# Algorithms for de novo genome assembly and disease analytics

Michael Schatz

March 11, 2014

University of Florida Genetics Institute

# Introductions

**Ke Jiang**

Transcriptomics and epigenetics

Tomato & Solanaceae

**Srividya "Sri" Ramakrishnan**

DOE Systems Biology Knowledgebase

Worlds fastest -omics pipelines

**Maria Nattestad**

Hi-C Chromatin Interactions

Plant Assembly & Analysis

**Tyler Garvin**

CNV analysis of single cells

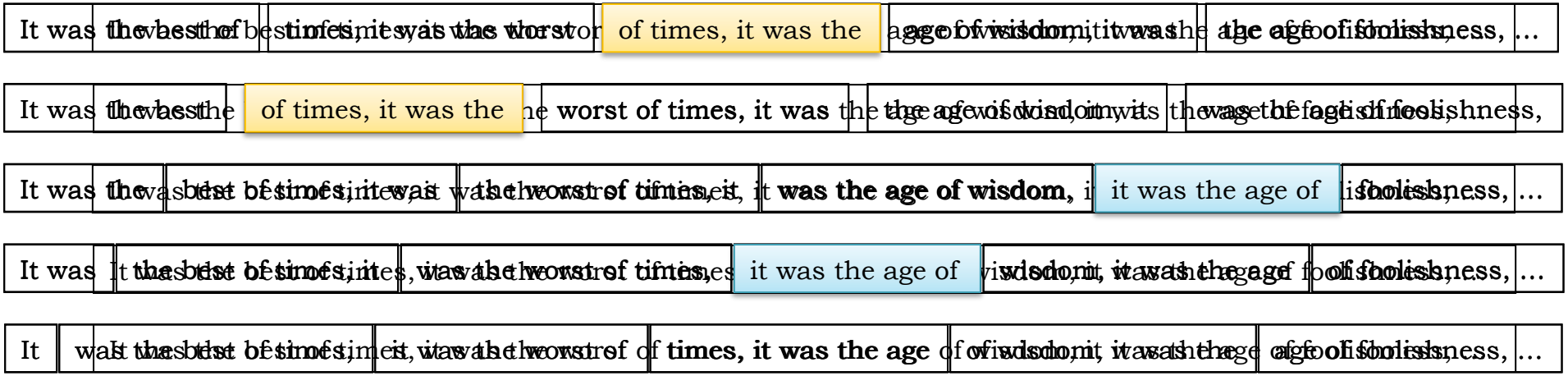
Breast & Prostate Cancer

# Outline

1. De novo assembly by analogy
2. Long Read Assembly
3. Disease Analytics

# Shredded Book Reconstruction

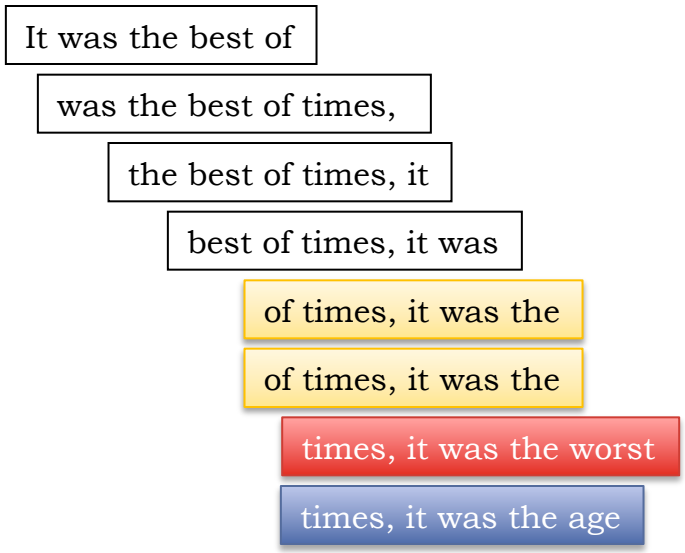
- Dickens accidentally shreds the first printing of A Tale of Two Cities
  - Text printed on 5 long spools




- 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



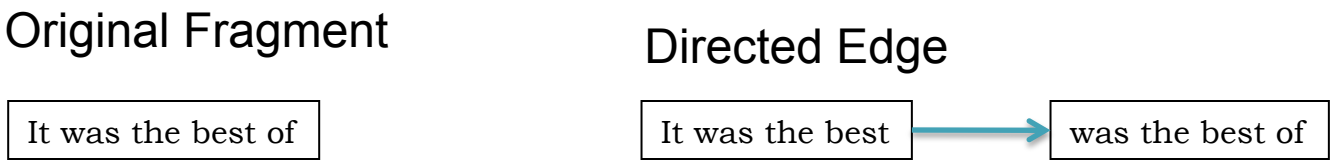

The repeated sequence make the correct reconstruction ambiguous

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

Model the assembly problem as a graph problem

# de Bruijn Graph Construction

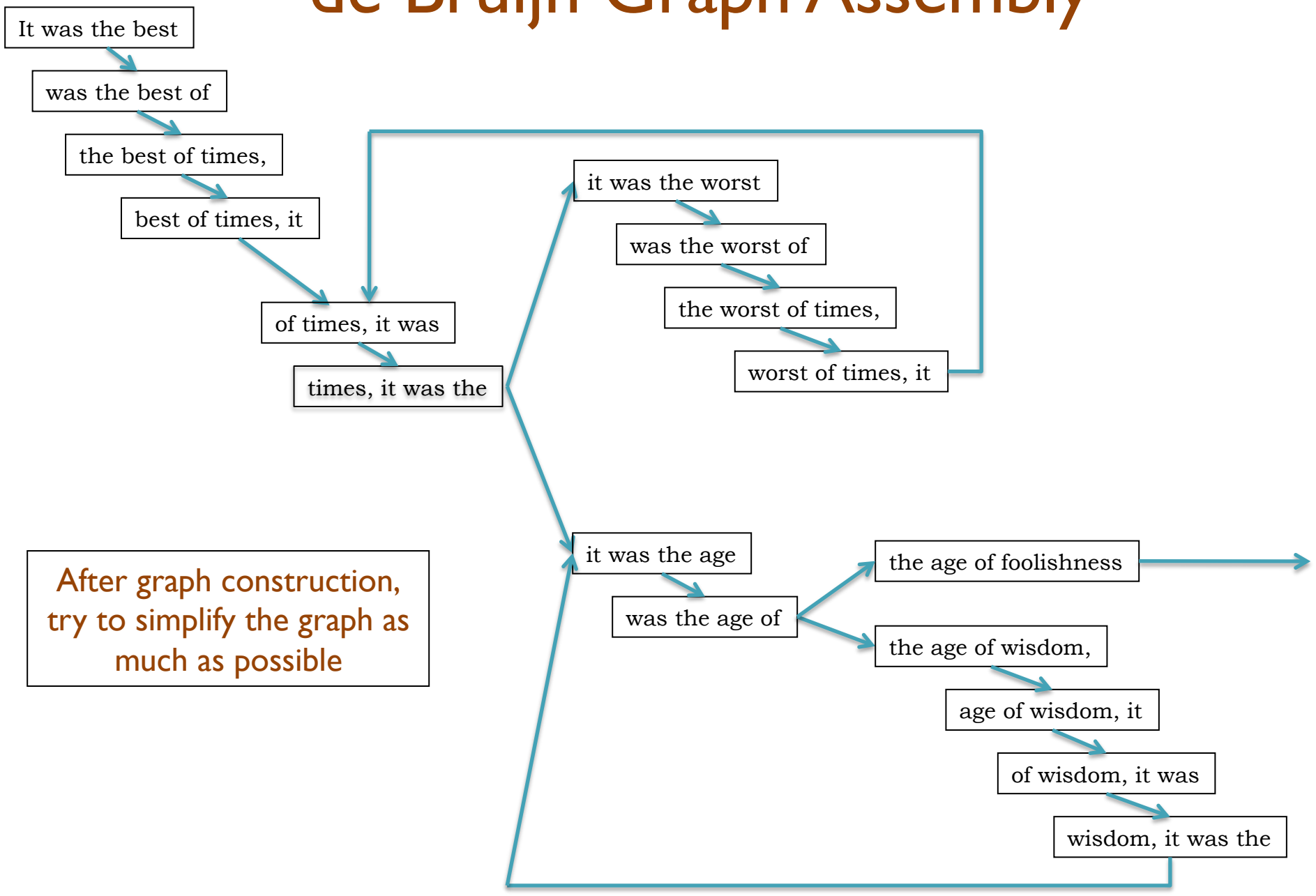
- $D_k = (V,E)$
  - V = All length-k subfragments (k < l)
  - E = Directed edges between consecutive subfragments
    - Nodes overlap by k-1 words




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

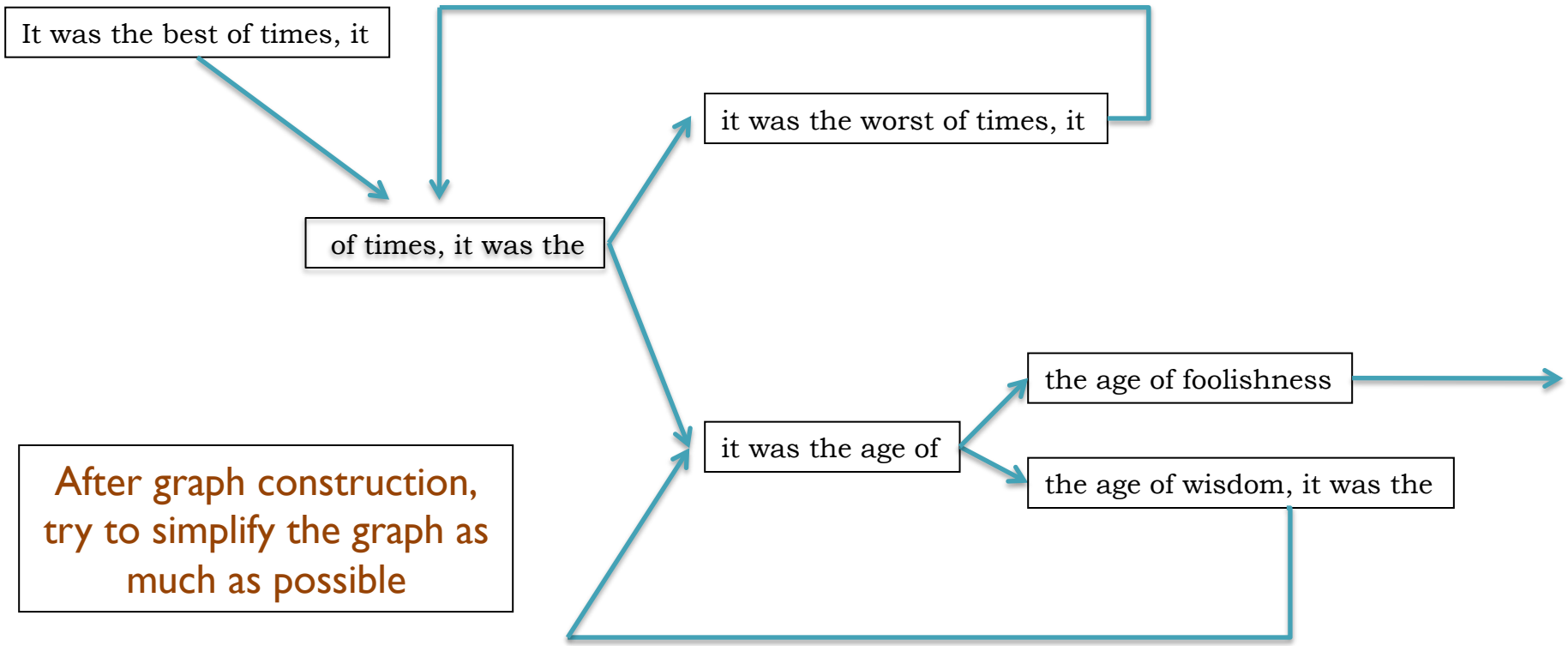
de Bruijn, 1946
Idury and Waterman, 1995
Pevzner, Tang, Waterman, 2001

# de Bruijn Graph Assembly



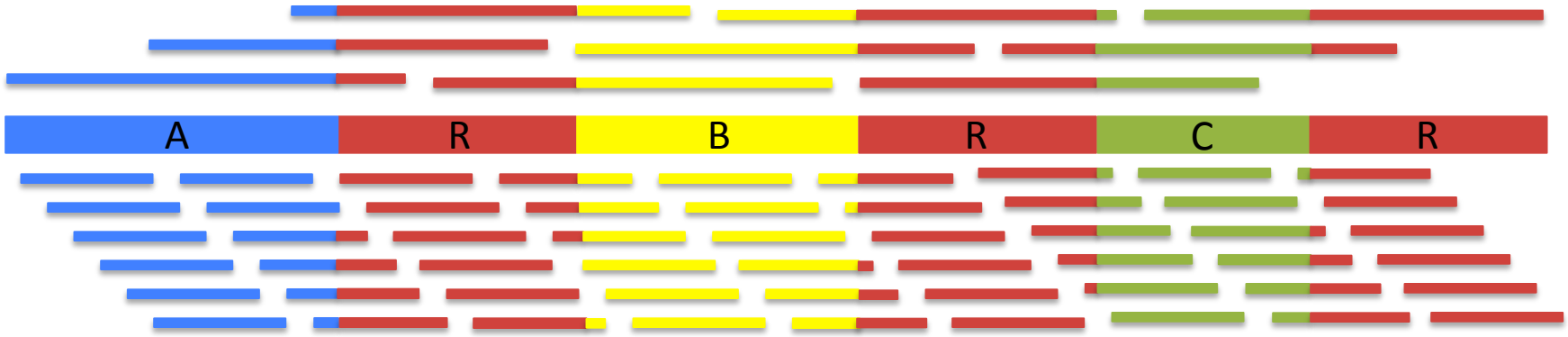

After graph construction, try to simplify the graph as much as possible

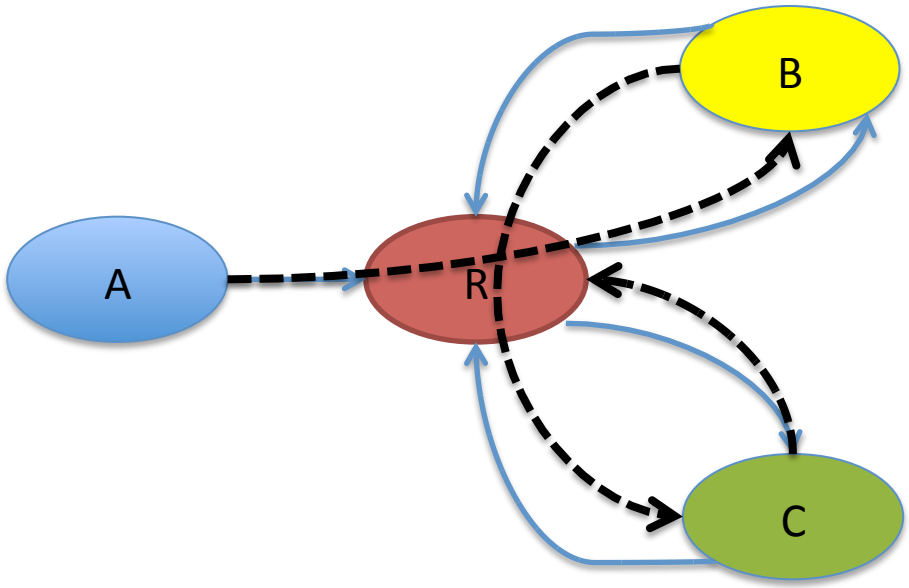
# de Bruijn Graph Assembly



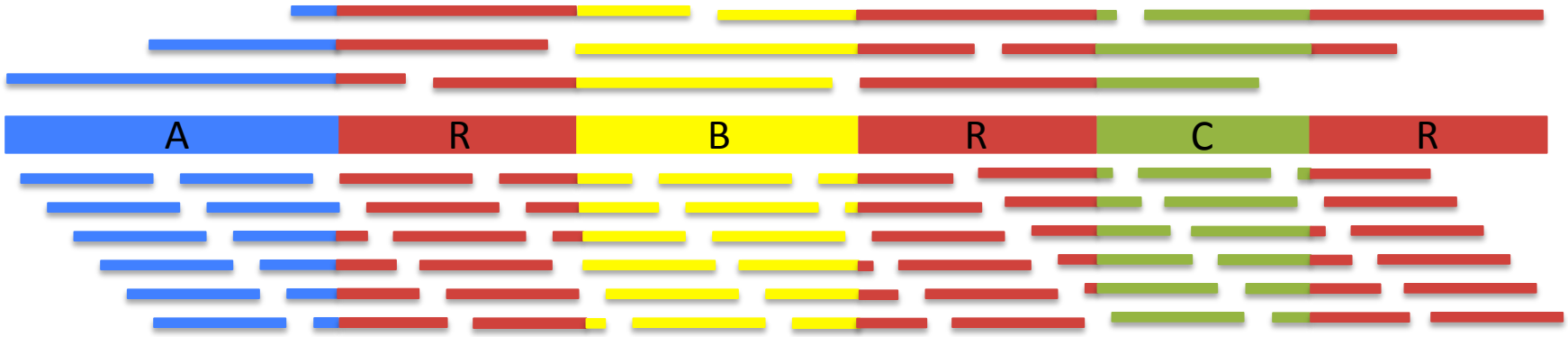

After graph construction, try to simplify the graph as much as possible

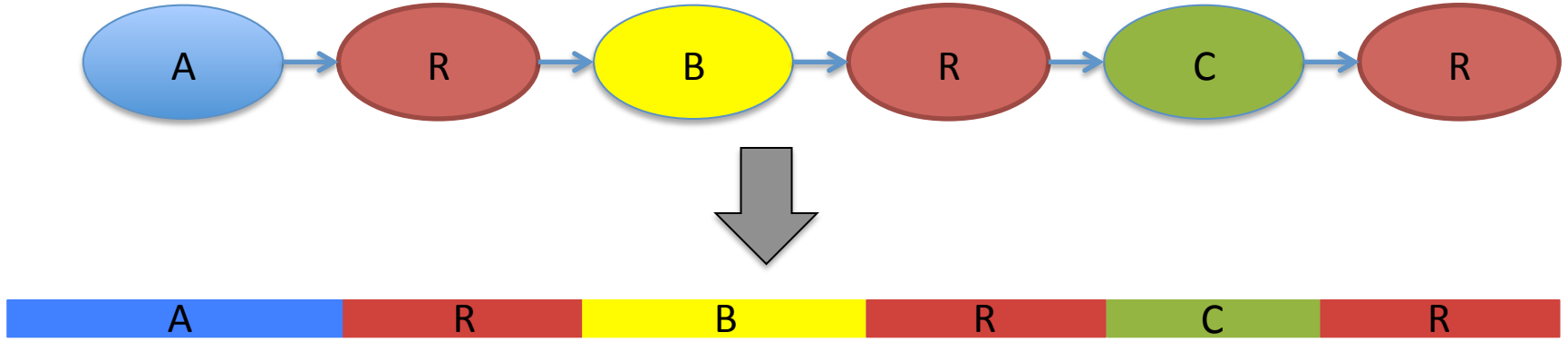
# Assembly Complexity



# Assembly Complexity

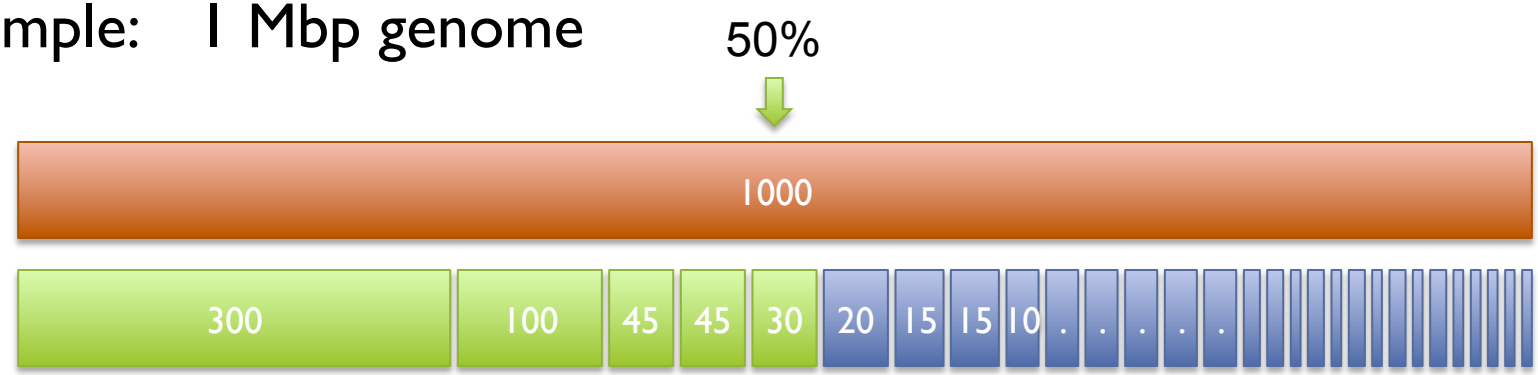


# N50 size

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)

Note:

A "good" N50 size is a moving target relative to other recent publications. 10-20kbp contig N50 is currently a typical value for most "simple" genomes.

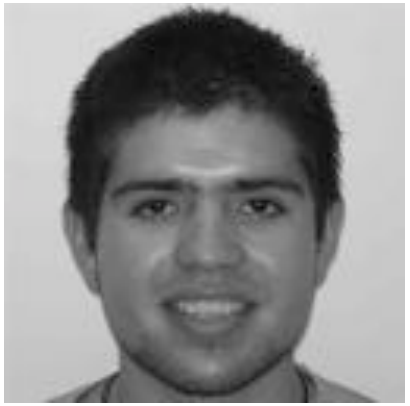
# Outline

1. De novo assembly by analogy
2. Long read assembly
3. Disease Analytics

# Population structure of *Oryza sativa*

**Conclusions**

- 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
  - Detailed analysis of agriculturally important loci
- Assemblies fragmented at (high copy) repeats
  - Missing regions have mean kmer coverage >10,000x
  - Difficult to identify full length gene models and regulatory features

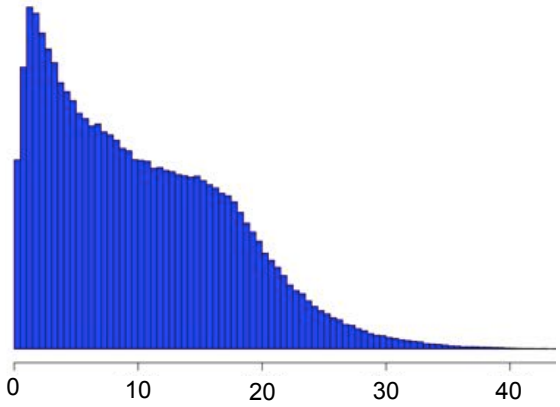


**New whole genome de novo assemblies of three divergent strains of rice (*O. sativa*) documents novel gene space of *aus* and *indica***

Schatz, MC, Maron, L, Stein, J, Ware, DW, McCouch, S, McCombie WR *et al* (2014) *In preparation*

# Long Read Sequencing Technology

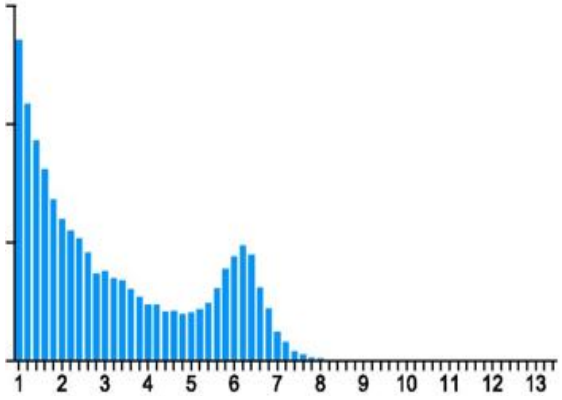
## PacBio RS II




CSHL/PacBio

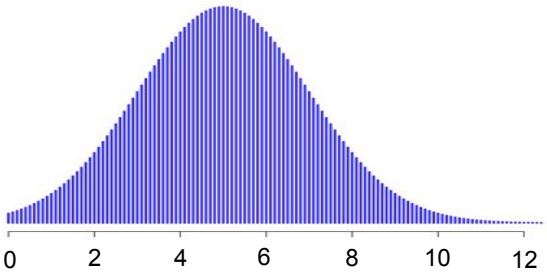
## Moleculo

illumina
moleculo




(Voskoboynik et al. 2013)

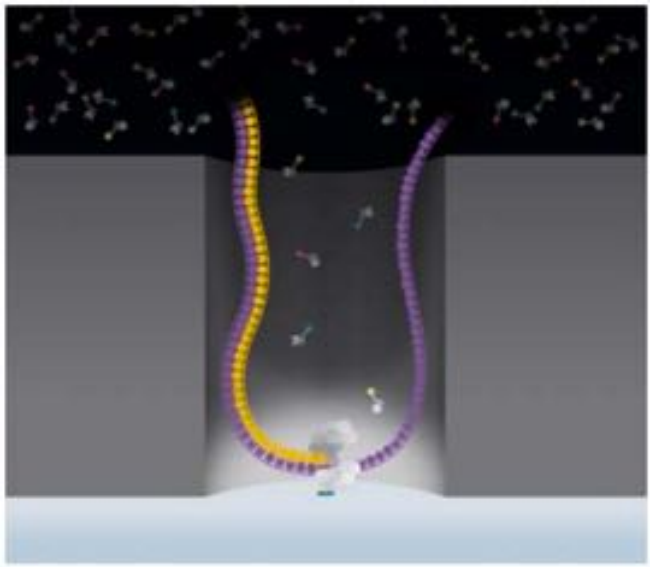
## Oxford Nanopore

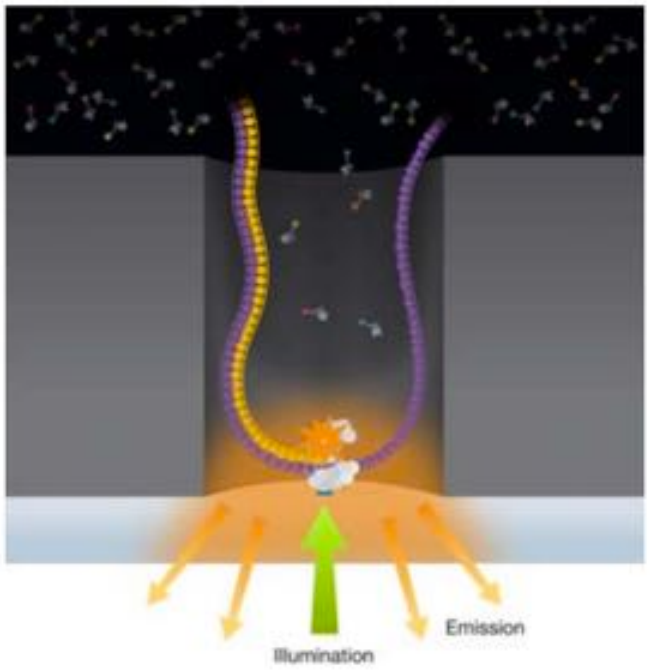


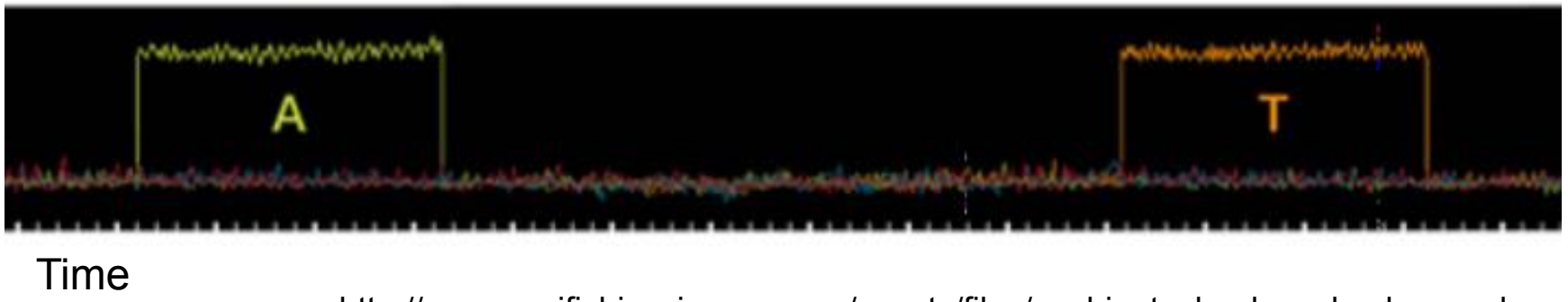

Broad/OxNano @ AGBT ***

# SMRT Sequencing

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

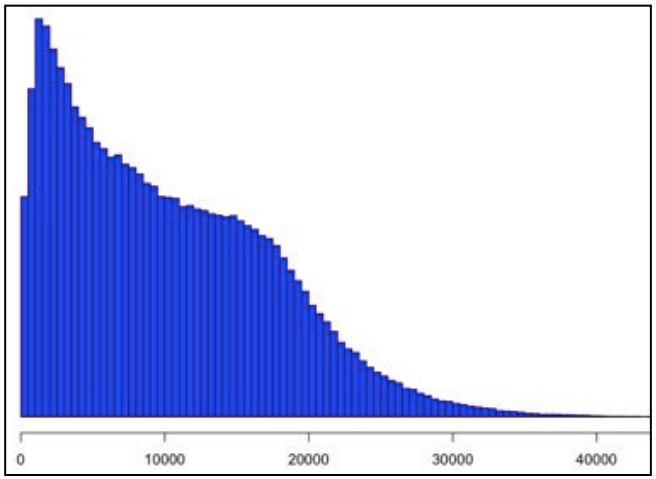





http://www.pacificbiosciences.com/assets/files/pacbio_technology_backgrounder.pdf

# SMRT Sequencing Data

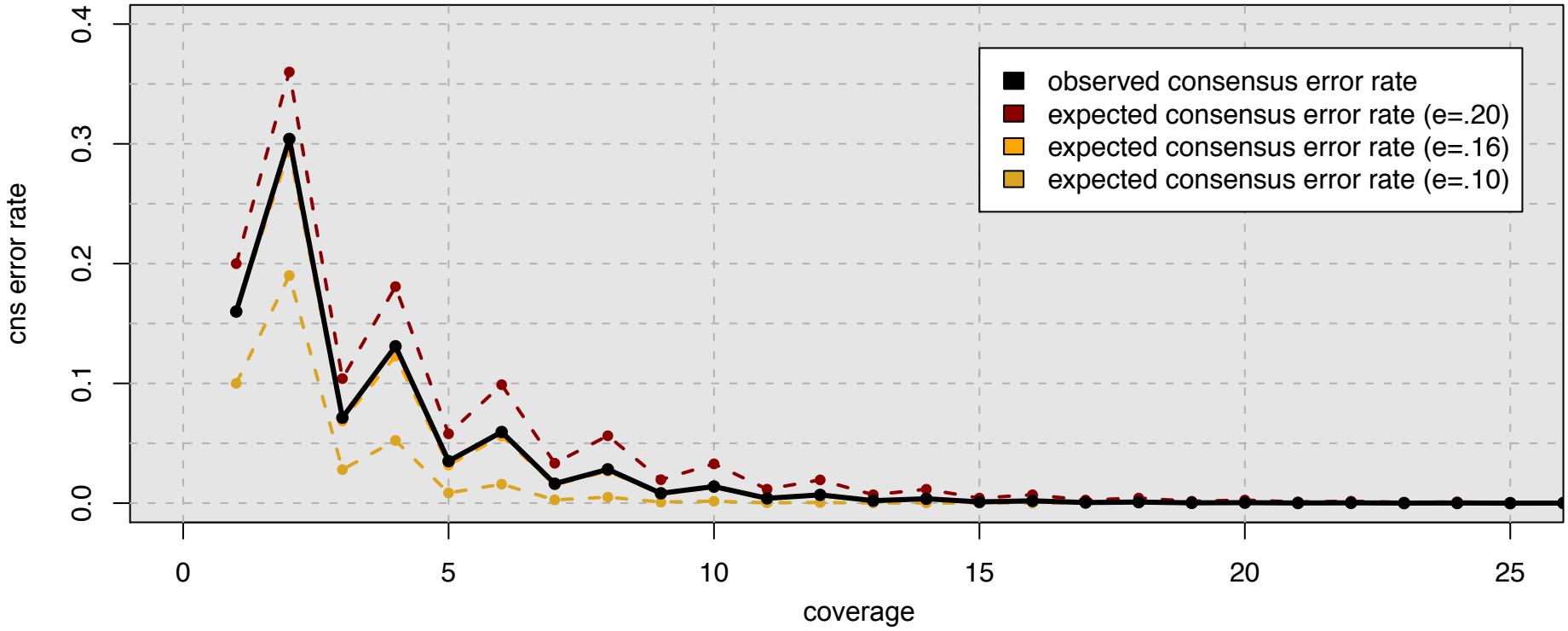



| | |
|---|---|
| Match | 83.7% |
| Insertions | 11.5% |
| Deletions | 3.4% |
| Mismatch | 1.4% |

```
TTGTAAGCAGTTGAAAACTATGTGTGGATTTAGAATAAAGAACATGAAAG
|||||||||||||||||||||||||| ||||||| ||||||||||||| |||
TTGTAAGCAGTTGAAAACTATGTGT-GATTTAG-ATAAAGAACATGGAAG

ATTATAAA-CAGTTGATCCATT-AGAAGA-AAACGCAAAAGGCGGCTAGG
| |||||| |||||||||||||| |||| | |||||| |||||| |||||||
A-TATAAATCAGTTGATCCATTAAGAA-AGAAACGC-AAAGGC-GCTAGG

CAACCTTGAATGTAATCGCACTTGAAGAACAAGATTTTATTCCGCGCCCG
| |||||| |||| ||  ||||||||||||||||||||||||||||||||
C-ACCTTG-ATGT-AT--CACTTGAAGAACAAGATTTTATTCCGCGCCCG

TAACGAATCAAGATTCTGAAAACACAT-ATAACAACCTCCAAAA-CACAA
| ||||||| |||||||||||||| || ||    |||||||||| |||||
T-ACGAATC-AGATTCTGAAAACA-ATGAT----ACCTCCAAAAGCACAA

-AGGAGGGGAAAGGGGGGAATATCT-ATAAAAGATTACAAATTAGA-TGA
 ||||||   ||     |||||||| || |||||||||||||| || |||
GAGGAGG---AA-----GAATATCTGAT-AAAGATTACAAATT-GAGTGA

ACT-AATTCACAATA-AATAACACTTTTA-ACAGAATTGAT-GGAA-GTT
|||  ||||||||| | |||||||||||| ||| ||||||| |||| |||
ACTAAATTCACAA-ATAATAACACTTTTAGACAAAATTGATGGGAAGGTT

TCGGAGAGATCCAAAACAATGGGC-ATCGCCTTTGA-GTTAC-AATCAAA
|| ||||||||| ||||||| ||| |||| |||||| ||||| |||||||
TC-GAGAGATCC-AAACAAT-GGCGATCG-CTTTGACGTTACAAATCAAA

ATCCAGTGGAAAATATAATTTATGCAATCCAGGAACTTATTCACAATTAG
||||||| ||||||||  |||||| ||||| ||||||||||||||||||
ATCCAGT-GAAAATATA--TTATGC-ATCCA-GAACTTATTCACAATTAG
```

Sample of 100k reads aligned with BLASR requiring >100bp alignment

# 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

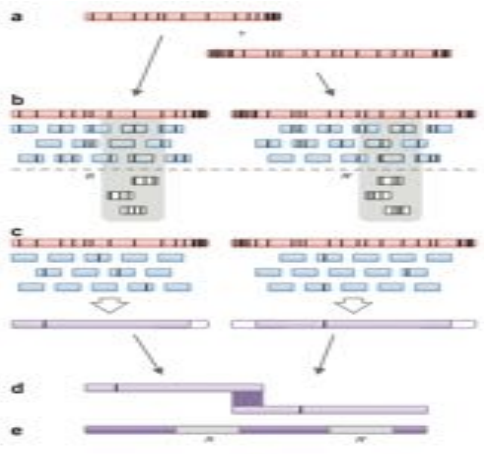
# PacBio Assembly Algorithms

## PBJelly


**Gap Filling and Assembly Upgrade**

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

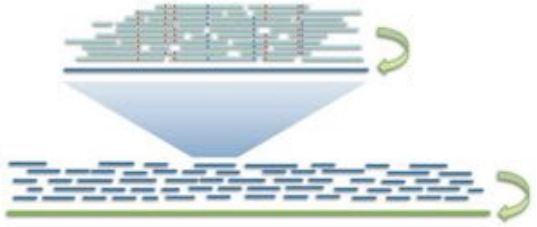
## PacBioToCA & ECTools




**Hybrid/PB-only Error Correction**

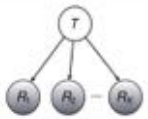
Koren, Schatz, *et al* (2012)
*Nature Biotechnology*. 30:693–700

## 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)*

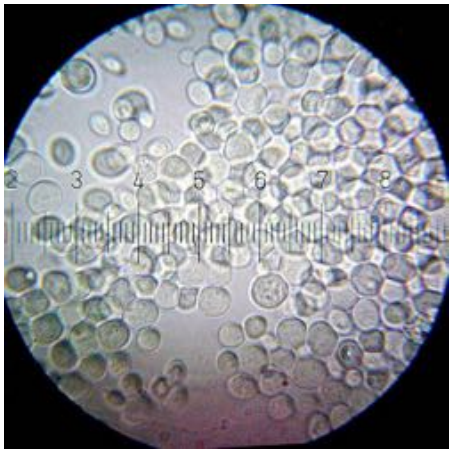
| | Initial Assembly | Quiver Consensus |
|---|---|---|
| QV | 43.4 | 54.5 |
| Accuracy | 99.99540% | 99.99964% |
| Differences | 141 | 11 |

**PB-only Correction & Polishing**

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

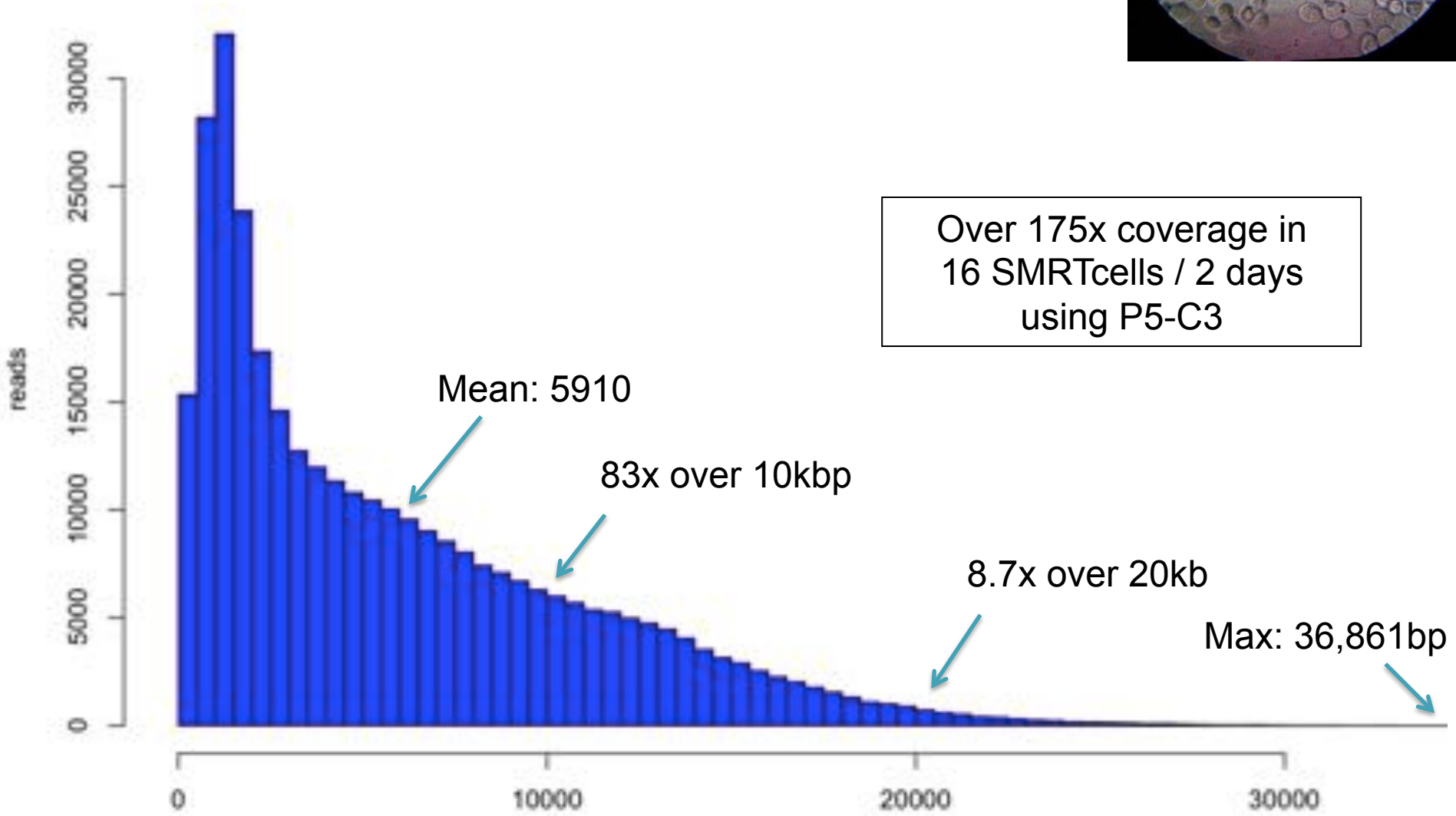
< 5x PacBio Coverage > 50x

# S. cerevisiae W303



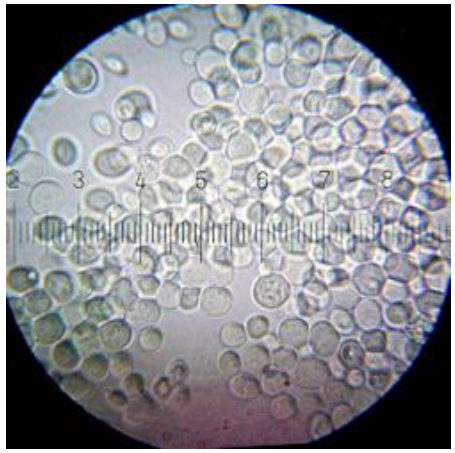
PacBio RS II sequencing at CSHL by Dick McCombie

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




Over 175x coverage in
16 SMRTcells / 2 days
using P5-C3

# S. cerevisiae W303

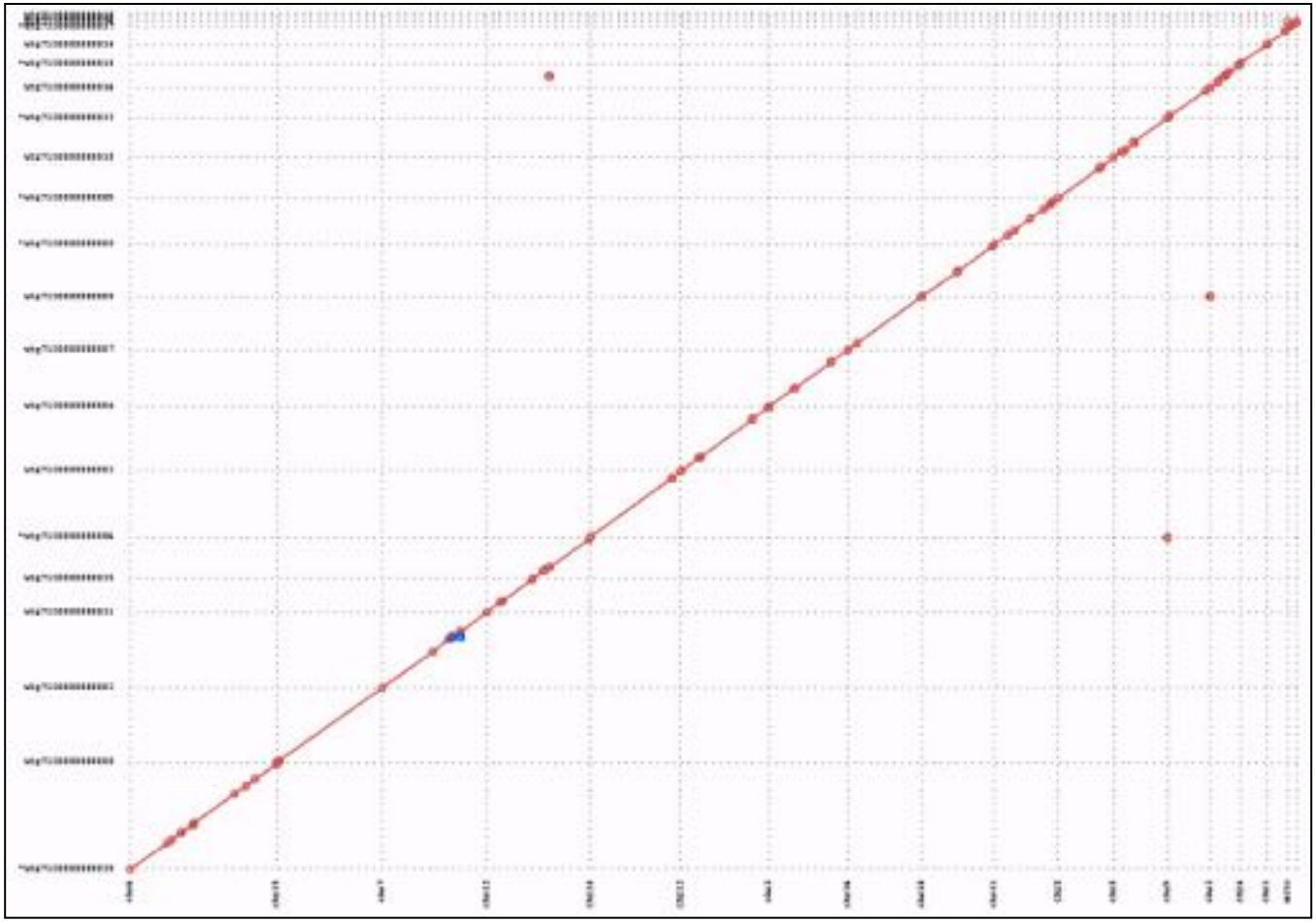


S288C Reference sequence

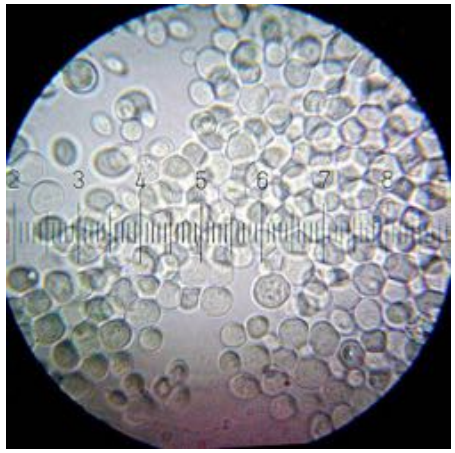
- 12.1Mbp; 16 chromo + mitochondria; N50: 924kbp

PacBio assembly using HGAP + Celera Assembler

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



# S. cerevisiae W303



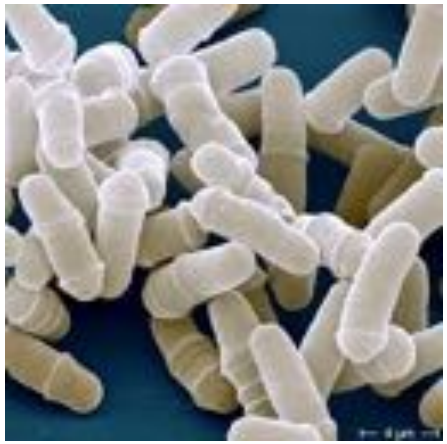
S288C Reference sequence

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

PacBio assembly using HGAP + Celera Assembler

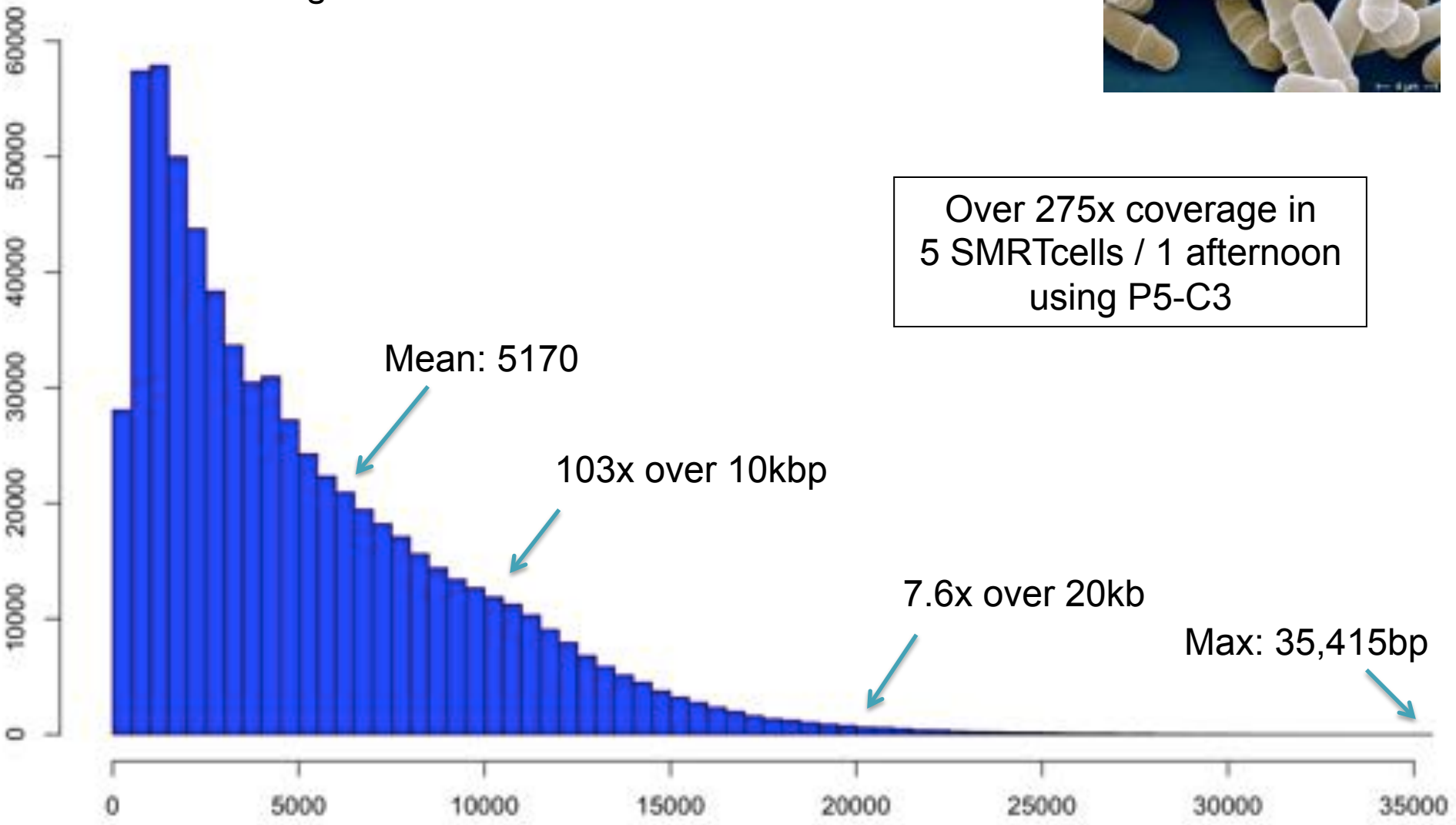
- 12.4Mbp; 21 non-redundant contigs; N50: 811kbp; >99.8% id

# S. pombe dg21



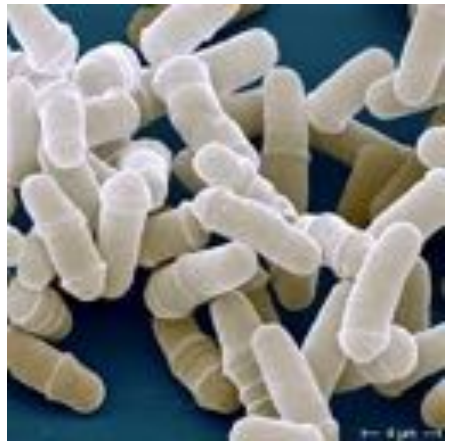
PacBio RS II sequencing at CSHL

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




Over 275x coverage in
5 SMRTcells / 1 afternoon
using P5-C3

# S. pombe dg21

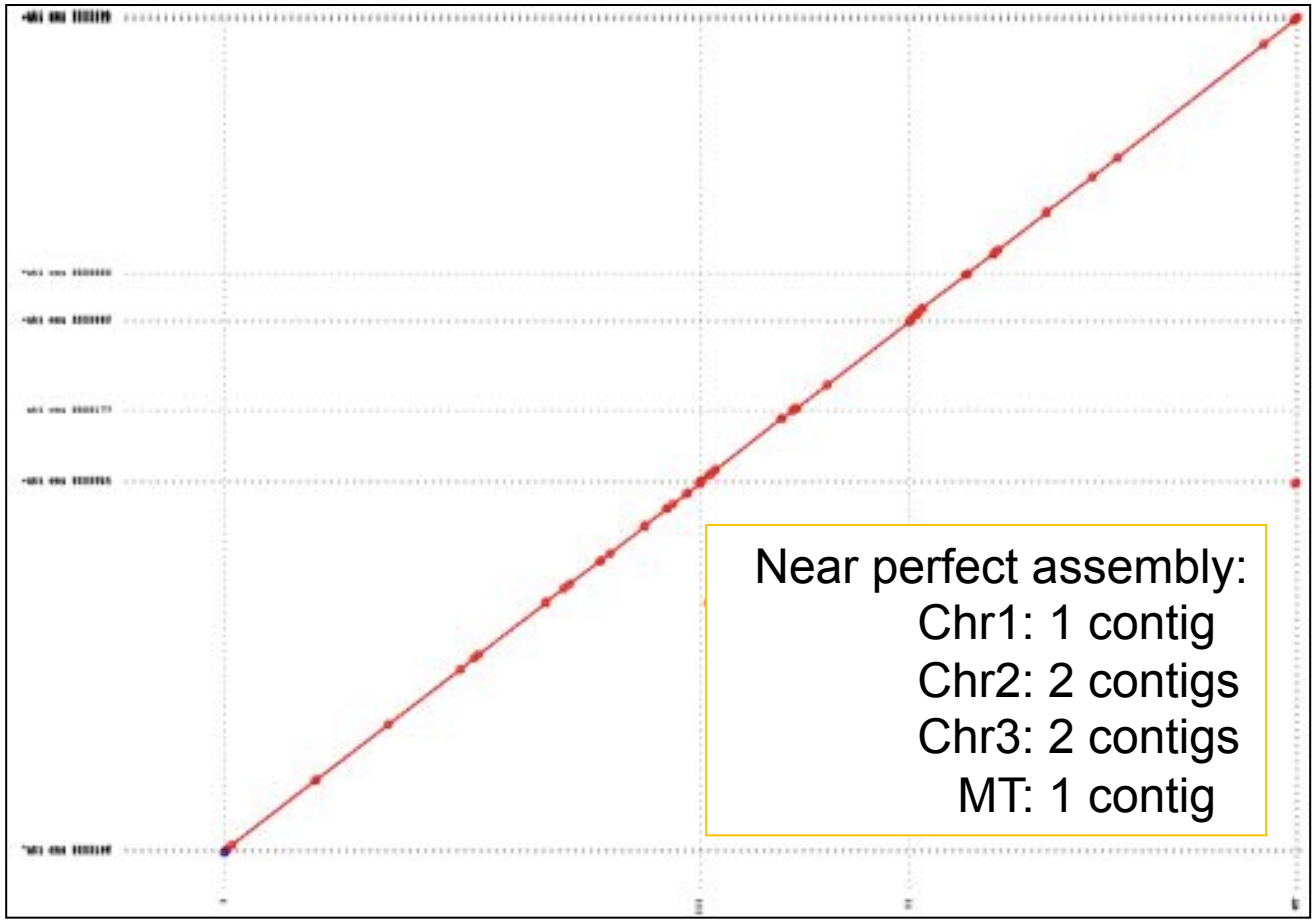


ASM294 Reference sequence

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

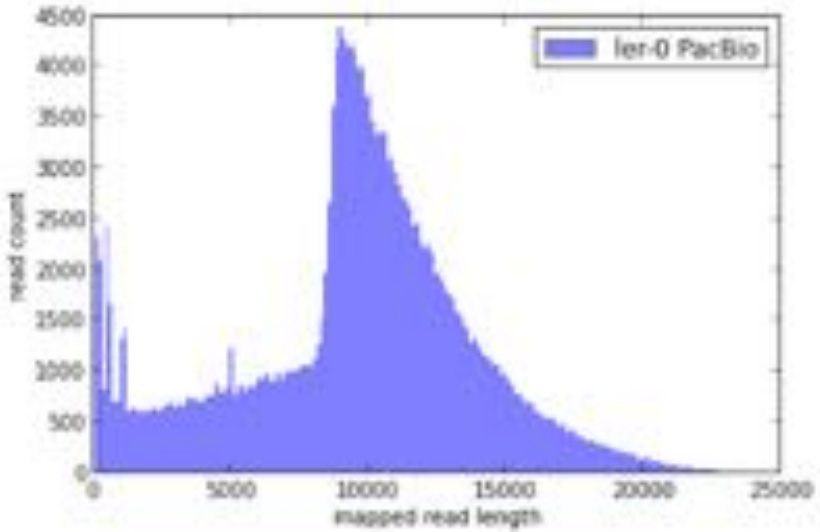
PacBio assembly using HGAP + Celera Assembler

- 12.7Mbp; 13 non-redundant contigs; N50: 3.83Mbp; >99.9% id



# A. thaliana Ler-0

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




*A. thaliana* Ler-0 sequenced at PacBio

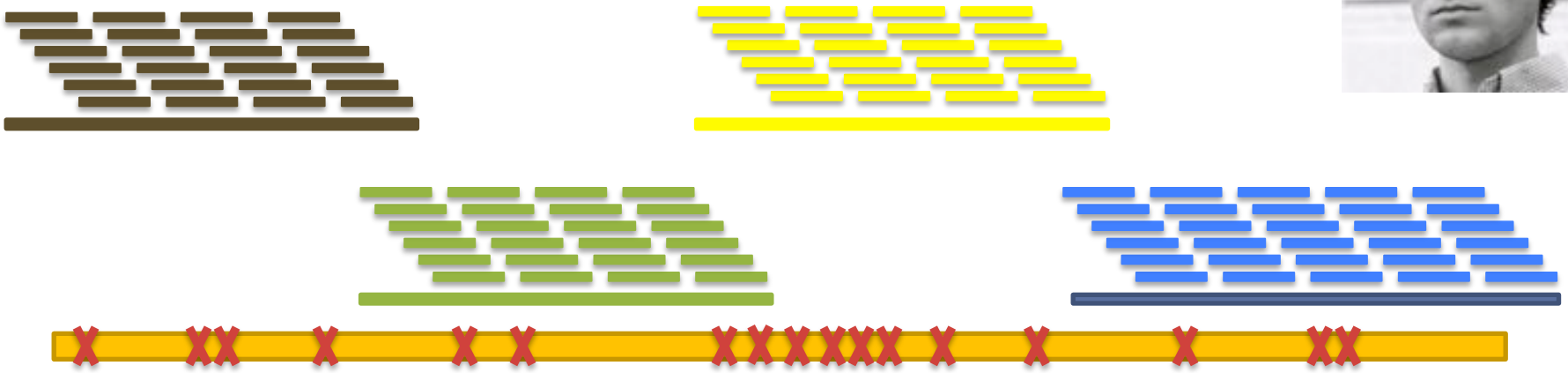
- Sequenced using the previous P4 enzyme and C2 chemistry
- Size selection using an 8 Kb to 50 Kb elution window on a BluePippin™ device from Sage Science
- Total coverage >119x

| | | | |
|---|---|---|---|
| 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**

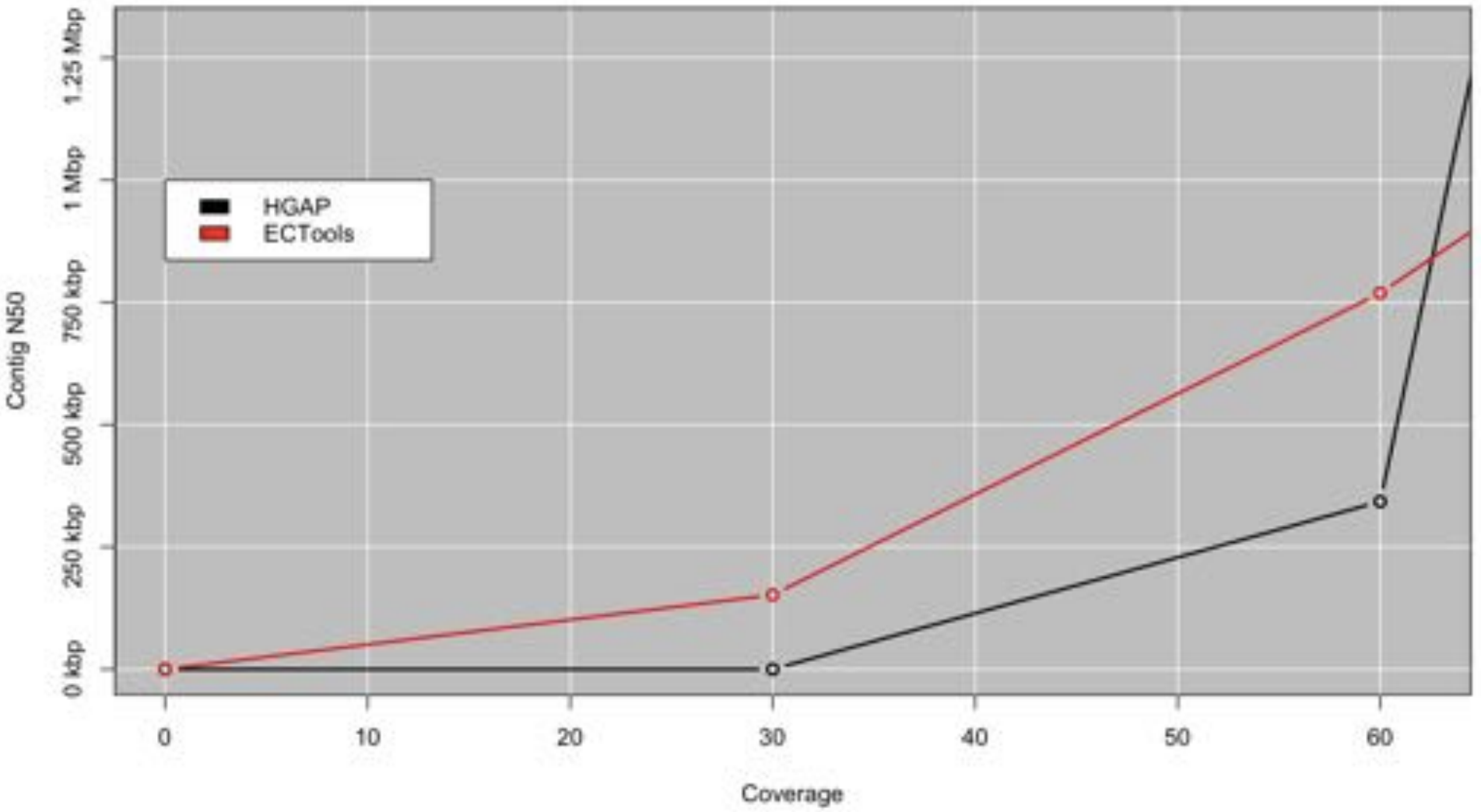
Can Help us overcome:

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

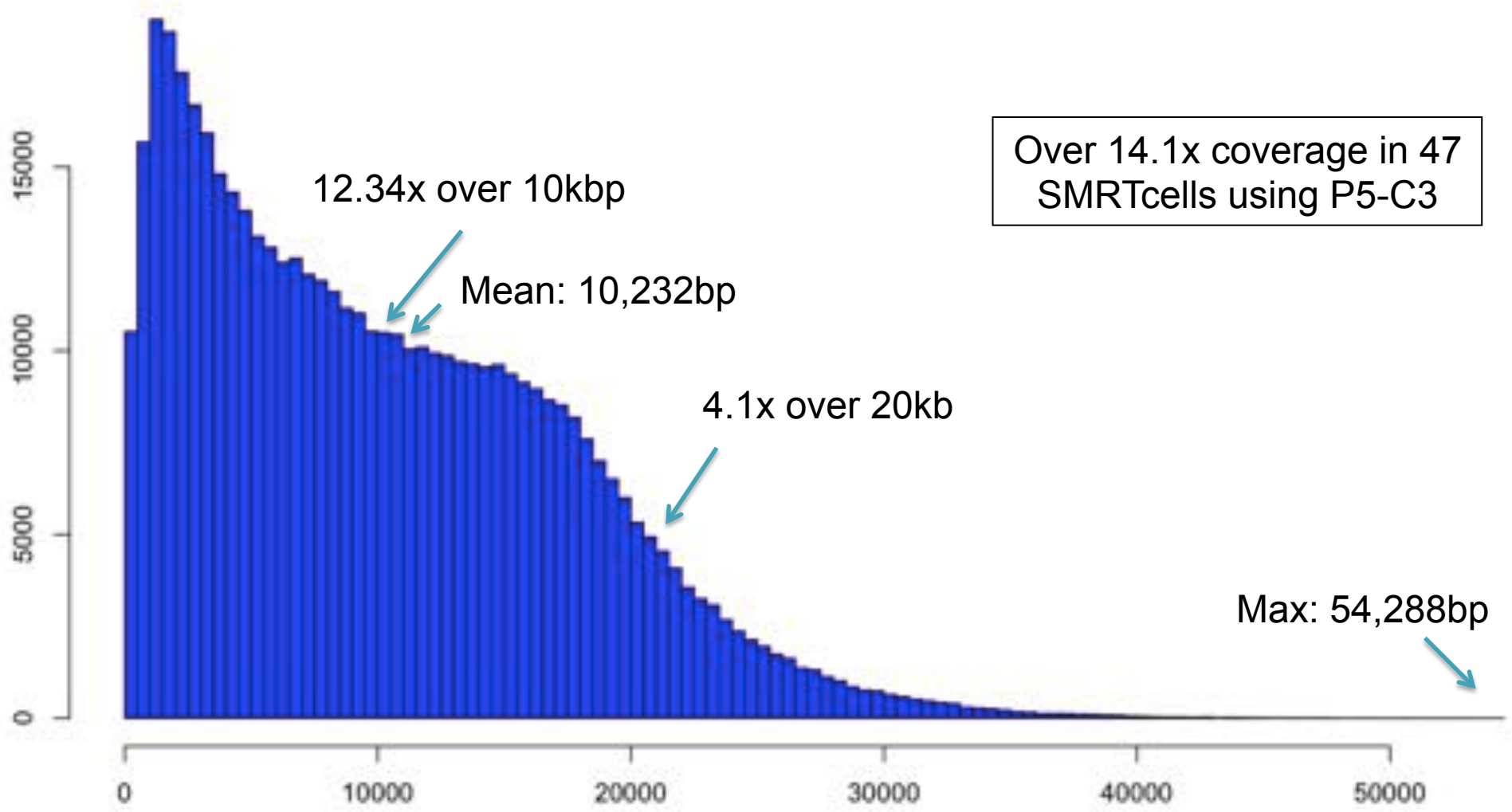
http://blog.pacificbiosciences.com/2013/08/new-data-release-arabidopsis-assembly.html



# 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

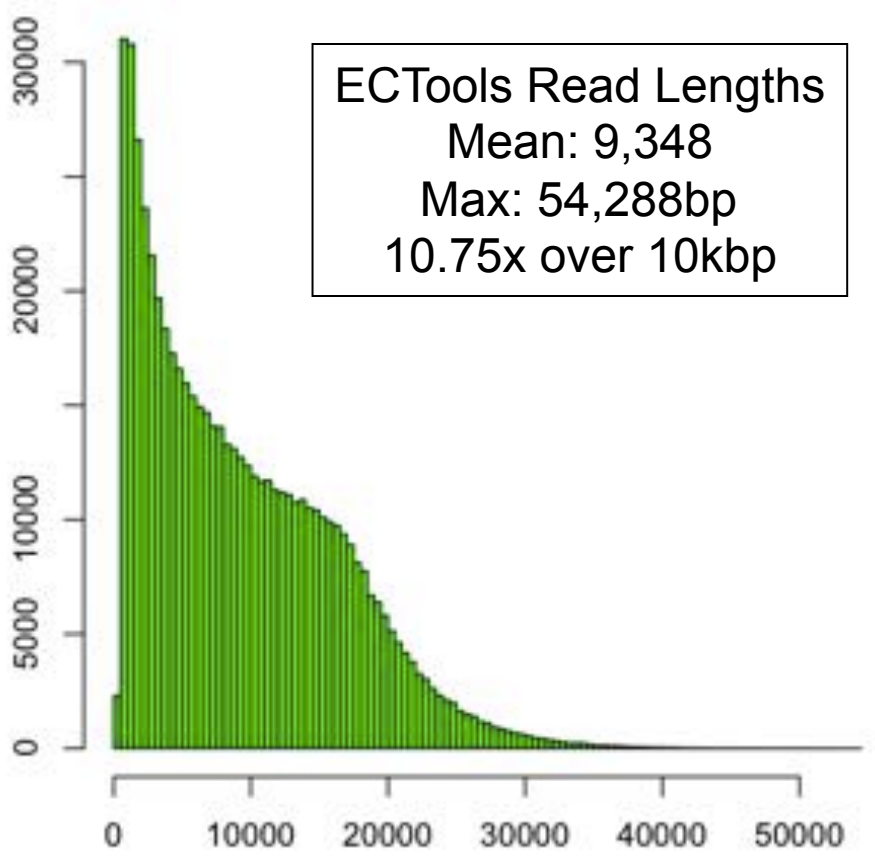



Over 14.1x coverage in 47 SMRTcells using P5-C3

# O. sativa pv Indica (IR64)

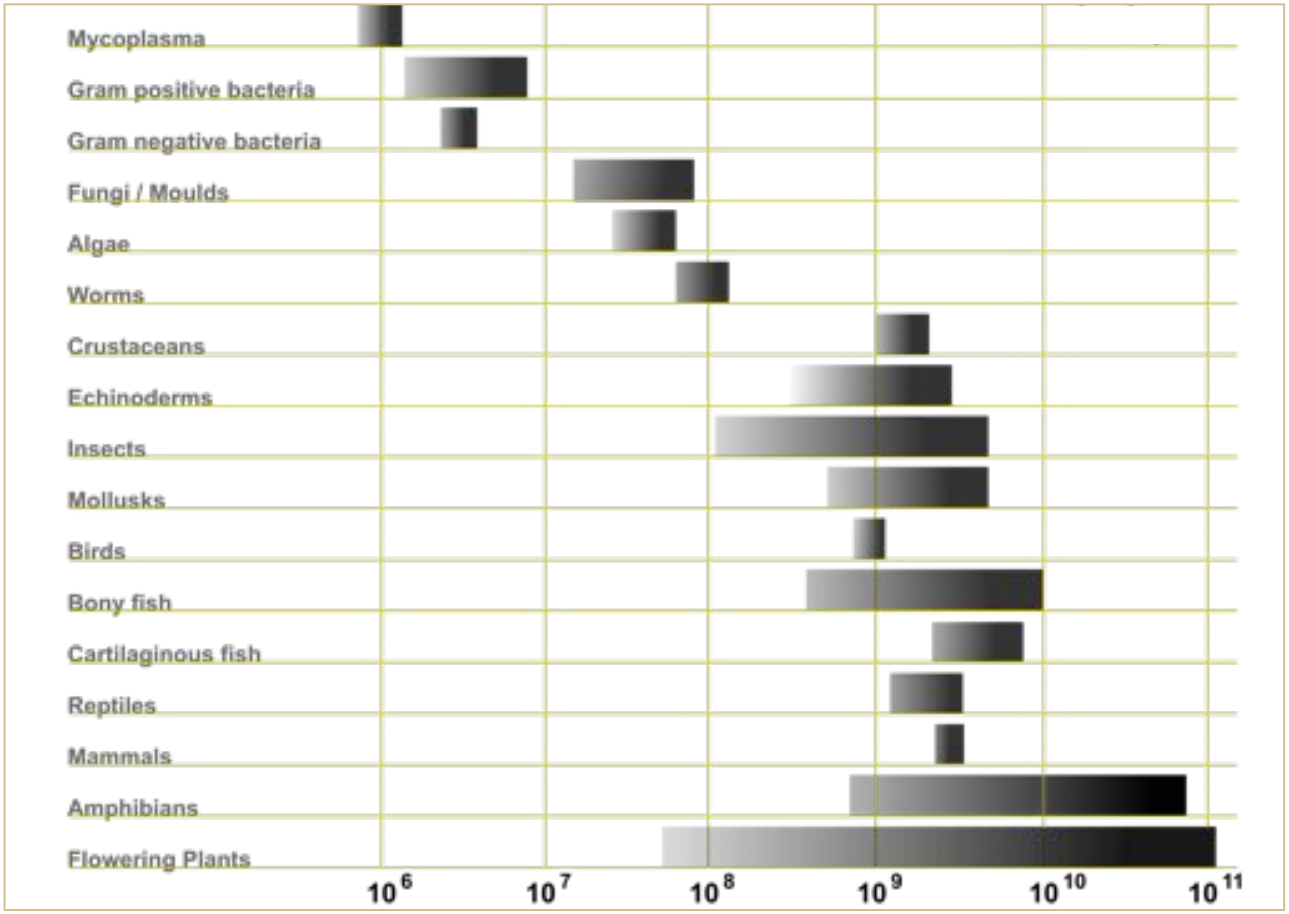
Genome size: ~370 Mb
Chromosome N50: ~29.7 Mbp

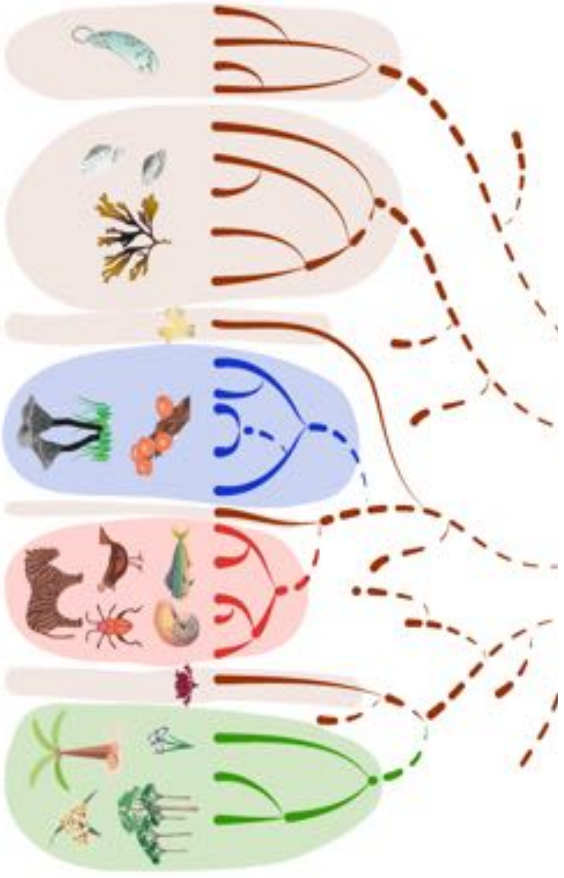
| Assembly | Contig NG50 |
|---|---|
| MiSeq Fragments<br>25x 456bp<br>(3 runs 2x300 @ 450 FLASH) | 19,078 |
| "ALLPATHS-recipe"<br>50x 2x100bp @ 180<br>36x 2x50bp @ 2100<br>51x 2x50bp @ 4800 | 18,450 |
| ECTools<br>10.7x @ 10kbp | 271,885 |




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

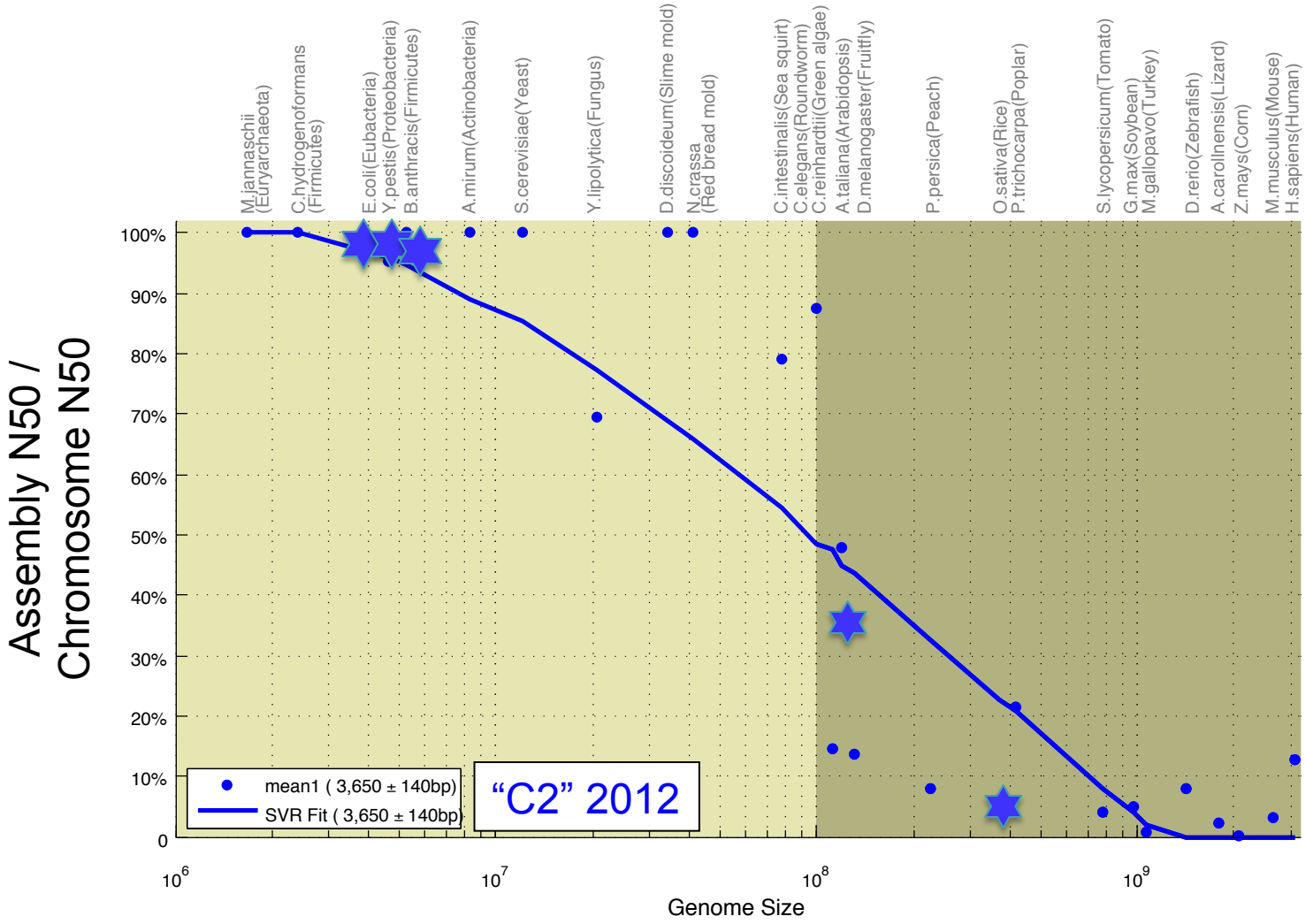
# What should we expect from an assembly?



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

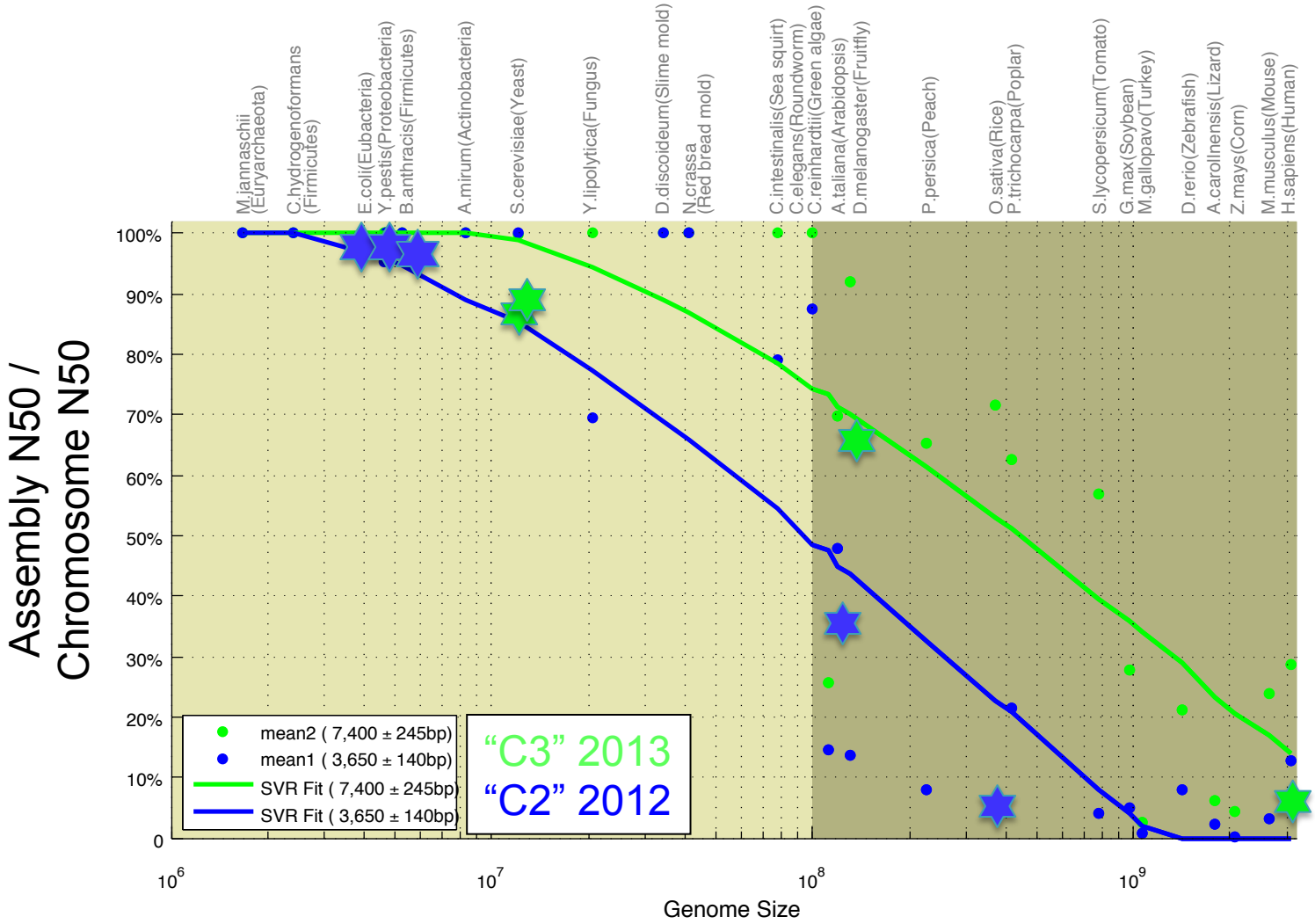
# Assembly Complexity of Long Reads




**Assembly complexity of long read sequencing**

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

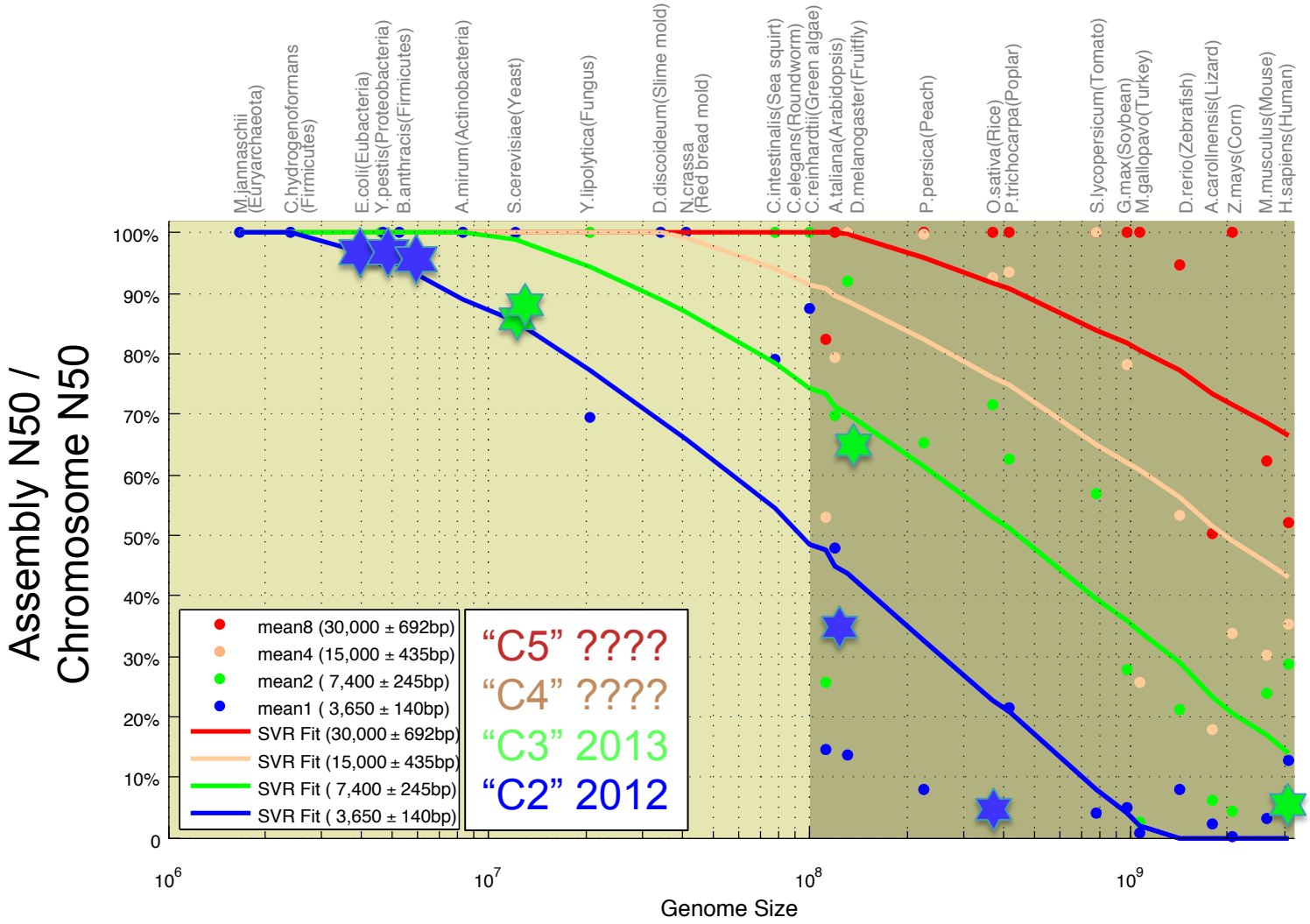
# Assembly Complexity of Long Reads




**Assembly complexity of long read sequencing**

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

# 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*

# Assembly Recommendations

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

  < 100 Mbp: HGAP/PacBio2CA @ 100x PB C3-P5
  expect near perfect chromosome arms

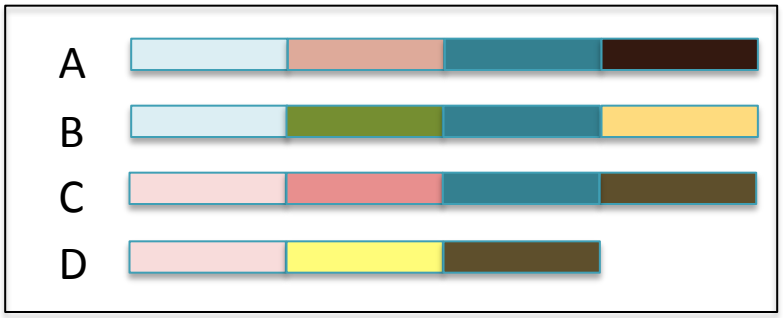
  < 1GB: HGAP/PacBio2CA @ 100x PB C3-P5
  expect high quality assembly: contig N50 over 1Mbp

  > 1GB: hybrid/gap filling
  expect contig N50 to be 100kbp – 1Mbp

  > 5GB: Email mschatz@cshl.edu

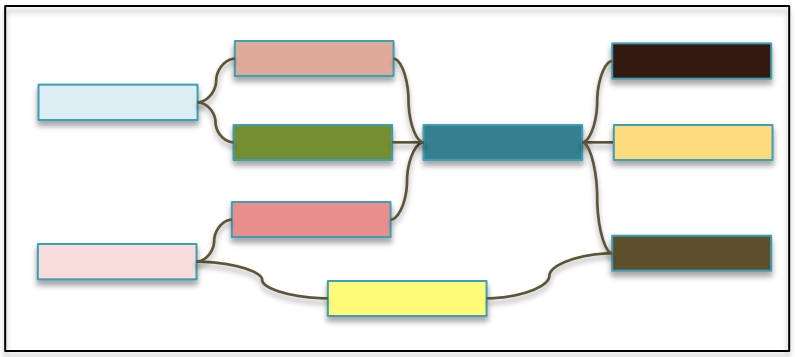
- **Caveats**
  - Model only as good as the available references (esp. haploid sequences)
  - Technologies are quickly improving, exciting new scaffolding technologies

# 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?

**Rapid pan genome analysis with suffix skips**
Marcus, S, Schatz, MC (2014) *In preparation*

# Outline

1. De novo assembly by analogy
2. Long Read Assembly
3. Disease Analytics

# 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

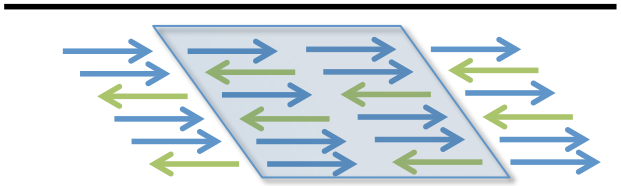
NRXN1 *de novo* SNP
(auSSC12501 chr2:50724605)

**Accurate detection of de novo and transmitted INDELs within exome-capture data using micro-assembly**

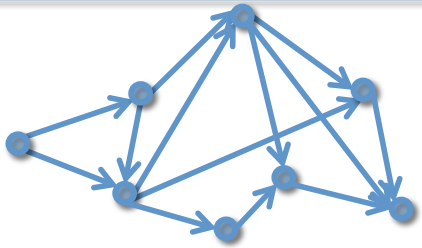
Narzisi, G, O'Rawe, J, Iossifov, I, Lee, Y, Wang, Z, Wu, Y, Lyon, G, Wigler, M, Schatz, MC (2014) *Under review.*

# Scalpel Pipeline

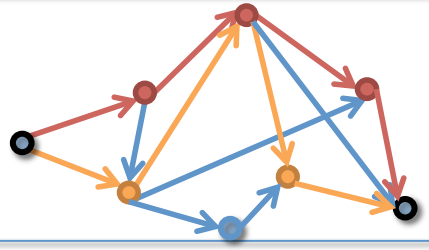
Extract reads mapping within the exon including (1) well-mapped reads, (2) soft-clipped reads, and (3) anchored pairs



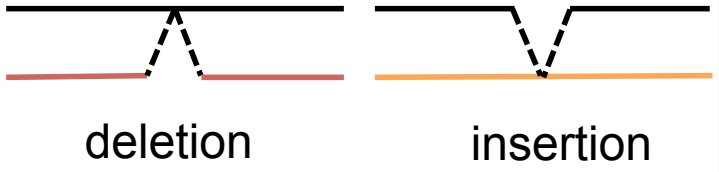
Decompose reads into overlapping *k*-mers and construct de Bruijn graph from the reads



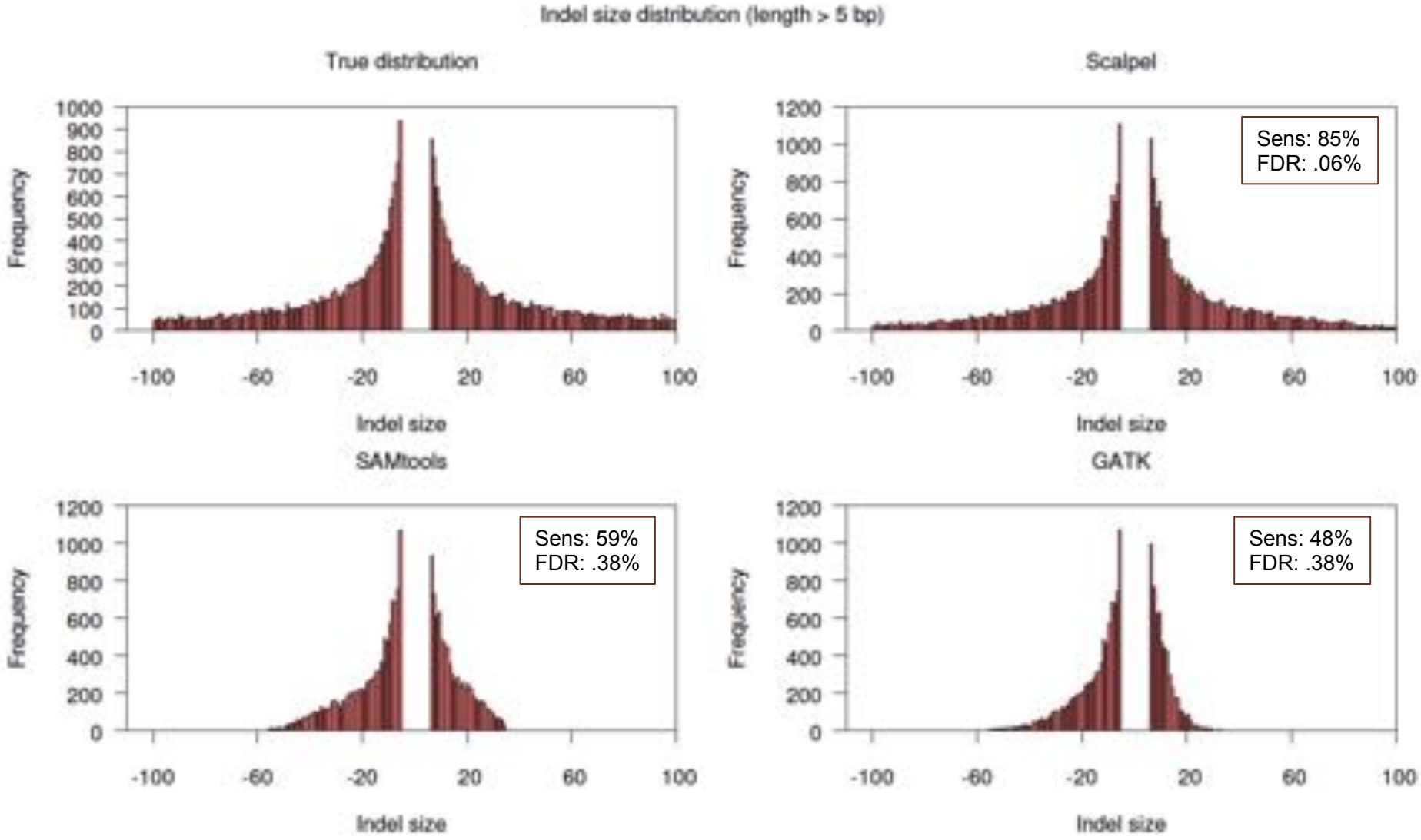
Find end-to-end haplotype paths spanning the region



Align assembled sequences to reference to detect mutations



# Simulation Analysis




Simulated 10,000 indels in a exome from a known log-normal distribution

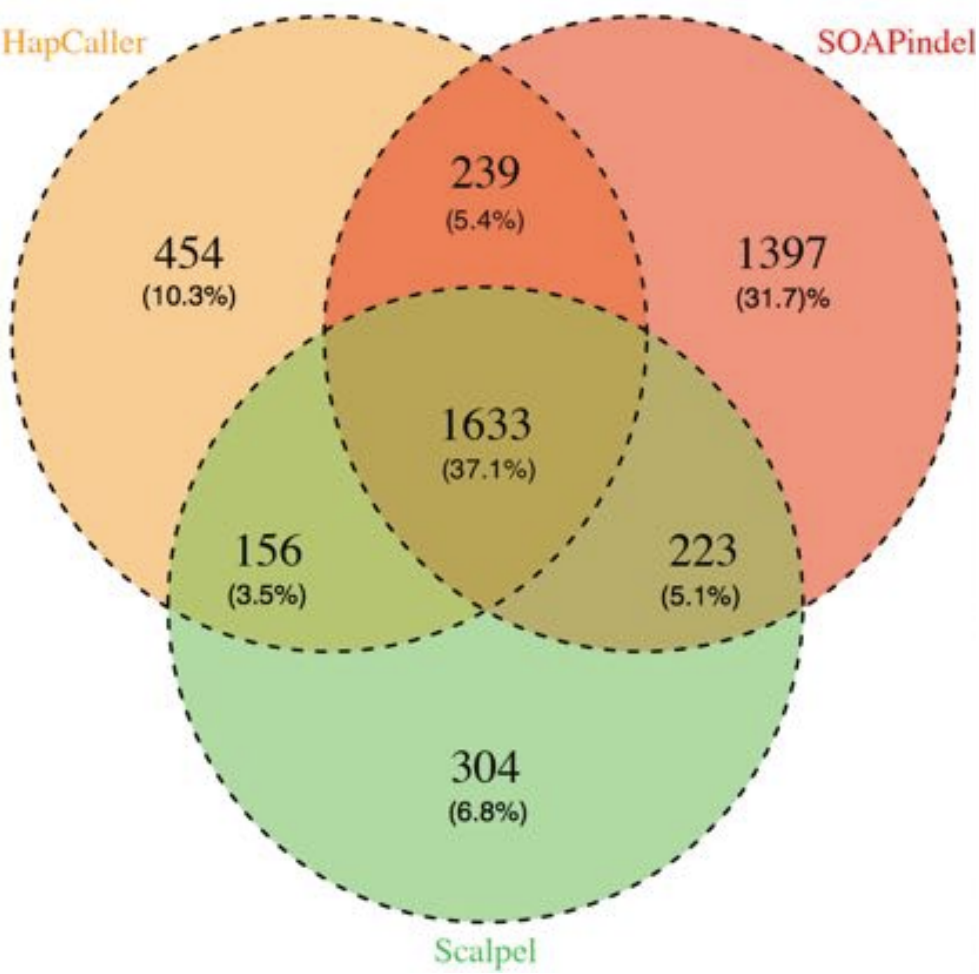
# Experimental Analysis & Validation

Selected one deep coverage exome for deep analysis

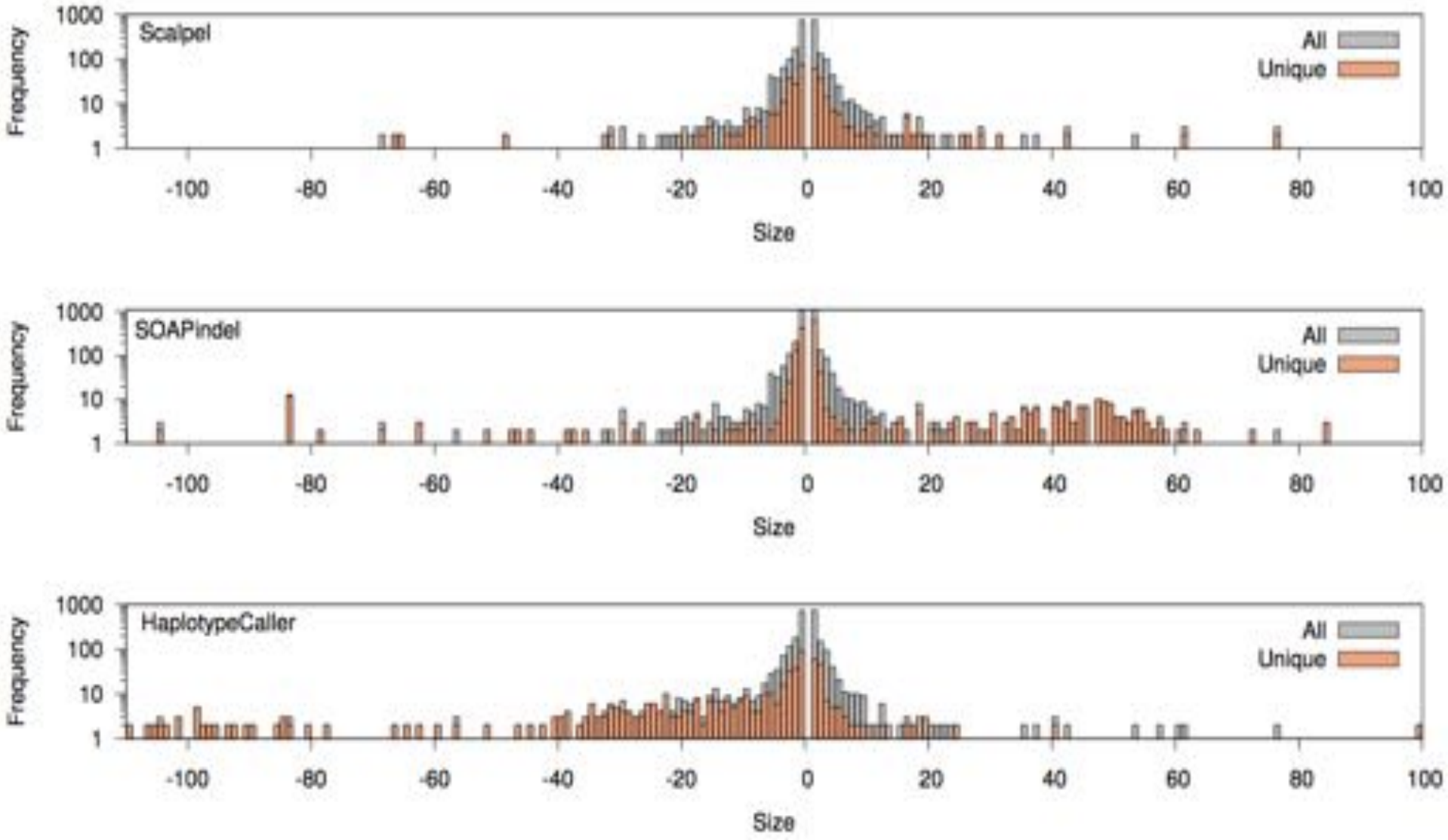
- Individual was diagnosed with ADHD and turrets syndrome
- 80% of the target at >20x coverage
- Evaluated with Scalpel, SOAPindel, and GATK Haplotype Caller

1000 indels selected for validation

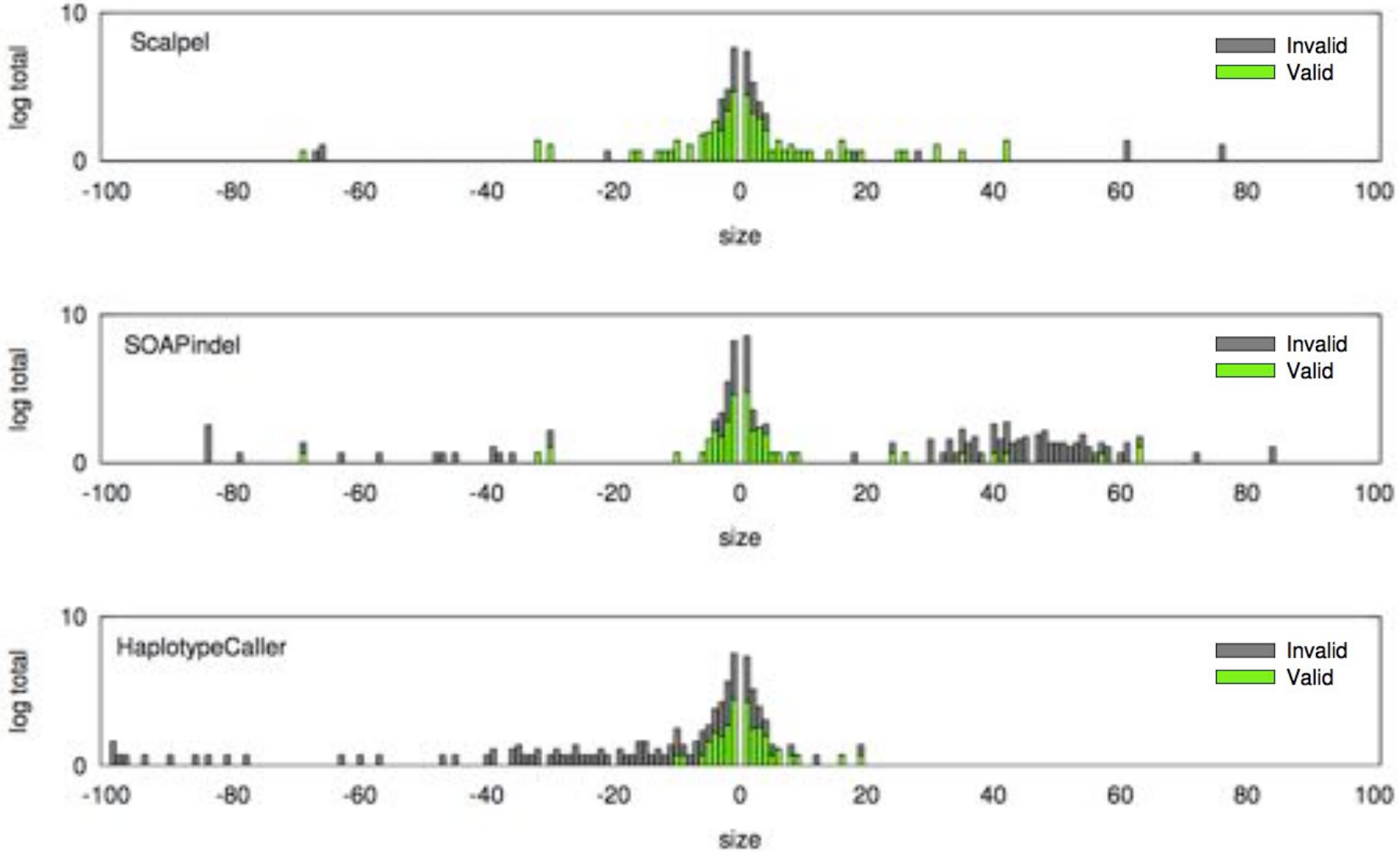
- 200 Scalpel
- 200 GATK Haplotype Caller
- 200 SOAPindel
- 200 within the intersection
- 200 long indels (>30bp)



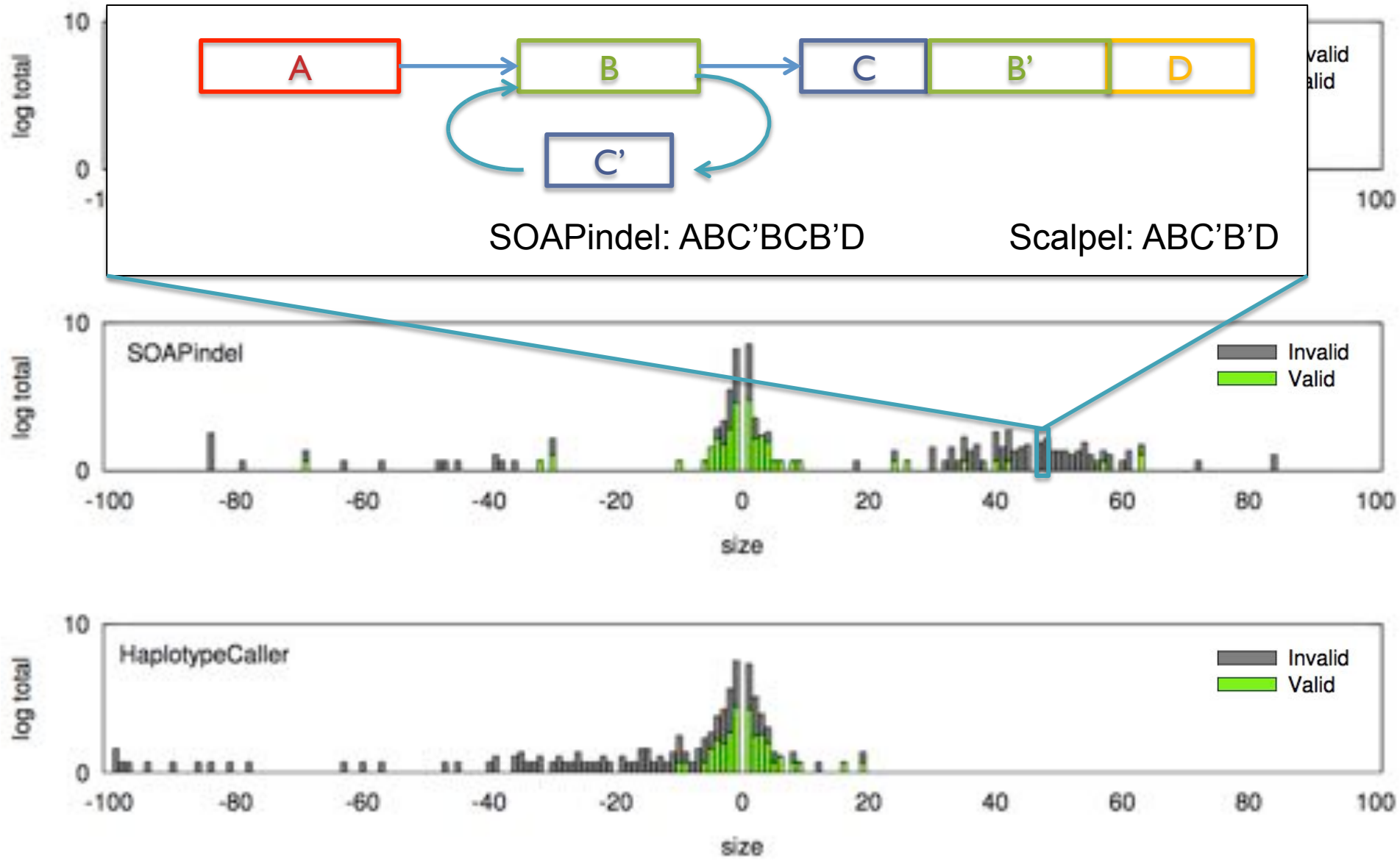
# Scalpel Indel Discovery



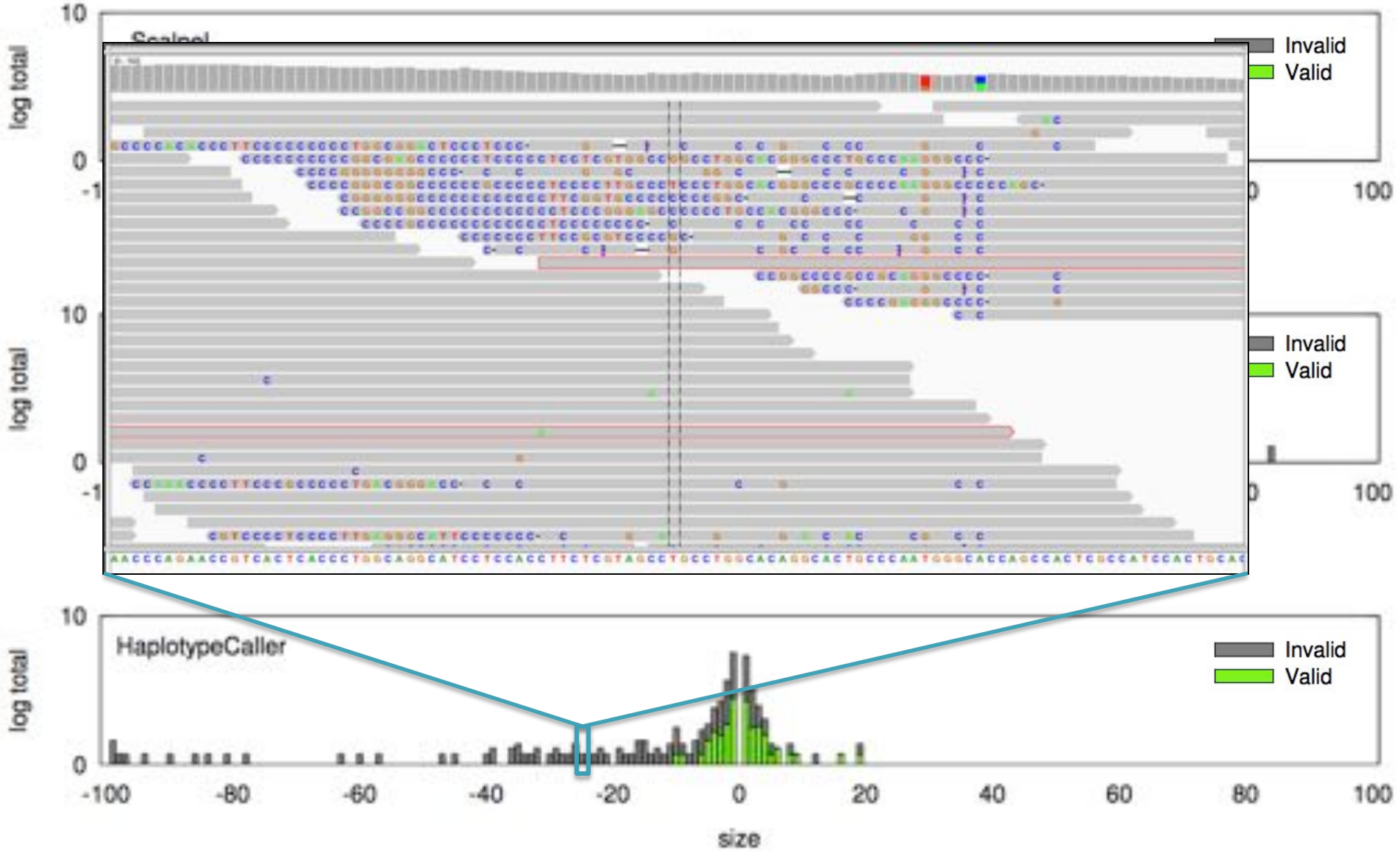
# Scalpel Indel Discovery



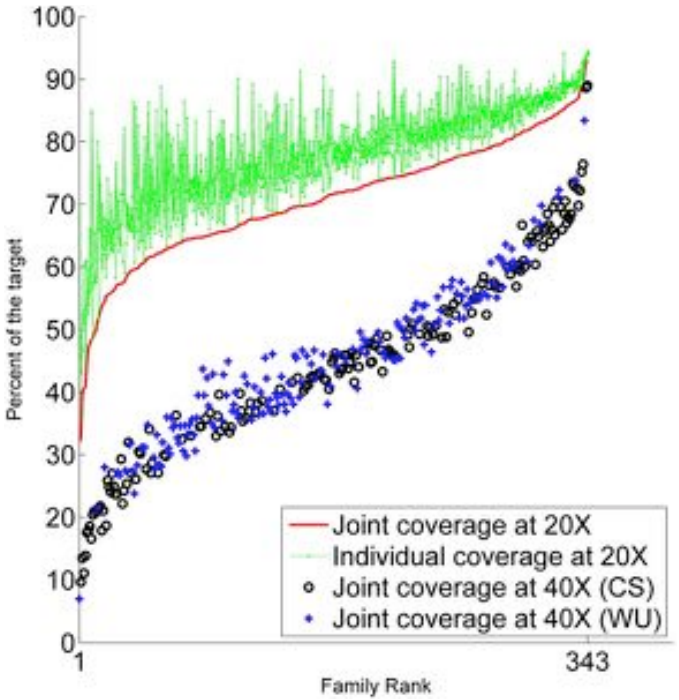
# Scalpel Indel Discovery



# Scalpel Indel Discovery



# Exome sequencing of the SSC




Last year saw 3 reports of >593 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
- Iossifov (343) and O'Roak (50) used GATK, Sanders (200) didn't attempt to identify indels

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

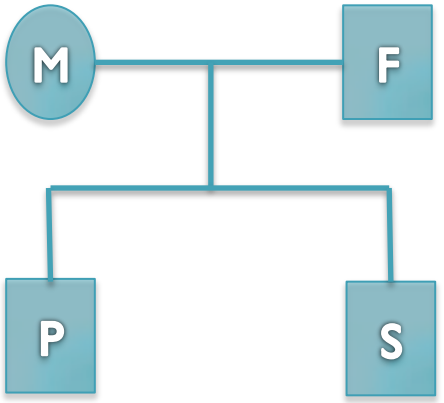
**De novo mutations revealed by whole-exome sequencing are strongly associated with autism**
Sanders *et al.* (2012) *Nature*. 485, 237–241.

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

# De novo mutation discovery and validation

**Concept**: Identify mutations not present in parents.

**Challenge**: Sequencing errors in the child or low coverage in parents lead to false positive de novos




```
Reference:   ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...

Father:      ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Mother:      ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Sibling:     ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Proband(1):  ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Proband(2):  ...TCAAATCCTTTTAAT****AAGAGCTGACA...
```

4bp heterozygous deletion at chr15:93524061 CHD2

# De novo Genetics of Autism

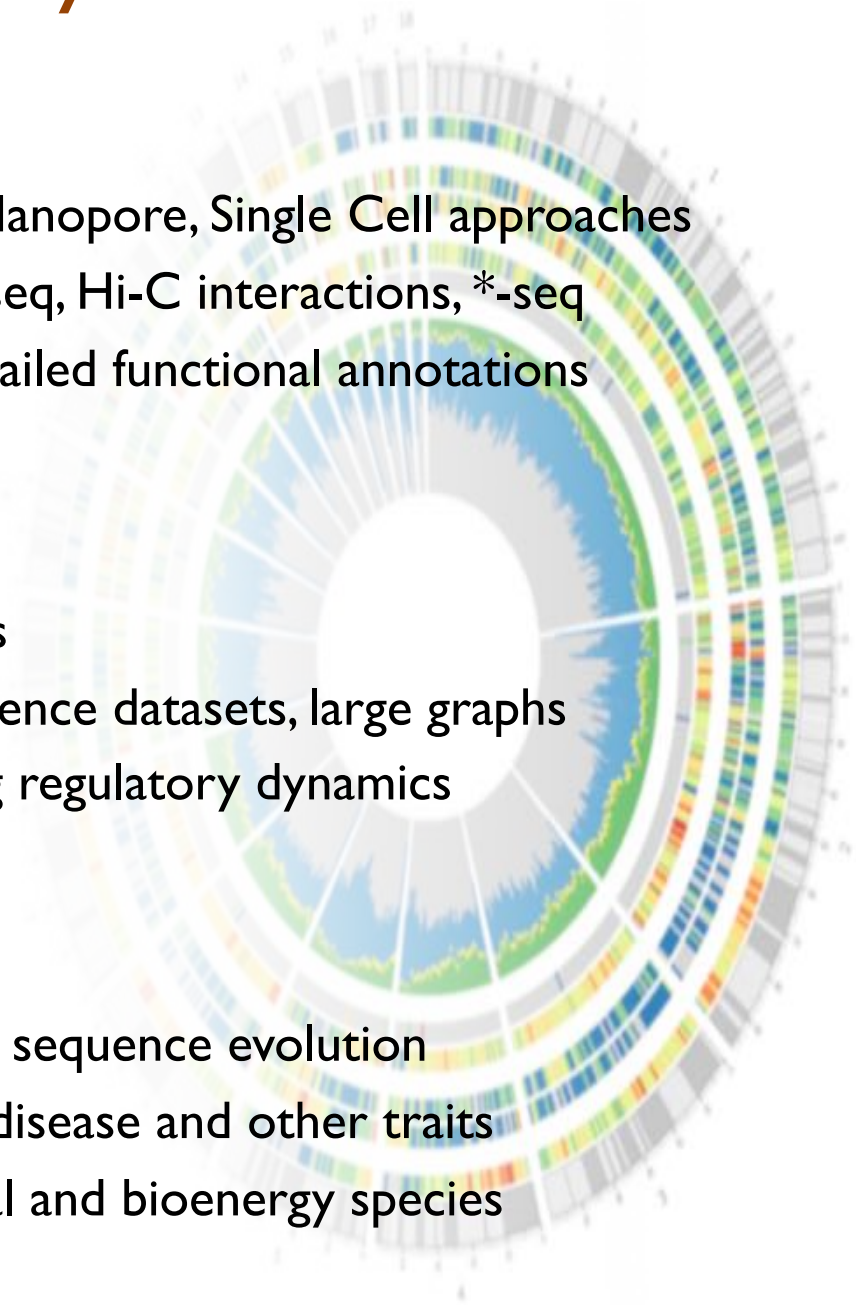
- In 593 family quads so far, we see significant enrichment in de novo ***likely gene killers*** in the autistic kids
  - Overall rate basically 1:1
  - 2:1 enrichment in nonsense mutations
  - 2:1 enrichment in frameshift indels
  - 4:1 enrichment in splice-site mutations
  - Most de novo originate in the paternal line in an age-dependent manner (56:18 of the mutations that we could determine)

- Observe strong overlap with fragile X protein (FMPR) network
  - Related to neuron development and synaptic plasticity
  - Also strong overlap with chromatin remodelers

**Accurate detection of de novo and transmitted INDELs within exome-capture data using micro-assembly**

Narzisi, G, O'Rawe, J, Iossifov, I, Lee, Y, Wang, Z, Wu, Y, Lyon, G, Wigler, M, Schatz, MC (2014) *Under review.*

# Summary



## *Biotechnology*

- Sequencing: Illumina, PacBio, Oxford Nanopore, Single Cell approaches
- Biochemical assays: RNA-seq, Methyl-seq, Hi-C interactions, *-seq
- More accurate assemblies & more detailed functional annotations

## *Algorithmics*

- Highly scalable algorithms and systems
- Indexing and analyzing very large sequence datasets, large graphs
- Constructing Pan-genomes & inferring regulatory dynamics

## *Comparative Genomics*

- Cross species comparisons, models of sequence evolution
- Identifying mutations associated with disease and other traits
- Genotype-to-phenotype of agricultural and bioenergy species

# Acknowledgements


Schatz Lab
**James Gurtowski**
**Hayan Lee**
**Giuseppe Narzisi**

Ke Jiang
Shoshana Marcus
Srividya Ramakrishnan
Rob Aboukhalil
Mitch Bekritsky
Charles Underwood
Tyler Gavin
Maria Nattestad
Alejandro Wences
Greg Vurture
Eric Biggers
Aspyn Palatnick

CSHL
**McCombie Lab**
**Wigler Lab**
**Lyon Lab**

Hannon Lab
Gingeras Lab
Jackson Lab
Hicks Lab
Iossifov Lab
Levy Lab
Lippman Lab
Martienssen Lab
Tuveson Lab
Ware Lab

Pacific Biosciences

SFARI
SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

National Human Genome Research Institute

U.S. DEPARTMENT OF ENERGY

NSF

**Biological Data Sciences**
**Cold Spring Harbor Laboratory, Nov 5 - 8, 2014**

# Thank you

http://schatzlab.cshl.edu

@mike_schatz