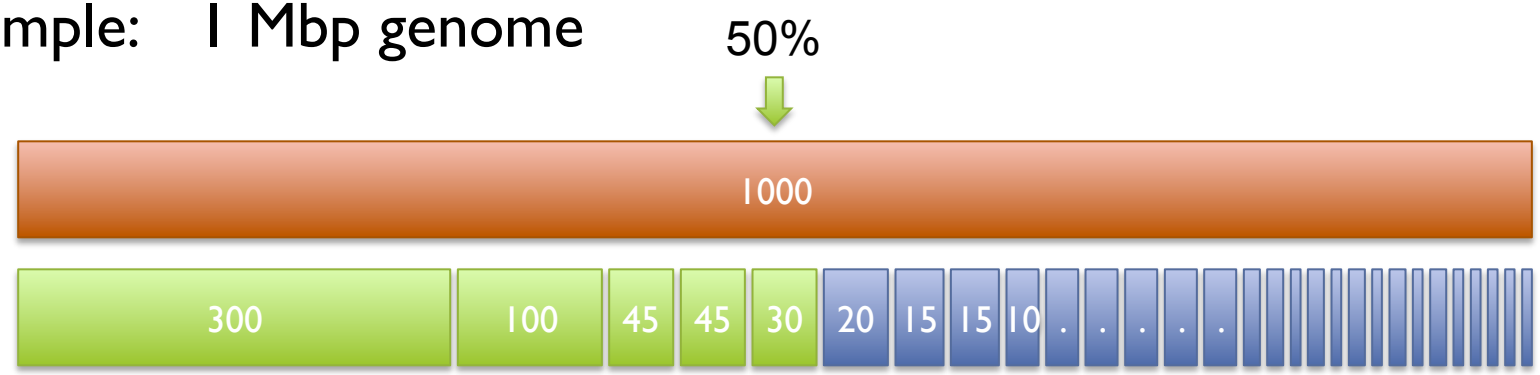
… it was the spring of hope it was the winder of despair …



N50 size

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

Example: 1 Mbp genome



N50 size = 30 kbp

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

A greater N50 is indicative of improvement in every dimension:

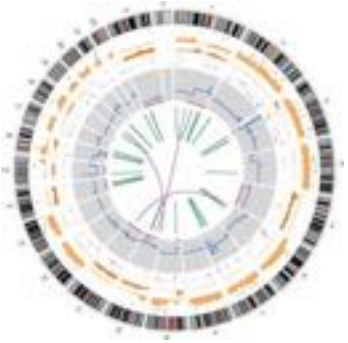
- Better resolution of genes and flanking regulatory regions
- Better resolution of transposons and other complex sequences
- Better resolution of chromosome organization
- Better sequence for all downstream analysis

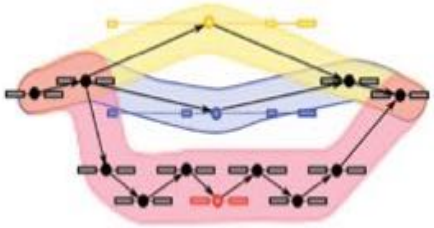
Assembly Applications

- Novel genomes

- Metagenomes

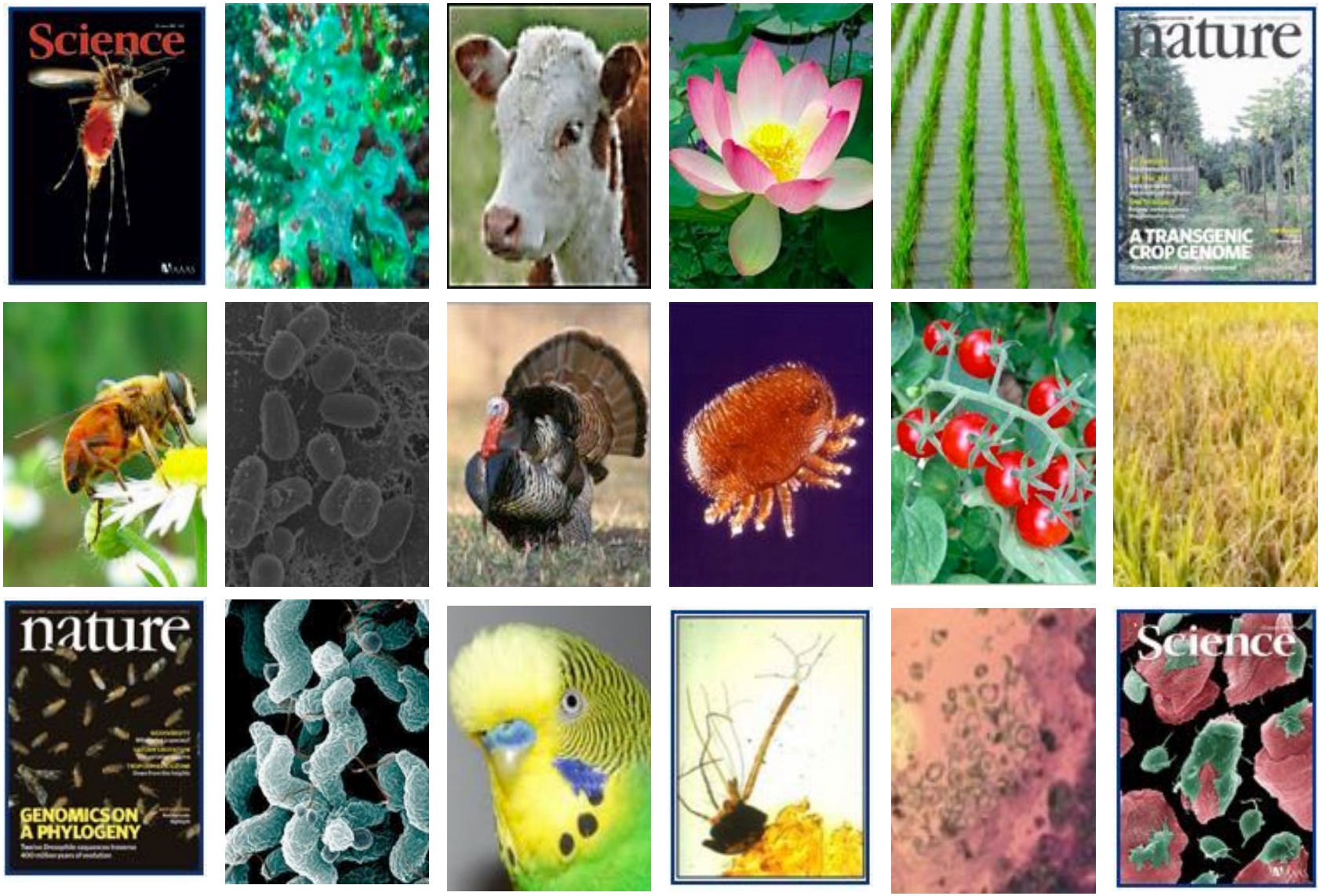
- Sequencing assays
  - Structural variations
  - Transcript assembly
  - …





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

Genomics Across the Tree of Life



The map-based sequence of the rice genome

International Rice Genome Sequencing Project*

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 gaps‡ (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)

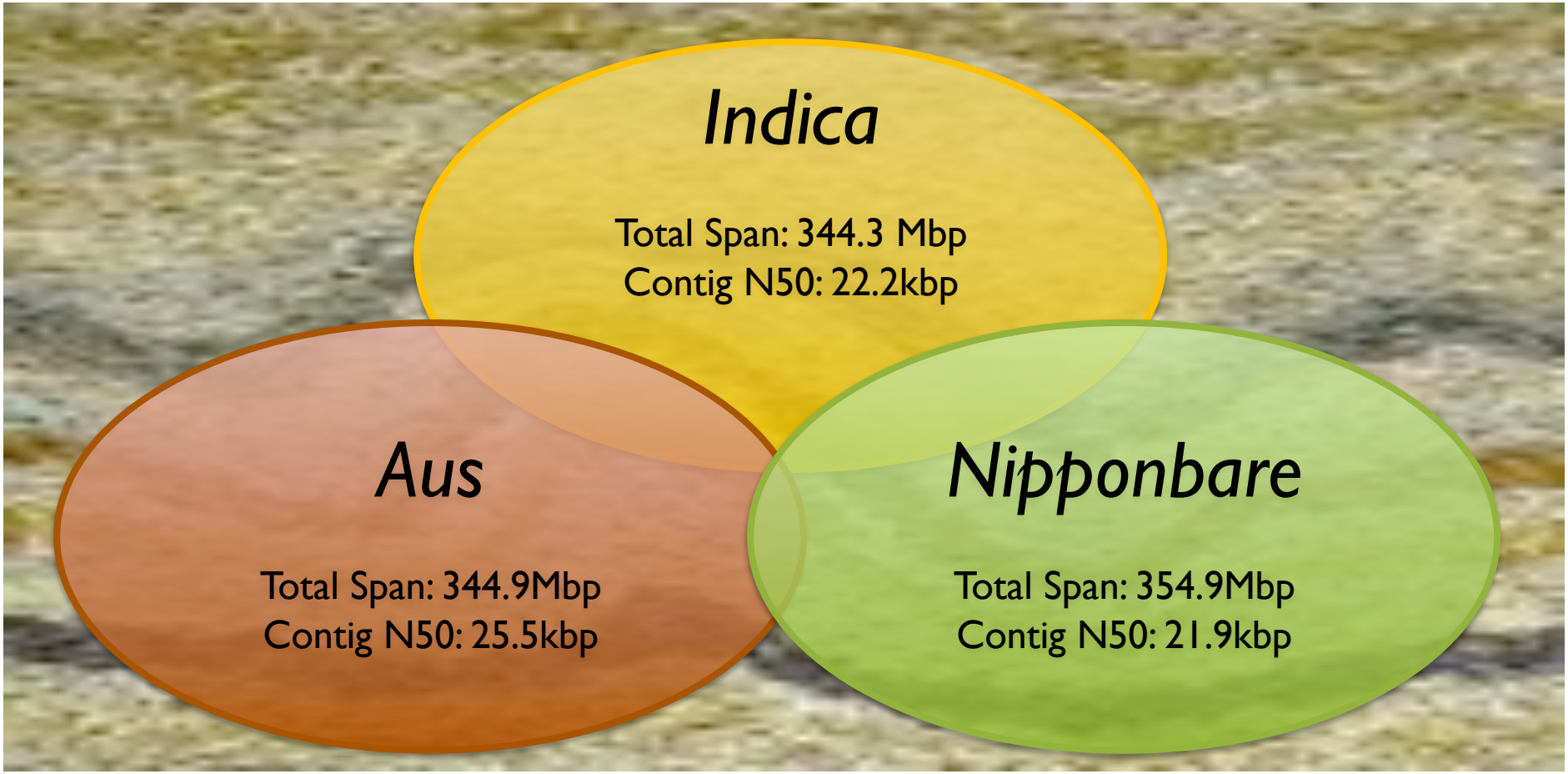
Assembler	Data set	N50 contig size	Max contig size	Total assembly size
Velvet	25X Nipponbare	1349bp	21833bp	325.8 Mbp
Velvet	50X Nipponbare	1719bp	23085bp	421.6 Mbp
Abyss	25X Nipponbare	1853bp	12588bp	288.4 Mbp
Abyss	50X Nipponbare	2947bp	34803bp	317.4 Mbp

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

W.R. McCombie

Genomics Arsenal in the year 2015

Population structure of *Oryza sativa*

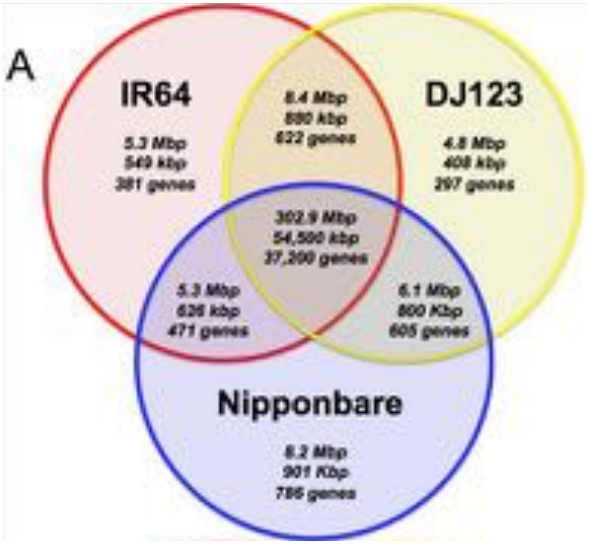


Whole genome de novo assemblies of three divergent strains of rice (*O. sativa*) documents novel gene space of *aus* and *indica*

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

Oryza sativa Gene Diversity

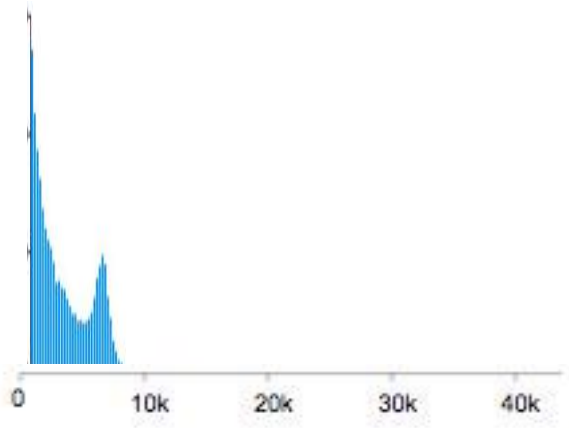
- Very high quality representation of the "gene-space"
  - Overall identity ~99.9%
  - Less than 1% of exonic bases missing
- Genome-specific genes enriched for disease resistance
  - Reflects their geographic and environmental diversity
- Assemblies fragmented at (high copy) repeats
  - Difficult to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

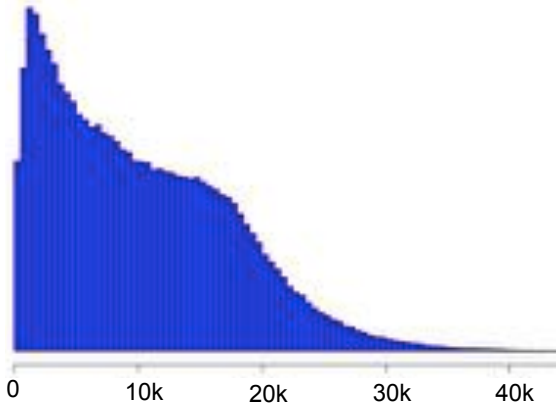
Long Read Sequencing Technology

Moleculo



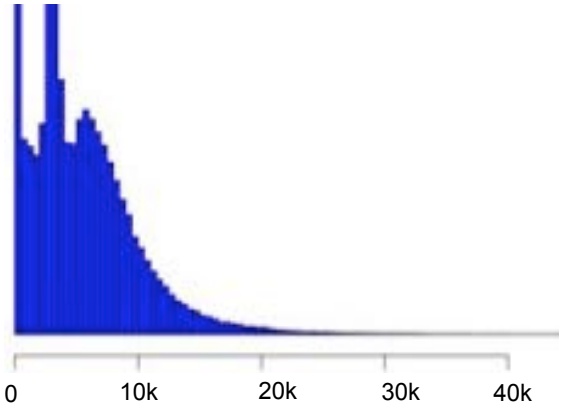
(Voskoboynik et al. 2013)

PacBio RS II



CSHL/PacBio

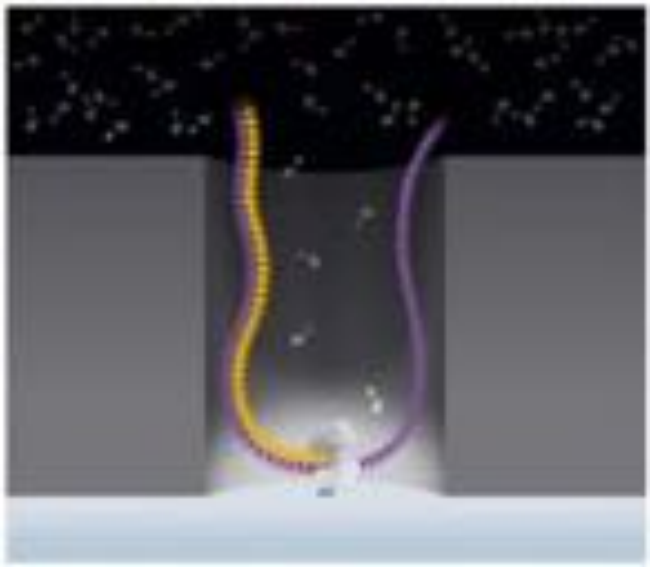
Oxford Nanopore

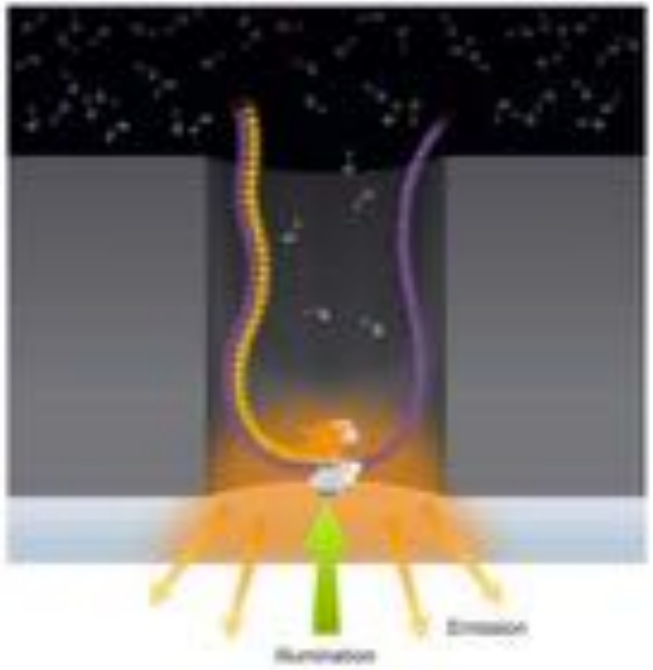


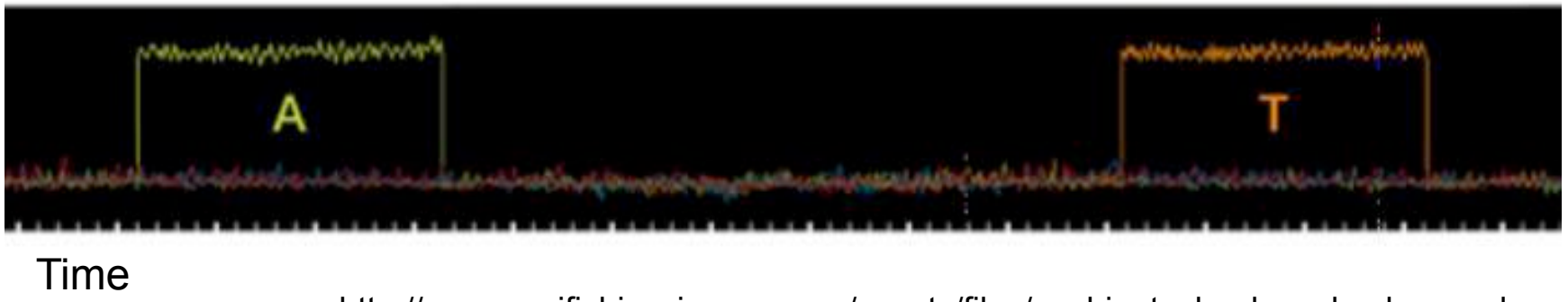
CSHL/ONT

PacBio SMRT Sequencing

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

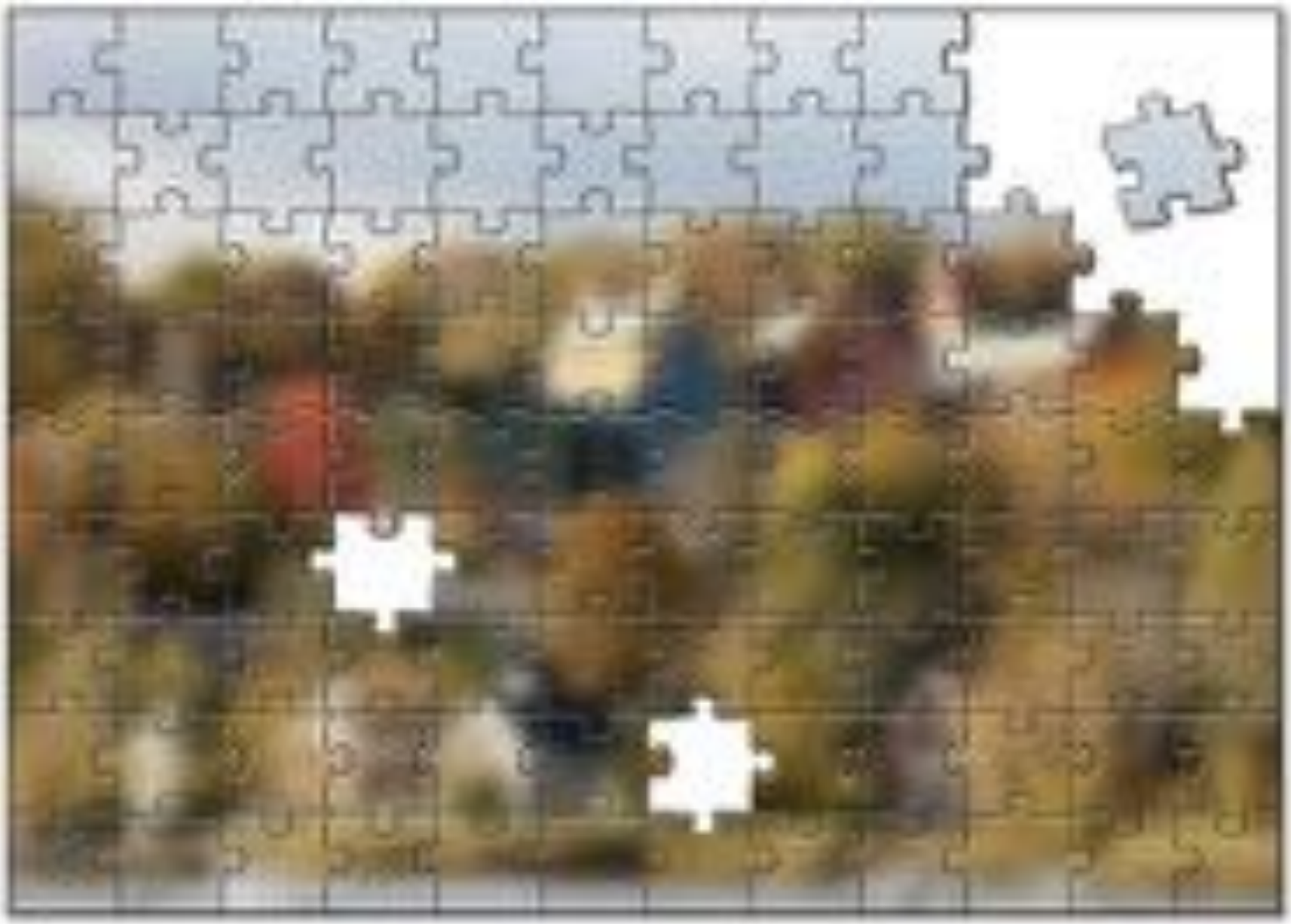




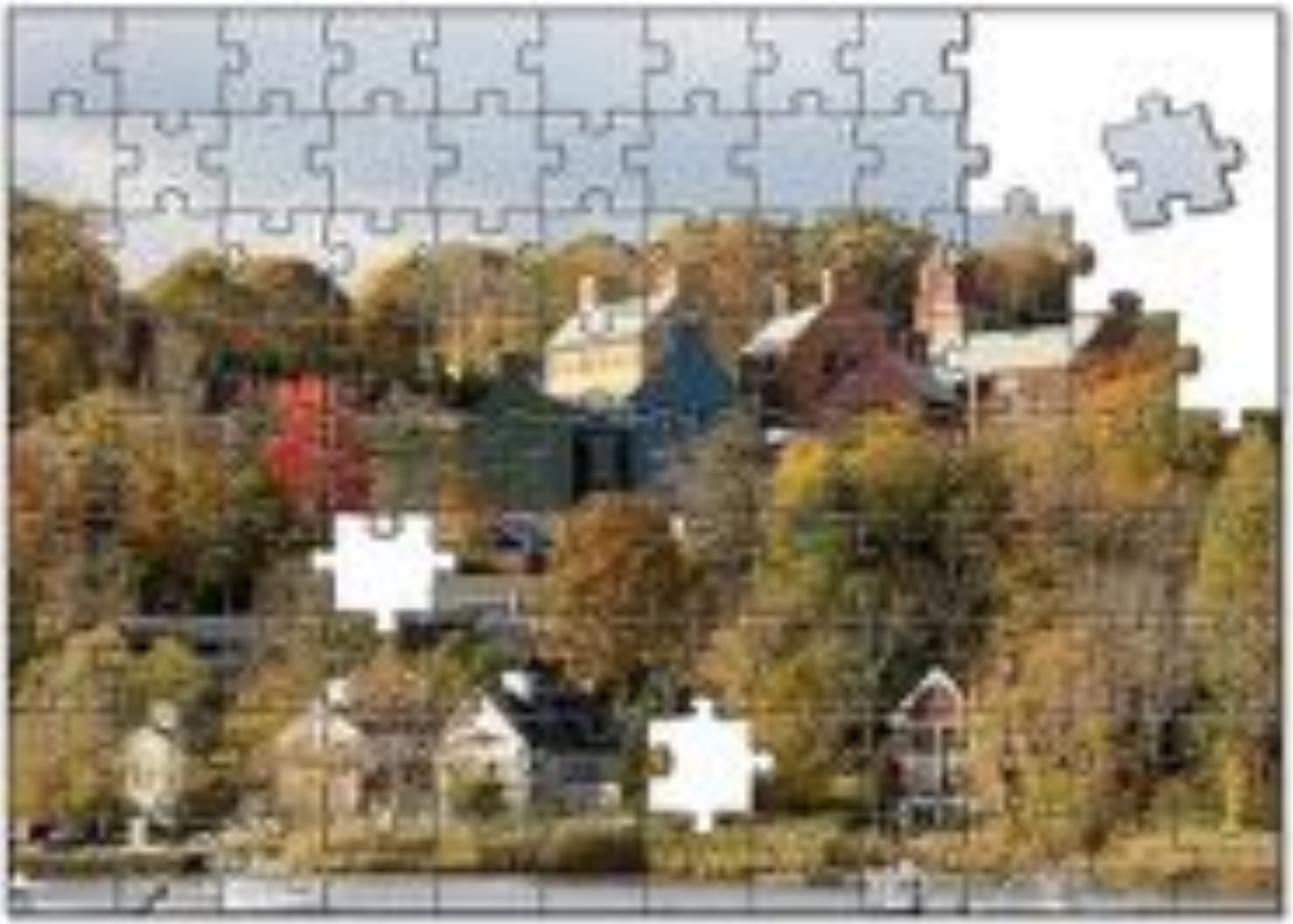


http://www.pacificbiosciences.com/assets/files/pacbio_technology_backgrounder.pdf

Single Molecule Sequences



“Corrective Lens” for Sequencing



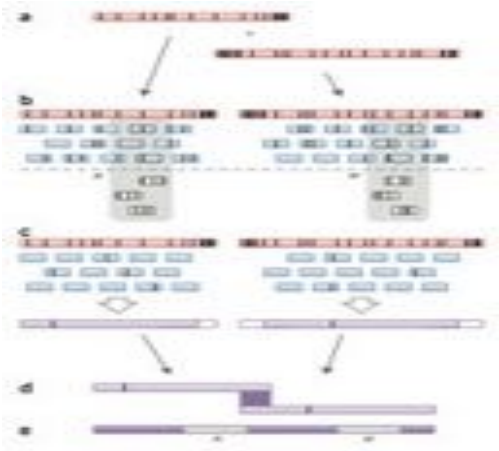
PacBio Assembly Algorithms

PBJelly

Gap Filling and Assembly Upgrade

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

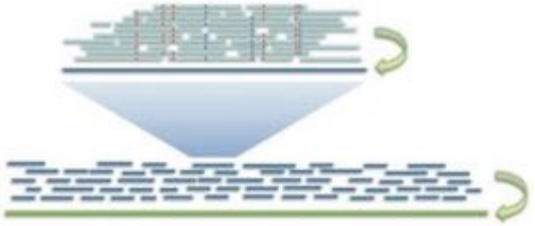
PacBioToCA & ECTools



Hybrid/PB-only Error Correction

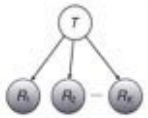
Koren, Schatz, *et al* (2012)
Nature Biotechnology. 30:693–700

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)

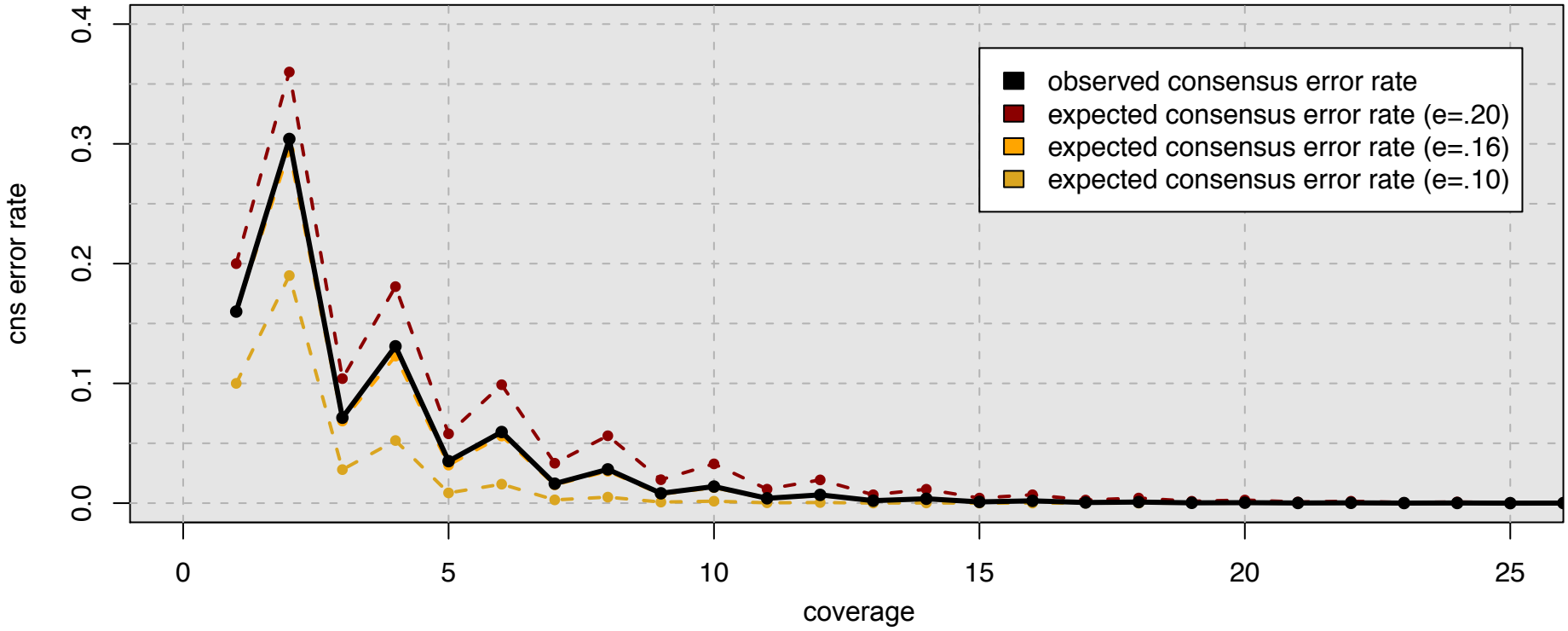
	Initial Assembly	Quiver Consensus
QV	43.4	54.5
Accuracy	99.99540%	99.99964%
Differences	141	11

PB-only Correction & Polishing

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

< 5x PacBio Coverage > 50x

Consensus Accuracy and Coverage



Coverage can overcome random errors

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

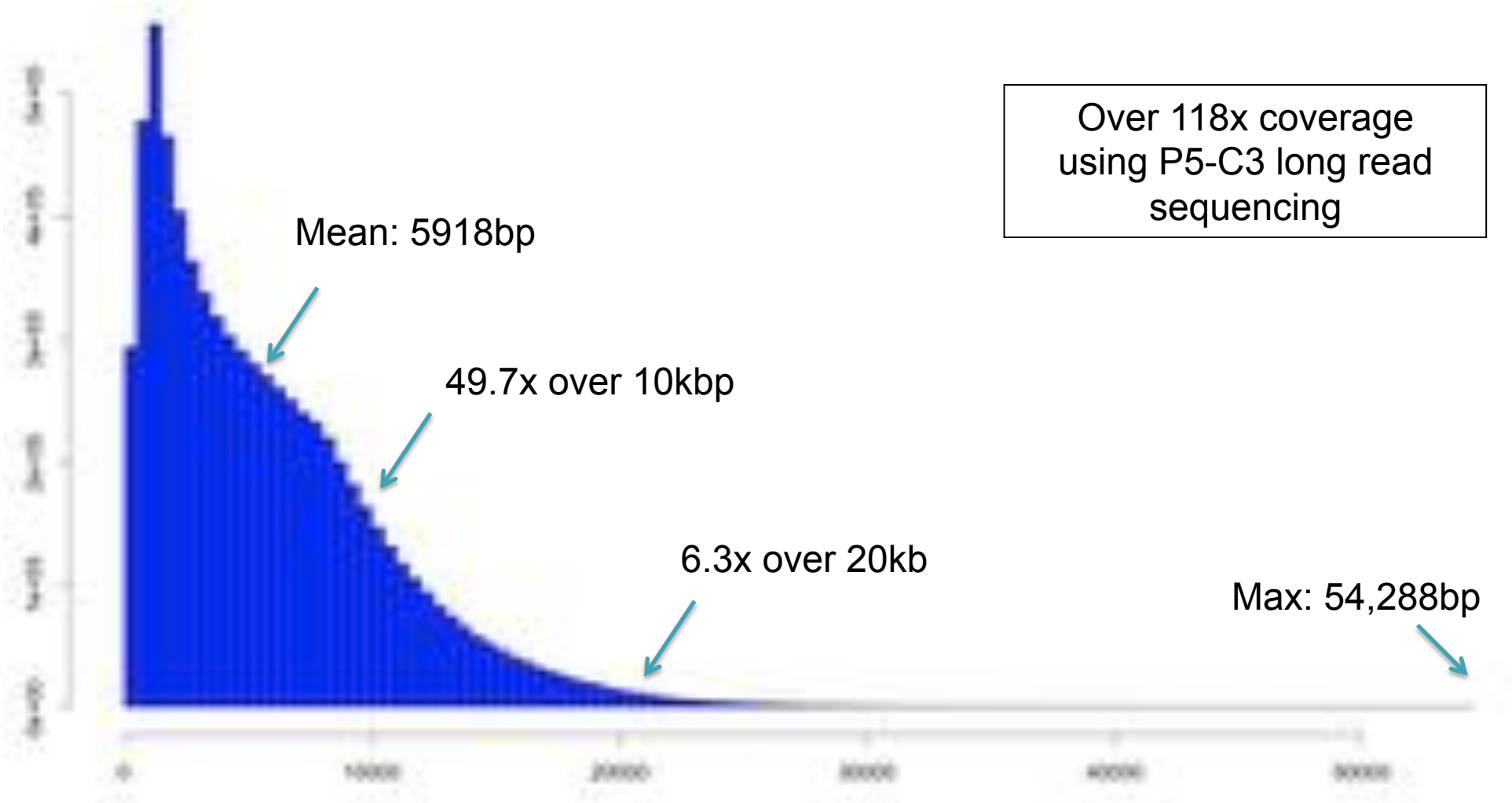
$$CNS\ Error = \sum_{i=\lceil c/2 \rceil}^{c} \binom{c}{i} (e)^{i} (1-e)^{n-i}$$

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

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

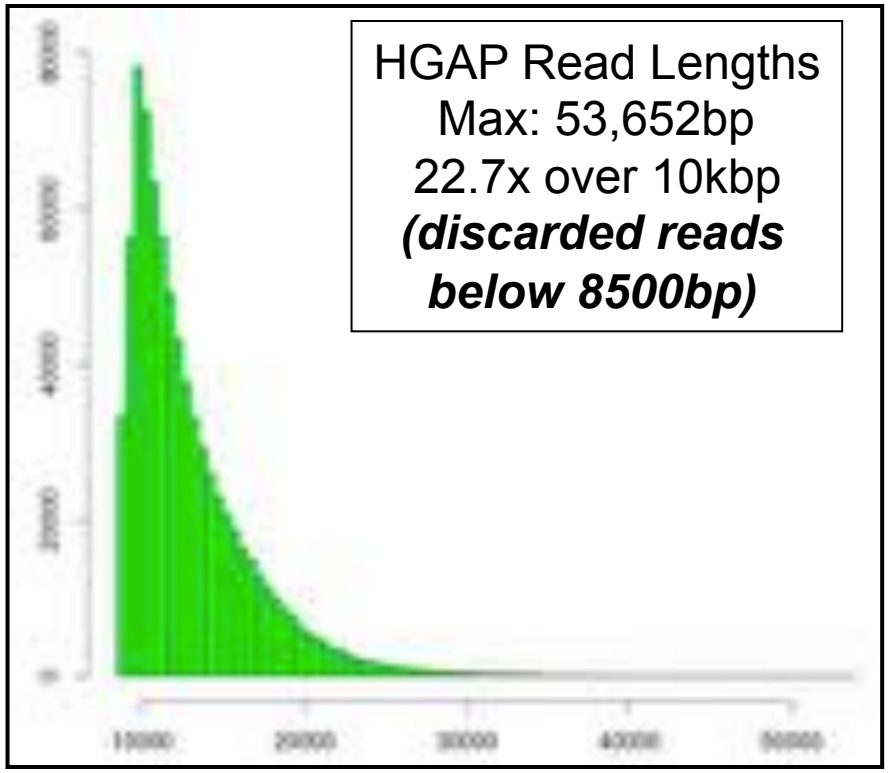


Over 118x coverage using P5-C3 long read sequencing

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

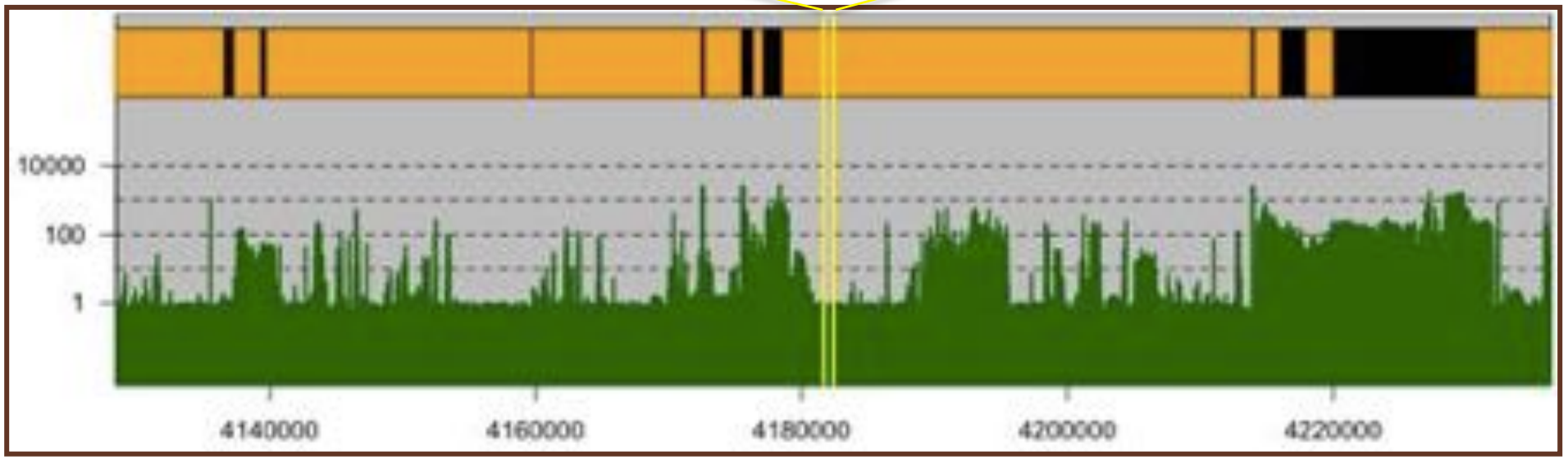
```
Sanger      …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
Illumina    …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
PacBio      …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
```

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

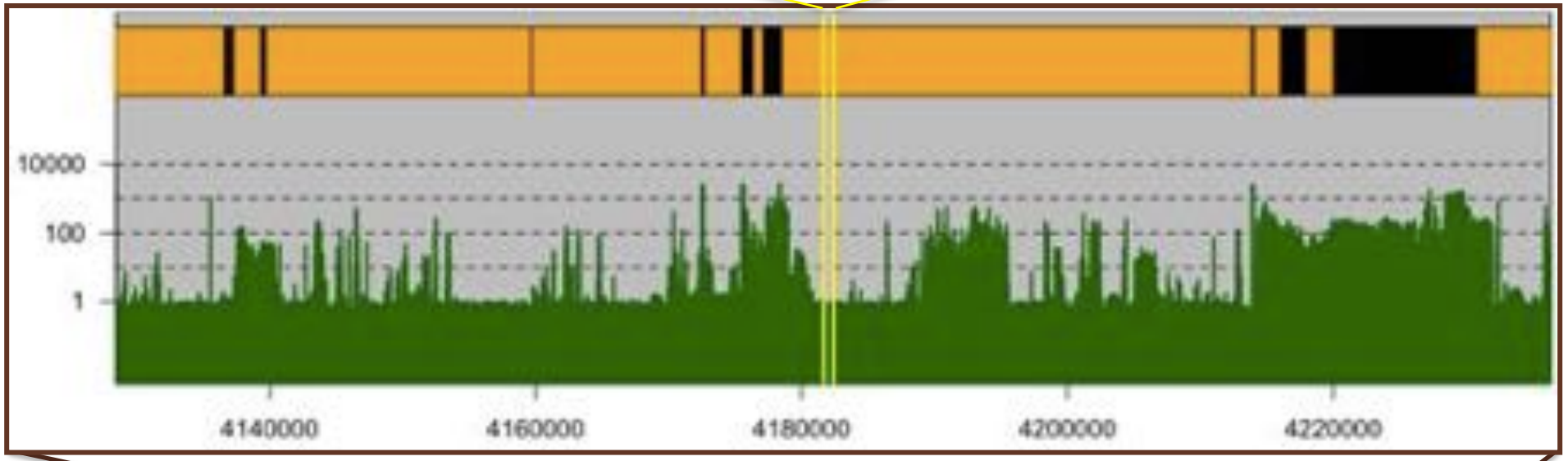
```
Sanger      …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
Illumina    …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
PacBio      …ACCCTGATATTCTGAGTTACAAGGCATTCAGCTACTGCTTGCCCACTGACGAGACC…
```

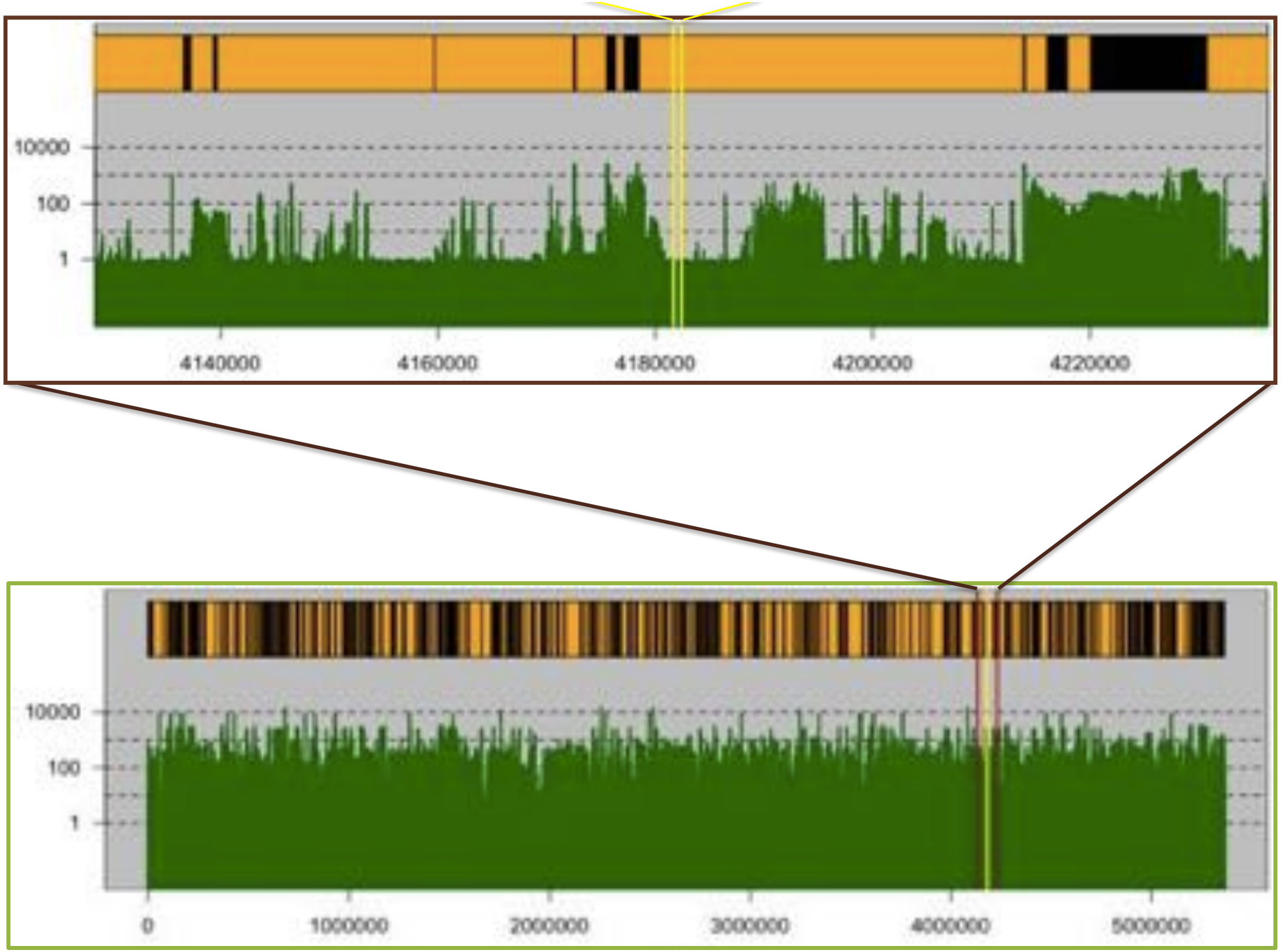


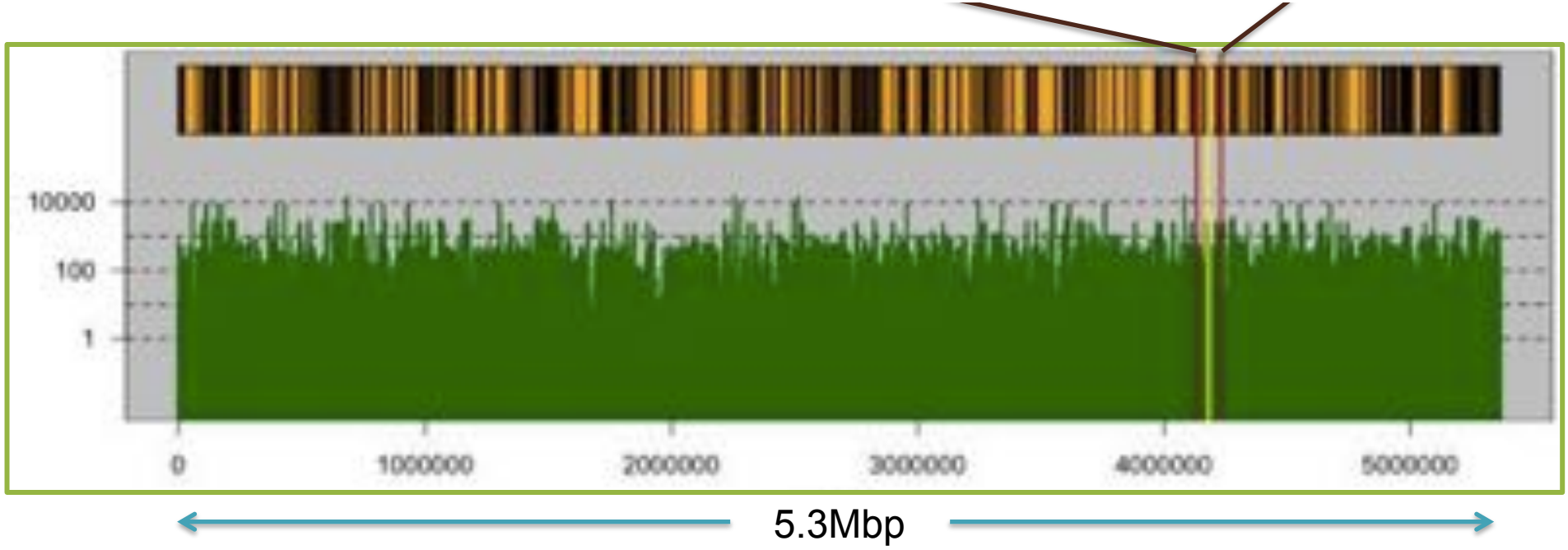
100kbp

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:

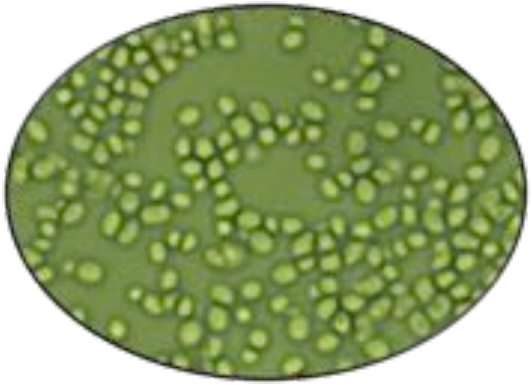
- Over 20 Megabases of additional sequence
  - Extremely high sequence identity (>99.9%)
  - Thousands of gaps filled, hundreds of mis-assemblies corrected
- Complete gene models, promoter regions for nearly every gene
  - True representation of transposons and other complex features
- Opportunities for studying large scale chromosome evolution
  - Largest contigs approach complete chromosome arms

Current Collaborations

Pineapple
UIUC

Human
CSHL/OICR

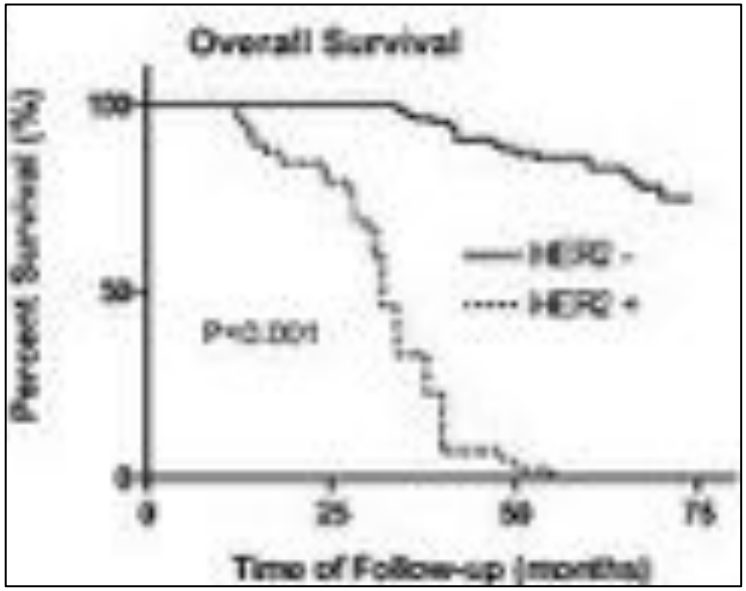
Asian Sea Bass
Temasek Life Sciences



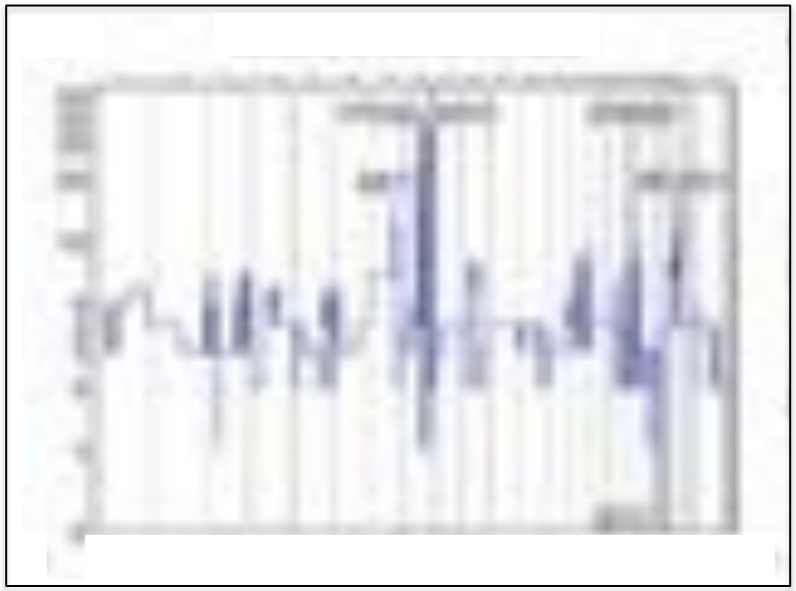
C. glabrata
JHU

T. vaginalis
NYU

Long Read Sequencing of SK-BR-3



(Wen-Sheng et al, 2009)



(Navin et al, 2011)

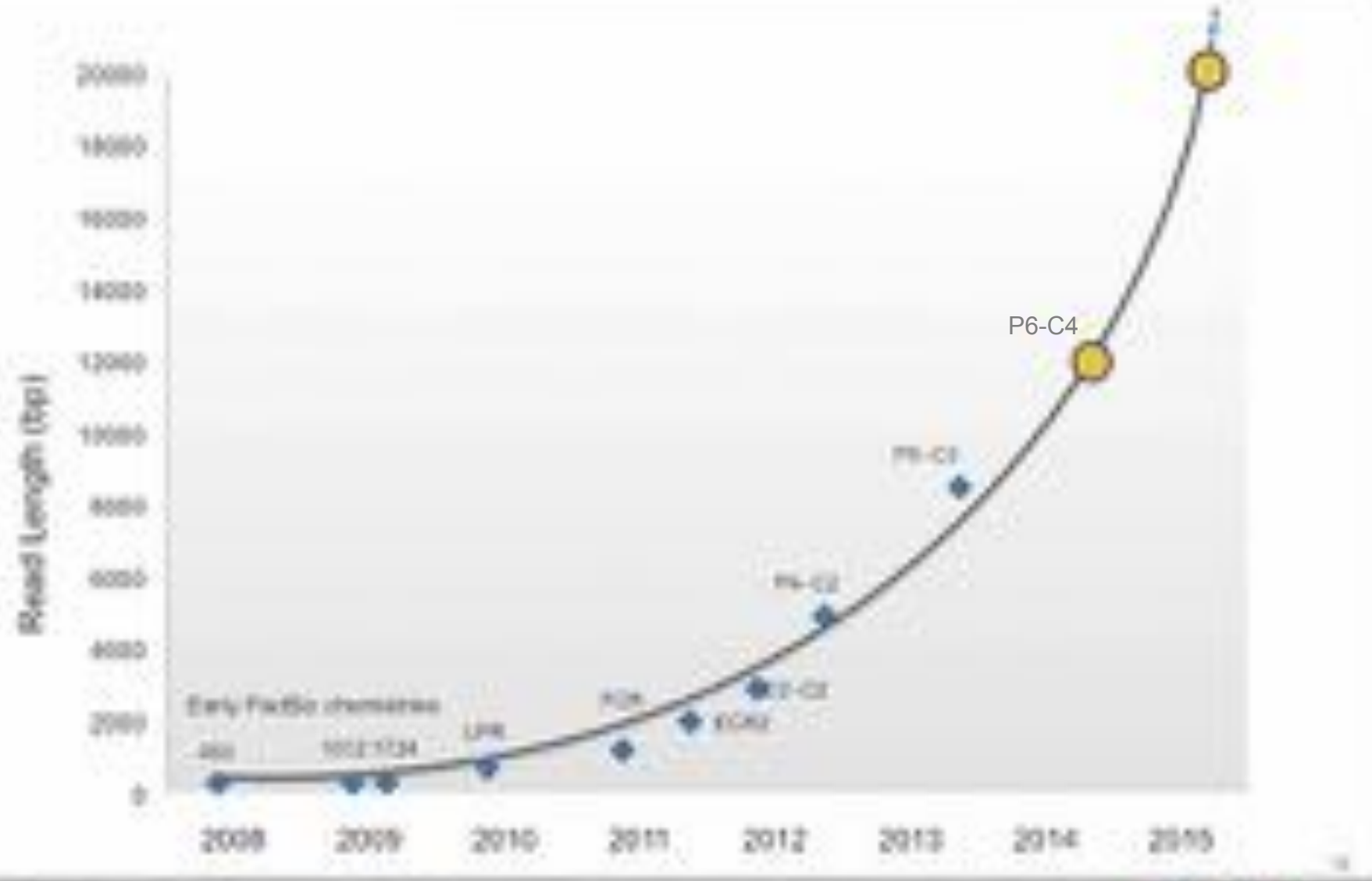
Long read PacBio sequencing of SK-BR-3 breast cancer cell line

- Her2+ breast cancer is one of the most deadly forms of the disease
  - SK-BR-3 is one of the most important models, known to have widespread CNVs

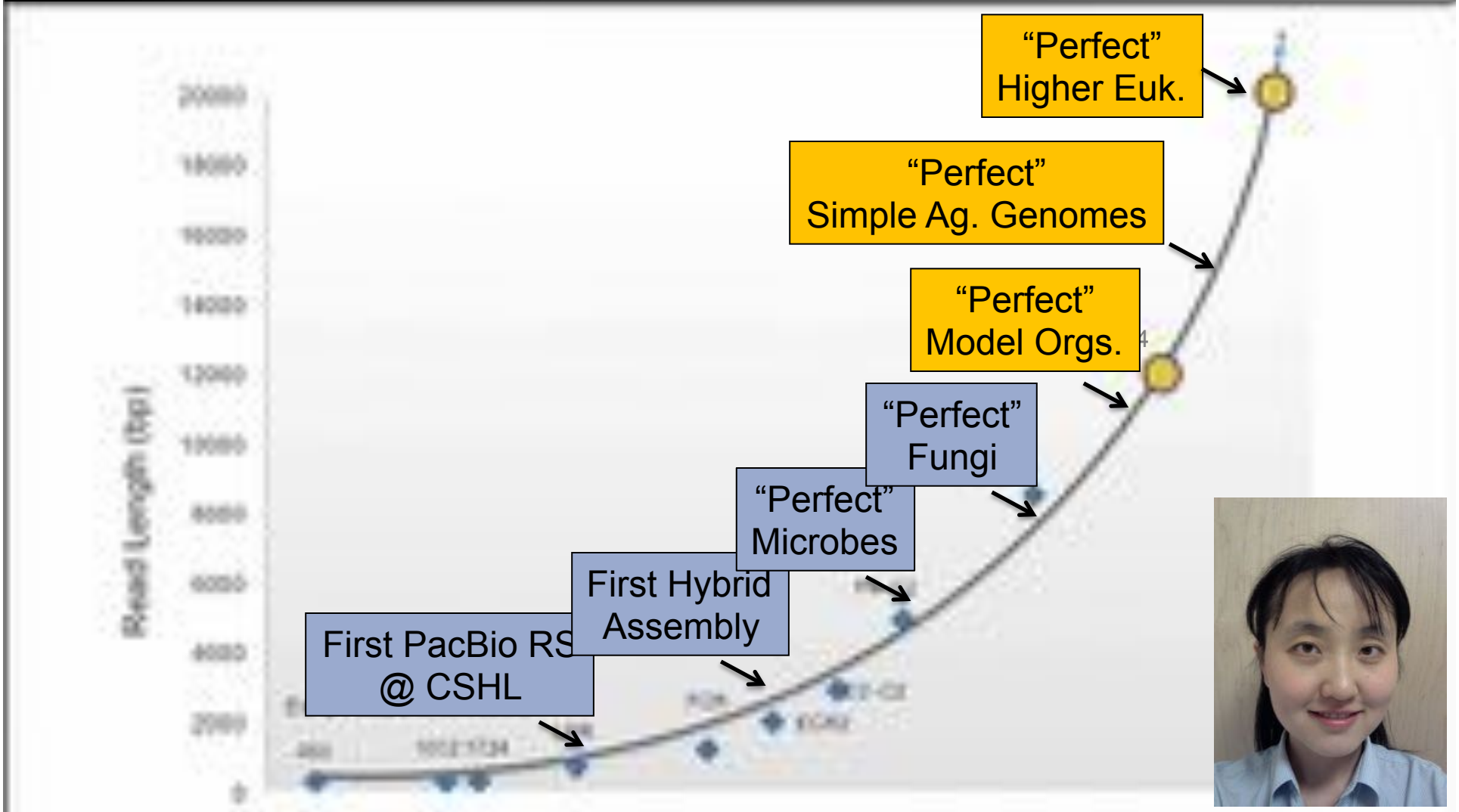
- Currently have 60x coverage with long read PacBio sequencing (mean: ~10kbp)
  - Discovered a complex series of nested duplications and translocations around HER2
  - Currently analyzing breakpoints in an attempt to infer the mutation history

In collaboration with McCombie (CSHL) and McPherson (OICR) labs

PacBio® Advances in Read Length



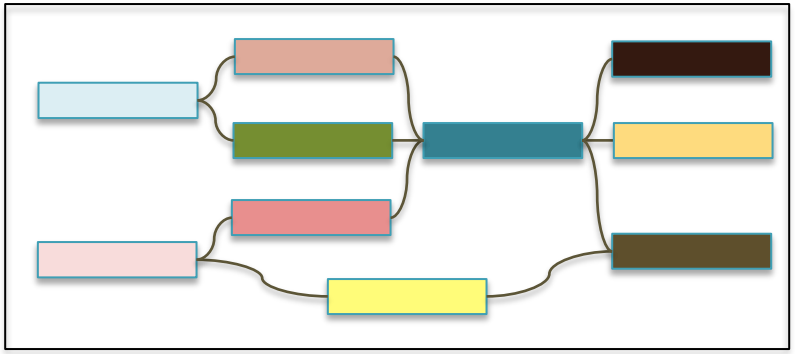
Advances in Assembly



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

Tomorrow at Noon

Oxford Nanopore Sequencing



Pan-Genomics

Oxford Nanopore Sequencing and de novo Assembly of a Eukaryotic Genome
Goodwin, S *et al.* (2015) bioRxiv doi: http://dx.doi.org/10.1101/013490

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

Genome Structure & Function

1. **Structure: Sequencing and Assembly**

   Long Read Single Molecule Sequencing

2. **Function: Disease Analytics**

   The role of indels in autism spectrum disorders

Genetic Basis of Autism Spectrum Disorders

Complex disorders of brain development

- Characterized by difficulties in social interaction, verbal and nonverbal communication and repetitive behaviors.
- Have their roots in very early brain development, and the most obvious signs of autism and symptoms of autism tend to emerge between 2 and 3 years of age.

U.S. CDC identify around 1 in 68 American children as on the autism spectrum

- Ten-fold increase in prevalence in 40 years, only partly explained by improved diagnosis and awareness.
- Studies also show that autism is four to five times more common among boys than girls.
- Specific causes remain elusive

What is Autism?
http://www.autismspeaks.org/what-autism

Searching for the genetics behind human disorders and plant phenotypes

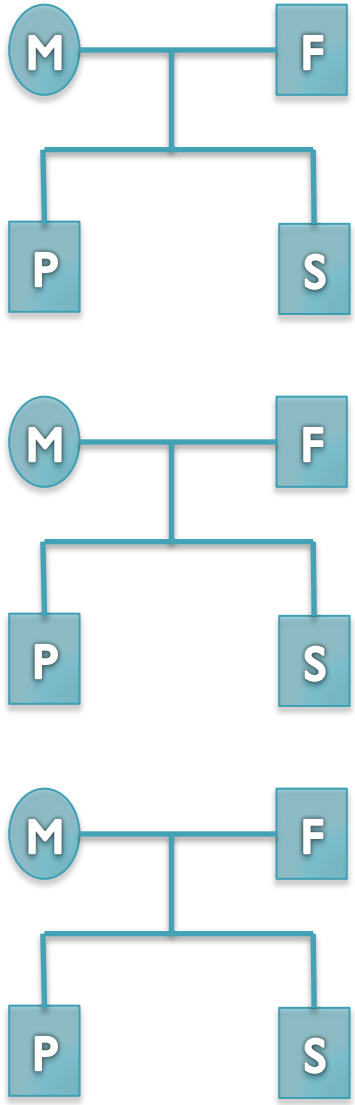
Search Strategy

- Currently uses WGS or WES short read resequencing for economic reasons

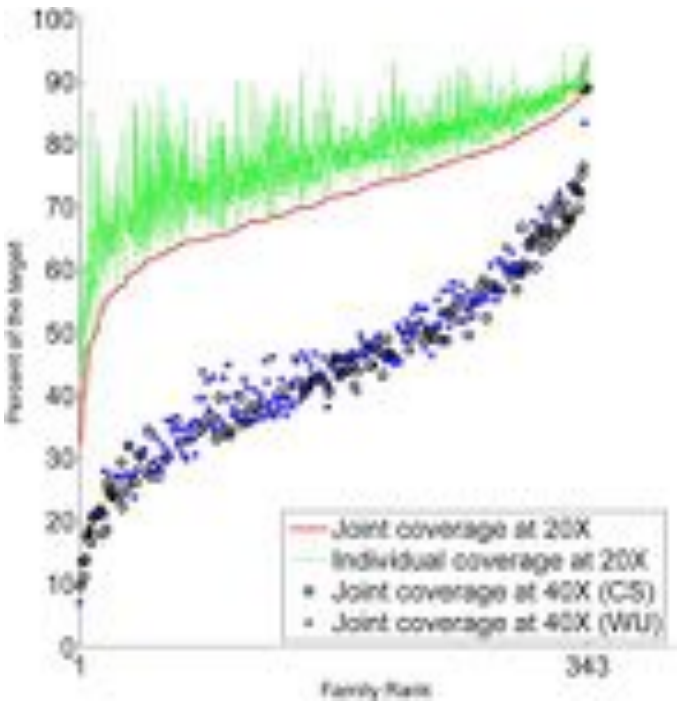
- Collaborate with Lyon, McCombie, Tuveson, and Wigler labs to examine the genetic basis of cancer, ASD, and other psychiatric disorders

- Also collaborating with the Lippman, Ware, and Gingeras labs to study high value crops

Are there any genetic variants present in affected individuals, that are not present or are present at a substantially reduced rate in their relatives?



Exome sequencing of the SSC



The year 2012 was an exciting year for autism genetics

- 3 reports of ~600 families from the Simons Simplex Collection (parents plus one child with autism and one non-autistic sibling)
- All attempted to find mutations enriched in the autistic children
- ***All used poor or no tools for indels:***
  - Iossifov (343 families) and O'Roak (50 families) used GATK UnifiedGenotype
  - Sanders (200 families) didn't attempt

De novo gene disruptions in children on the autism spectrum
Iossifov *et al.* (2012) *Neuron*. 74:2, 285-299.

De novo mutations revealed by whole-exome sequencing are strongly associated with autism
Sanders *et al.* (2012) *Nature*. 485, 237–241.

Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations
O'Roak *et al.* (2012) *Nature*. 485, 246–250.

Scalpel: Haplotype Microassembly

DNA sequence **micro-assembly** pipeline for accurate detection and validation of *de novo* mutations (SNPs, indels) within exome-capture data.

Features

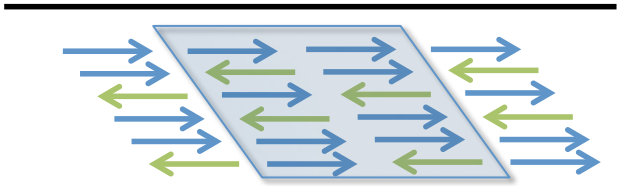
1. Combine mapping and assembly
2. Exhaustive search of haplotypes
3. De novo mutations

NRXN1 *de novo* SNP
(auSSC12501 chr2:50724605)

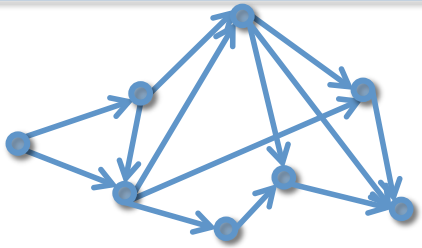
Accurate de novo and transmitted indel detection in exome-capture data using microassembly.
Narzisi, G, O'Rawe, JA, Iossifov, I, Fang, H, Lee, YH, Wang, Z, Wu, Y, Lyon, G, Wigler, M, Schatz MC
*Nature Methods (*2014) doi:10.1038/nmeth.3069

Scalpel Algorithm

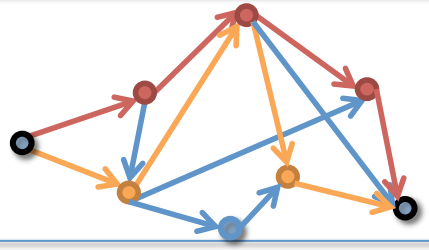
Extract reads mapping within the exon including (1) well-mapped reads, (2) soft-clipped reads, and (3) anchored pairs



Decompose reads into overlapping *k*-mers and construct de Bruijn graph from the reads



Find end-to-end haplotype paths spanning the region



Align assembled sequences to reference to detect mutations

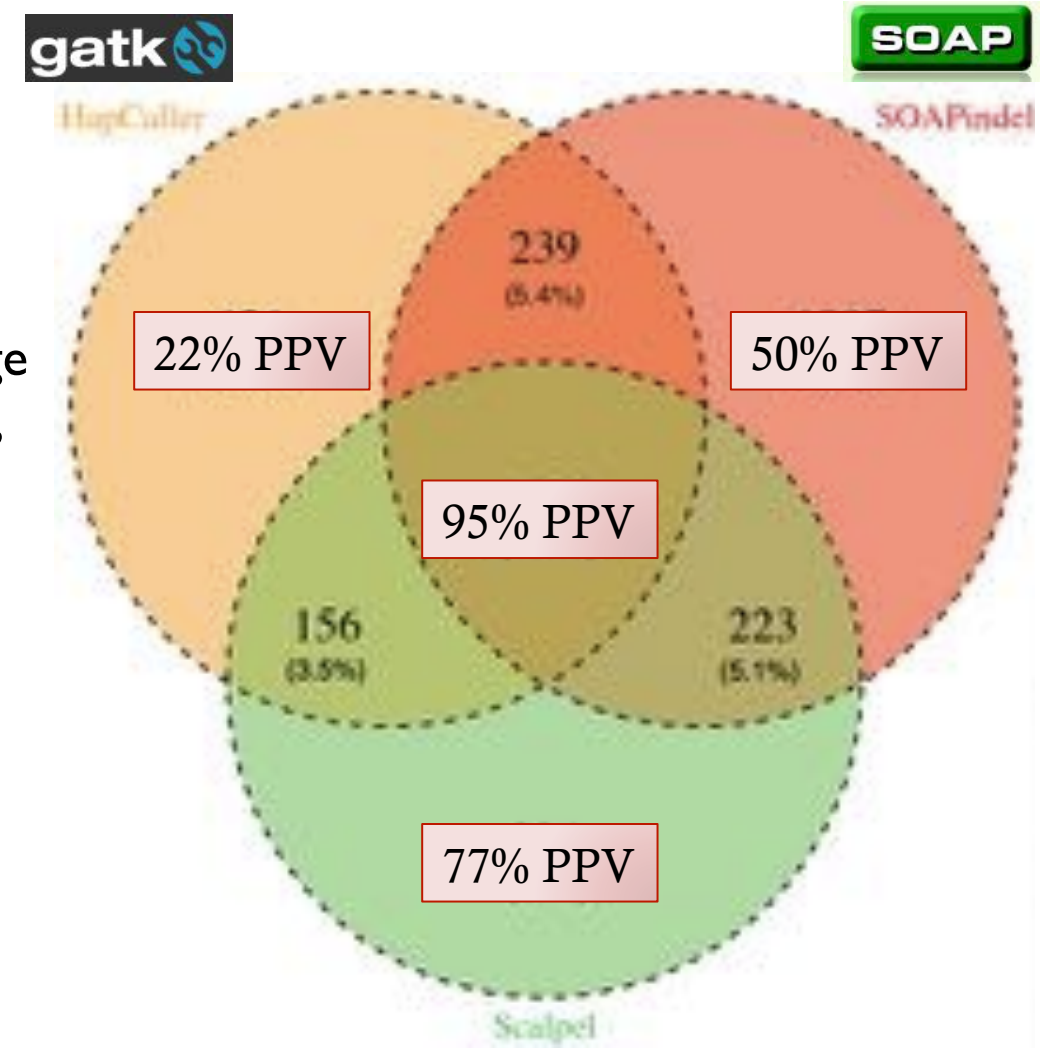
Experimental Analysis & Validation

Selected one deep coverage exome for deep analysis

- Individual was diagnosed with ADHD and turrets syndrome
- 80% of the target at >20x coverage
- Evaluated with Scalpel, SOAPindel, and GATK Haplotype Caller

1000 indels selected for validation

- 200 Scalpel
- 200 GATK Haplotype Caller
- 200 SOAPindel
- 200 within the intersection
- 200 long indels (>30bp)



CSH

Refined indel analysis

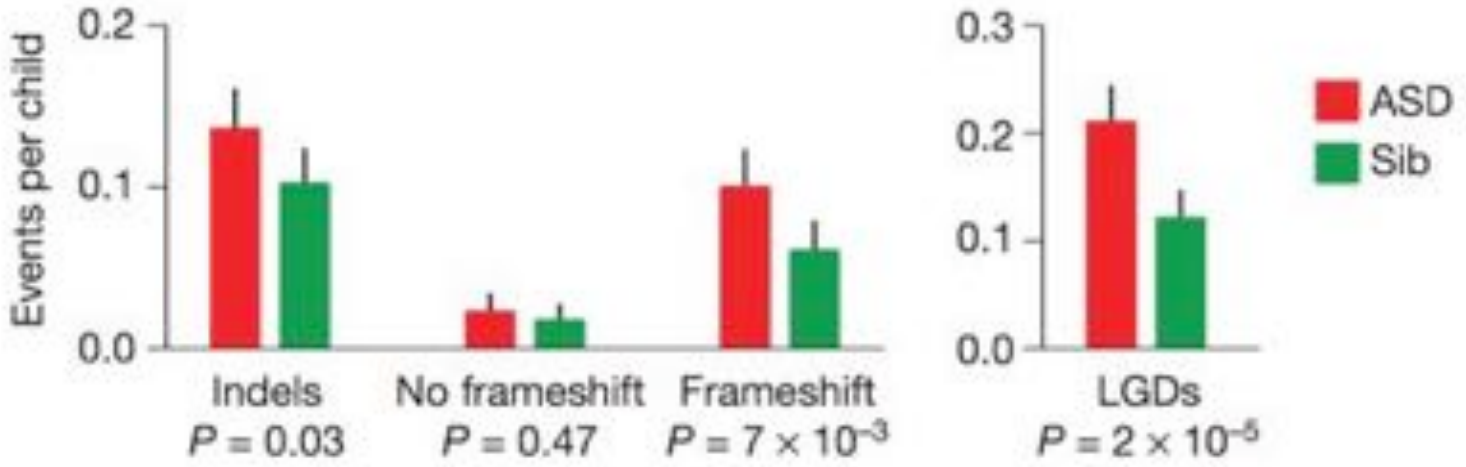
Examine sources of indel errors

- Experimental validation of indels called from 30x whole genome vs. 110x whole exome
- Most of the errors due to microsatellite slippage introduced during exome capture, also missing most long indels
- Recommend PCR-free WGS if at all possible

	All INDELs	Valid	PPV	INDELs >5bp	Valid (>5bp)	PPV (>5bp)
Intersection	160	152	95.0%	18	18	100%
WGS	145	122	84.1%	33	25	75.8%
WES	161	91	56.5%	1	1	100%

Reducing INDEL calling errors in whole-genome and exome sequencing data
Fang, H, Wu, Y, Narzisi, G, O'Rawe, JA, Jimenez Barrón LT, Rosenbaum, J, Ronemus, M, Iossifov I, Schatz, MC§, Lyon, GL§
Genome Medicine. doi: 10.1186/s13073-014-0089-z

De novo Genetics of Autism



- In 2,500 family quads we see significant enrichment in de novo ***likely gene disruptions (LGDs)*** in the autistic kids
  - Overall rate basically 1:1
  - 2:1 enrichment in frameshift indels
  - Confirmed trends observed in previous studies, contributed dozens of new autism candidate genes.

The burden of de novo coding mutations in autism spectrum disorders.
Iossifov *et al* (2014) *Nature*. doi:10.1038/nature13908

What's next?

Giuseppe Narzisis

Somatic mutation detection

Coding and non-coding mutations in cancer and autism

Srividya "Sri" Ramakrishnan

DOE Systems Biology Knowledgebase

Worlds fastest -omics pipelines

Maria Nattestad

Hi-C Chromatin Interactions

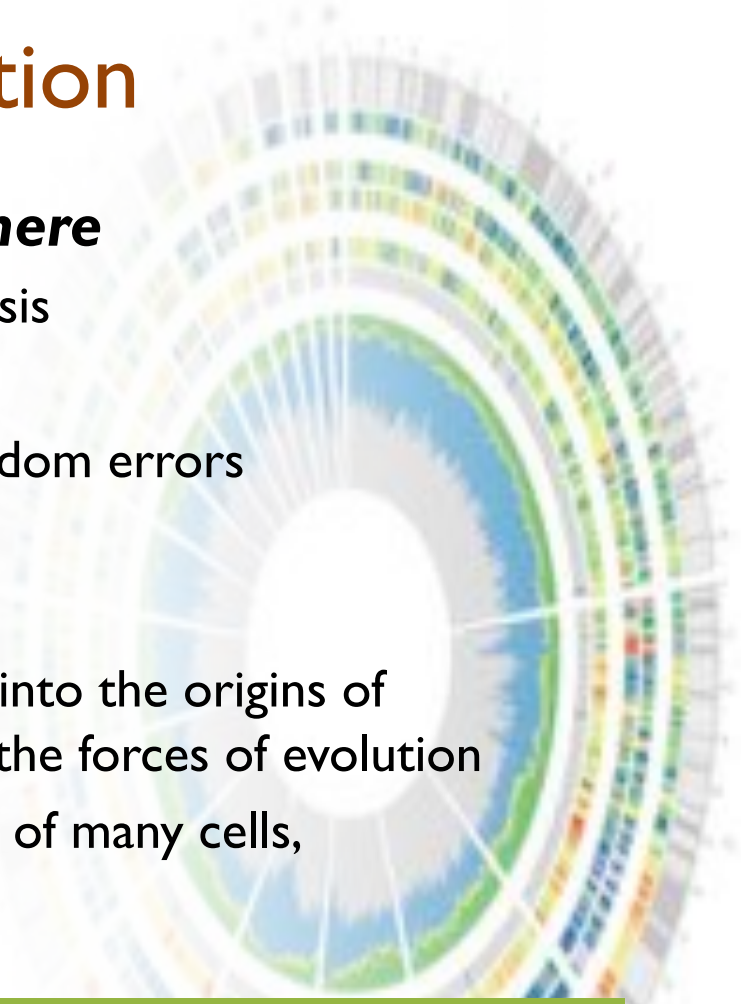
Plant Assembly & Analysis

Tyler Garvin

Single Cell CNV

Tumor and Somatic Heterogeneity

Understanding Genome Structure & Function



Reference quality genome assembly is here

- Use the longest possible reads for the analysis
- Don't fear the error rate
  - Coverage and algorithmics conquer random errors

Population analysis

- Large scale sequencing give us new insights into the origins of disease, the processes of development, and the forces of evolution
- See similar trends in the population analysis of many cells, integration of multiple assays

Also very interested in teaching the next generation of undergraduate and graduate students

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