

Algorithms for de novo genome assembly and disease analytics

Michael Schatz

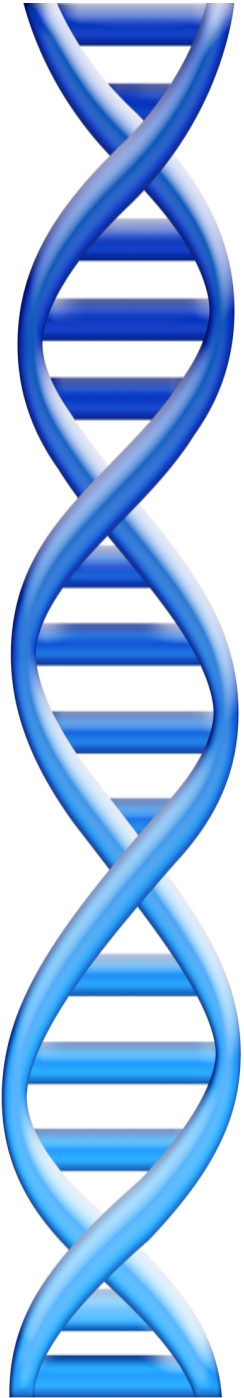
Feb 11, 2014

IDIES Seminar, Johns Hopkins University

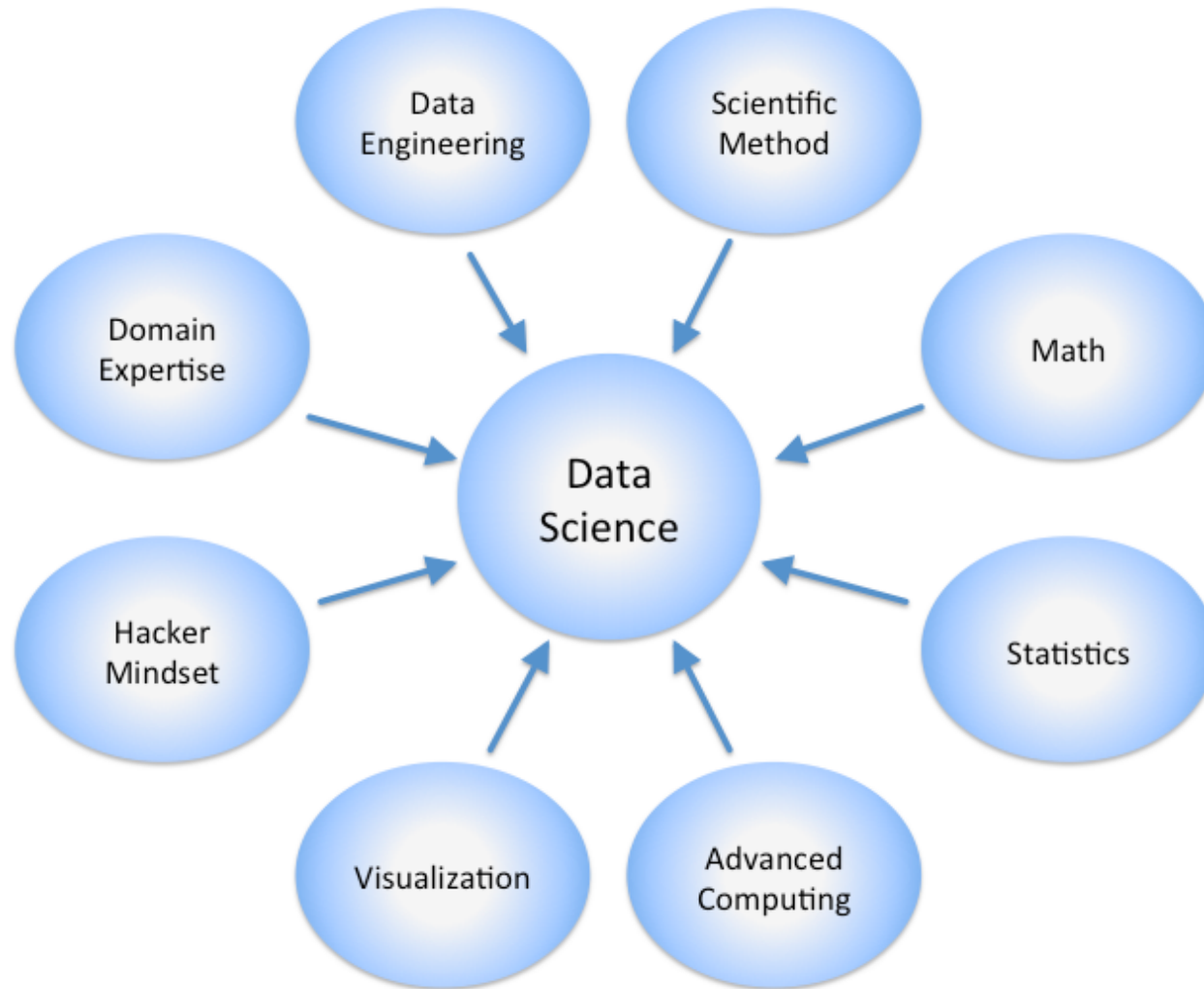


Outline

1. Biological Data Science
2. De novo genome assembly
3. Disease Analytics



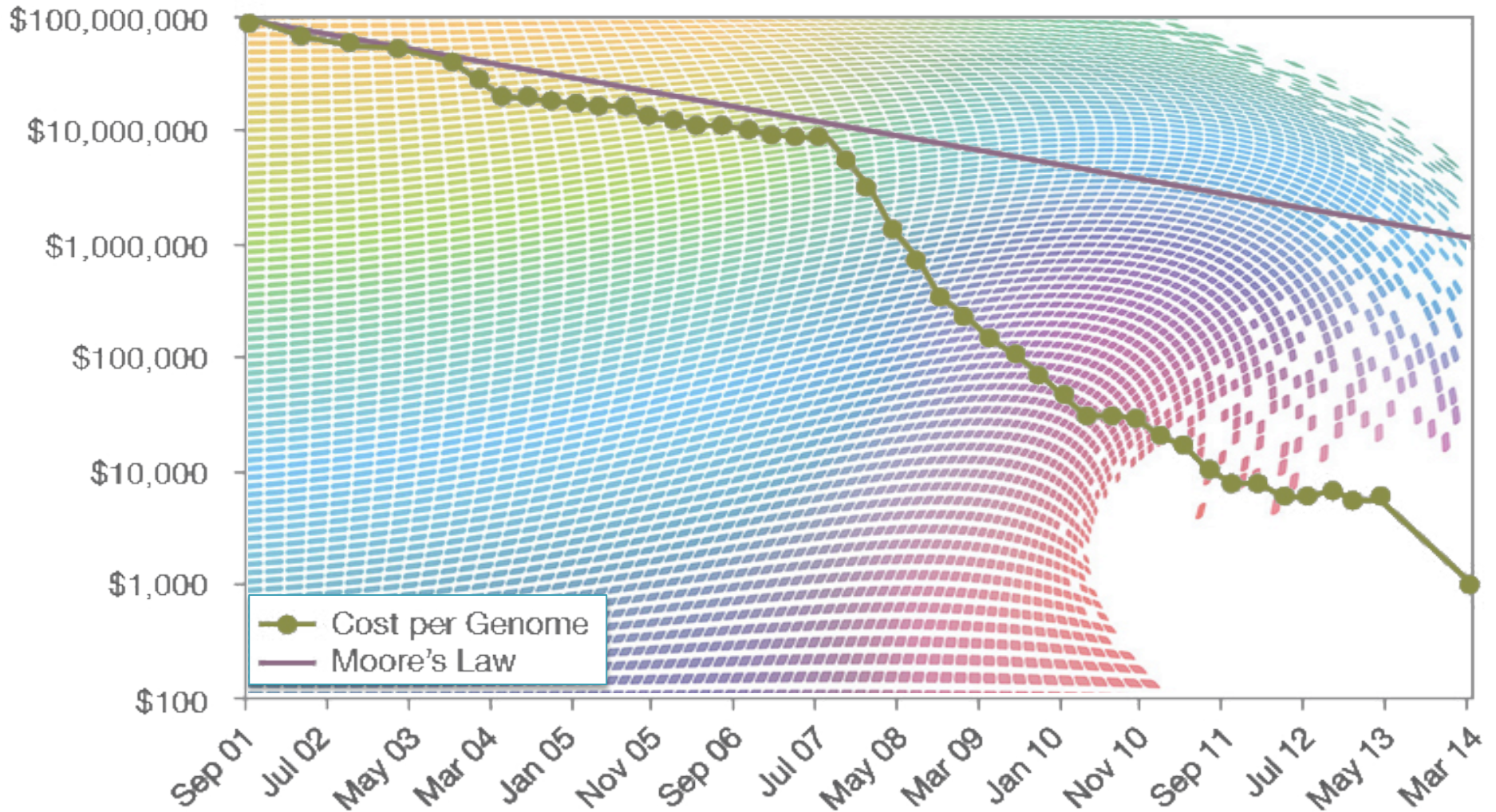
What is Data Science?



http://en.wikipedia.org/wiki/Data_science

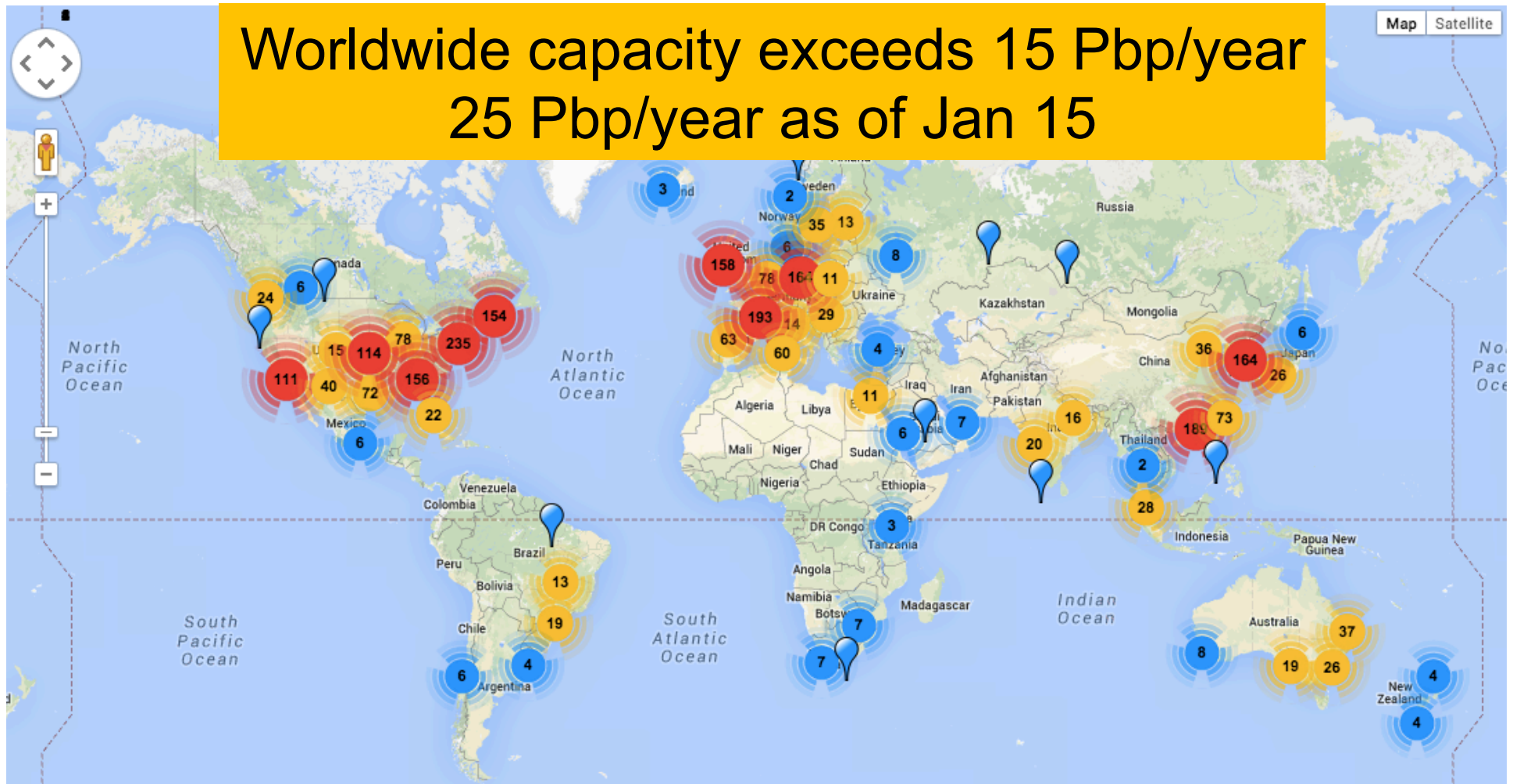
<http://simplystatistics.org/2013/12/12/the-key-word-in-data-science-is-not-data-it-is-science/>

Cost per Genome



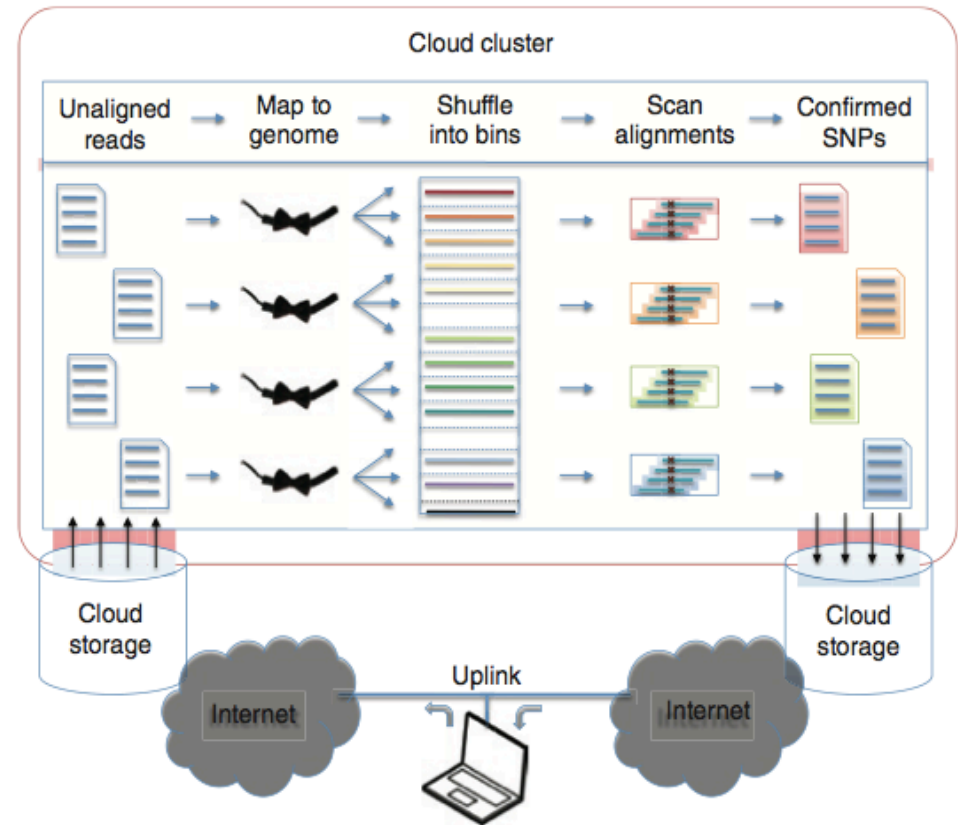
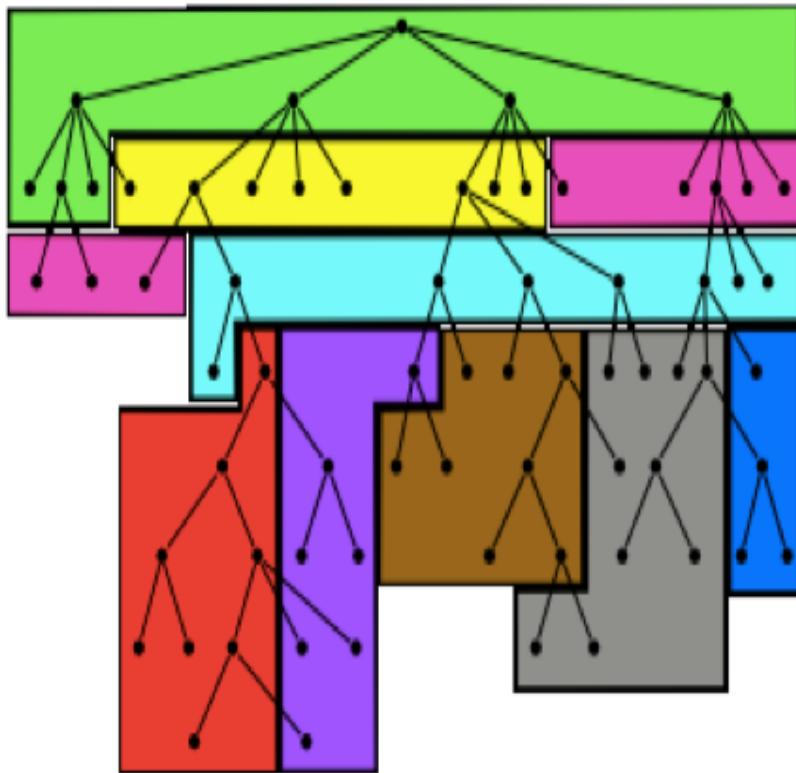
<http://www.genome.gov/sequencingcosts/>
<http://res.illumina.com/documents/products/datasheets/datasheet-hiseq-x-ten.pdf>

Sequencing Centers



Next Generation Genomics: World Map of High-throughput Sequencers
<http://omicsmaps.com>

System Level Advances



Optimizing data intensive GPGPU computations for DNA sequence alignment
Trapnell, C, Schatz, MC (2009) *Parallel Computing*. 35(8-9):429-440.

The DNA Data Deluge

Schatz, MC and Langmead, B (2013) *IEEE Spectrum*. July, 2013

Unsolved Questions in Biology

There is tremendous interest to sequence:

- What is your genome sequence?
- How does your genome compare to my genome?

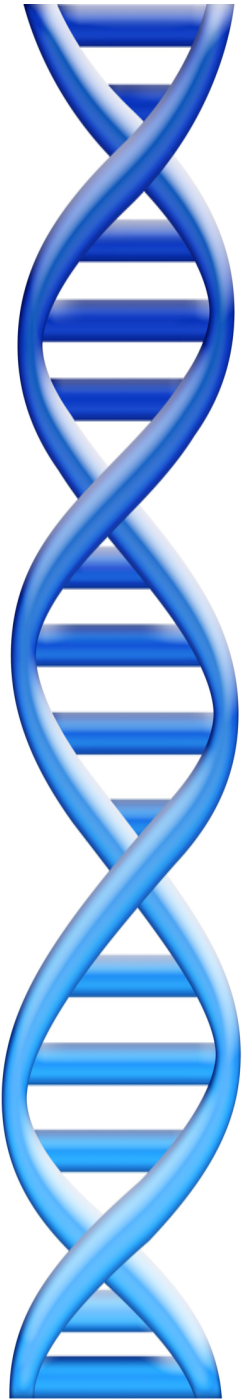
Biological Data Sciences:

- Widely Distributed Sensors
- Substantial Data Volumes
- Diverse Data Types
- Diverse Research questions
- Continuum of domain expertise
 - “Pure CS Problems” through “Pure Bio Problems”
- What virus and microbes are living inside you.
- How do your mutations relate to disease?
- What drugs should we give you?
- Plus hundreds and hundreds more



Outline

1. Biological Data Science
2. De novo genome assembly
3. Disease Analytics



De Novo Assembly Applications

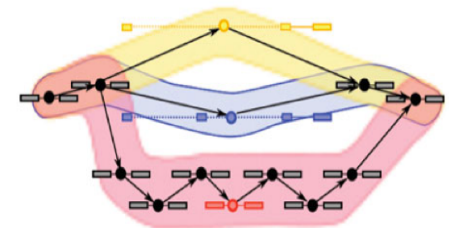
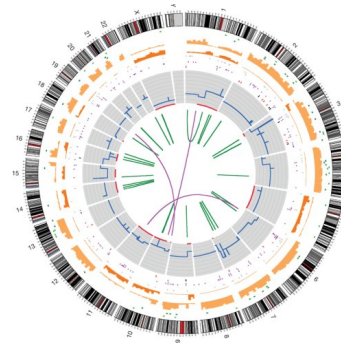
- Novel genomes



- Metagenomes

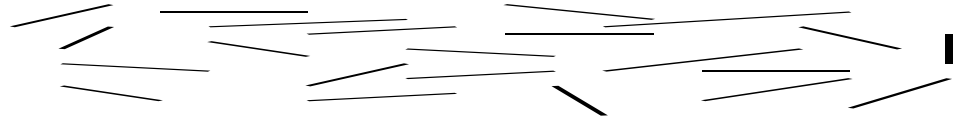


- Sequencing assays
 - Structural variations
 - Transcript assembly
 - ...



Assembling a Genome

1. Shear & Sequence DNA



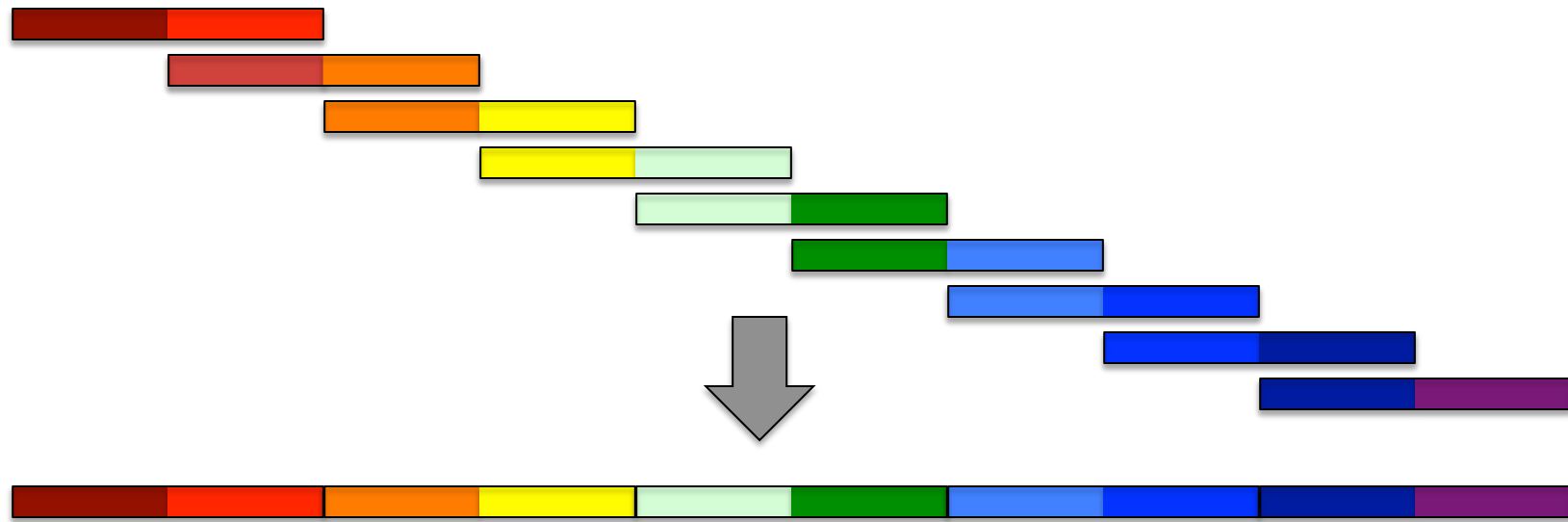
2. Construct assembly graph from overlapping reads

...AGCCTAGGGATGCGCGACACGT

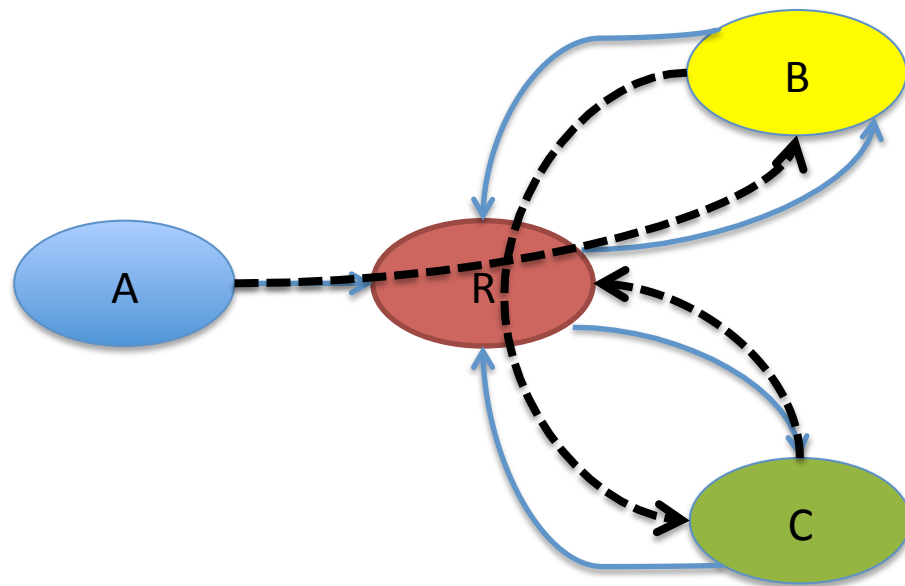
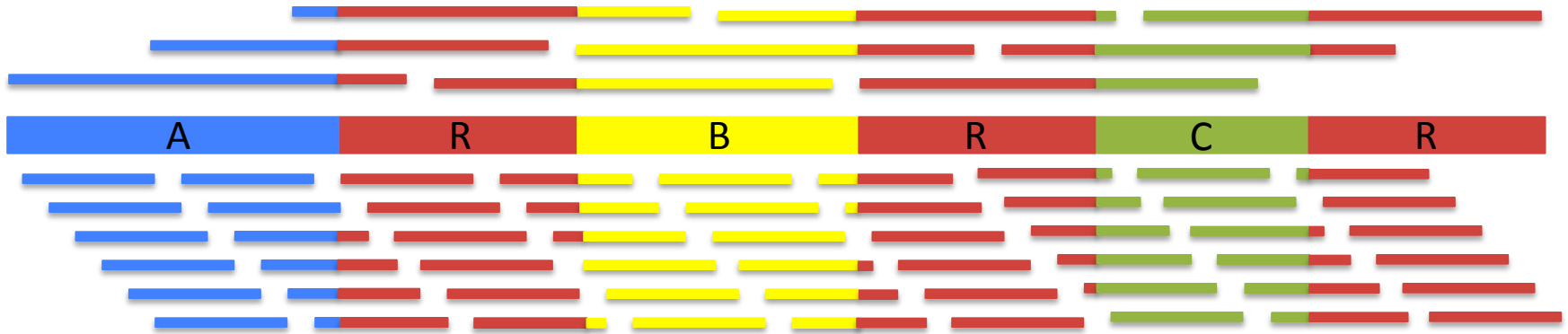
GGATGCGCGACACGTGCATATCCGGTTTGGTCAACCTCGGACGGAC

CAACCTCGGACGGACCTCAGCGAA...

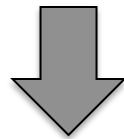
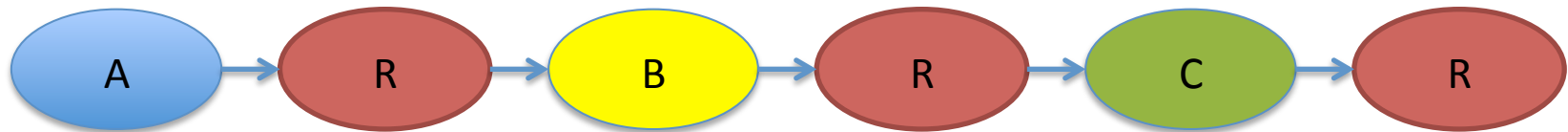
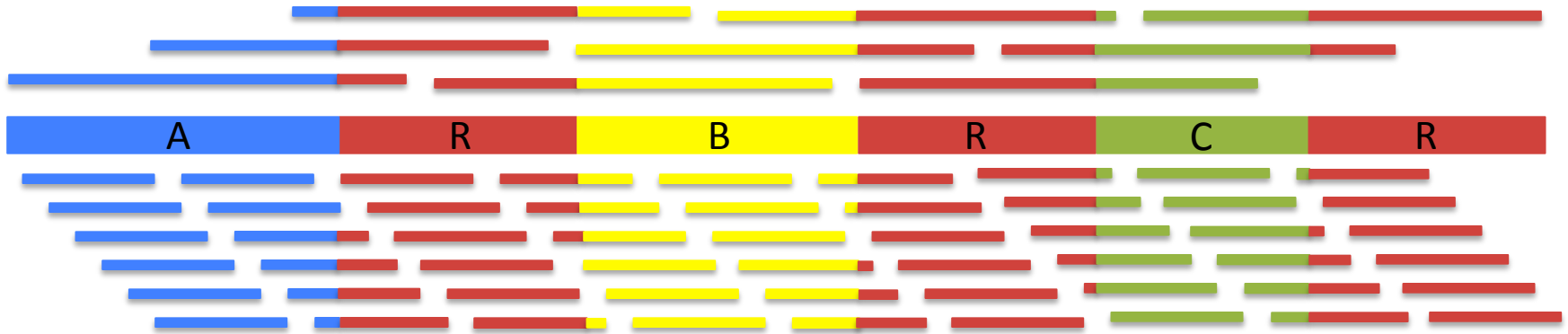
3. Simplify assembly graph



Assembly Complexity



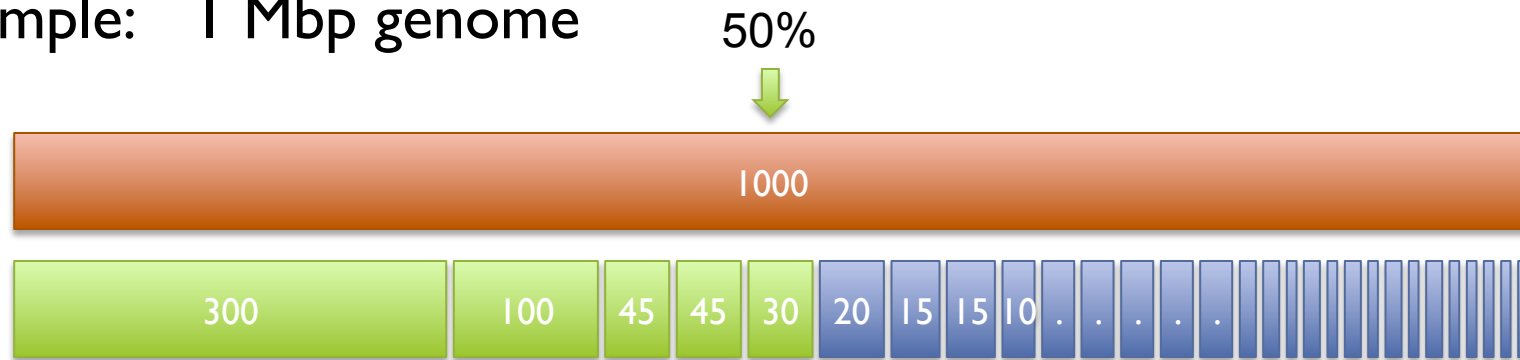
Assembly Complexity



N50 size

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

Example: 1 Mbp genome



N50 size = 30 kbp

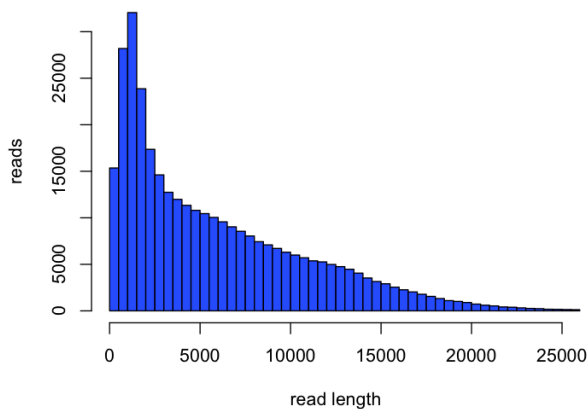
(300k+100k+45k+45k+30k = 520k \geq 500kbp)

Note:

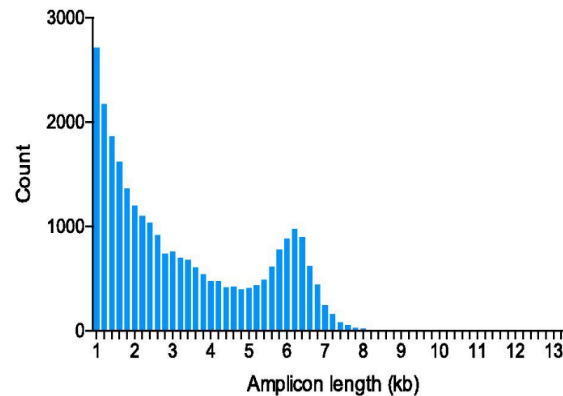
A “good” N50 size is a moving target relative to other recent publications. 10-20kbp contig N50 is currently a typical value for most “simple” genomes.

Long Read Sequencing Technology

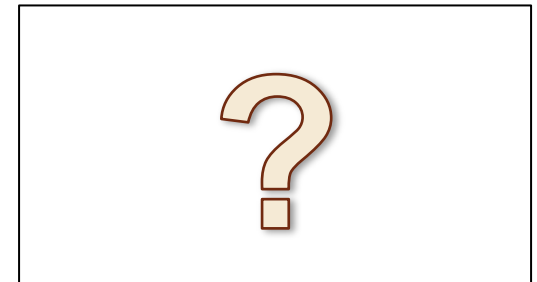
PacBio RS II



Moleculo



Oxford Nanopore

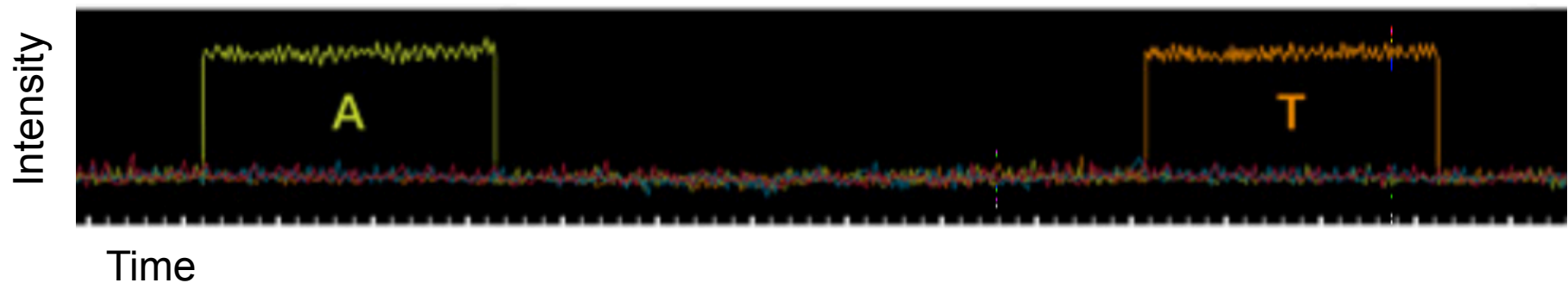
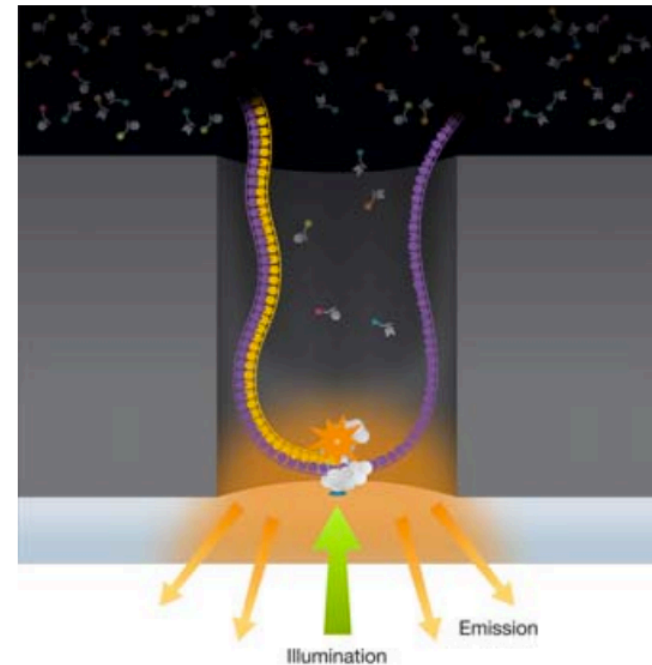
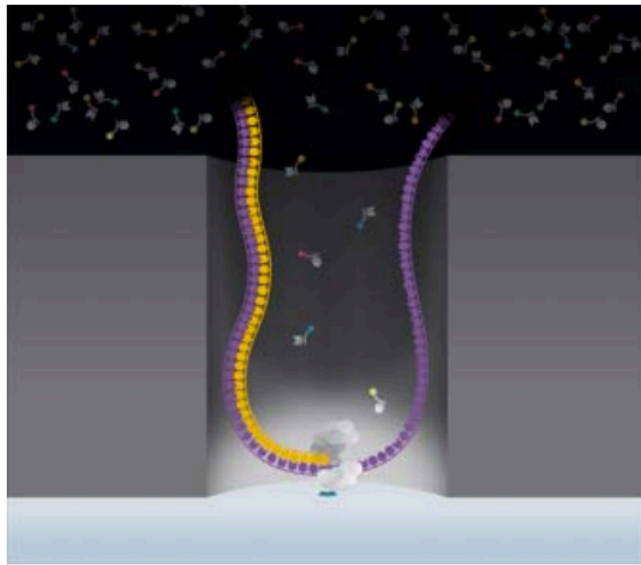


Oxford Nanopore @nanopore
Happy New Year! Registration for the MinION Access Programme
will close at 5pm GMT on Wed 22nd January 2014.

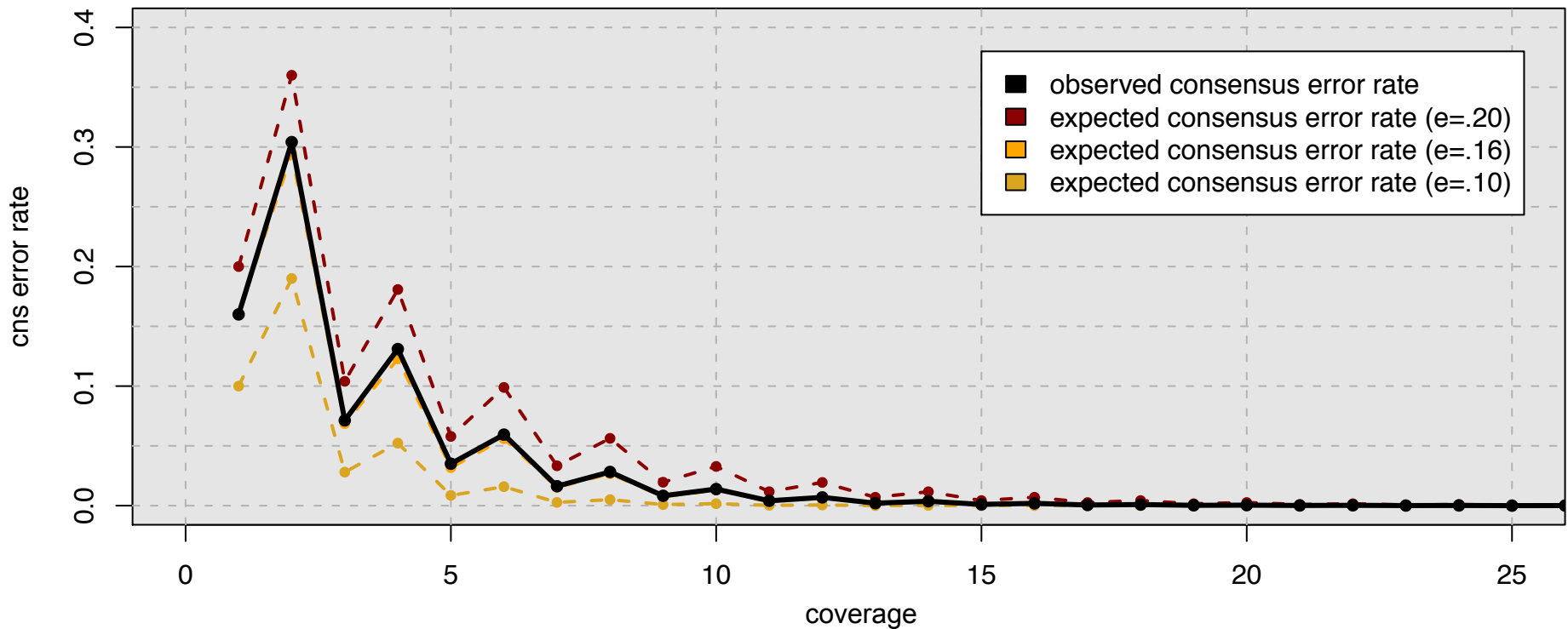
9 Jan

SMRT Sequencing

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



Consensus Accuracy and Coverage



Coverage can overcome random errors

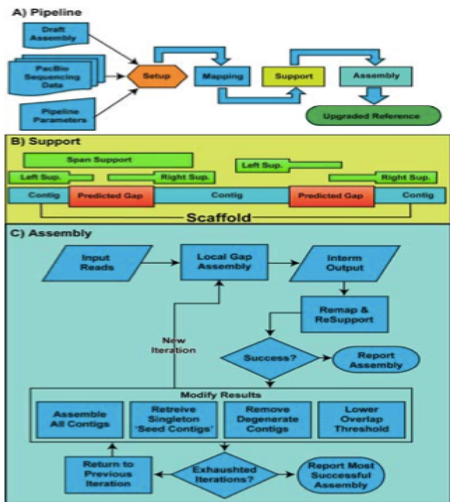
- Dashed: error model from binomial sampling
- Solid: observed accuracy

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

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

PacBio Assembly Algorithms

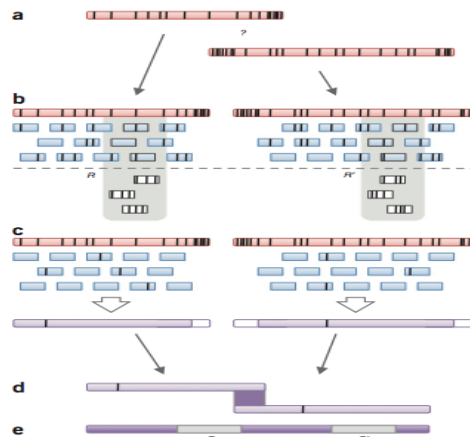
PBJelly



Gap Filling and Assembly Upgrade

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

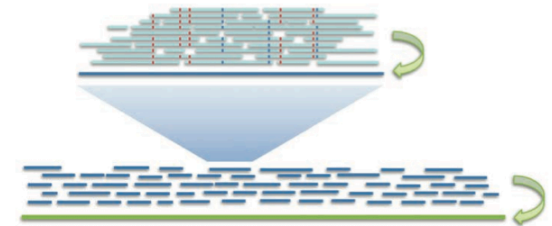
PacBioToCA & ECTools



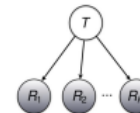
Hybrid/PB-only Error Correction

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

HGAP & Quiver



$$\Pr(\mathbf{R} | T) = \prod_k \Pr(R_k | T)$$



Quiver Performance Results Comparison to Reference Genome (<i>M. ruber</i> ; 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

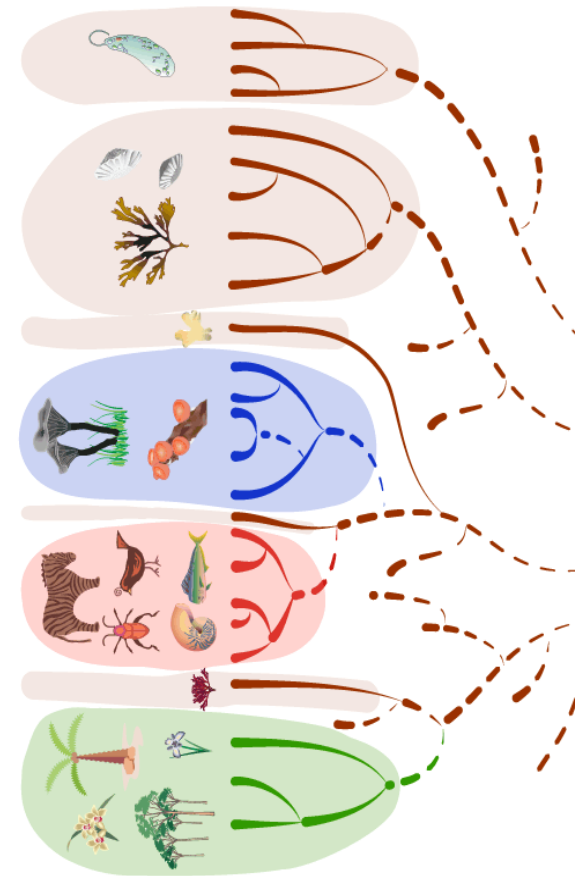
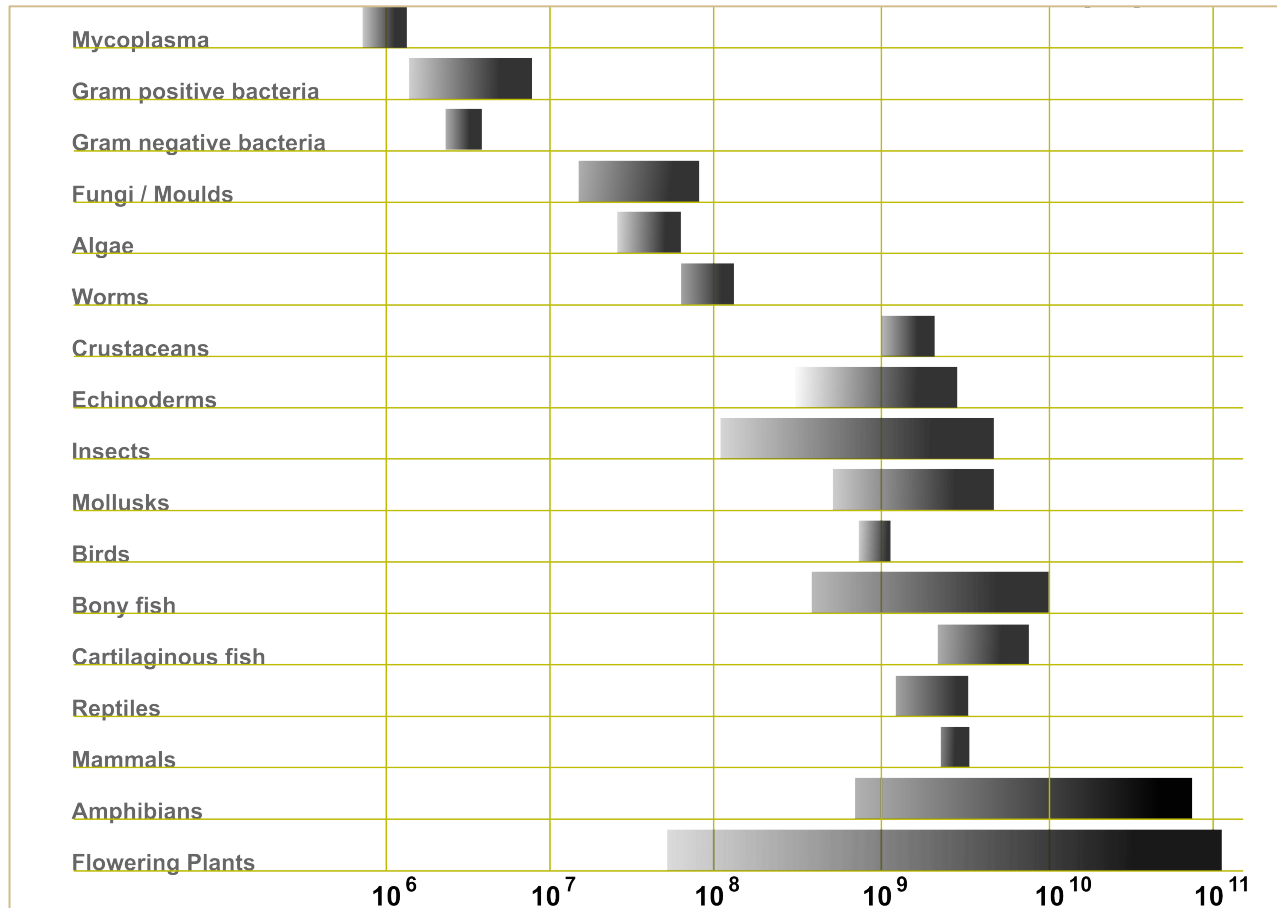
Chin *et al* (2013)
Nature Methods. 10:563–569

< 5x

PacBio Coverage

> 50x

What should we expect from an assembly?

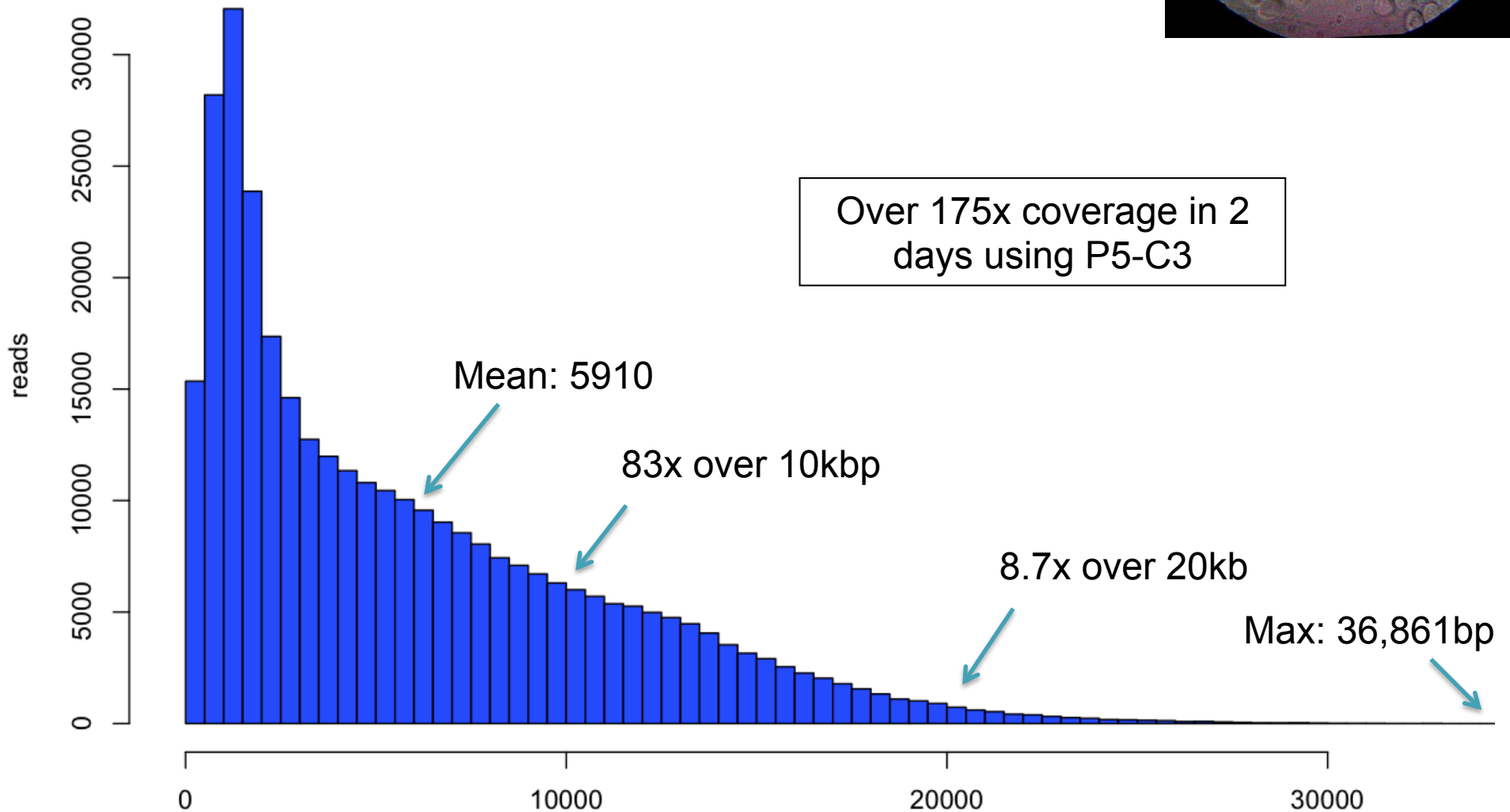
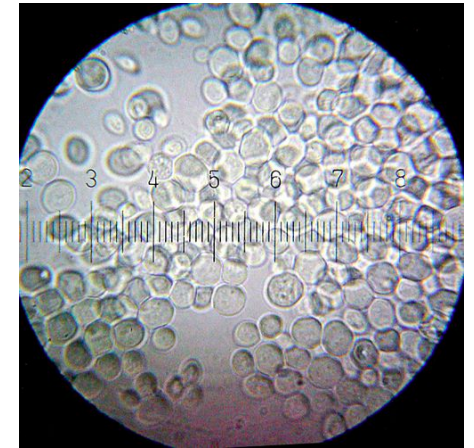


https://en.wikipedia.org/wiki/Genome_size

S. cerevisiae W303

PacBio RS II sequencing at CSHL by Dick McCombie

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



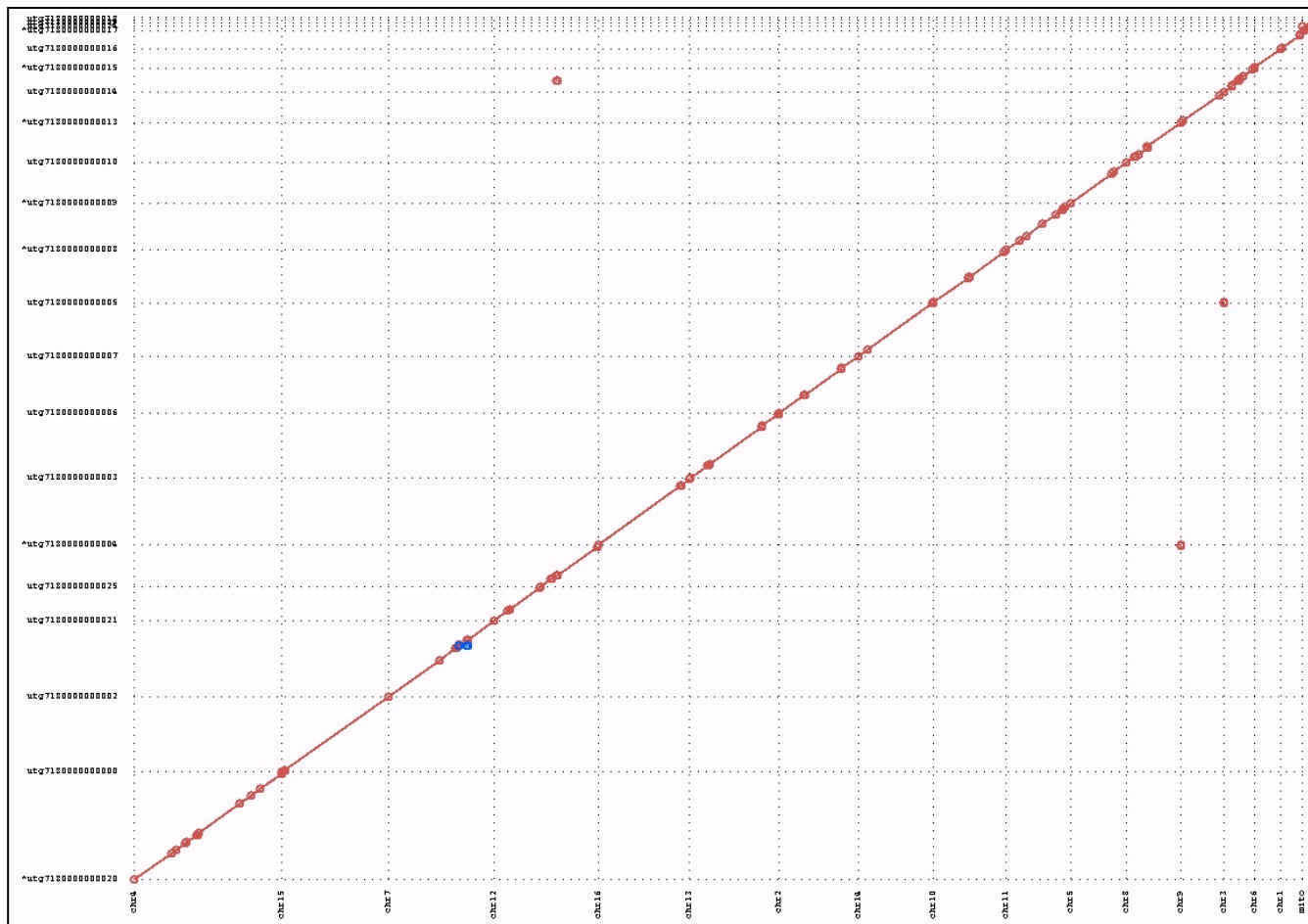
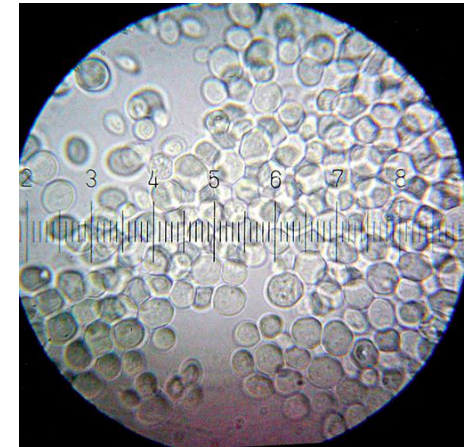
S. cerevisiae W303

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



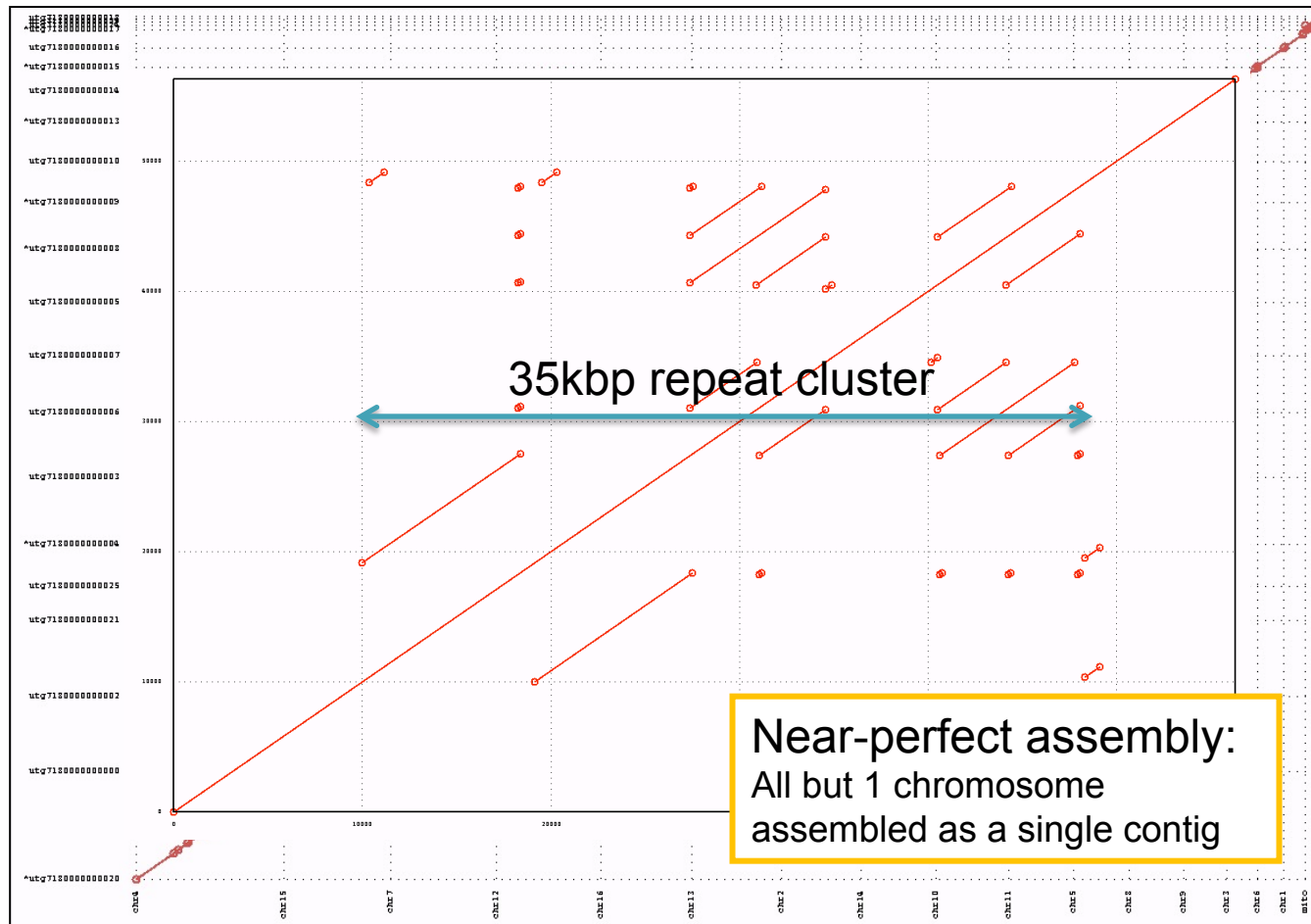
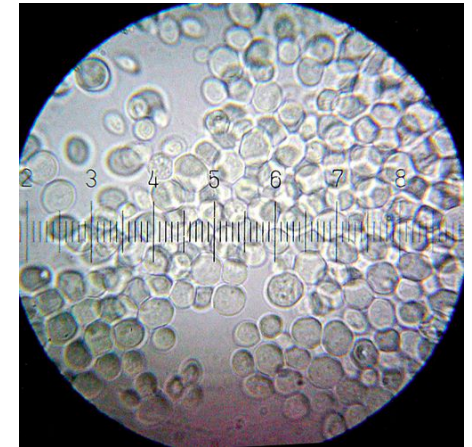
S. cerevisiae W303

S288C Reference sequence

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

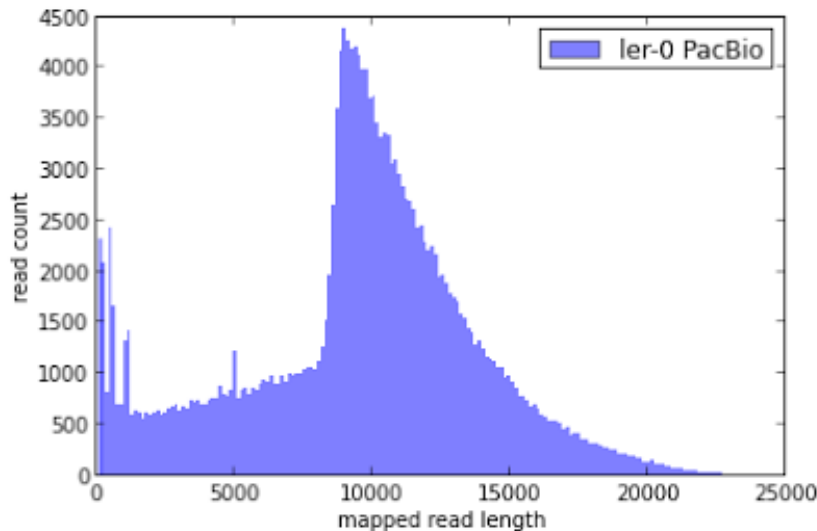
PacBio assembly using HGAP + Celera Assembler

- 12.4Mbp; 21 non-redundant contigs; N50: 811kbp; >99.8% 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
Chromosome N50: 23.0 Mbp
Corrected coverage: 20x over 10kb

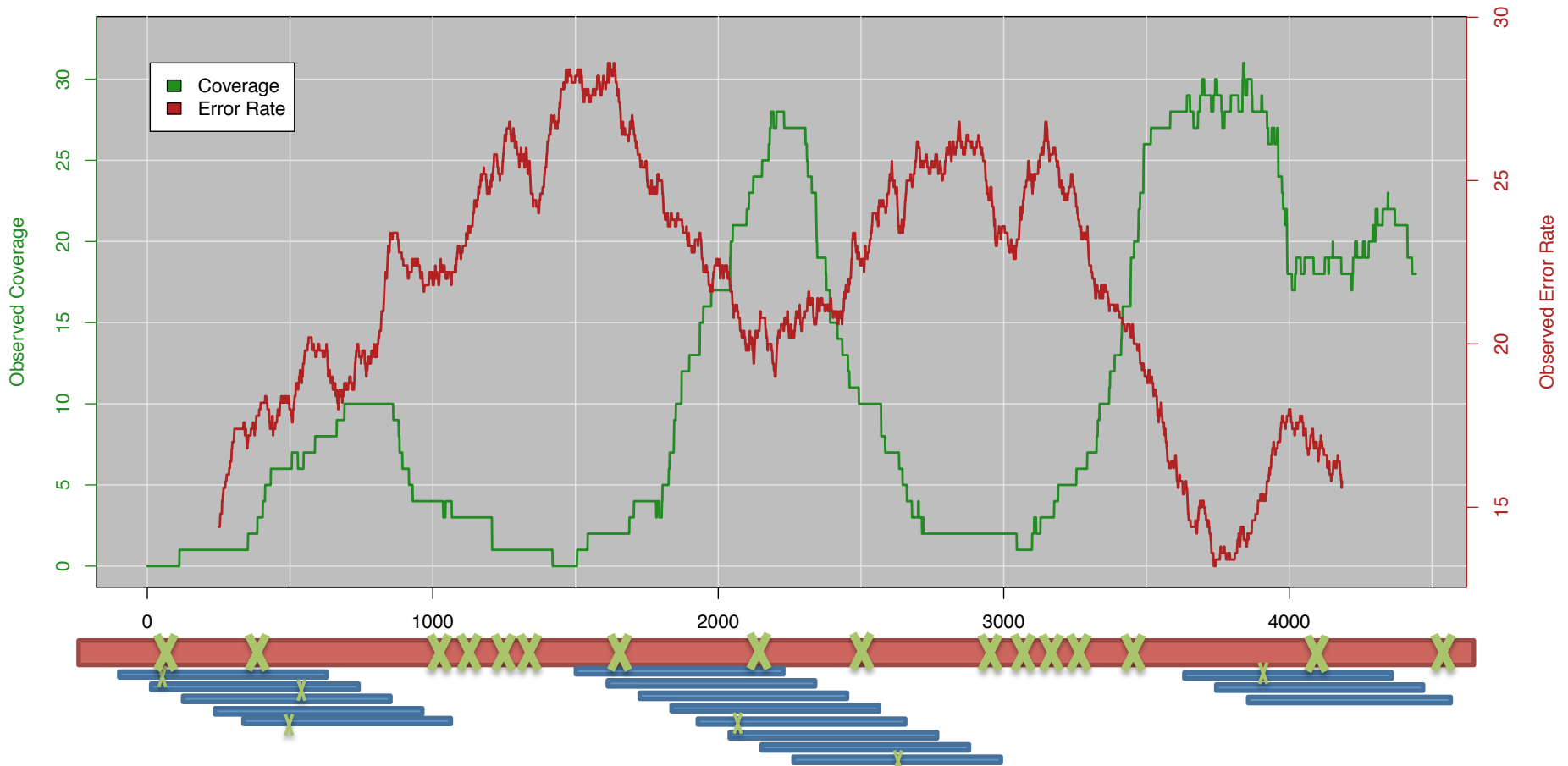
Sum of Contig Lengths: 149.5Mb
N50 Contig Length: 8.4 Mb
Number of Contigs: 1788

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

Hybrid Approaches for Larger Genomes

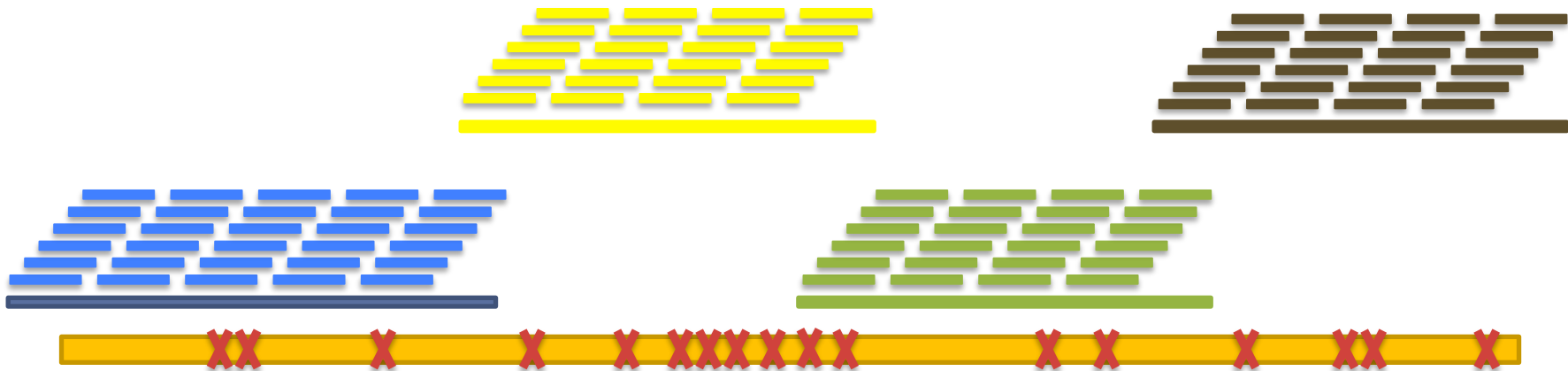
PacBioToCA fails in complex regions

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

O. sativa pv Nipponbare

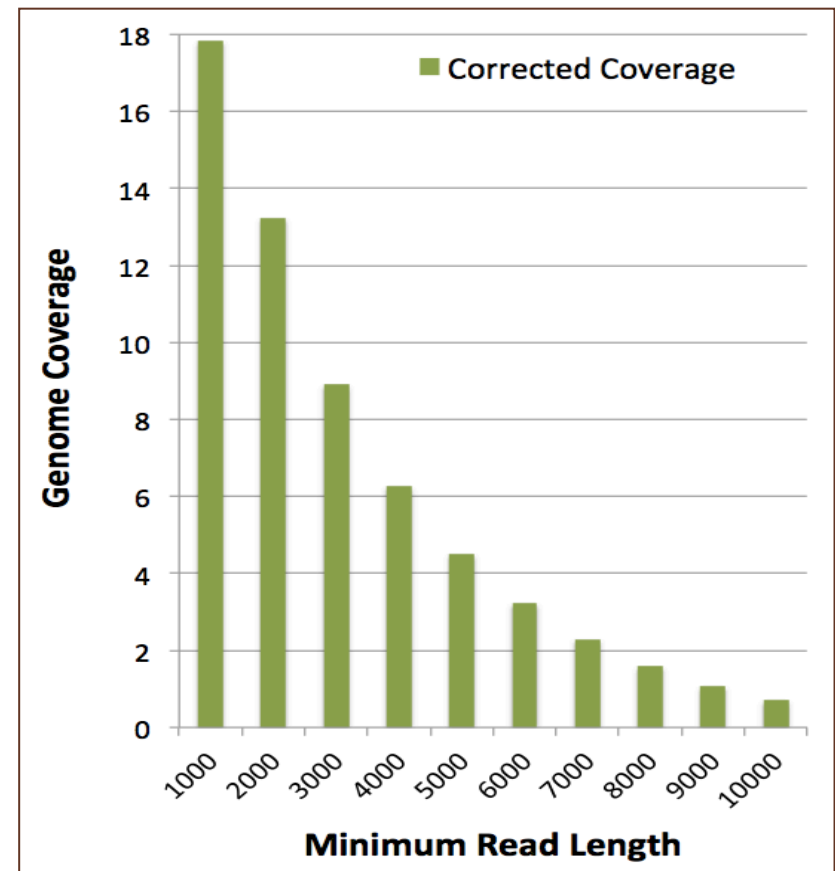
Genome size: 370 Mb

Chromosome N50: 29.7 Mbp

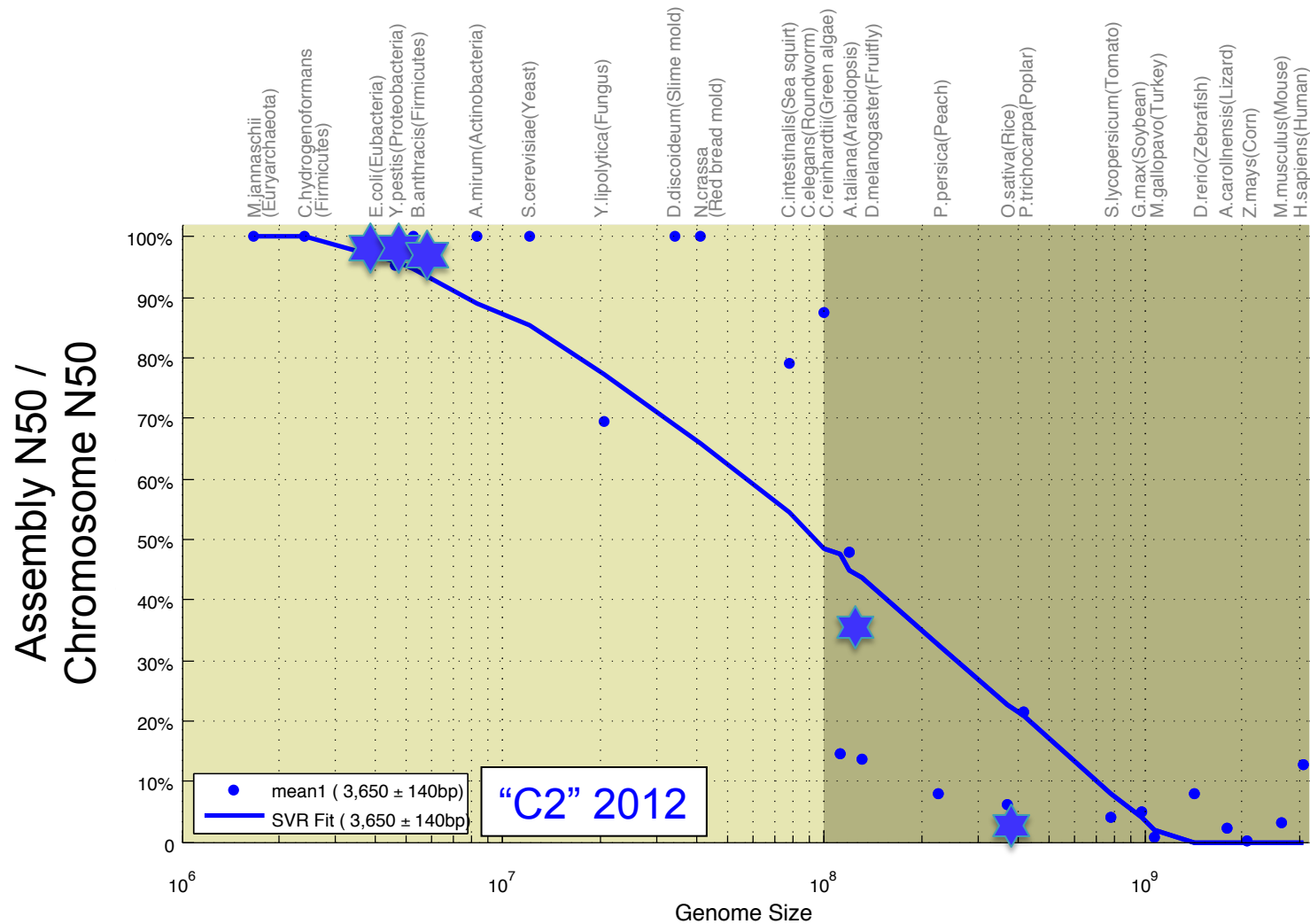
19x PacBio C2XL sequencing at CSHL from Summer 2012



Assembly	Contig NG50
MiSeq Fragments 23x 459bp 8x 2x251bp @ 450	6,332
“ALLPATHS-recipe” 50x 2x100bp @ 180 36x 2x50bp @ 2100 51x 2x50bp @ 4800	18,248
PacBioToCA 19x @ 3500 ** MiSeq for correction	50,995
ECTools 19x @ 3500 ** MiSeq for correction	155,695



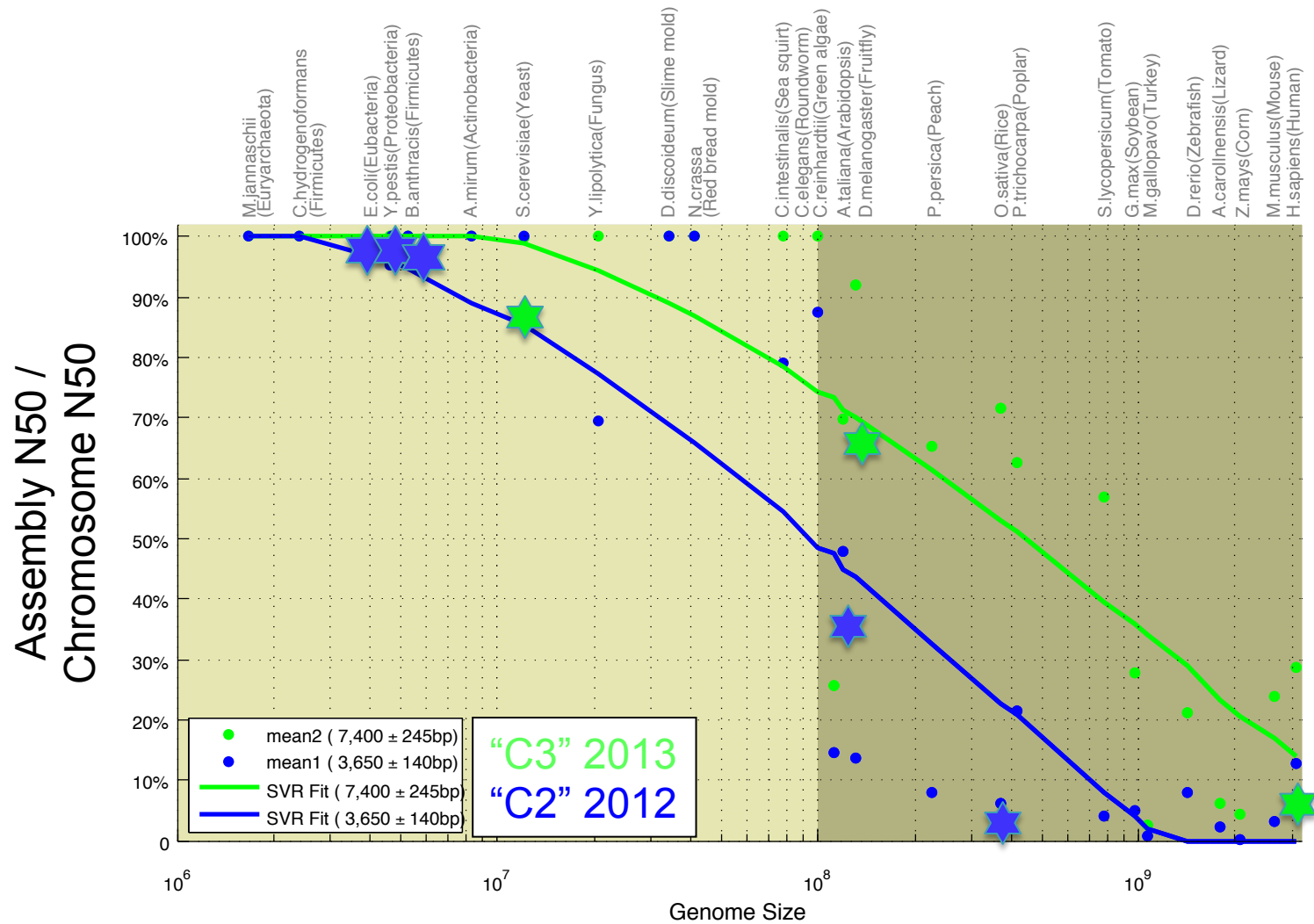
Assembly Complexity of Long Reads



Assembly complexity of long read sequencing

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

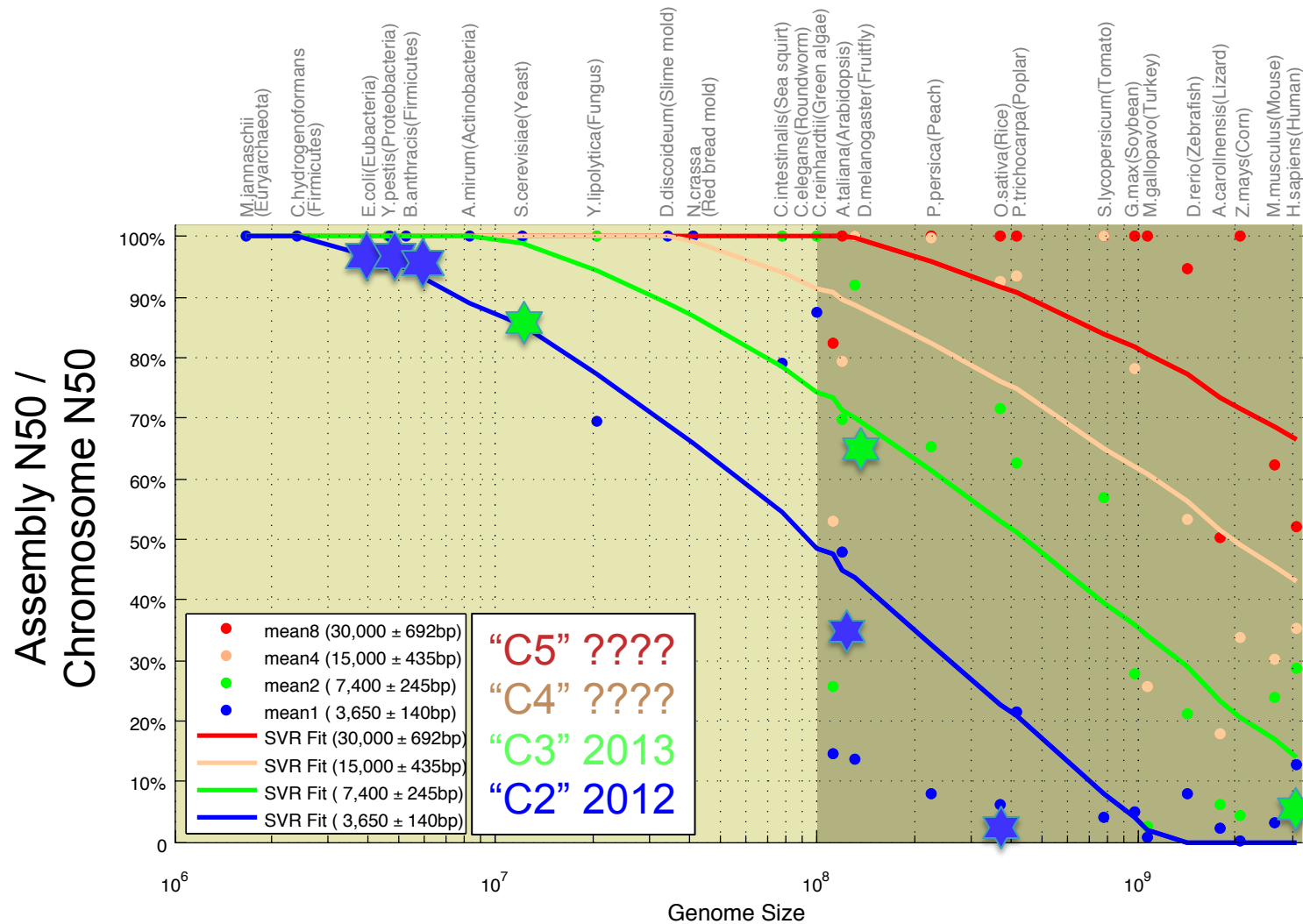
Assembly Complexity of Long Reads



Assembly complexity of long read sequencing

Lee, H*, Gurtowski, J*, Yoo, S, Marcus, S, McCombie, WR, Schatz MC et al. (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 et al. (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

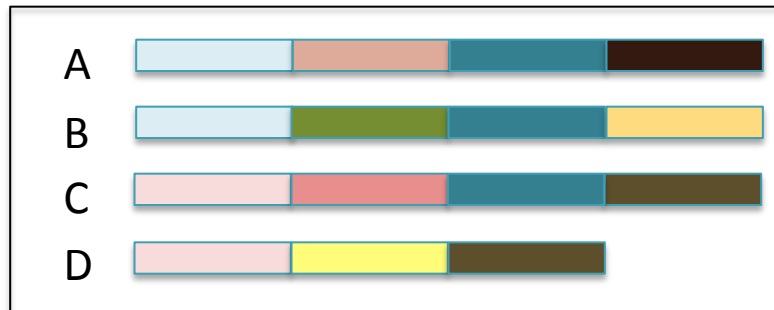
- > 1GB: hybrid/gap filling
expect contig N50 to be 100kbp – 1Mbp

- > 5GB: Email mschatz@cshl.edu

- **Caveats**

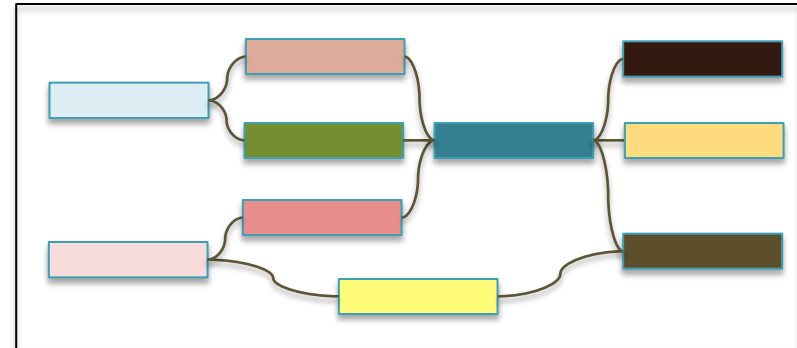
- Model only as good as the available references (esp. haploid sequences)
 - Technologies are quickly improving, exciting new scaffolding technologies

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

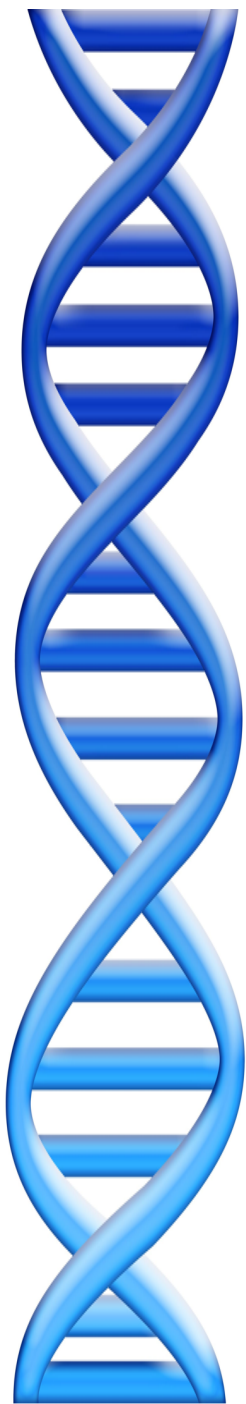
- Encodes all the sequence relationships between the genomes
- How well conserved is a given sequence?
- What are the pan-genome network properties?

Rapid pan genome analysis with augmented suffix trees

Marcus, S, Schatz, MC (2014) *In preparation*

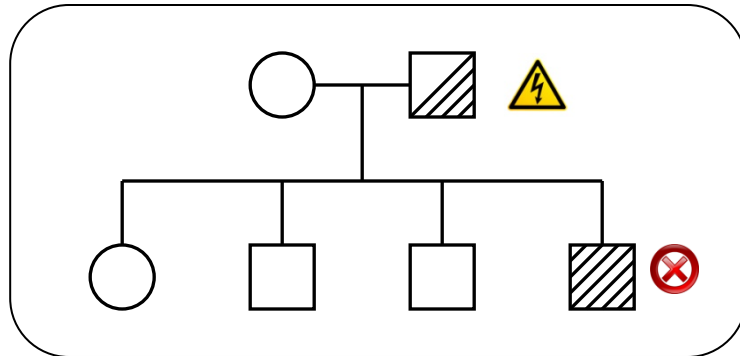
Outline

1. Biological Data Science
2. De novo genome assembly
3. **Disease Analytics**



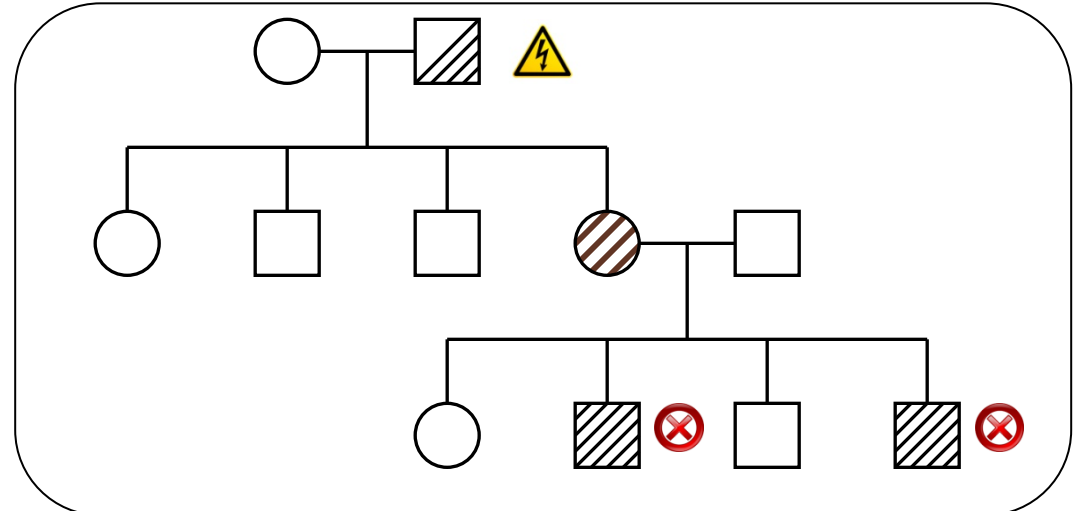
Unified Model of Autism

Sporadic Autism: 1 in 100



Prediction: De novo mutations of high penetrance contributes to autism, especially in low risk families with no history of autism.

Familial Autism: 90% concordance in twins



Legend



Sporadic mutation



Fails to procreate

A unified genetic theory for sporadic and inherited autism

Zhao *et al.* (2007) *PNAS*. 104(31)12831-12836.

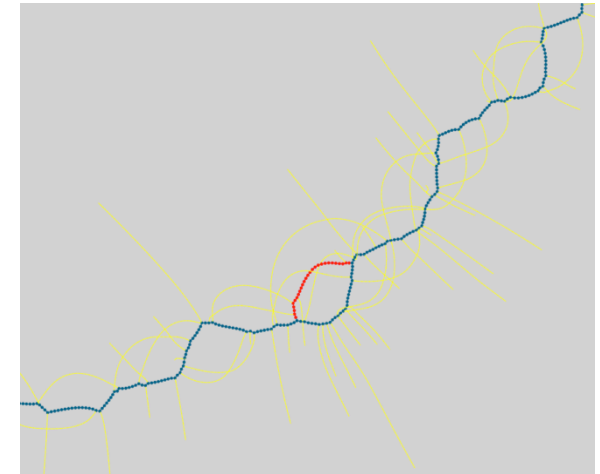
Scalpel: Haplotype Microassembly

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**



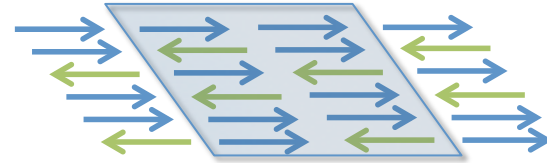
NRXN1 *de novo* SNP
(auSSC12501 chr2:50724605)

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

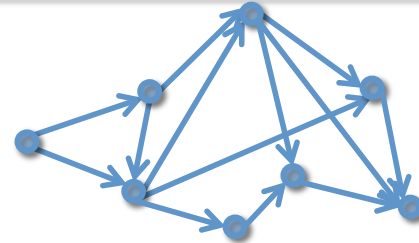
Narzisi, G, O’Rawe, J, Iossifov, I, Lee, Y, Wang, Z, Wu, Y, Lyon, G, Wigler, M, Schatz, MC (2014) *Under review.*

Scalpel Pipeline

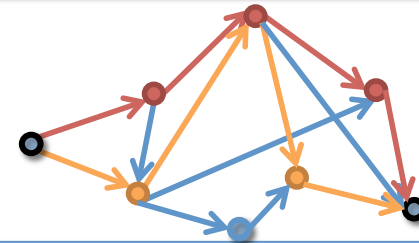
Extract reads mapping within the exon including (1) well-mapped reads, (2) soft-clipped reads, and (3) anchored pairs



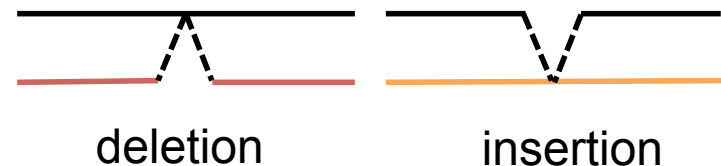
Decompose reads into overlapping k -mers and construct de Bruijn graph from the reads



Find end-to-end haplotype paths spanning the region

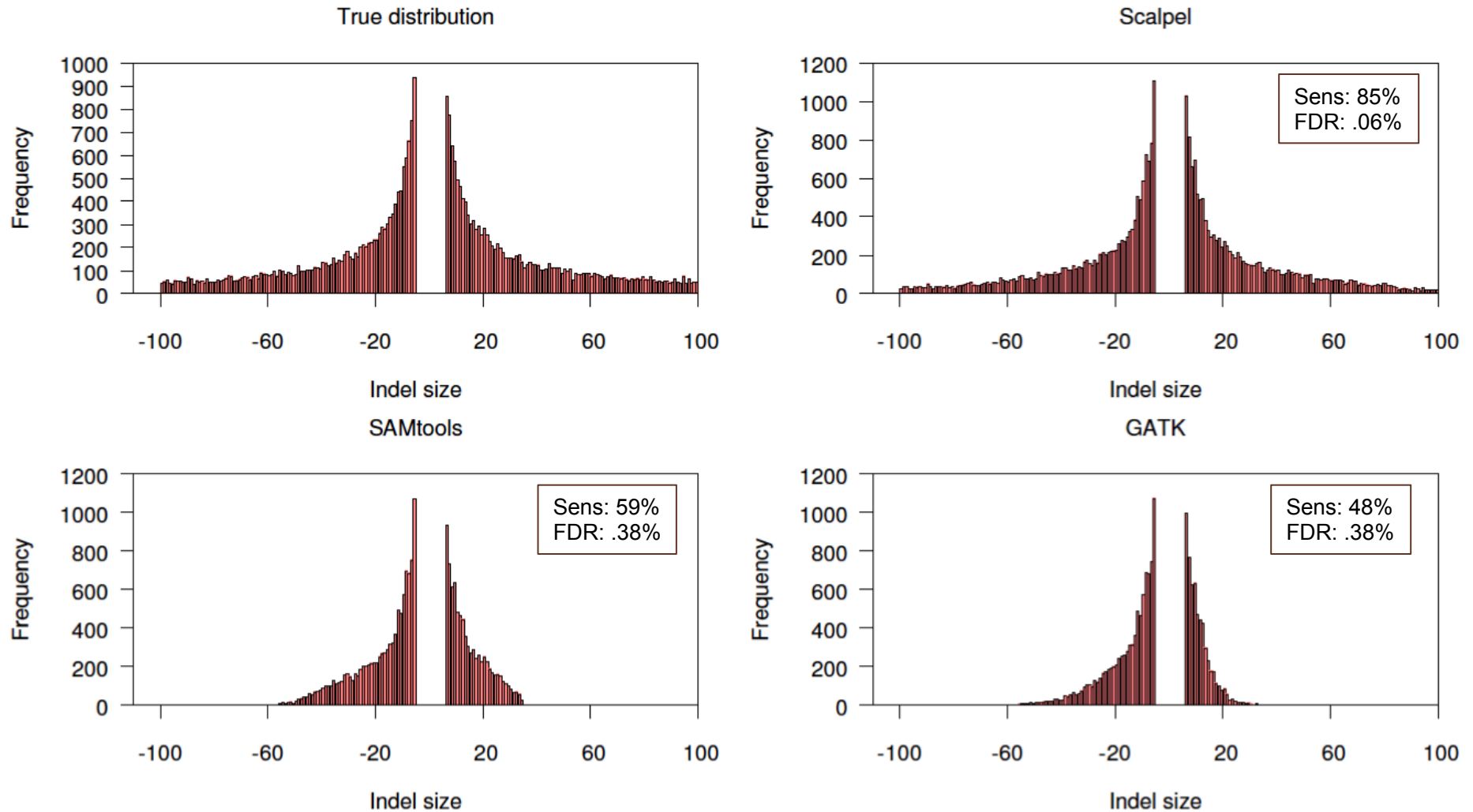


Align assembled sequences to reference to detect mutations



Simulation Analysis

Indel size distribution (length > 5 bp)



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

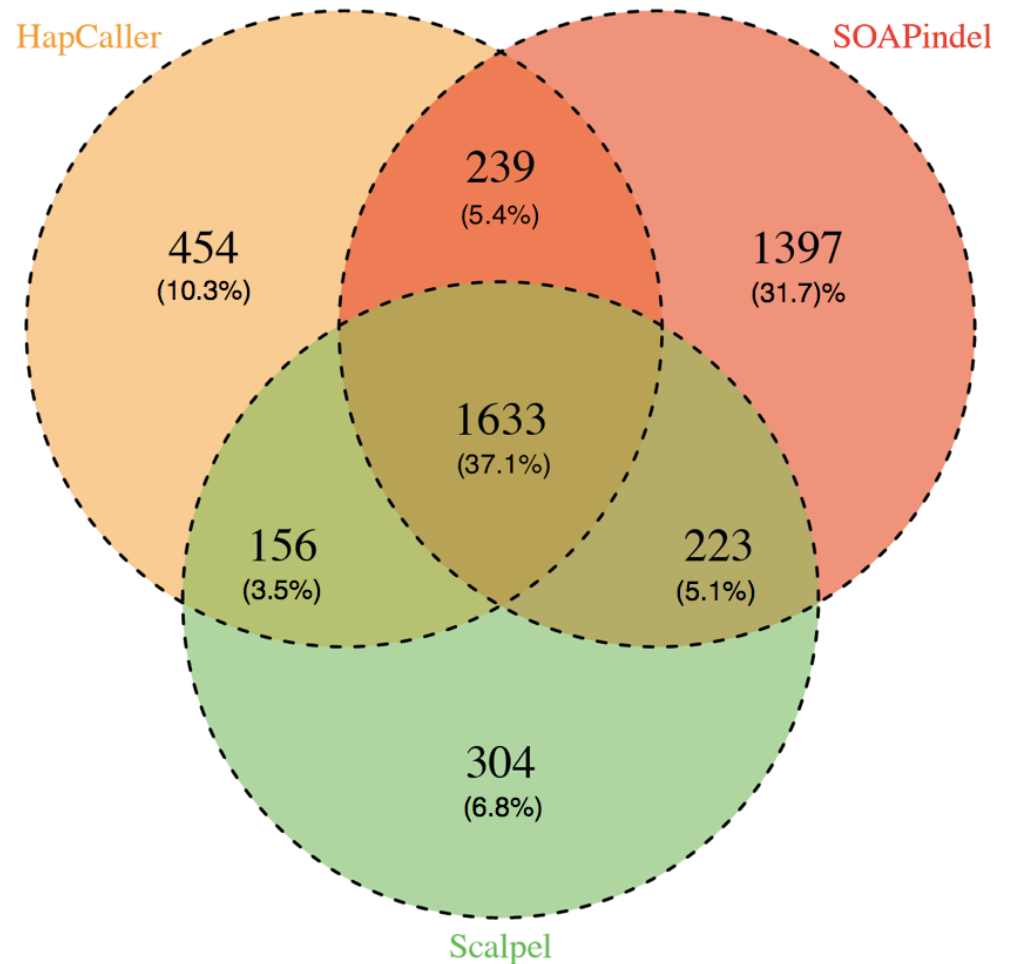
Experimental Analysis & Validation

Selected one deep coverage exome for deep analysis

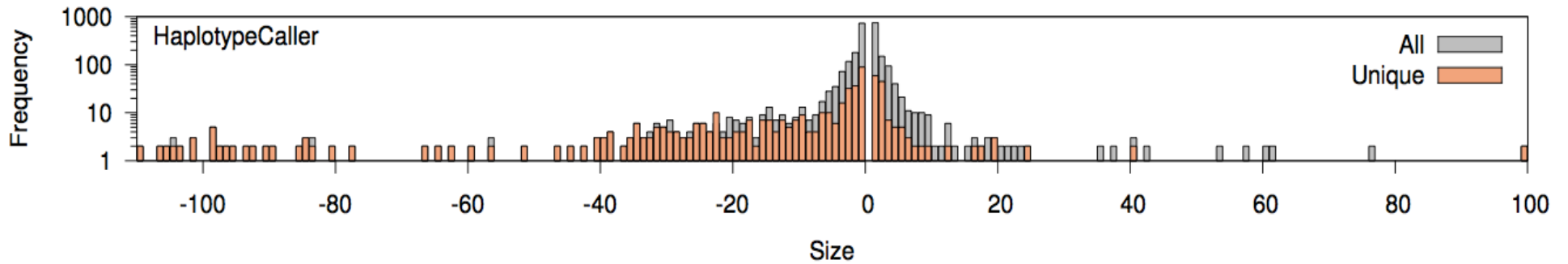
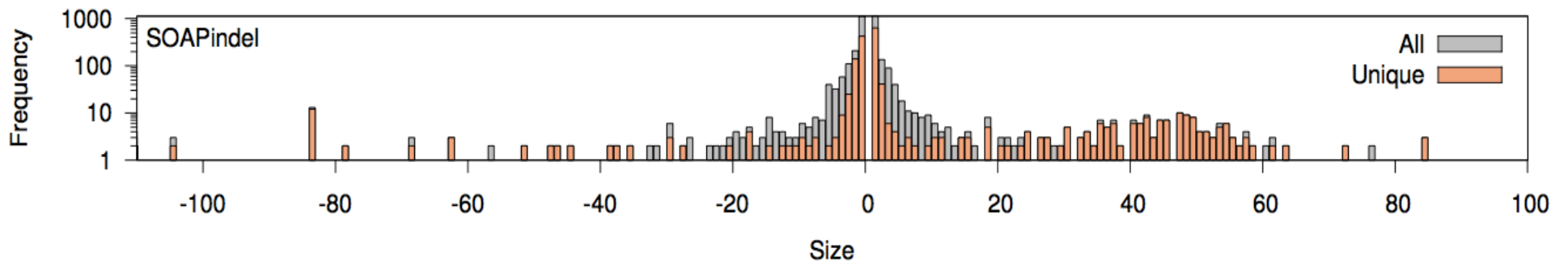
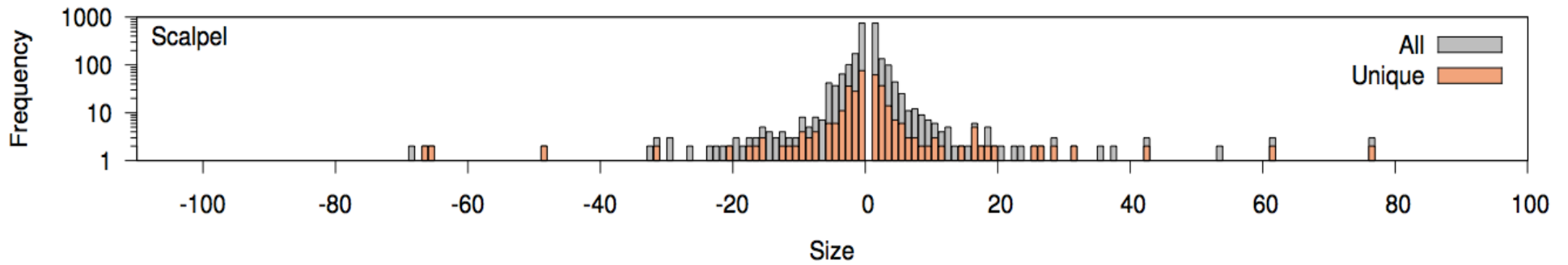
- Individual was diagnosed with ADHD and turrets syndrome
- 80% of the target at >20x coverage
- Evaluated with Scalpel, SOAPindel, and GATK Haplotype Caller

1000 indels selected for validation

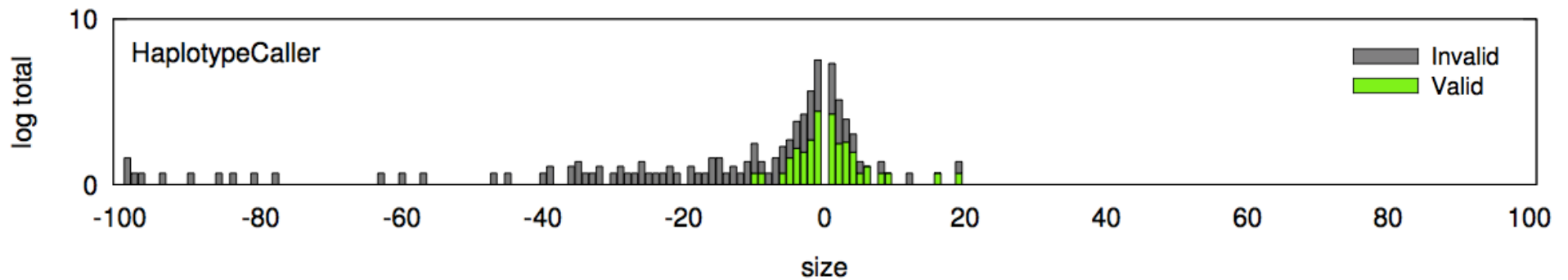
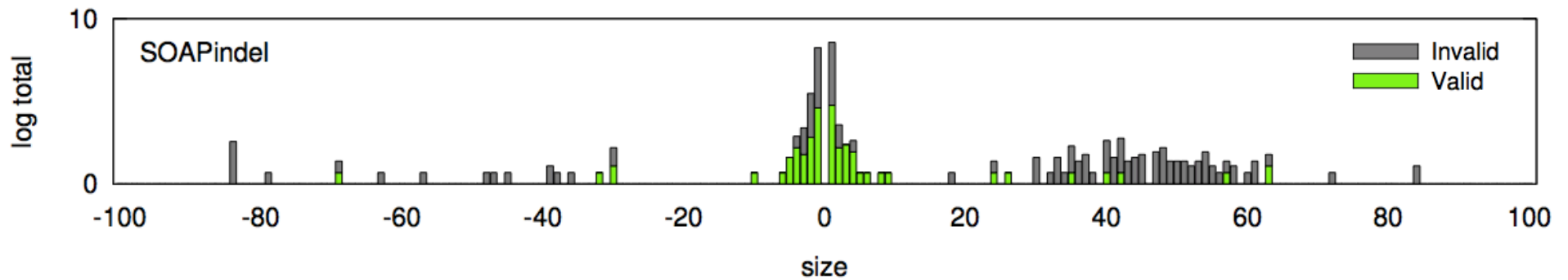
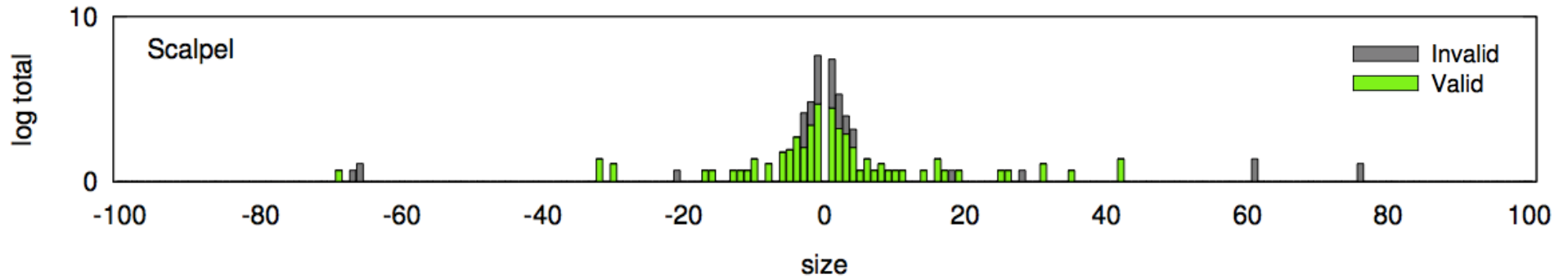
- 200 Scalpel
- 200 GATK Haplotype Caller
- 200 SOAPindel
- 200 within the intersection
- 200 long indels (>30bp)



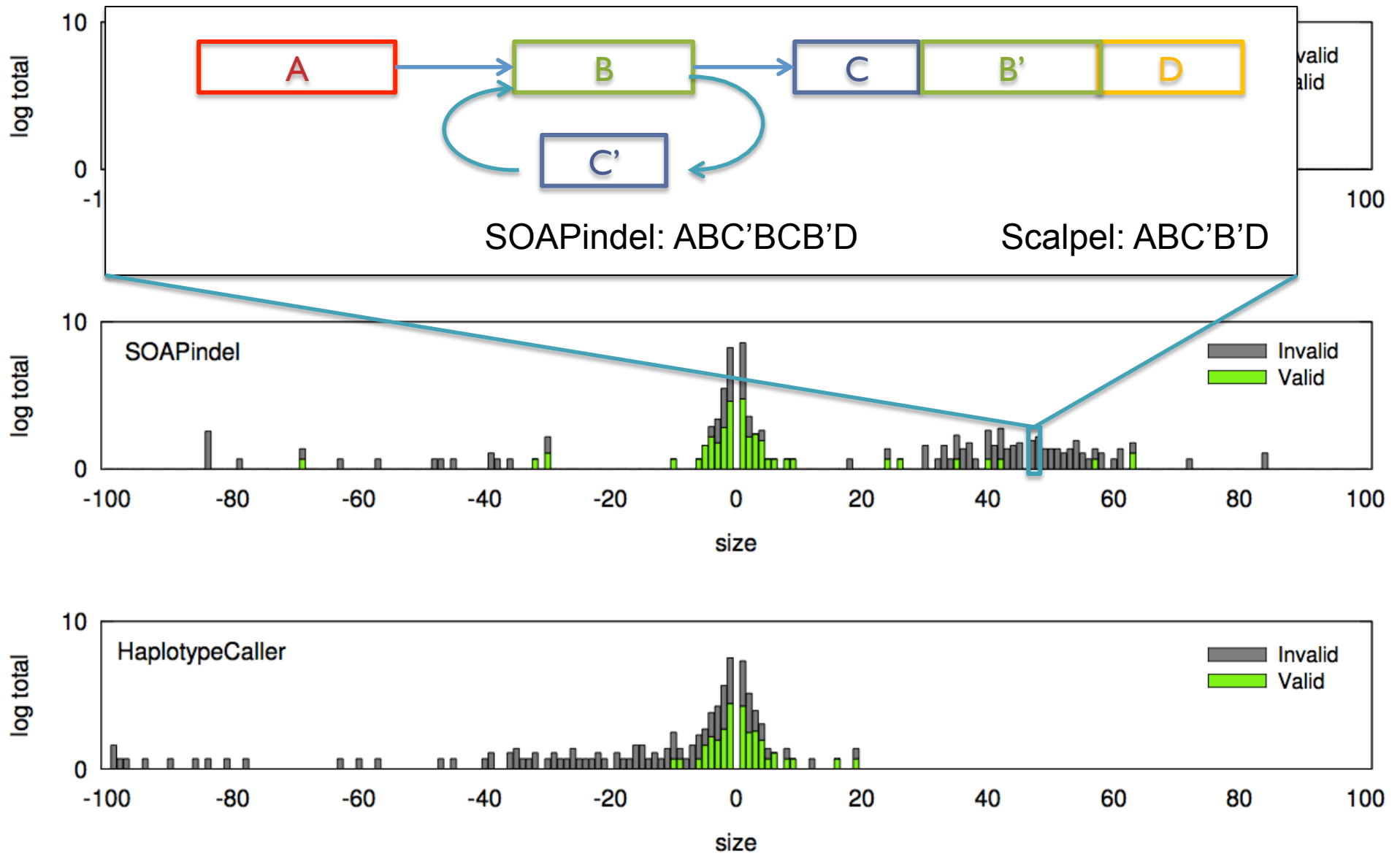
Scalpel Indel Discovery



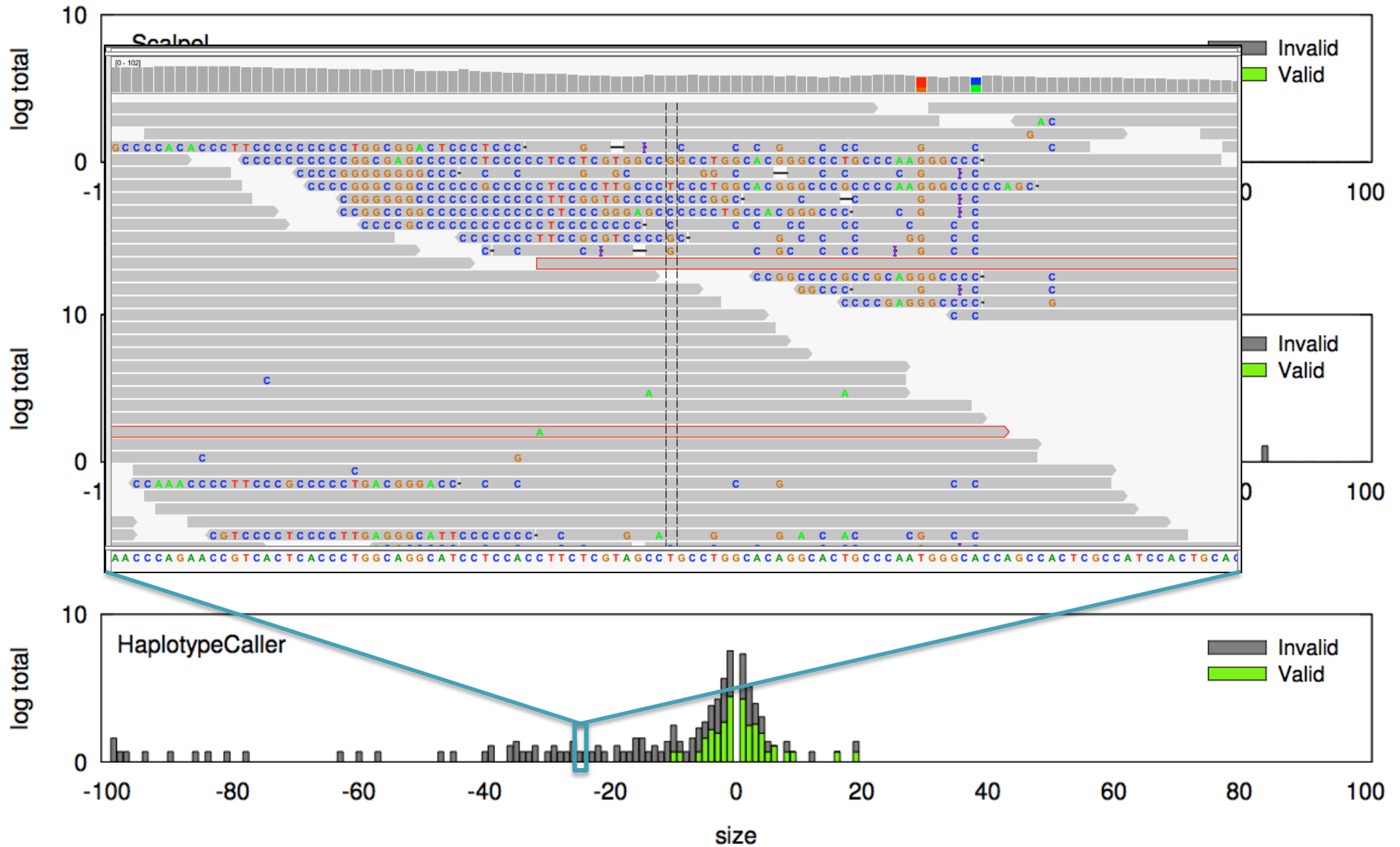
Scalpel Indel Discovery



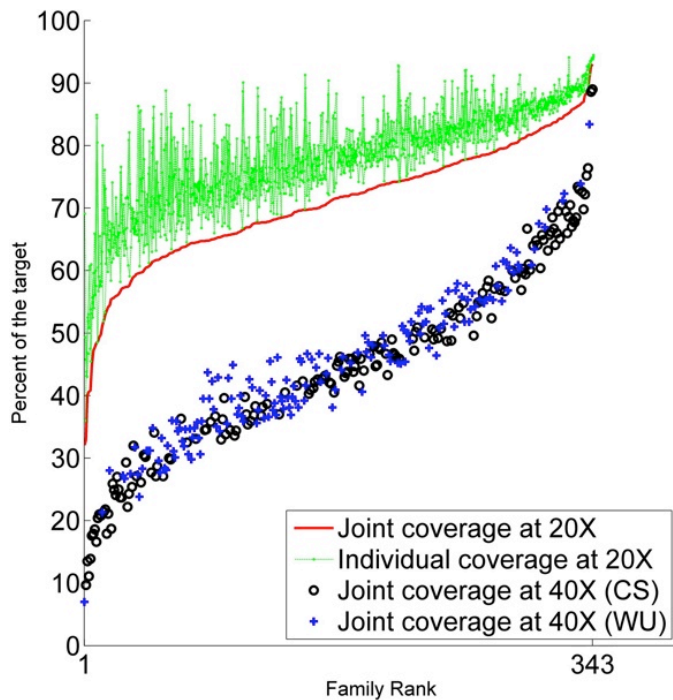
Scalpel Indel Discovery



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
- Iossifov (343) and O’Roak (50) used GATK, Sanders (200) didn’t attempt to identify indels
- All attempted to find “gene killing mutations” specific to the autistic children to find genes associated with the disease

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.

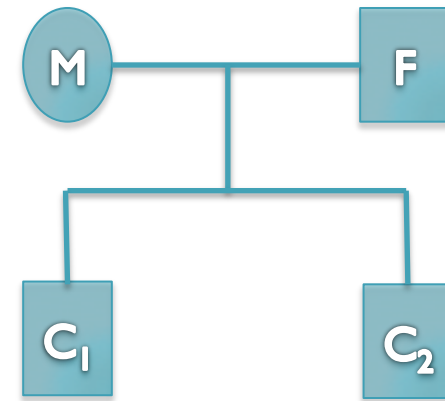
Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations

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



Ref: ...TCAGAACAGCTGGATGAGATCTTAGCCAACCTACCAGGAGATTGTCTTTGCCCGGA...

Father: ...TCAGAACAGCTGGATGAGATCTTAGCCAACCTACCAGGAGATTGTCTTTGCCCGGA...

Mother: ...TCAGAACAGCTGGATGAGATCTTAGCCAACCTACCAGGAGATTGTCTTTGCCCGGA...

Sib: ...TCAGAACAGCTGGATGAGATCTTAGCCAACCTACCAGGAGATTGTCTTTGCCCGGA...

Aut(1): ...TCAGAACAGCTGGATGAGATCTTAGCCAACCTACCAGGAGATTGTCTTTGCCCGGA...

Aut(2): ...TCAGAACAGCTGGATGAGATCTTACC-----CCGGGAGATTGTCTTTGCCCGGA...

6bp heterozygous deletion at chr13:25280526 ATP12A

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 the 842 genes known to be associated with fragile X protein FMR1
 - Related to neuron development and synaptic plasticity
 - Also strong overlap with chromatin remodelers

De novo gene disruptions in children on the autism spectrum

Iossifov *et al.* (2012) *Neuron*. 74:2 285-299

Summary

New Biotechnology

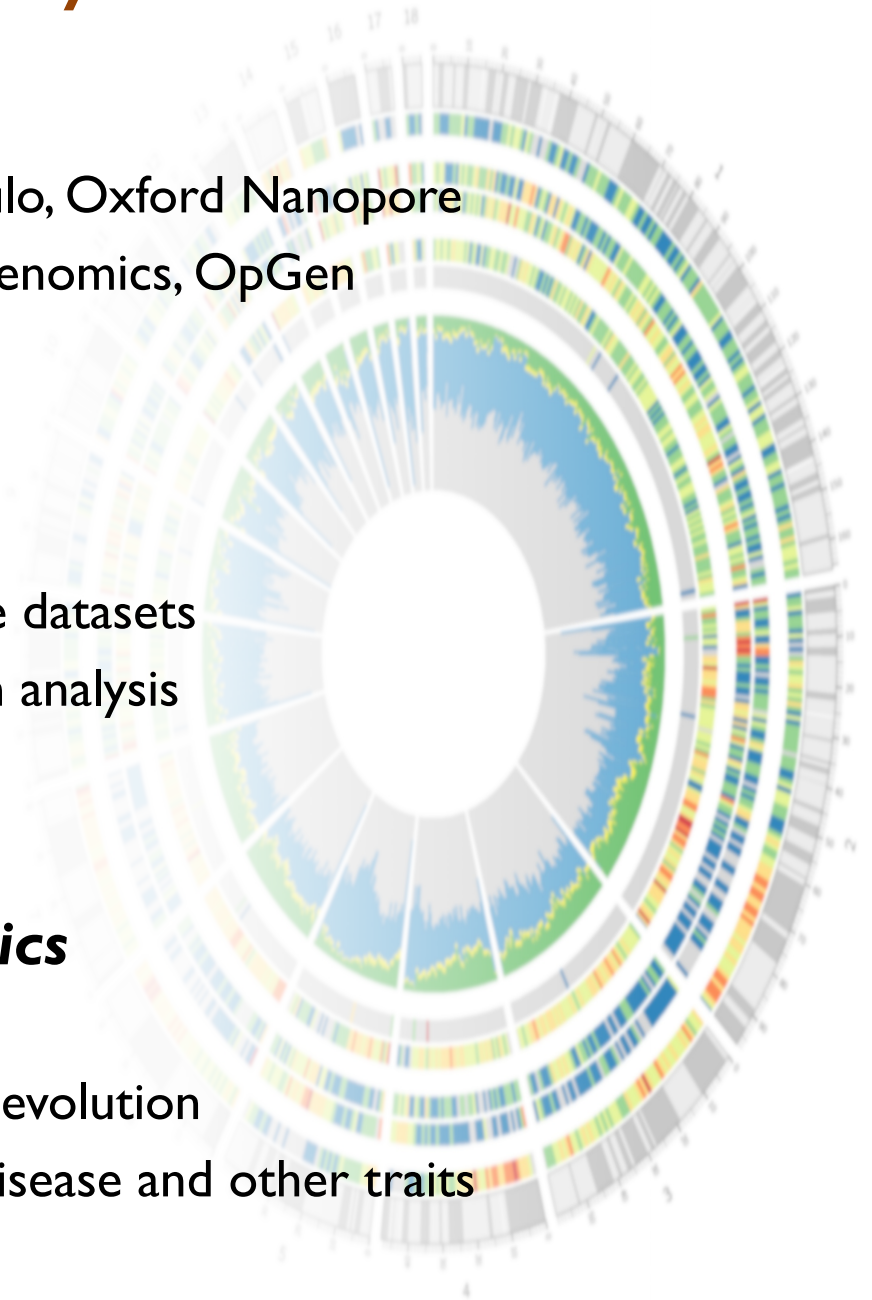
- Sequencing: Pacific Biosciences, MolecuLo, Oxford Nanopore
- Mapping: Hi-C interactions, BioNanoGenomics, OpGen
- Faster/Cheaper/Better assemblies

Algorithmics

- Indexing and compressing of very large datasets
- Improved error correction, large graph analysis
- Analyzing populations of genomes

Annotation & Comparative Genomics

- Identifying functional elements
- Cross species comparisons, models of evolution
- Identifying mutations responsible for disease and other traits



Acknowledgements

Schatz Lab

Giuseppe Narzisi
Shoshana Marcus
James Gurtowski
Srividya
Ramakrishnan
Hayan Lee
Rob Aboukhalil
Mitch Bekritsky
Charles Underwood
Tyler Gavin
Alejandro Wences
Greg Vurture
Eric Biggers
Aspyn Palatnick

CSHL

Hannon Lab
Gingeras Lab
Jackson Lab
Iossifov Lab
Levy Lab
Lippman Lab
Lyon Lab
Martienssen Lab
McCombie Lab
Tuveson Lab
Ware Lab
Wigler Lab
IT Department



National Human
Genome Research
Institute



U.S. DEPARTMENT OF
ENERGY



Thank you

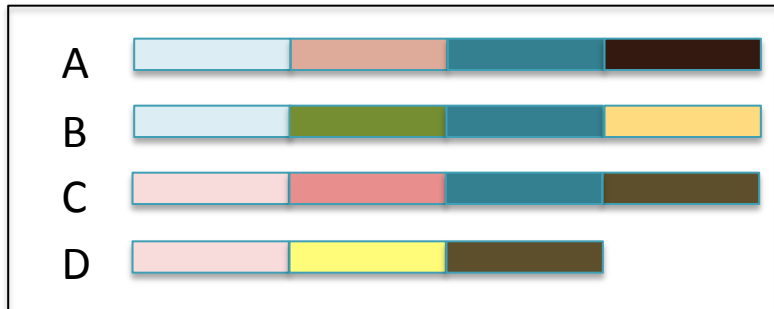
<http://schatzlab.cshl.edu>

@mike_schatz

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

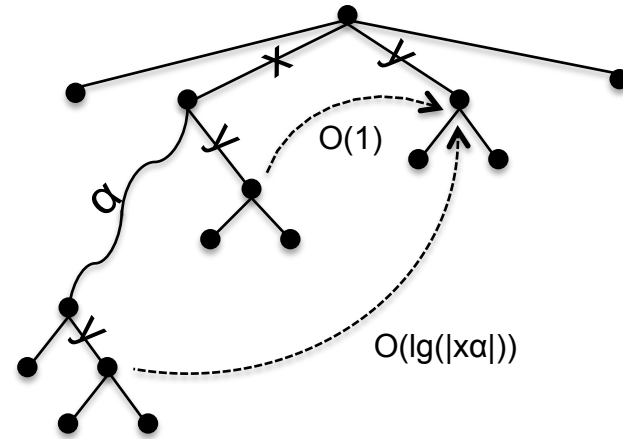


Pan-Genome Alignment & Assembly



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



Align the genomes using a suffix tree augmented with “suffix skips”

- Similar to suffix links, but navigate between distant suffixes in $O(\lg |p|)$
- Uses pointer doubling techniques to rapidly add additional links

Rapid pan genome analysis with augmented suffix trees

Marcus, S, Schatz, MC (2014) *In preparation*