

# Algorithms for genome sequencing and disease analytics

Michael Schatz

Sept 9, 2014

CSHL Special Seminar



# Unsolved Questions in Biology

The instruments provide the data, but none of the answers to any of these questions.

***What software and systems will?***

***And who will create them?***

What drugs and treatments should we give you.

• ***Plus thousands and thousands more***



# Introductions



**Tyler Garvin**

CNV and  
transcriptome  
analysis of single cells

Tomorrow @ noon!



**Srividya "Sri"  
Ramakrishnan**

DOE Systems Biology  
Knowledgebase

Worlds fastest -omics  
pipelines



**Maria Nattestad**

Hi-C Chromatin  
Interactions

Plant Assembly &  
Analysis



**Aspyn Palatnick**

Mobile Genotype  
Analysis

Flu antiviral analysis

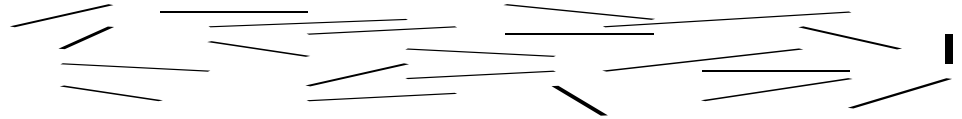


# Genome Structure & Function

- 1. Structure: Sequencing and Assembly**  
*“A tale of two sequencers”*
  
- 2. Function: Disease Analytics**
  1. Pan-genome analysis
  2. The role of indels in human diseases

# Sequencing a Genome

## 1. Shear & Sequence DNA



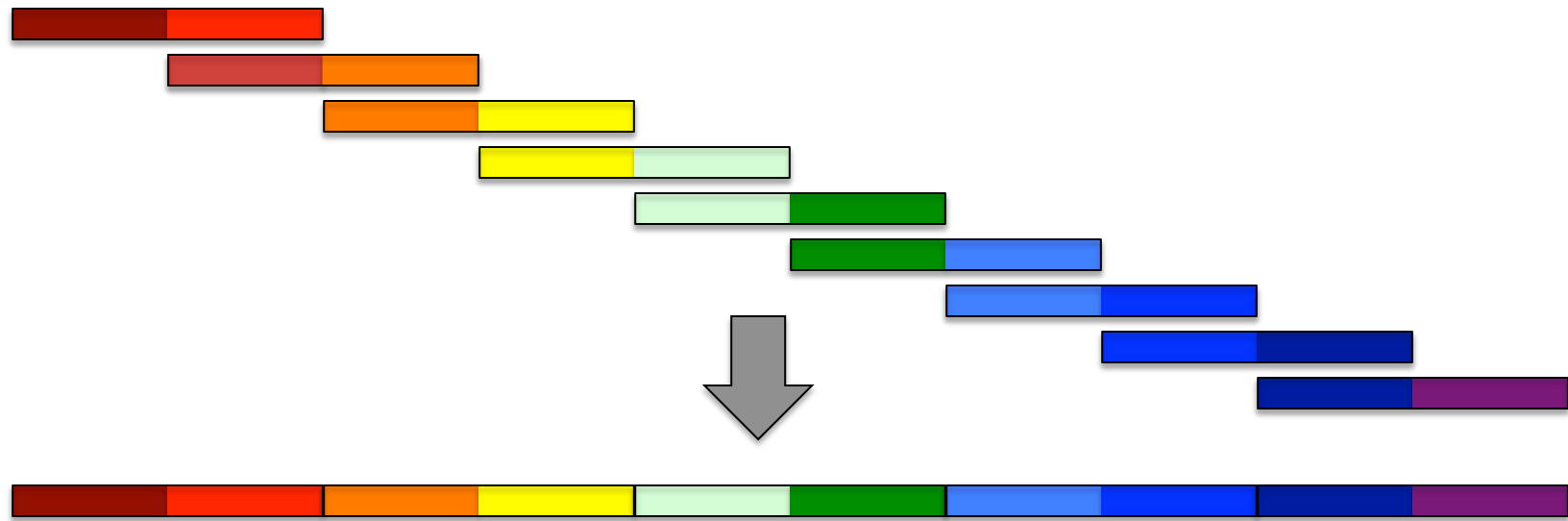
## 2. Construct assembly graph from overlapping reads

...AGCCTAGGGATGCGCGACACGT

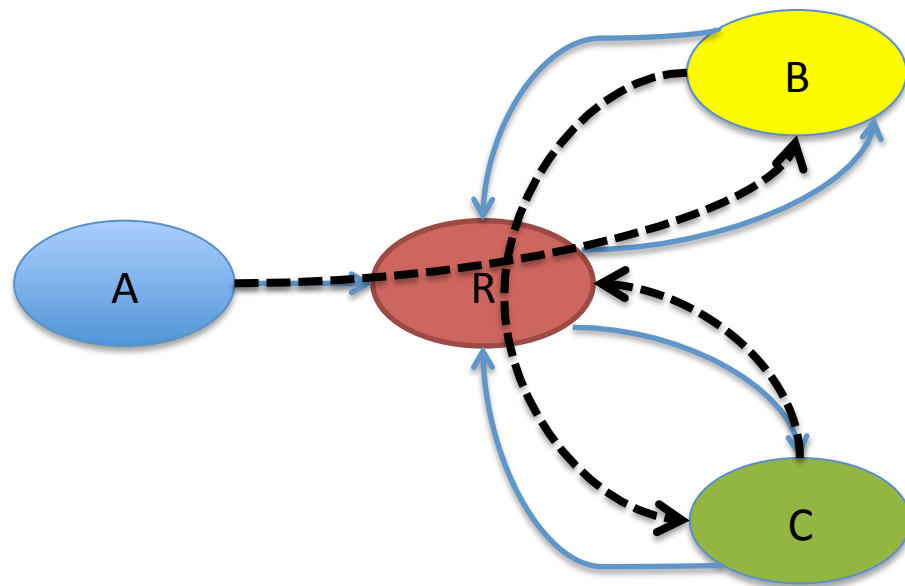
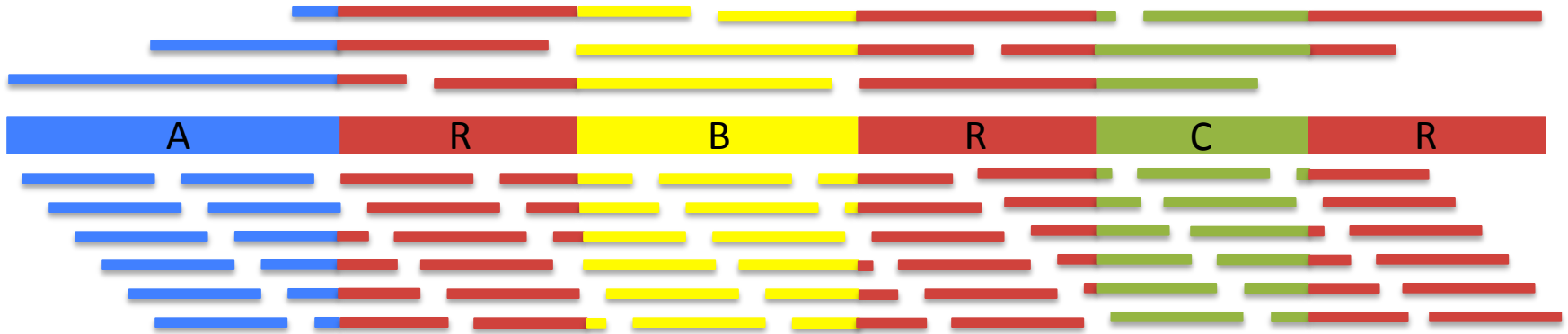
GGATGCGCGACACGTGCATATCCGGTTTGGTCAACCTCGGACGGAC

CAACCTCGGACGGACCTCAGCGAA...

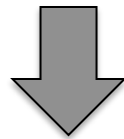
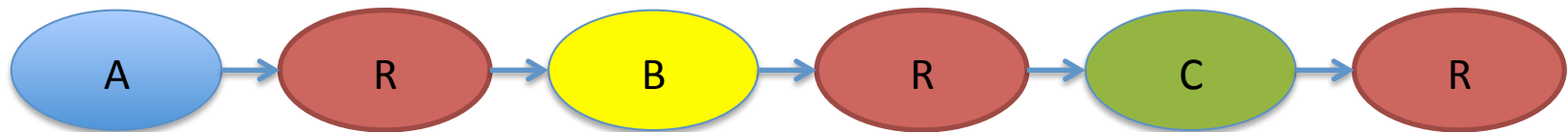
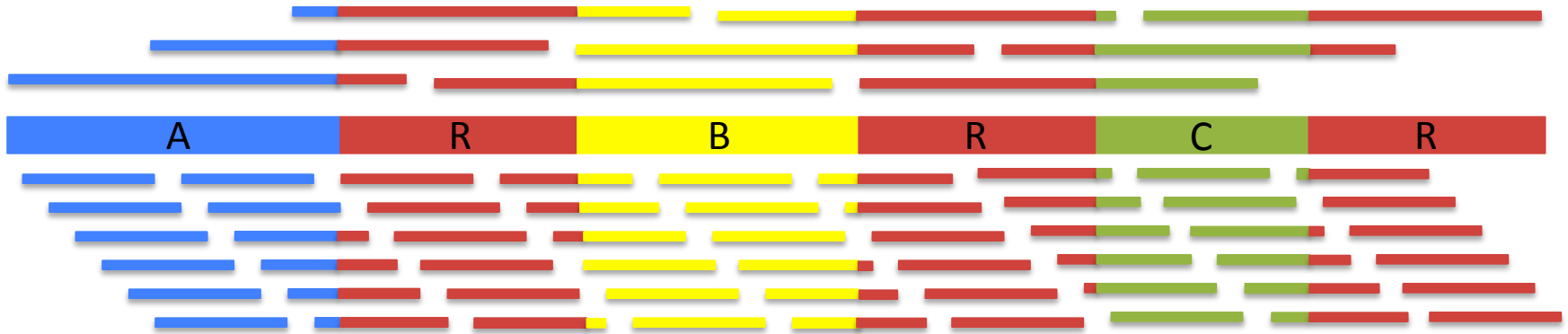
## 3. Simplify assembly graph



# Assembly Complexity



# Assembly Complexity



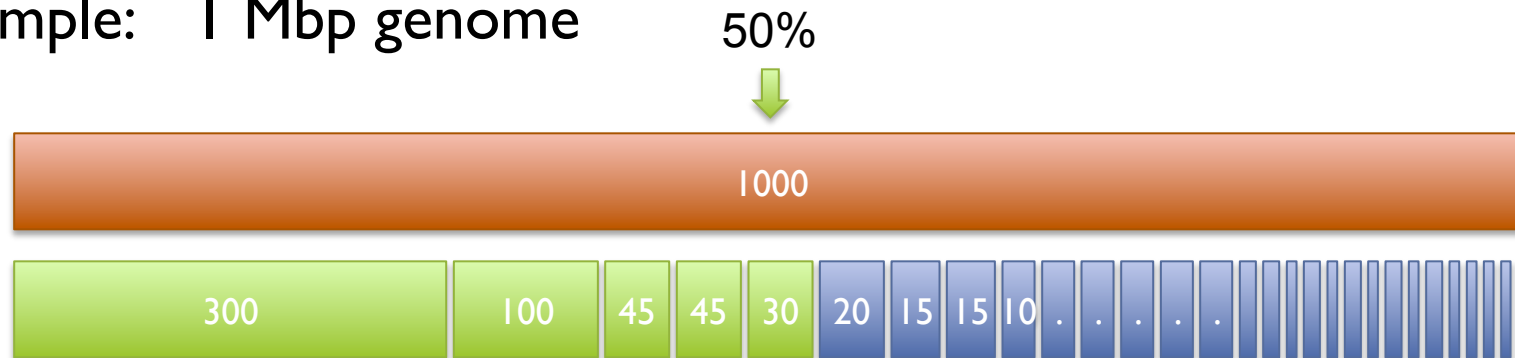
**The advantages of SMRT sequencing**

Roberts, RJ, Carneiro, MO, Schatz, MC (2013) *Genome Biology*. 14:405

# 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)

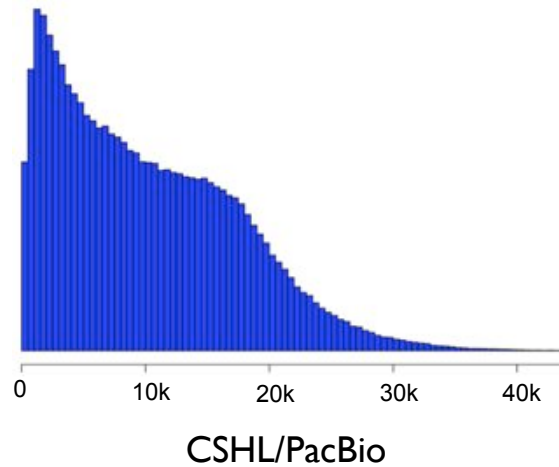
***A greater N50 is indicative of improvement in every dimension:***

- 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

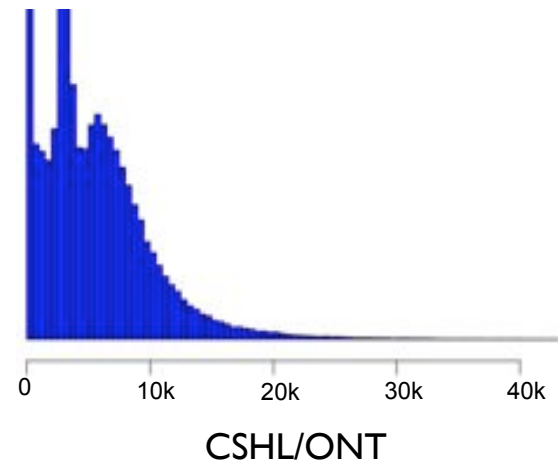


# 3<sup>rd</sup> Gen Long Read Sequencing

## PacBio RS II

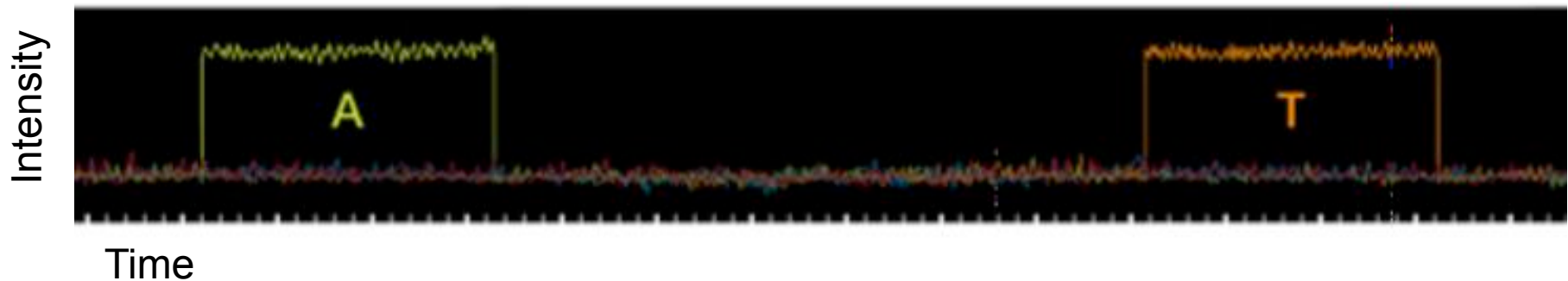
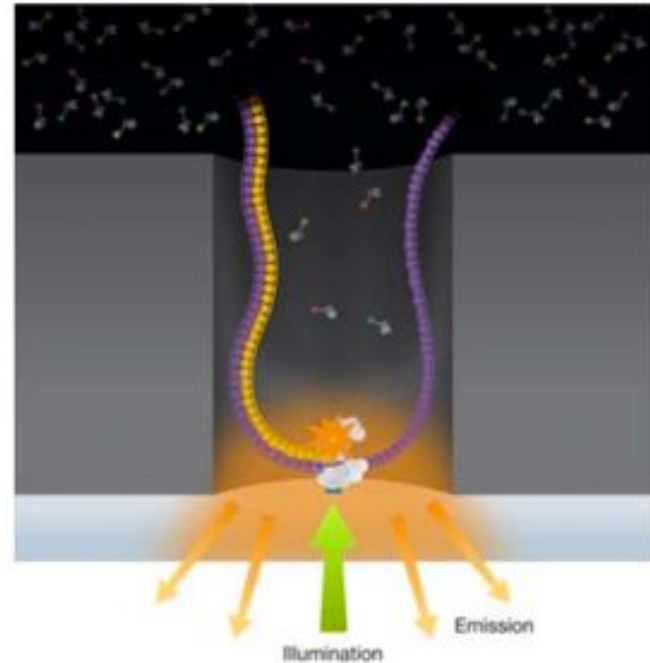
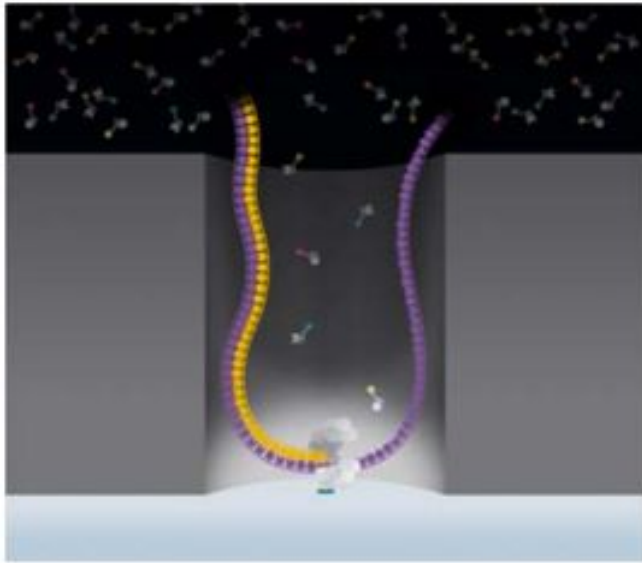


## Oxford Nanopore

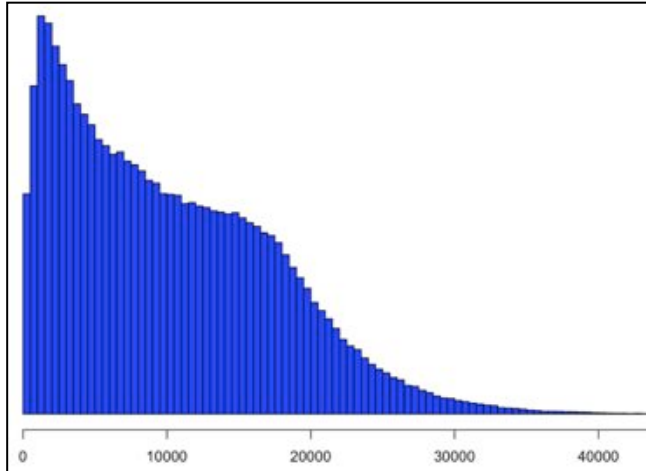


# PacBio SMRT Sequencing

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



# SMRT Sequencing Data



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

TTGTAAGCAGTTGAAAACATATGTGTGGATTTAGATAAAGAACATGAAAG  
 |||  
 TTGTAAGCAGTTGAAAACATATGTGT-GATTTAG-ATAAAGAACATGGAAG

ATTATAAA-CAGTTGATCCATT-AGAAGA-AAACGCAAAGGCCGCTAGG  
 |  
 A-TATAAATCAGTTGATCCATTAGAA-AGAAACGC-AAAGGC-GCTAGG

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

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

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

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

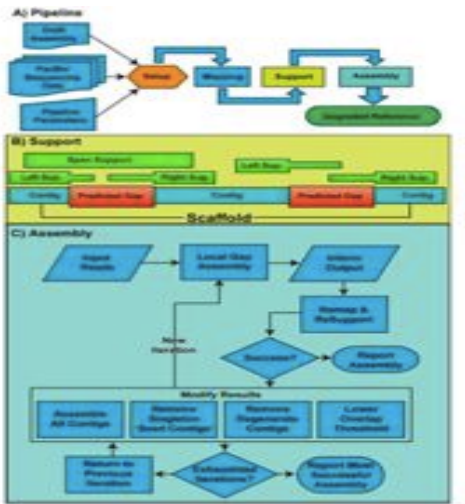
TCGGAGAGATCCAAAACAATGGGC-ATCGCCTTTGA-GTTAC-AATCAAA  
 |||  
 TC-GAGAGATCC-AAACAAT-GGCGATCG-CTTTGACGTTACAATCAAA

ATCCAGTGAAAAATATAATTTATGCAATCCAGGAACCTTATTCACAATTAG  
 |||  
 ATCCAGT-GAAAAATATA--TTATGC-ATCCA-GAECTTATTCACAATTAG

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

# PacBio Assembly Algorithms

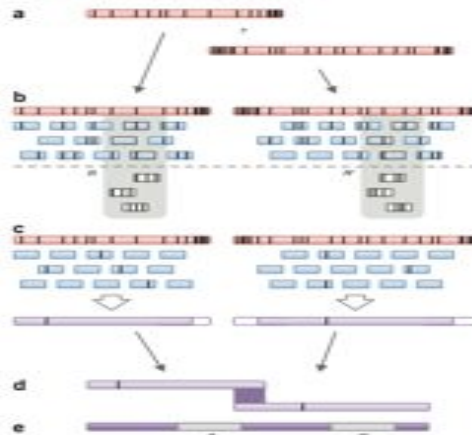
## PBJelly



**Gap Filling  
and Assembly Upgrade**

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

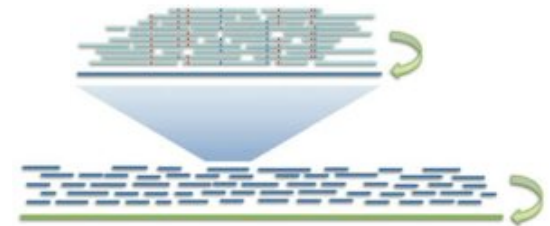
## PacBioToCA & ECTools



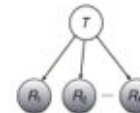
**Hybrid/PB-only Error  
Correction**

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**

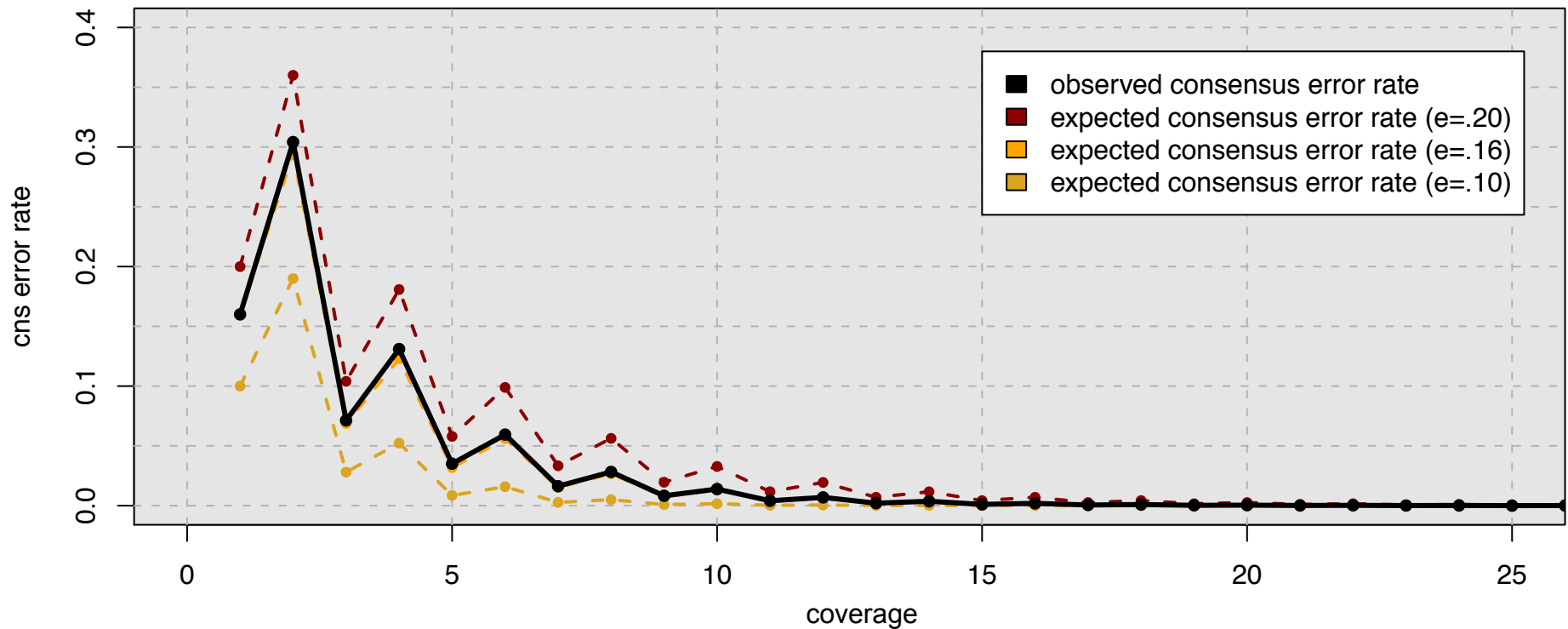
Chin *et al* (2013)  
*Nature Methods*. 10:563–569

< 5x

PacBio Coverage

> 50x

# Consensus Accuracy and Coverage



## Coverage can overcome random errors

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

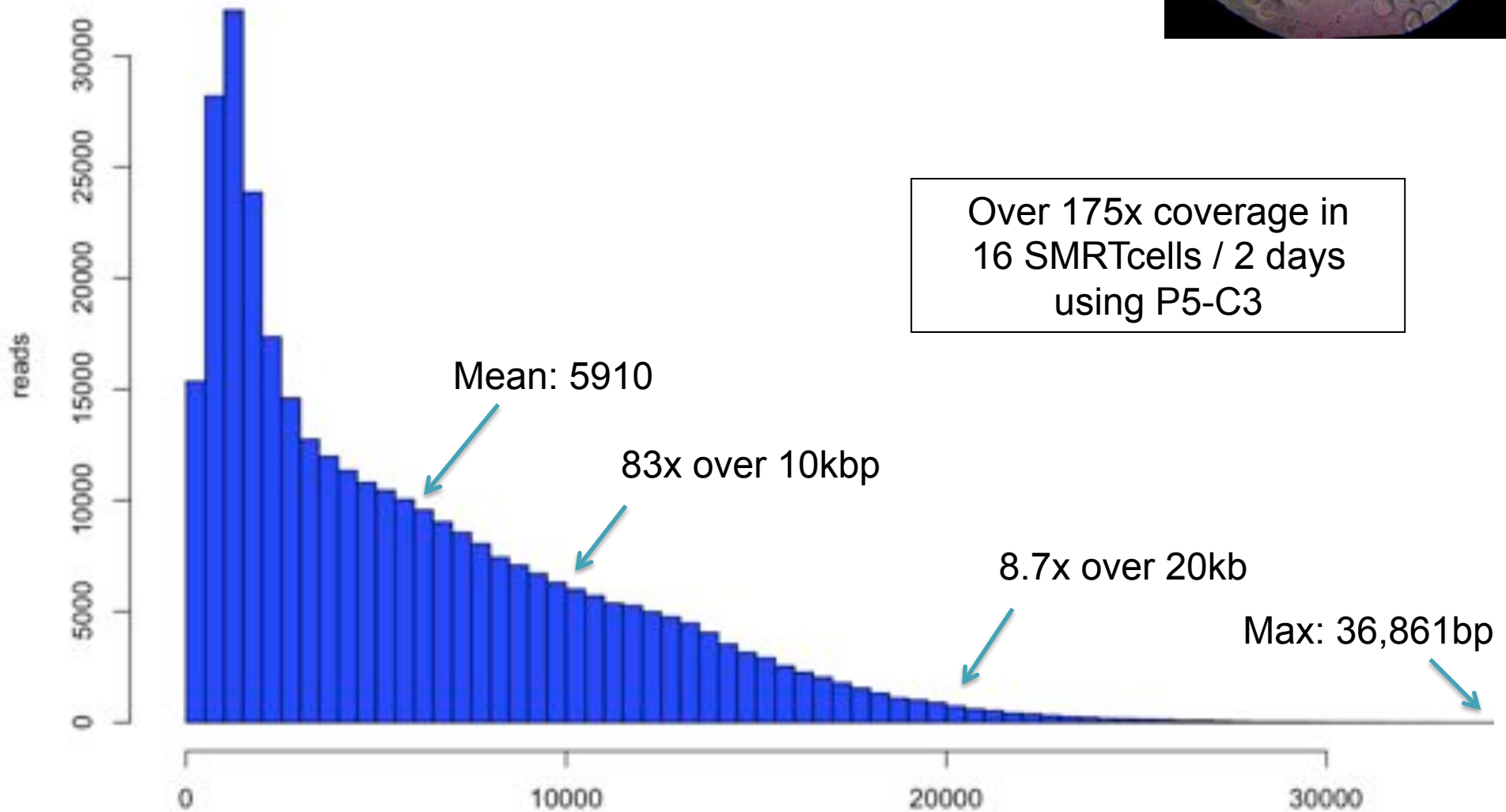
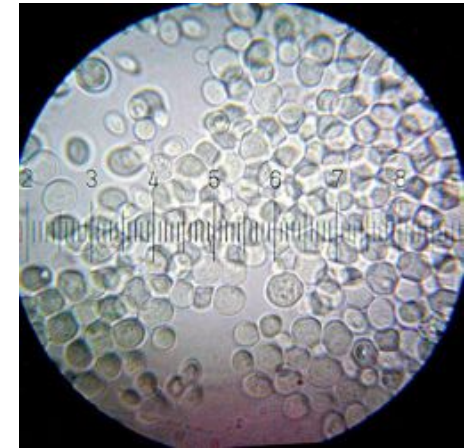
Koren, Schatz, et al (2012)  
*Nature Biotechnology*. 30:693–700

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

# S. cerevisiae W303

PacBio RS II sequencing at CSHL in the McCombie Lab

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



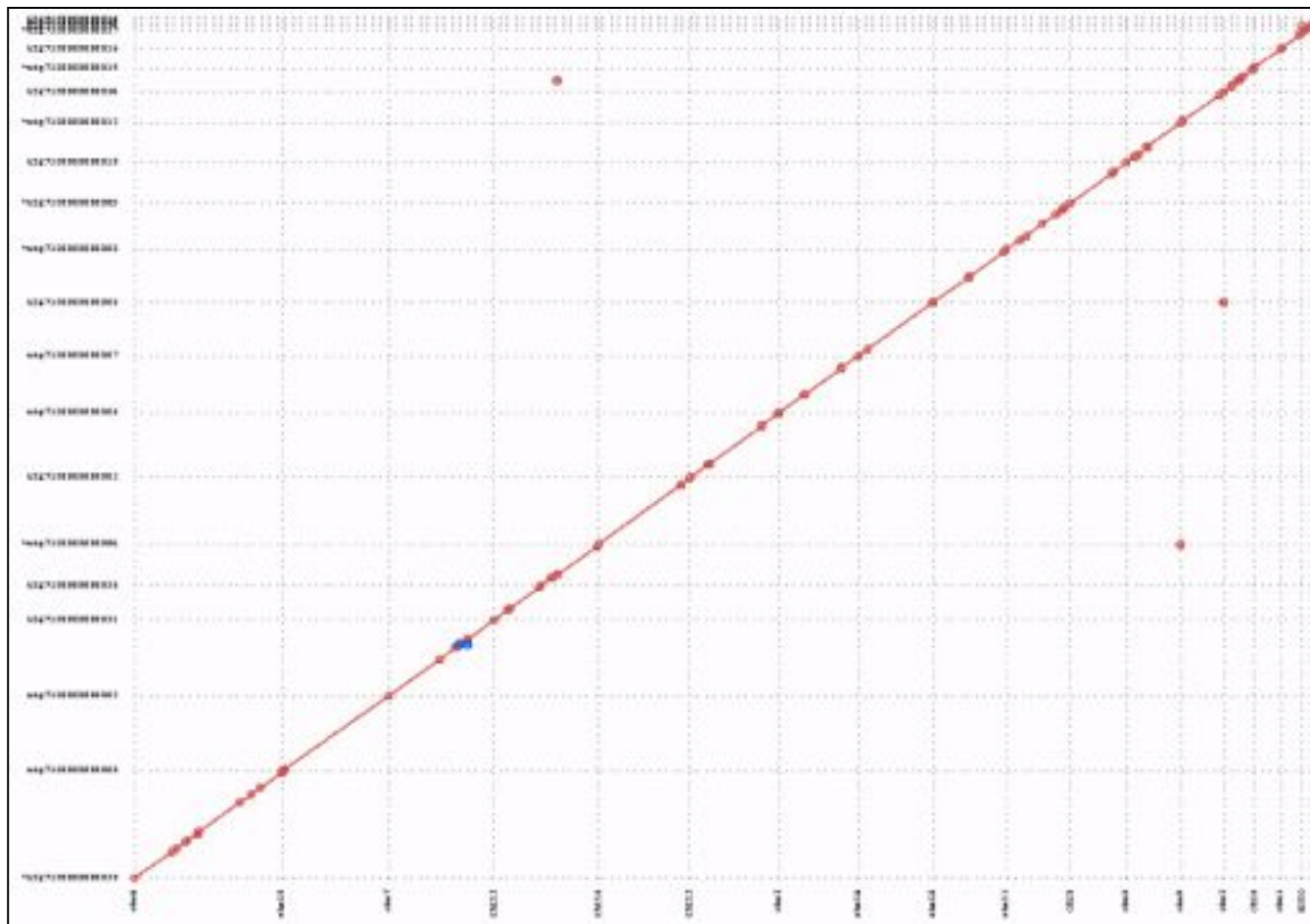
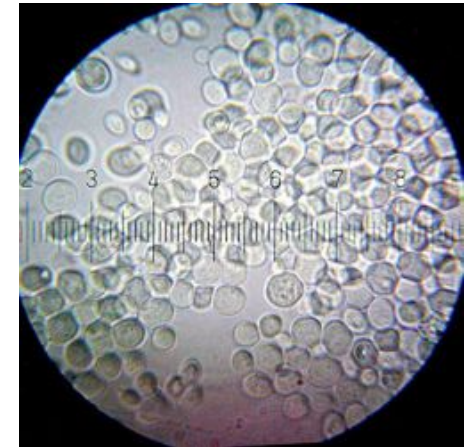
# S. cerevisiae W303

S288C Reference sequence

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

PacBio assembly using HGAP + Celera Assembler

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



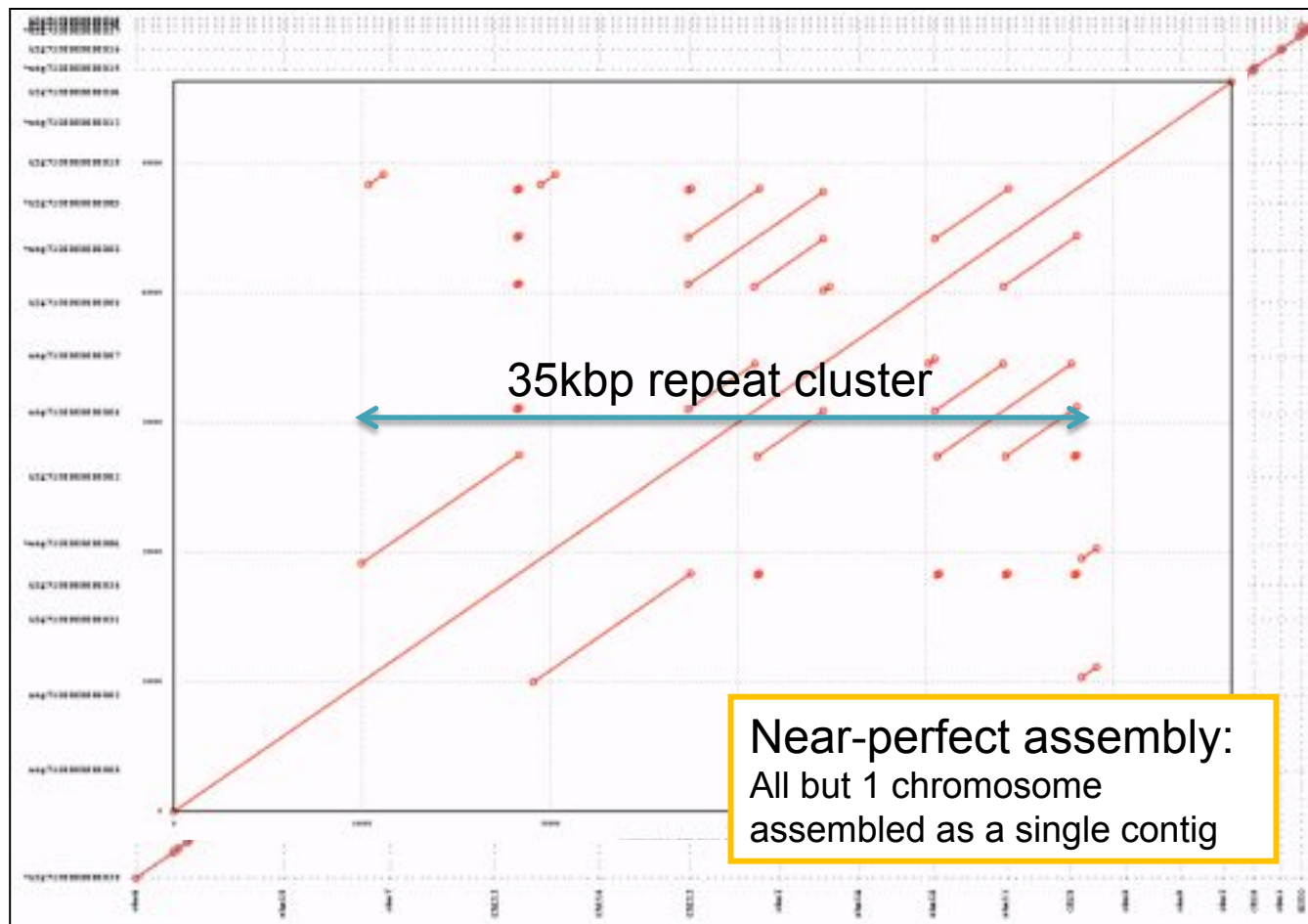
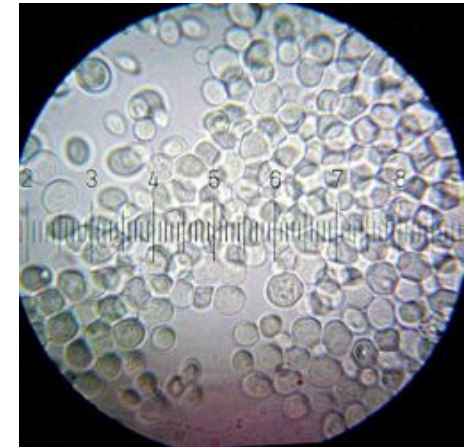
# S. cerevisiae W303

S288C Reference sequence

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

PacBio assembly using HGAP + Celera Assembler

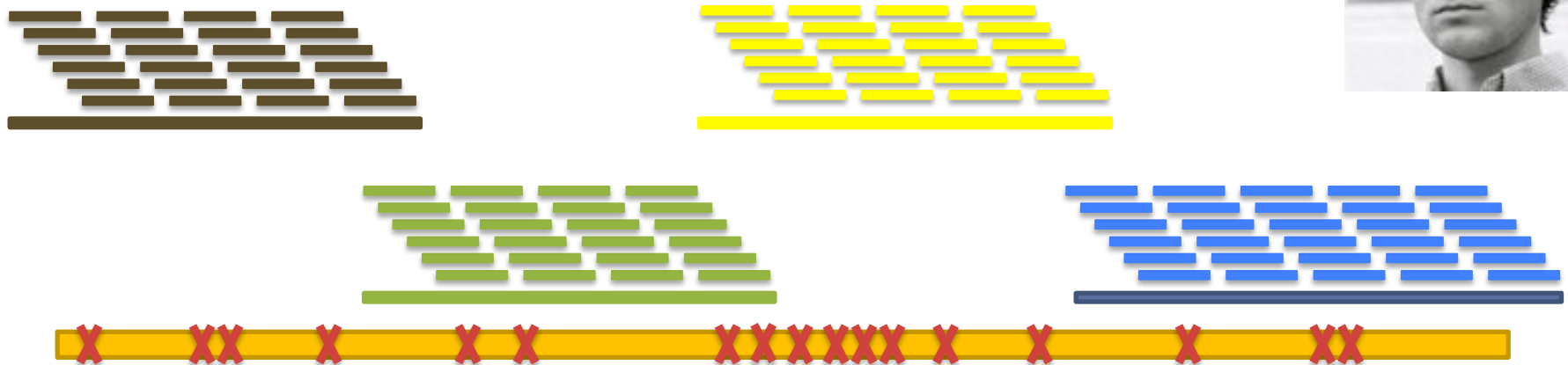
- 12.4Mbp; 21 non-redundant contigs; N50: 811kbp; >99.8% id





# ECTools: Hybrid Error Correction for large genomes

<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**

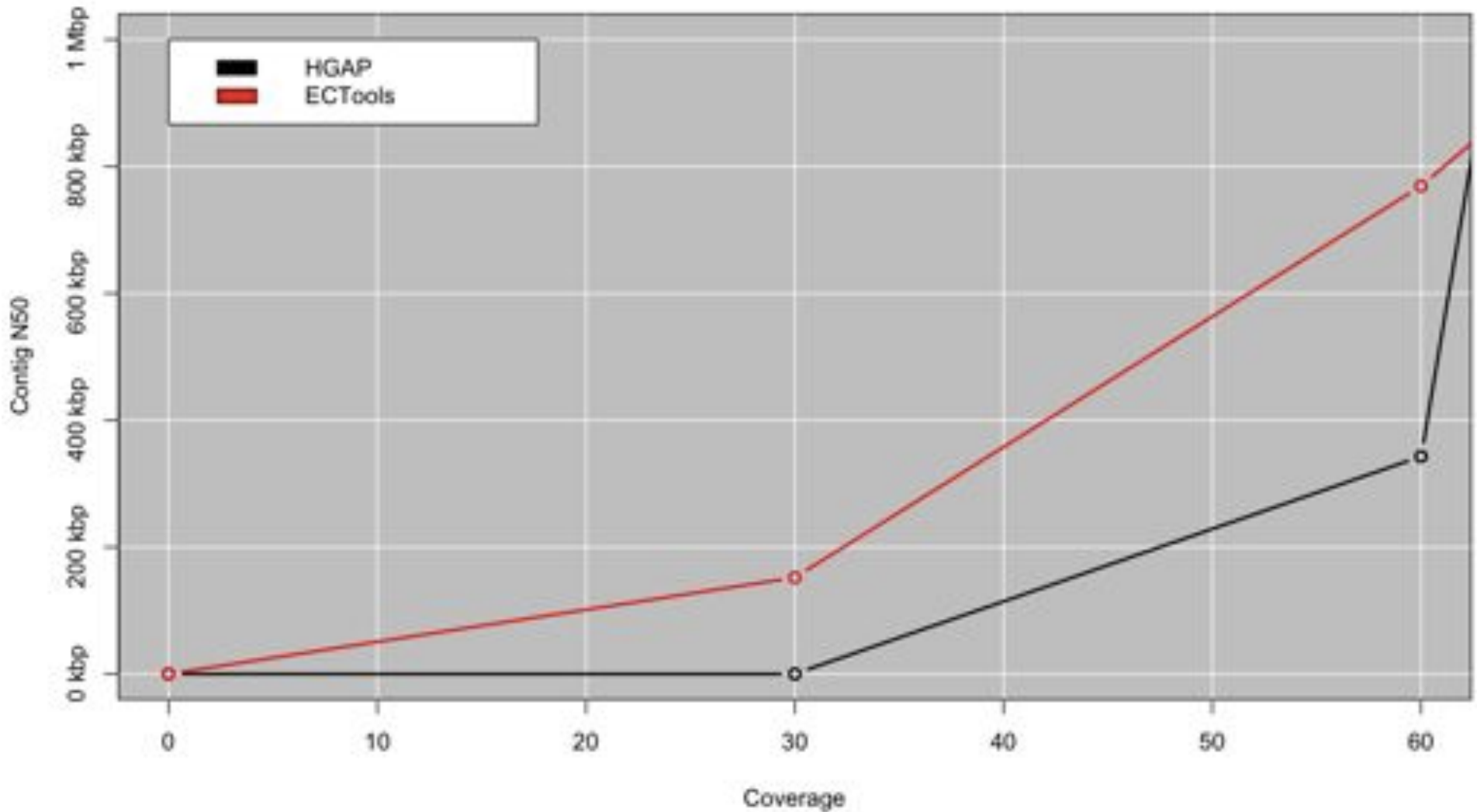
# A. thaliana Ler-0

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



## Downsampling experiment

Accuracy approaches 100% but limited by repeats



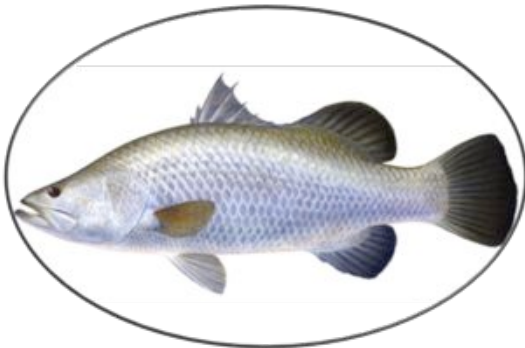
# Current Collaborations



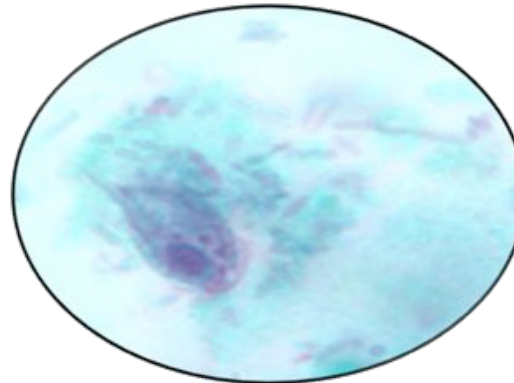
***Indica & Aus Rice***  
McCombie/Ware/McCouch



***Pinapple***  
UIUC



***Asian Sea Bass***  
Temasek Life Sciences Laboratory

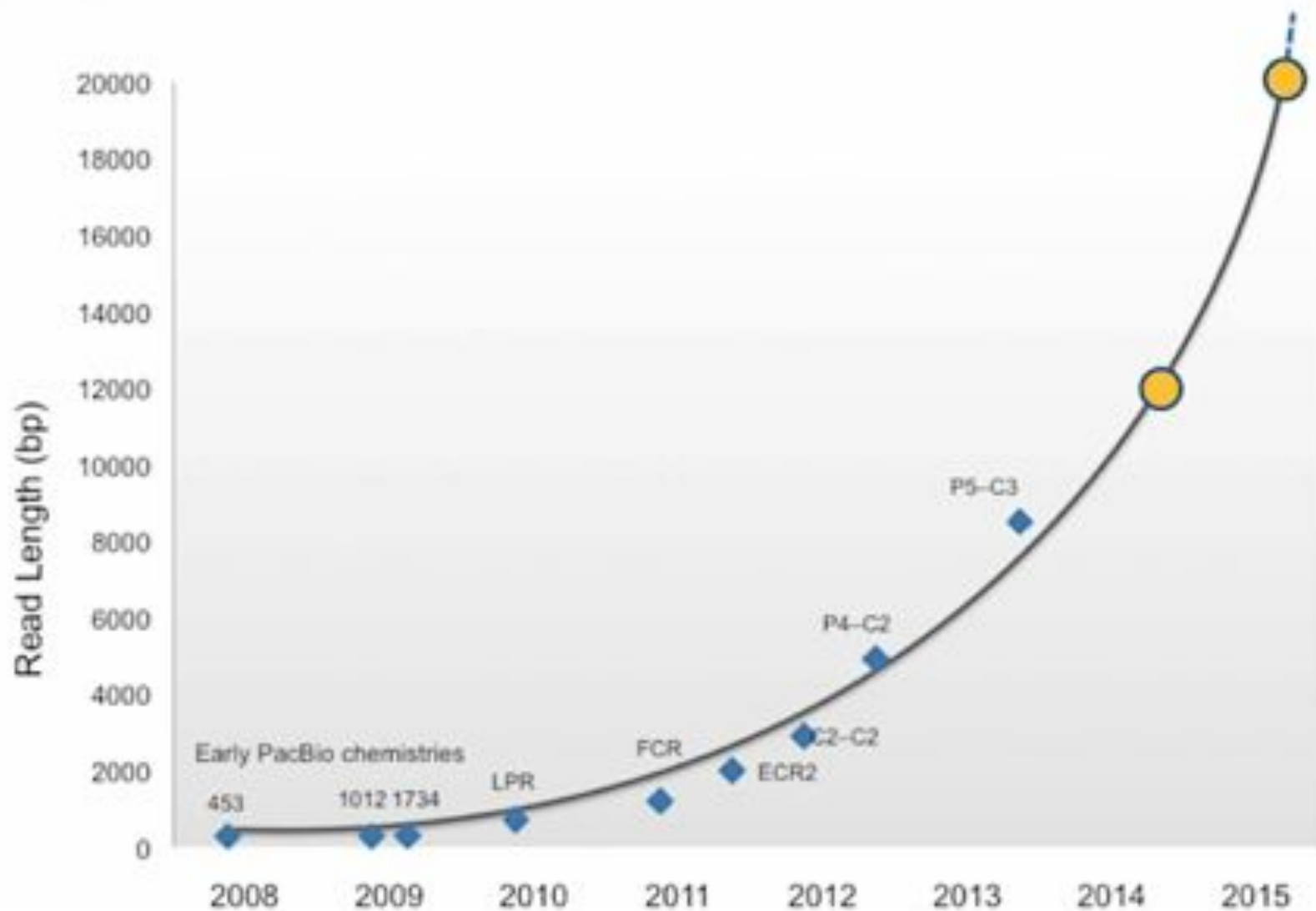


***P. hominis***  
NYU



***M. ligano***  
Hannon

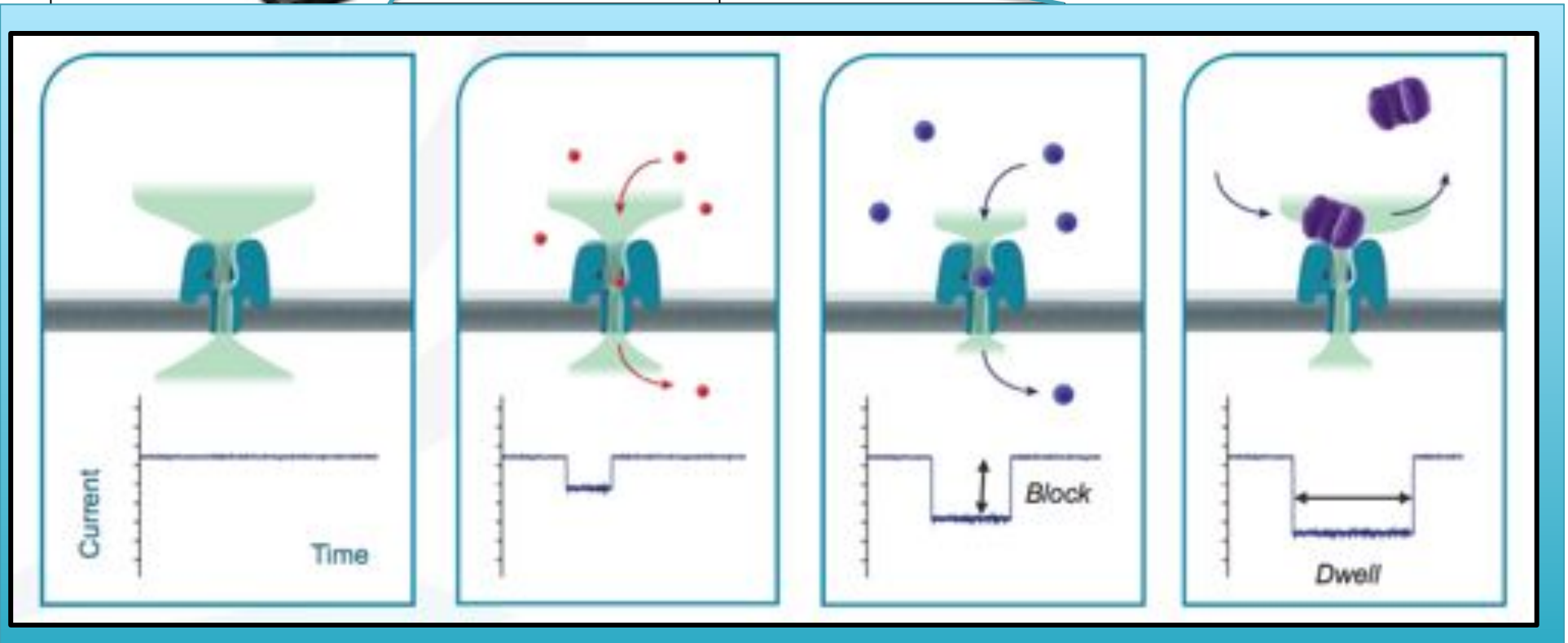
# PacBio® Advances in Read Length



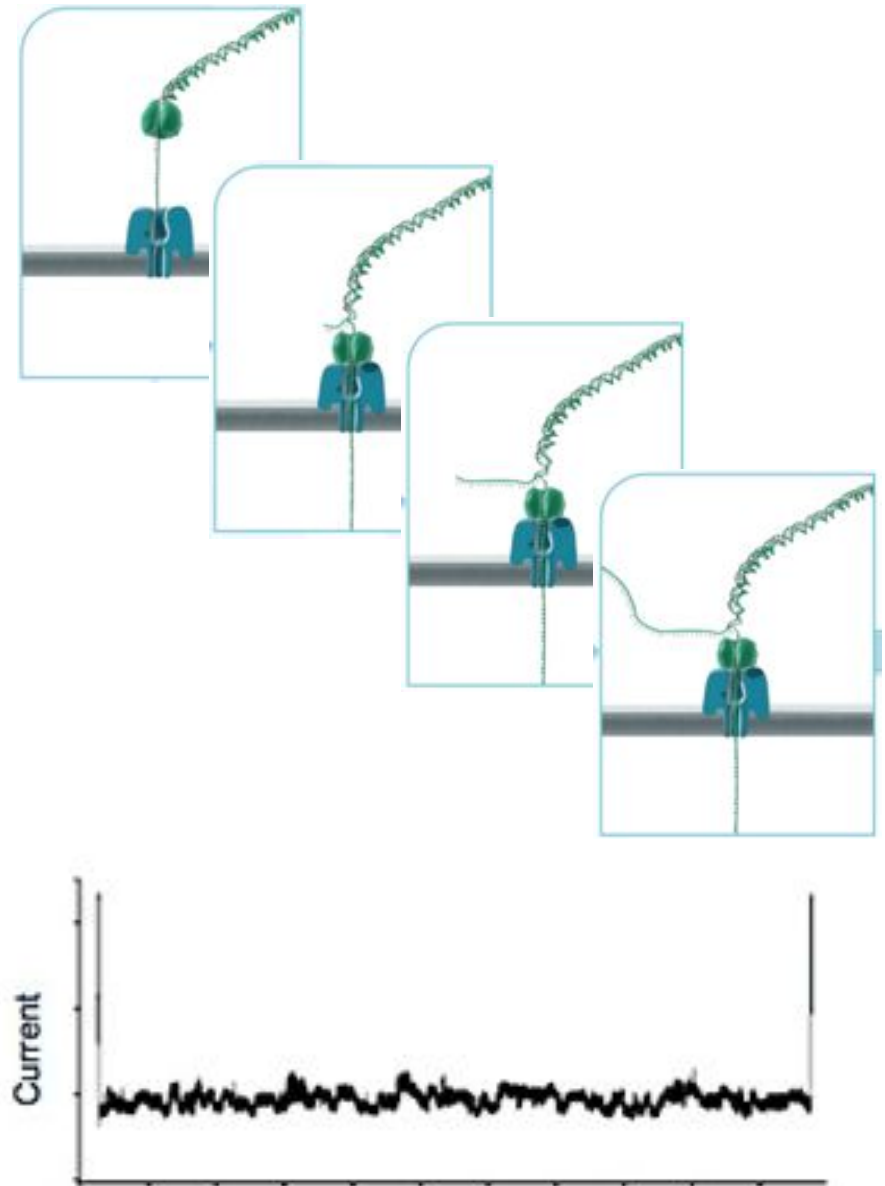
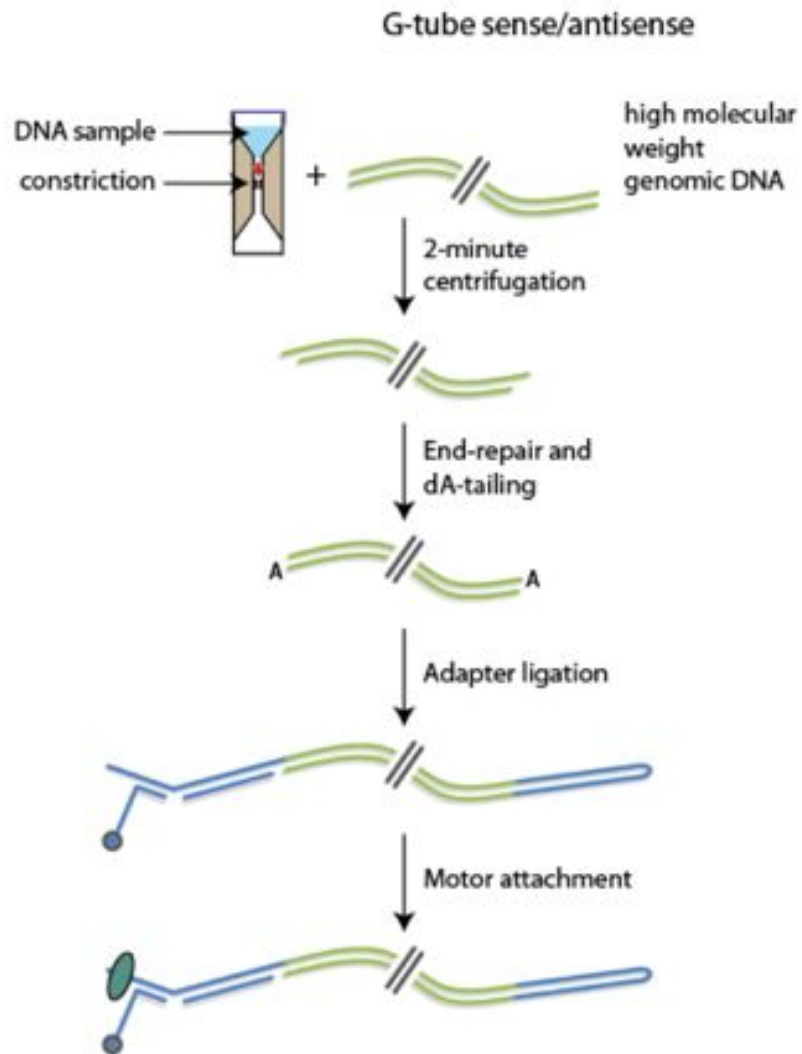
# Oxford Nanopore MinION



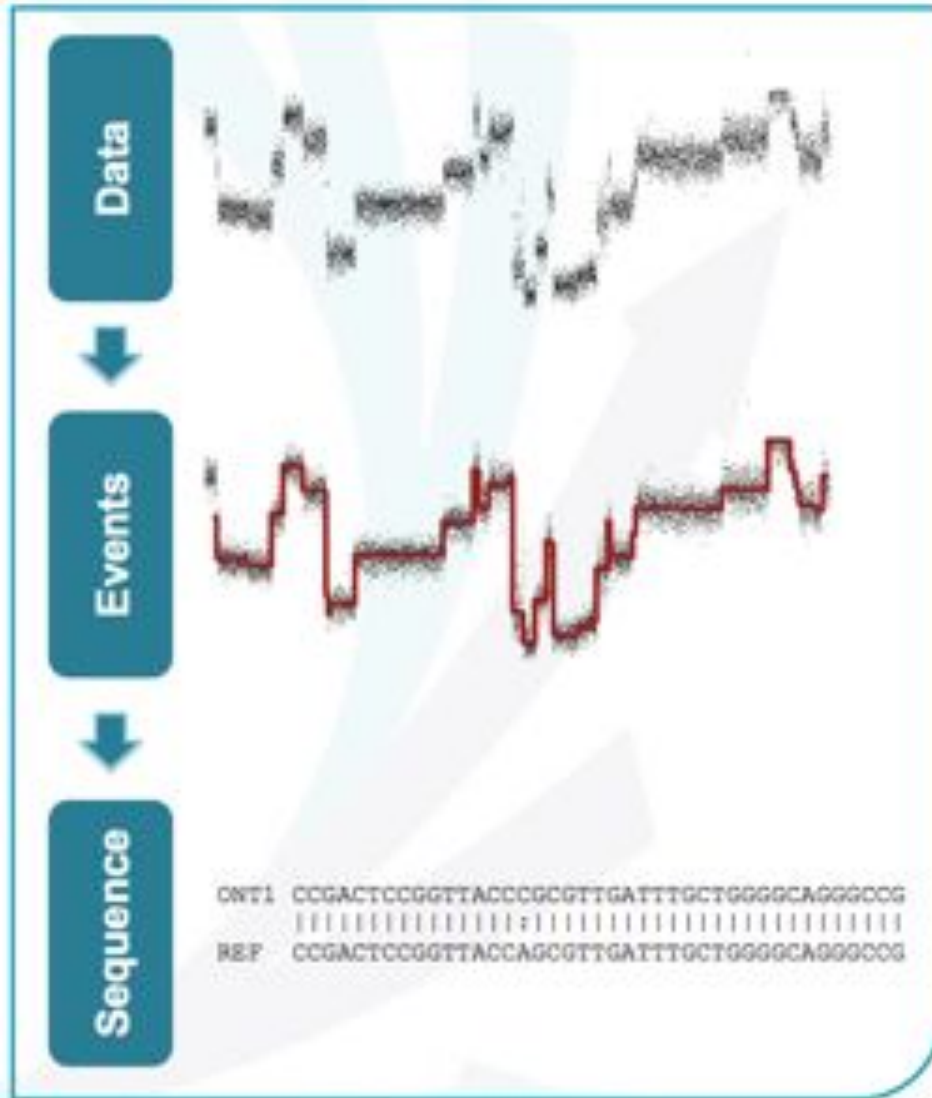
- Thumb drive sized sequencer powered over USB
- Capacity for 512 reads at once
- Senses DNA by measuring changes to ion flow



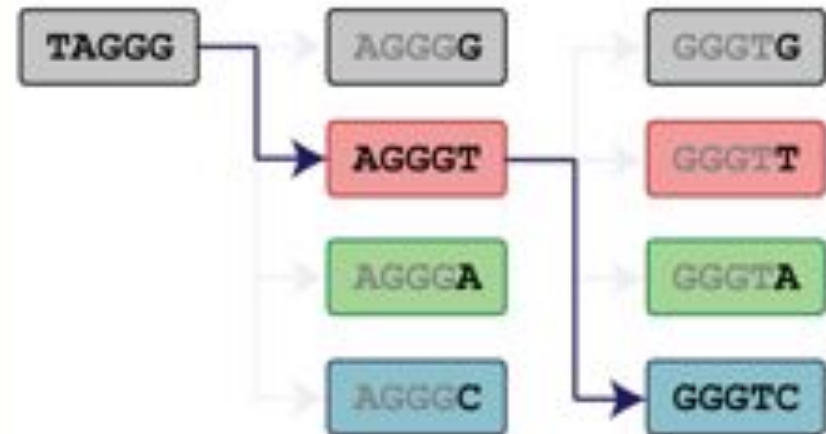
# Nanopore Sequencing



# Nanopore Basecalling



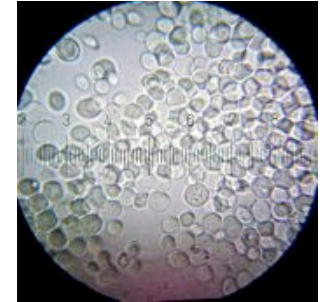
- Hidden Markov model
- Only four options per transition
- Pore type = distinct kmer length



- Form probabilistic path through measured states currents and transitions
  - e.g. Viterbi algorithm

Basecalling currently performed at Amazon with frequent updates to algorithm

# Nanopore Readlengths



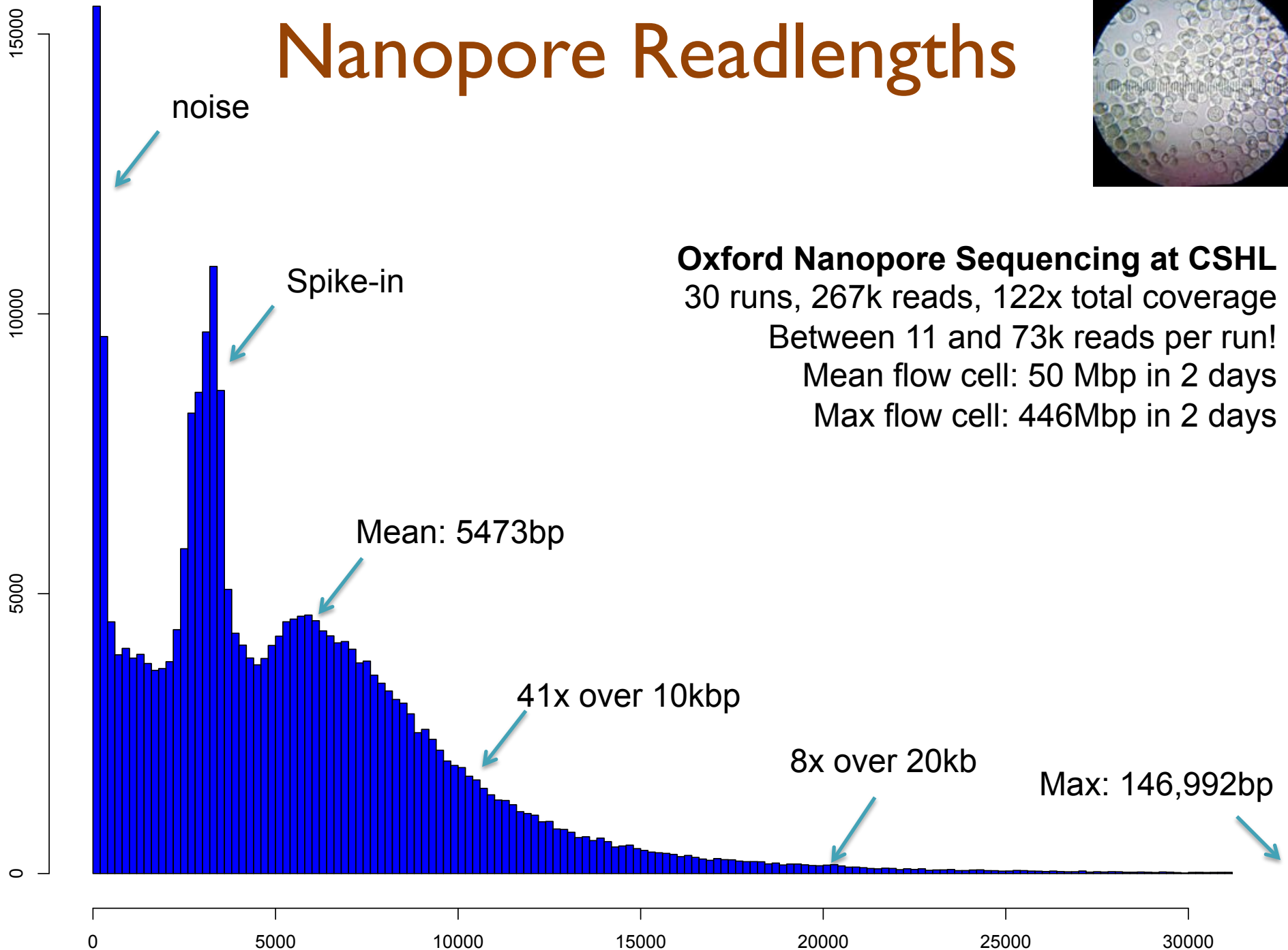
## Oxford Nanopore Sequencing at CSHL

30 runs, 267k reads, 122x total coverage

Between 11 and 73k reads per run!

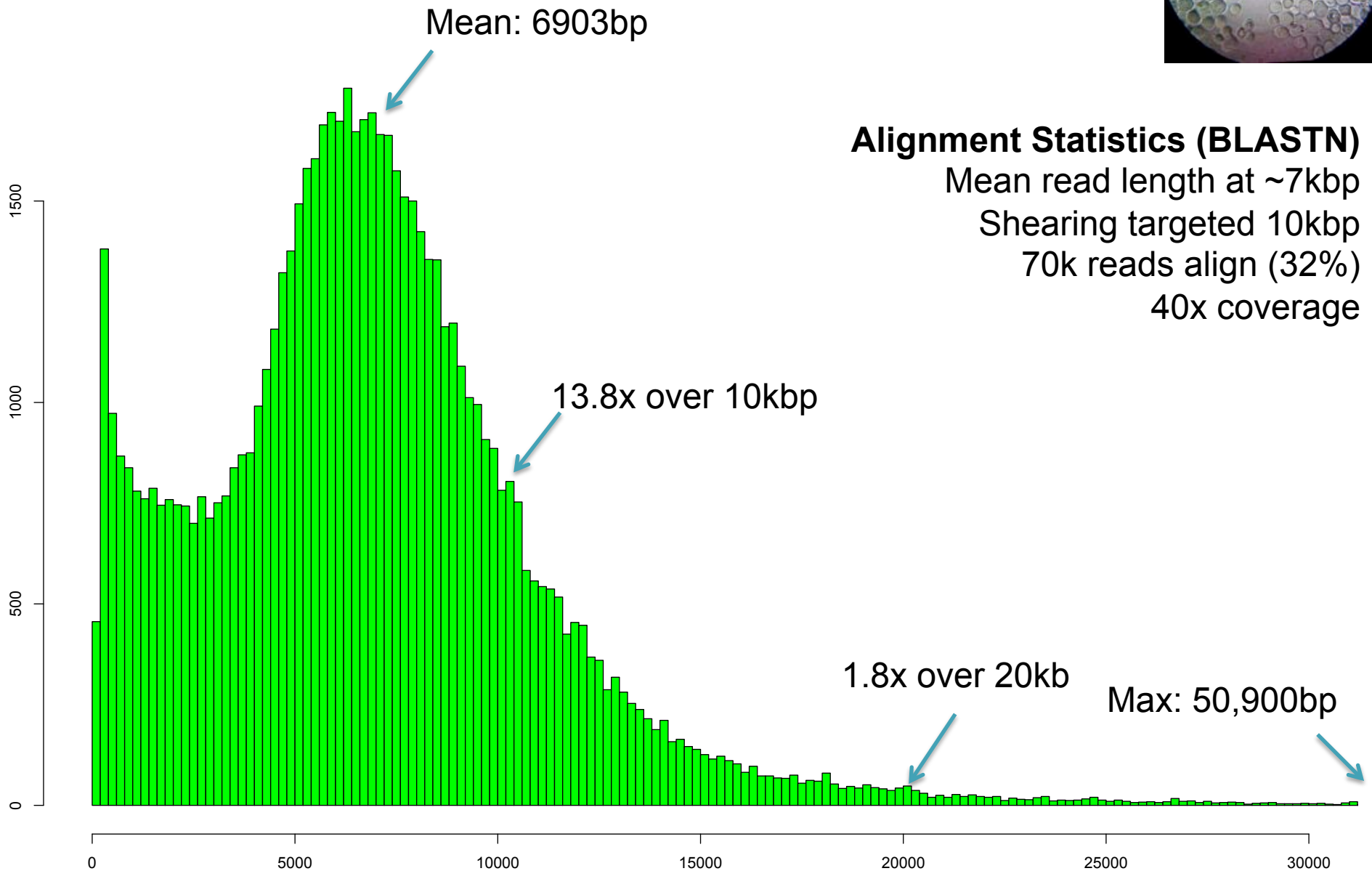
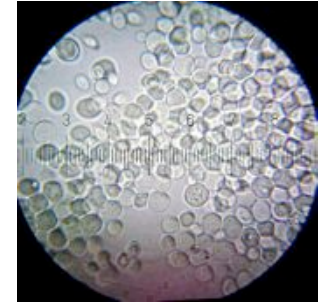
Mean flow cell: 50 Mbp in 2 days

Max flow cell: 446Mbp in 2 days





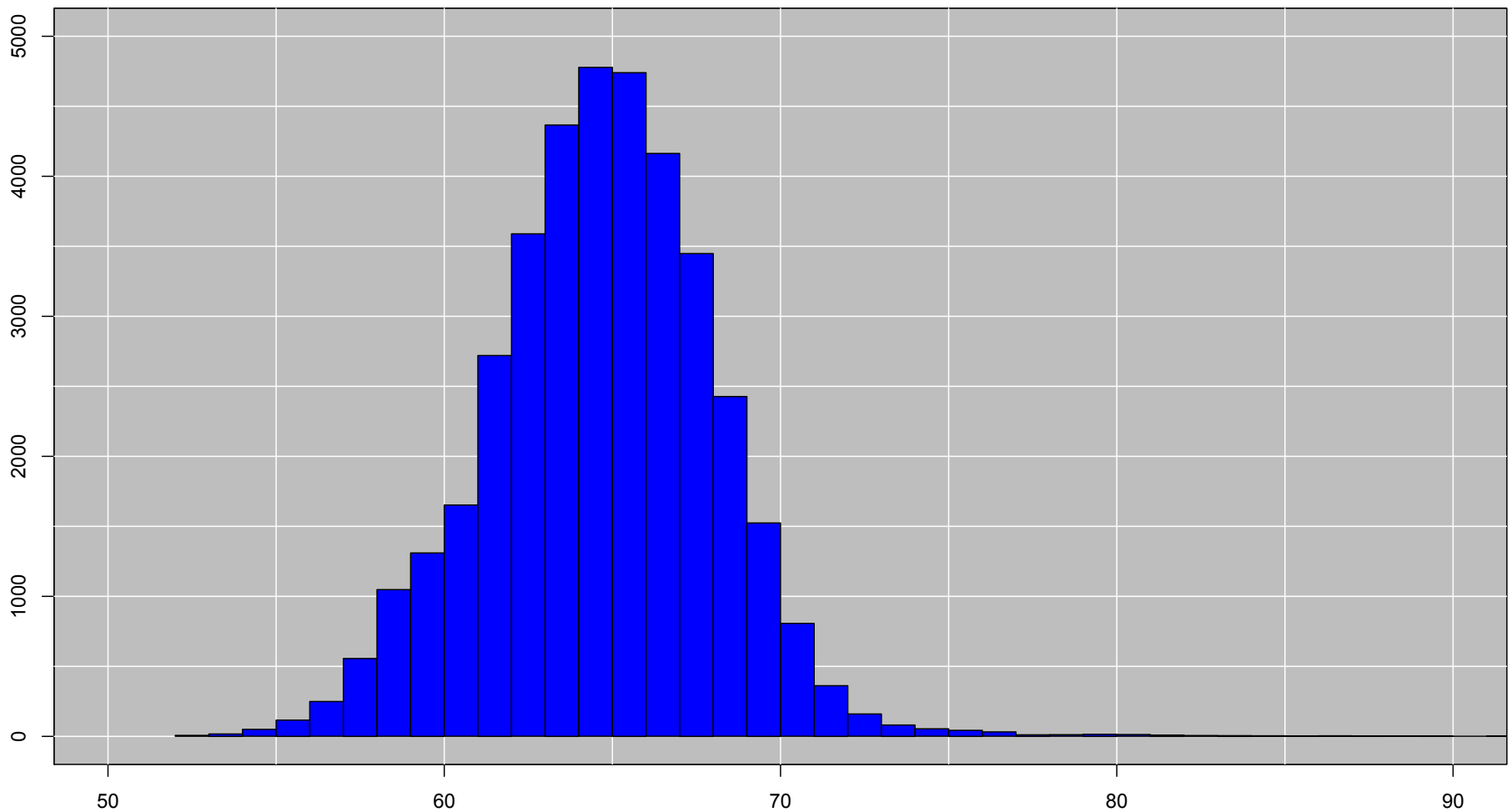
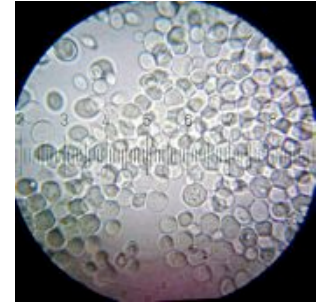
# Nanopore Alignments



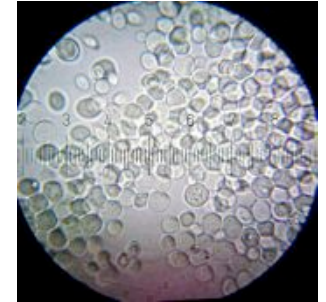
# Nanopore Accuracy

## Alignment Quality (BLASTN)

Of reads that align, average ~64% identity



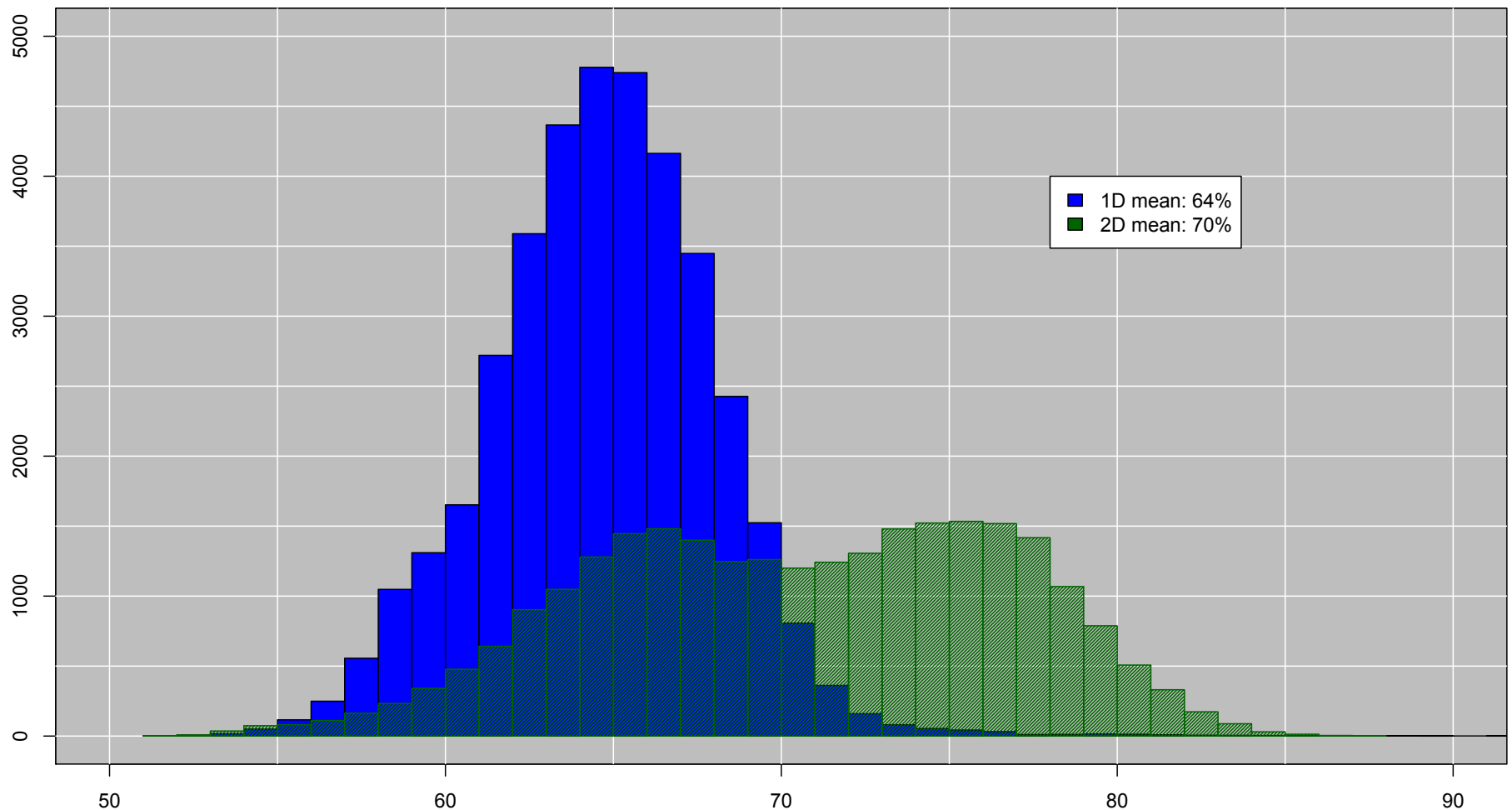
# Nanopore Accuracy



## Alignment Quality (BLASTN)

Of reads that align, average ~64% identity

“2D base-calling” improves to ~70% identity

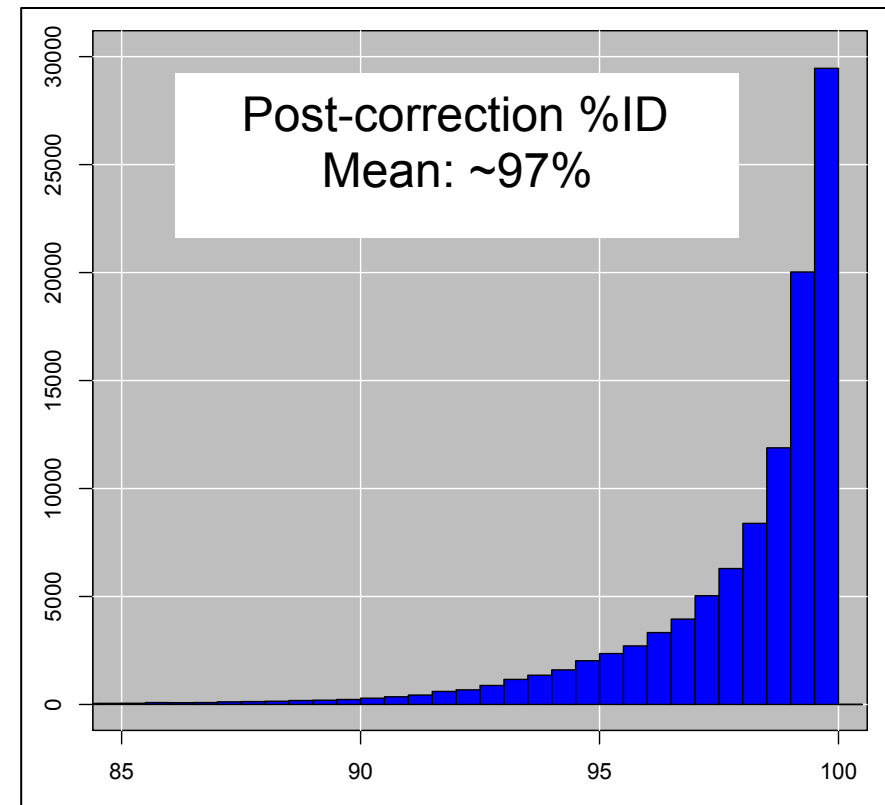


# NanoCorr: Nanopore-Illumina Hybrid Error Correction

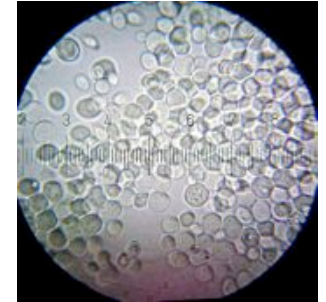


<https://github.com/jgurtowski/nanocorr>

1. BLAST Miseq reads to all raw Oxford Nanopore reads
2. Select non-repetitive alignments
  - First pass scans to remove “contained” alignments
  - Second pass uses Dynamic Programming (LIS) to select set of high-identity alignments with minimal overlaps
3. Compute consensus of each Oxford Nanopore read
  - Currently using Pacbio’s pbdagcon



# Long Read Assembly



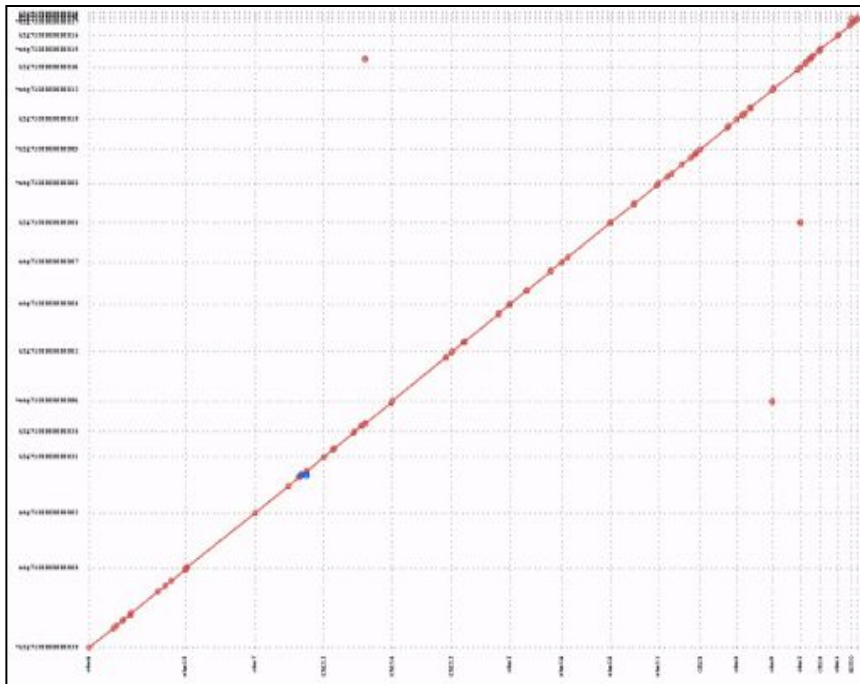
S288C Reference sequence

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

## **Pacific Biosciences**

HGAP + Celera Assembler

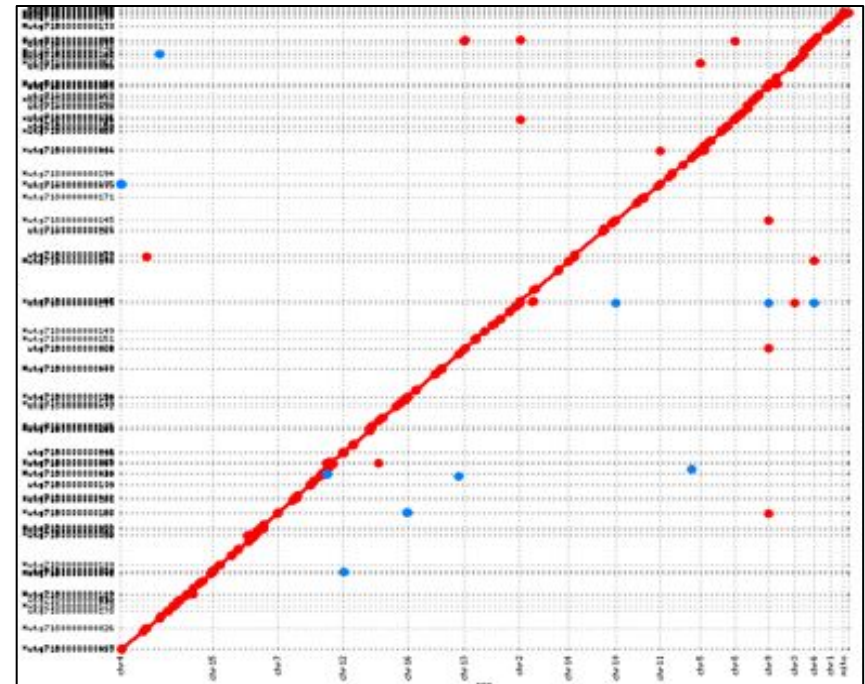
- 21 non-redundant contigs
- N50: 811kbp >99.8% id



## **Oxford Nanopore**

NanoCorr + Celera Assembler

- 234 non-redundant contigs
- N50: 362kbp >99.78% id



# Platform Technology:

Instruments for scaled nanopore analysis

MinION



Miniaturised USB device

Entry / Field / Small runs



GridION



One instrument:  
Scalable, groupable, networked, real-time

Desktop



Mid range Installation



Supercenter



# What should we expect from an assembly?

**Analysis of dozens of genomes from across the tree of life with real and simulated data**

## ***Summary & Recommendations***

- < 100 Mbp: HGAP/PacBio2CA @ 100x PB C3-P5  
expect near perfect chromosome arms
- < 1GB: HGAP/PacBio2CA @ 100x PB C3-P5  
high quality assembly: contig N50 over 1Mbp
- > 1GB: hybrid/gap filling  
expect contig N50 to be 100kbp – 1Mbp
- > 5GB: Email [mschatz@cshl.edu](mailto:mschatz@cshl.edu)



**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>

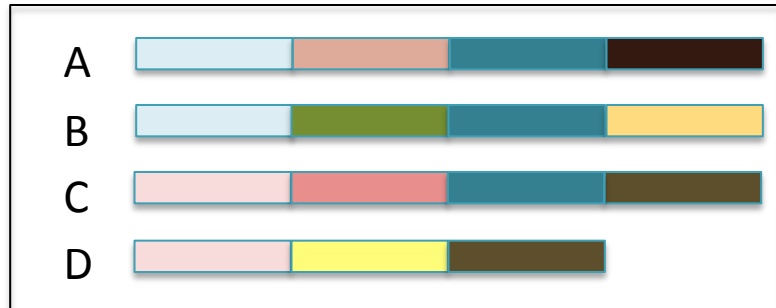


# Genome Structure & Function

- 1. Structure: Sequencing and Assembly**  
*“A tale of two sequencers”*
  
- 2. Function: Disease Analytics**
  1. Pan-genome analysis
  2. The role of indels in human diseases

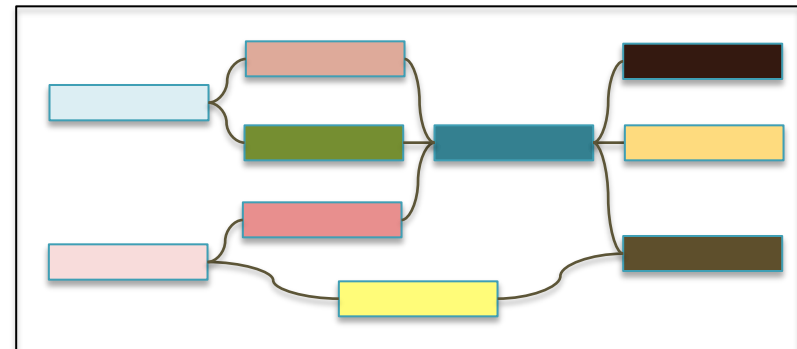


# Pan-Genome Alignment & Assembly



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: Graphical pan-genome analysis with suffix skips

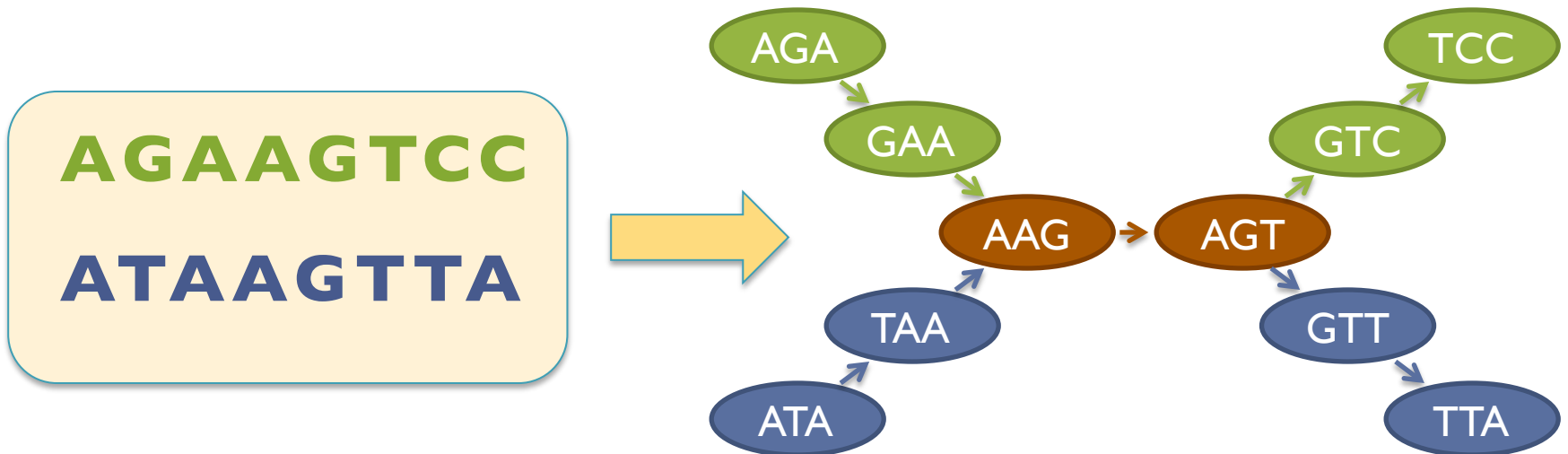
Marcus, S, Lee, H, Schatz, MC

<http://biorxiv.org/content/early/2014/04/06/003954>

# Graphical pan-genome analysis

## Colored de Bruijn graph

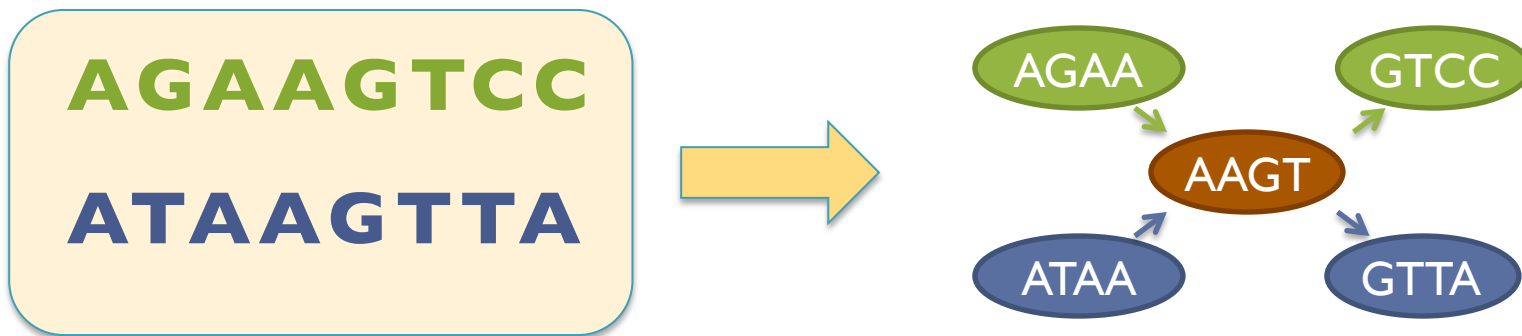
- Node for each distinct kmer
- Directed edge connects consecutive kmers
- Nodes overlap by  $k-1$  bp



# Graphical pan-genome analysis

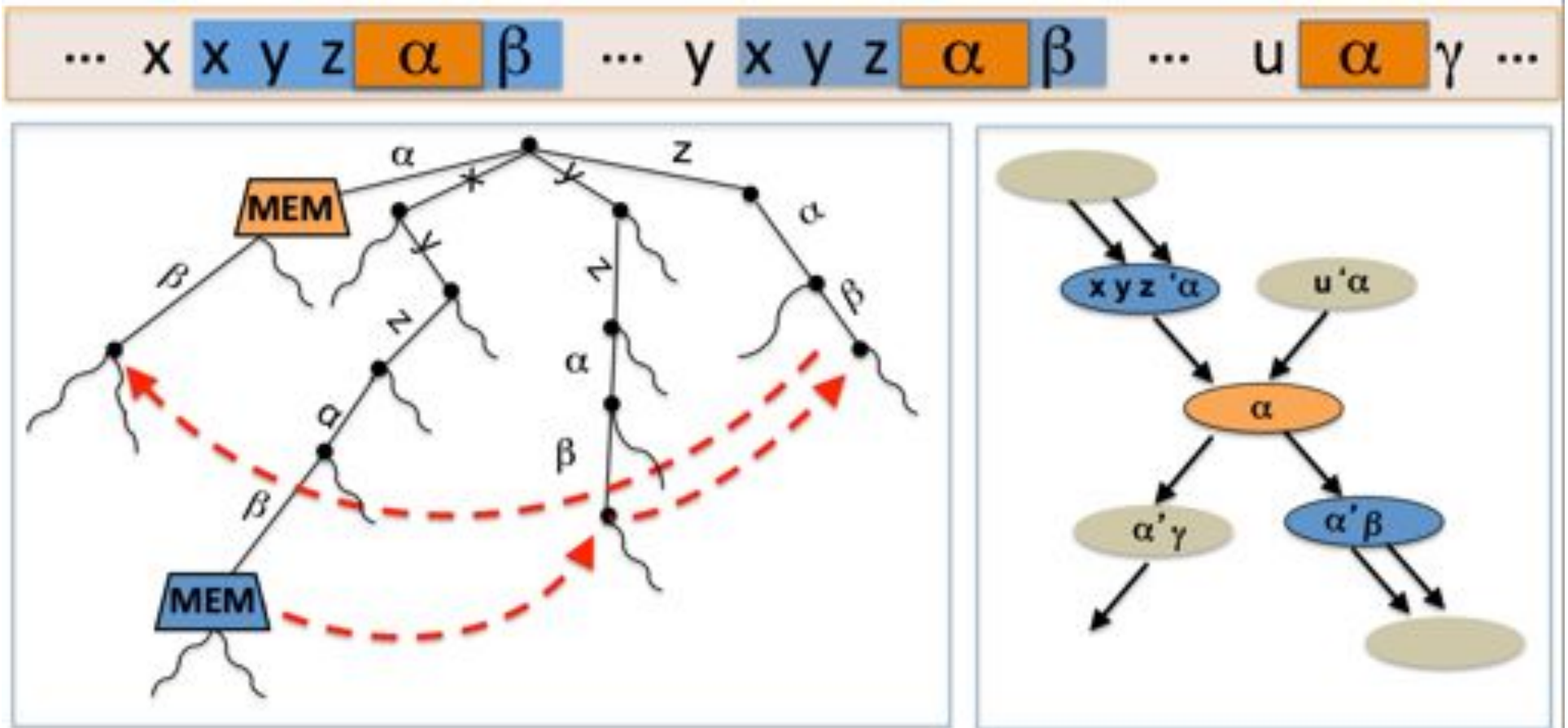
## Colored de Bruijn graph

- Node for each distinct kmer
- Directed edge connects consecutive kmers
- Nodes overlap by  $k-1$  bp



Other approaches all start from the raw de Bruijn graph, we aim to directly build the compressed graph as quickly as possible

# Suffix Trees & de Bruijn Graphs



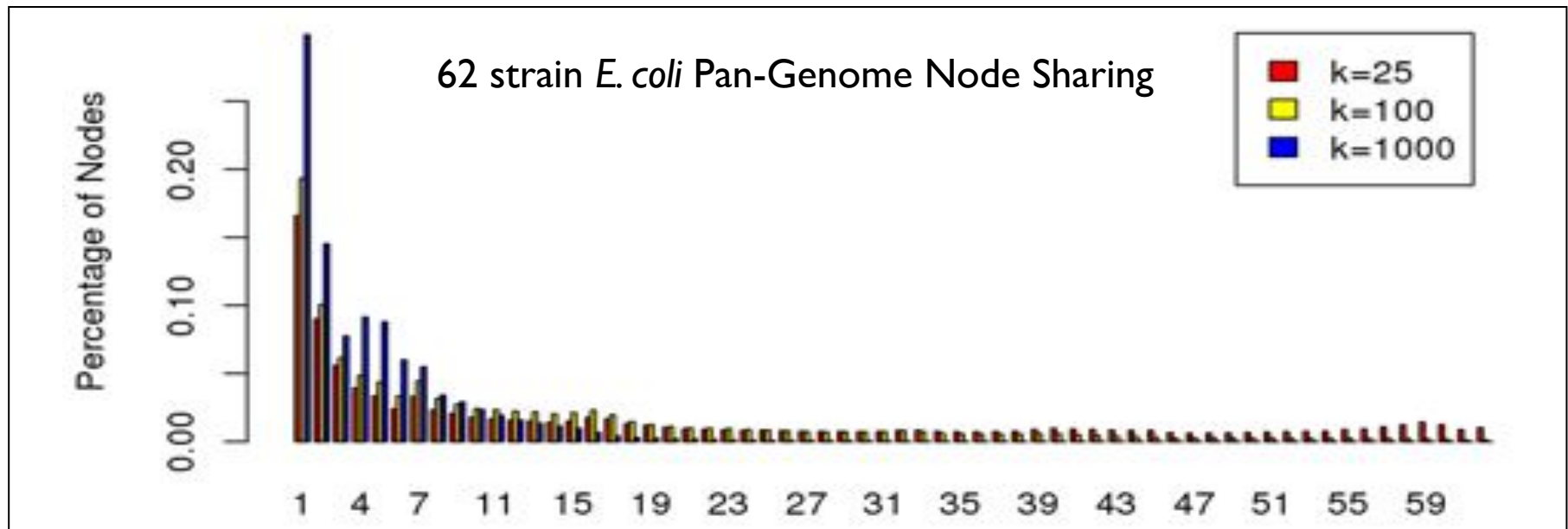
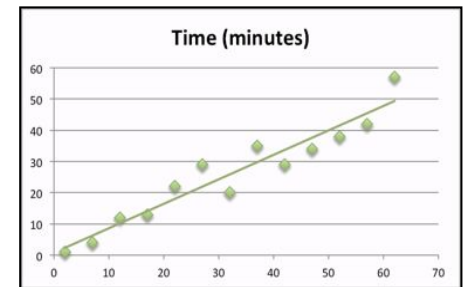
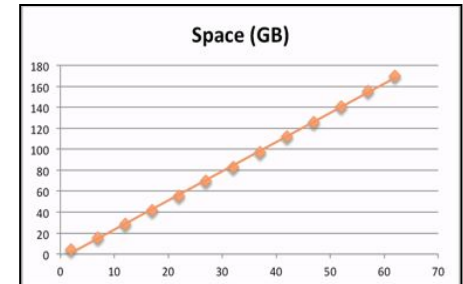
## Key concepts:

- Shared sequences form repeats called “maximal exact matches” (MEM)
- Easy to identify MEMs in a suffix tree, but may be nested within other MEMs
- Use “suffix skips” to quickly decompose MEMs, add in the missing nodes and edges

# Microbial Pan-Genomes

## **E. coli (62) and B. anthracis (9) pan-genome analysis**

- Analyzed all available strains in Genbank
- Space and time are linear in the number of genomes
  - $O(n \log g)$  where  $g$  is the length of the longest genome
- Many possible applications:
  - Identifying “core” genes present in all strains
  - Characterizing highly variable regions
  - Cataloging sequences shared by pathogenic varieties

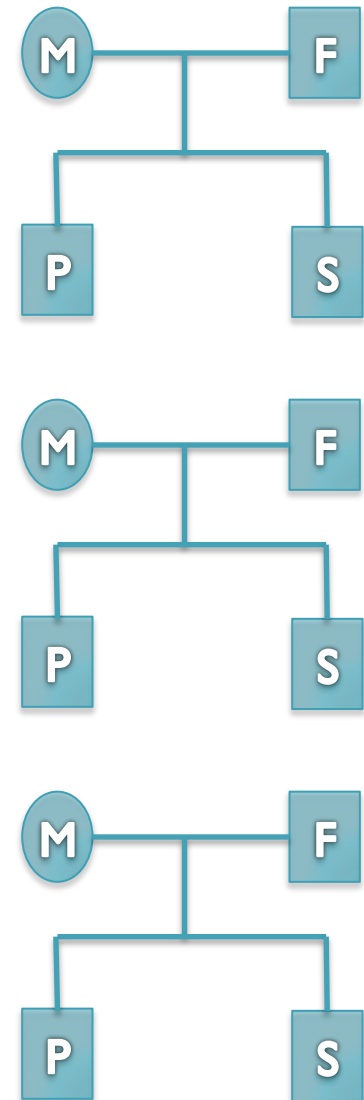


# Searching for the genetics behind human disorders and plant phenotypes

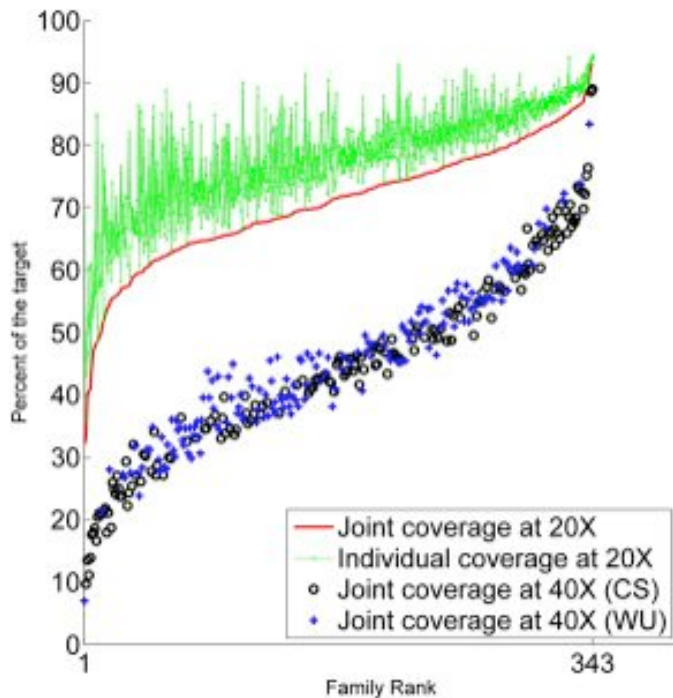
## **Search Strategy**

- 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 >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
- ***All used poor or no tools for indels:***
  - lossifov (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**

lossifov *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)

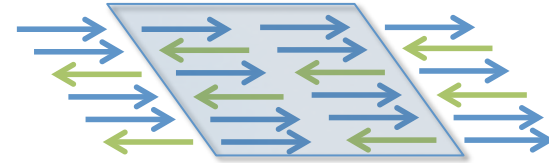
**Accurate *de novo* and transmitted indel detection in exome-capture data using microassembly.**

Narzisi, G, O'Rawe, JA, Iossifov, I, Fang, H, Lee, YH, Wang, Z, Wu, Y, Lyon, G, Wigler, M, Schatz MC (2014) *Nature Methods*. doi:10.1038/nmeth.3069

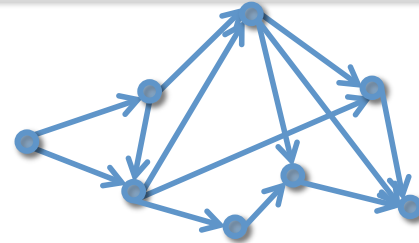


# Scalpel Algorithm

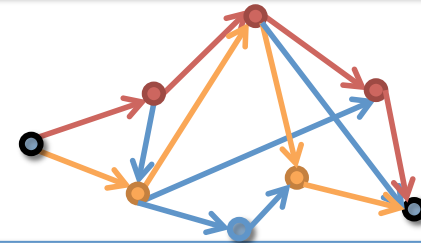
Extract reads mapping within the exon including (1) well-mapped reads, (2) soft-clipped reads, and (3) anchored pairs



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



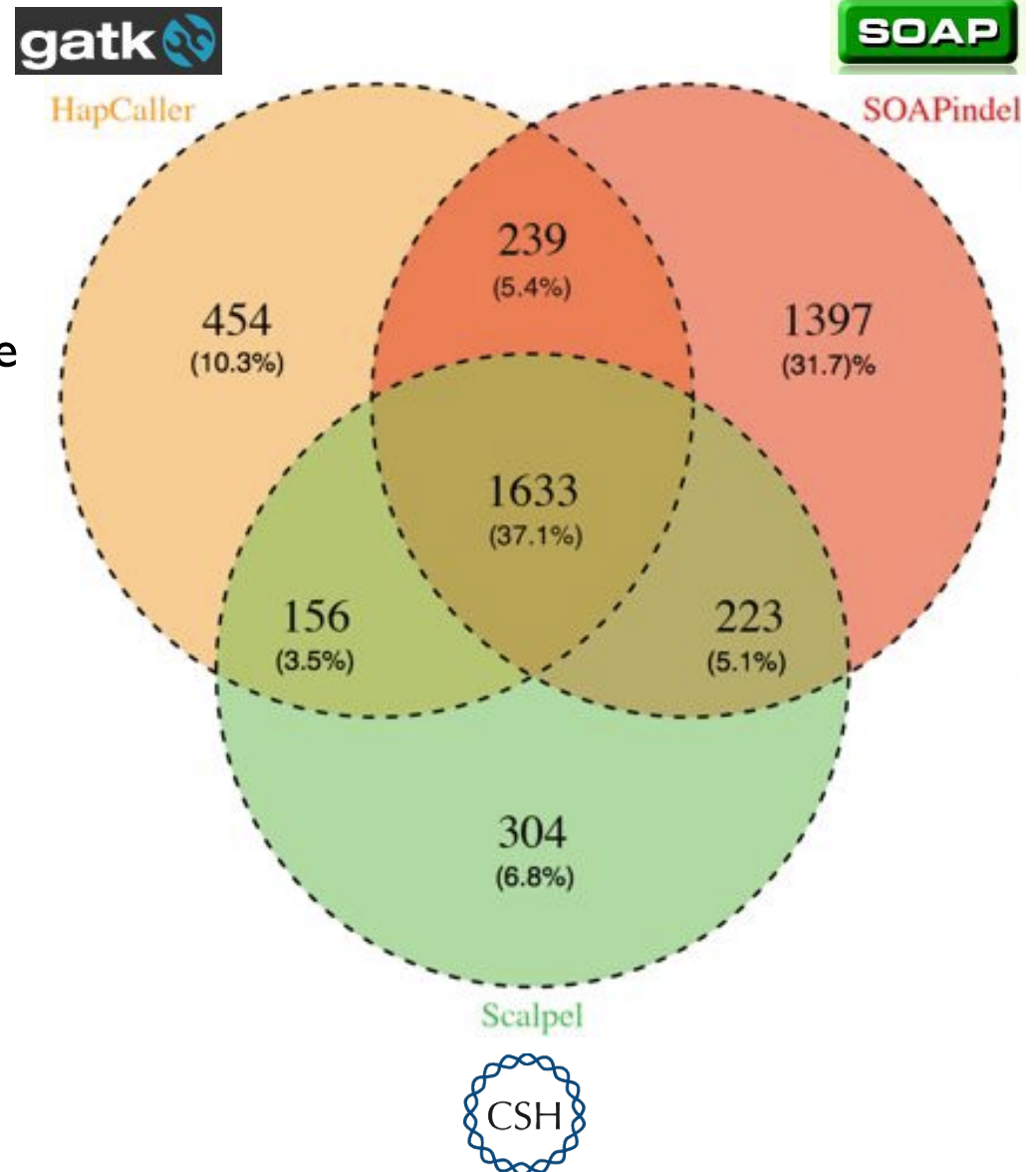
# Experimental Analysis & Validation

Selected one deep coverage exome for deep analysis

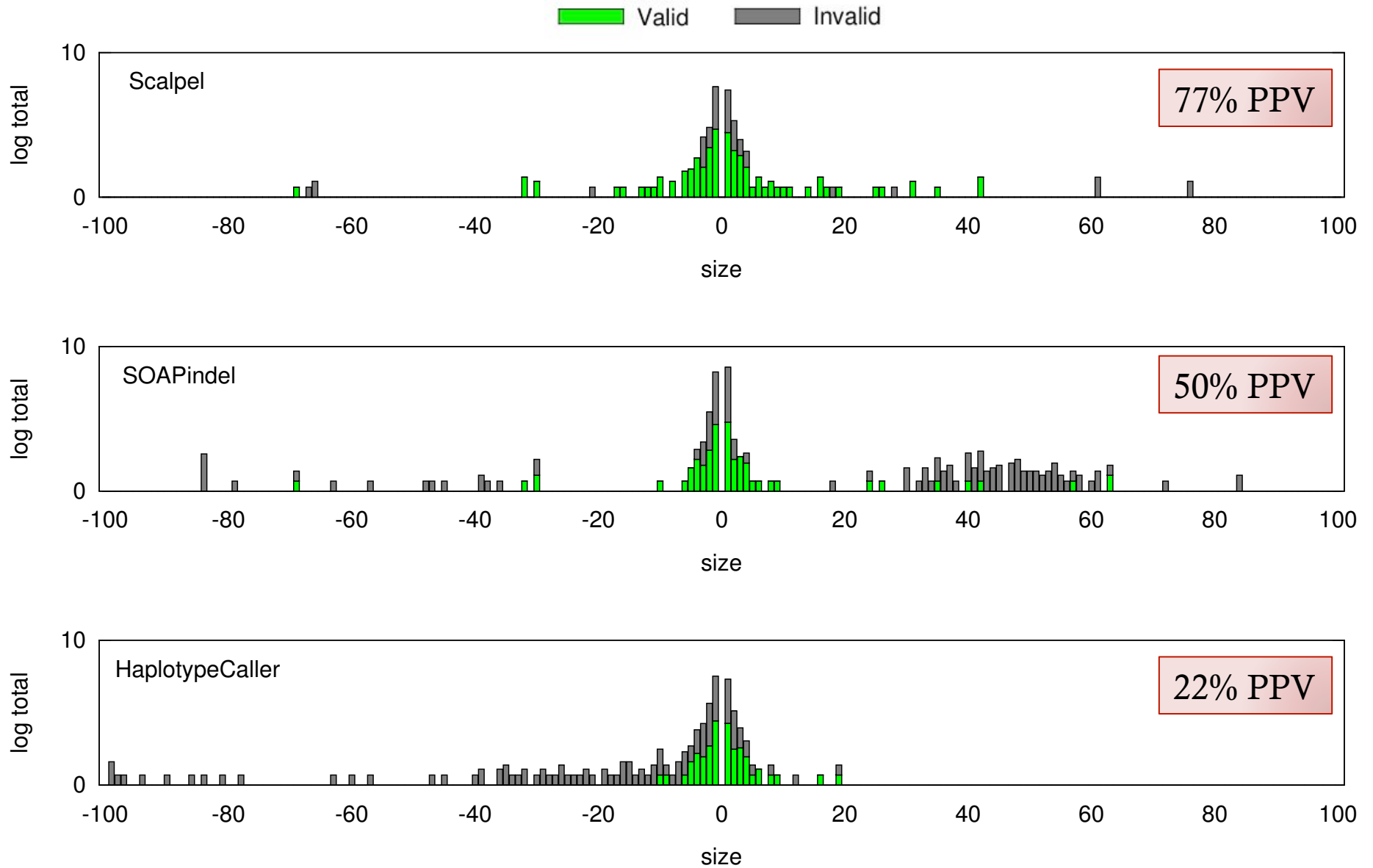
- Individual was diagnosed with ADHD and turrets syndrome
- 80% of the target at >20x coverage
- Evaluated with Scalpel, SOAPindiel, and GATK Haplotype Caller

1000 indels selected for validation

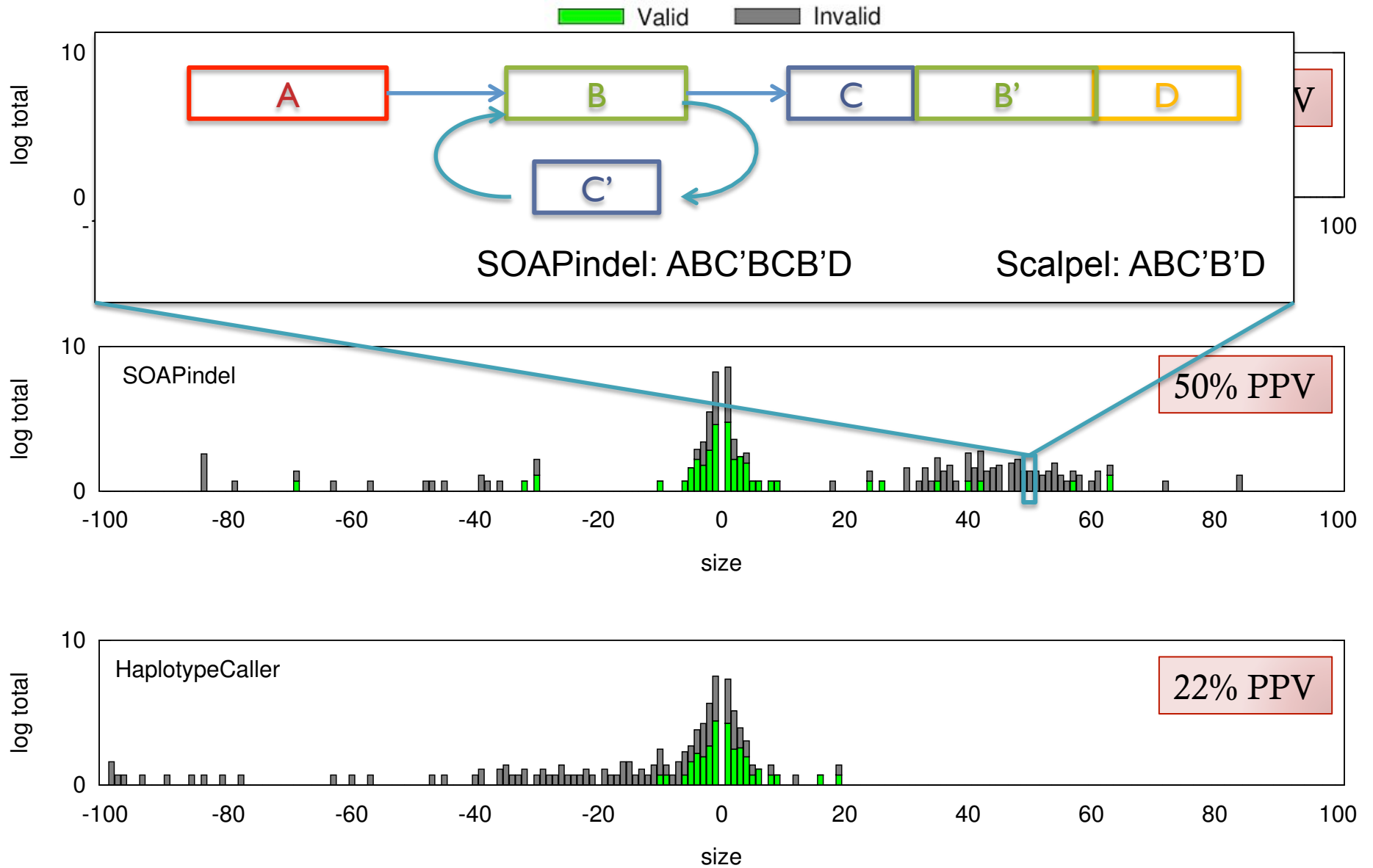
- 200 Scalpel
- 200 GATK Haplotype Caller
- 200 SOAPindiel
- 200 within the intersection
- 200 long indels (>30bp)



# Scalpel Indel Validation

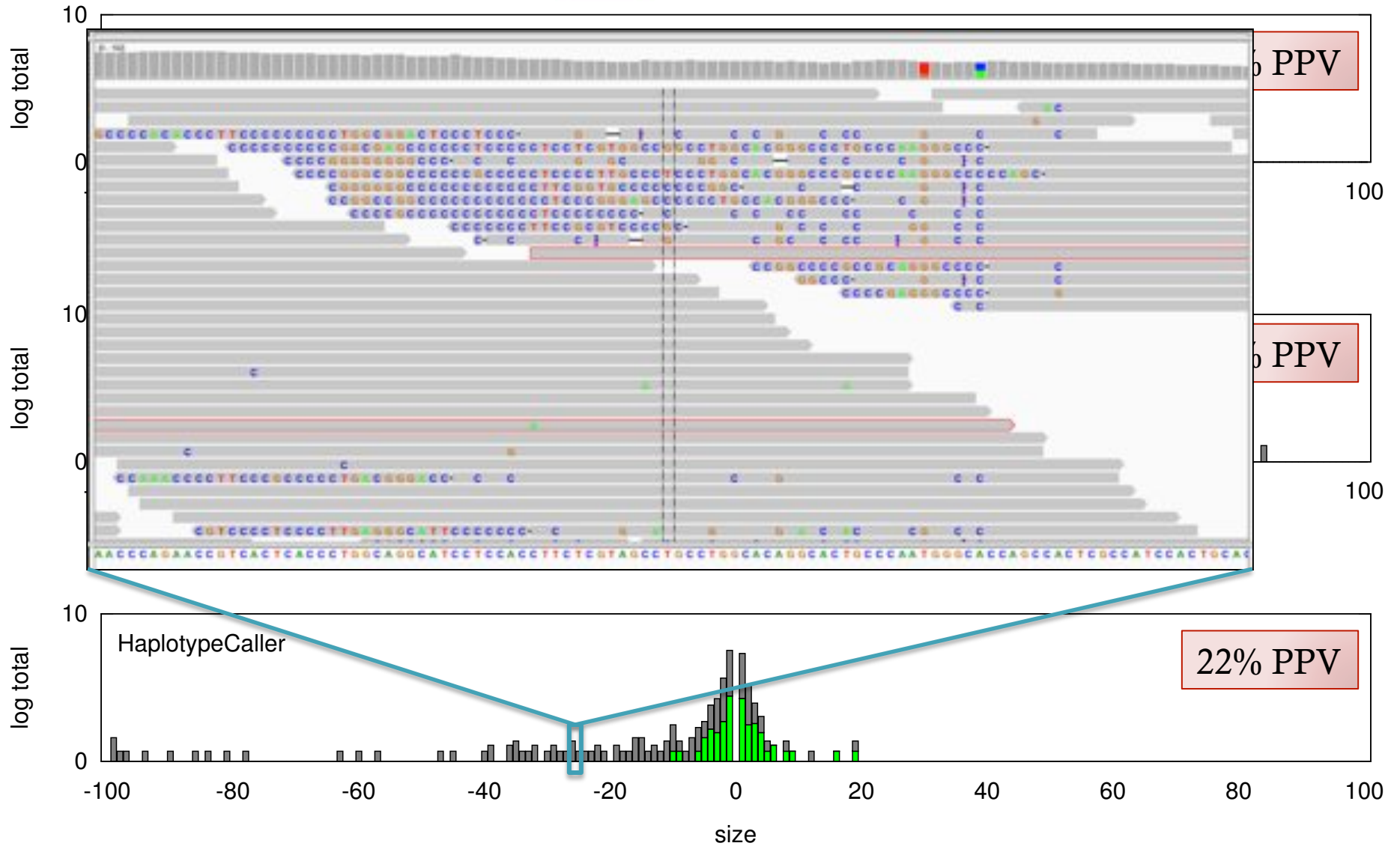


# Scalpel Indel Validation



# Scalpel Indel Validation

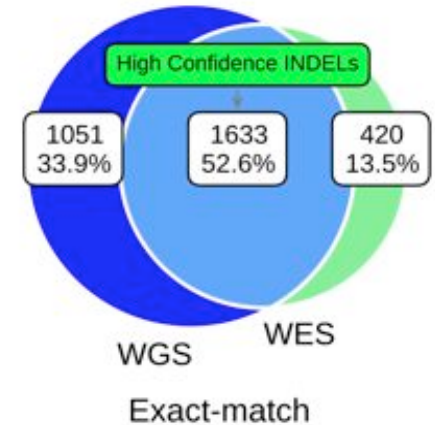
Valid Invalid



# Refined indel analysis

## Examine sources of indel errors

- Experimental validation of indels called from 30x whole genome vs. 110x whole exome of the same sample
- Most of the errors due to short microsatellite errors introduced during exome capture, also misses most long indels
- Recommend WGS for indel analysis instead



	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, Lossifov I, Schatz, MC<sup>§</sup>, Lyon, GL<sup>§</sup>  
<http://www.biorxiv.org/content/early/2014/06/10/006148>

# De novo Genetics of Autism

- In 593 family quads so far, 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 (35:16)
- Confirmed trends observed in previous studies, contributed dozens of new autism candidate genes.
  - 8 out of 35 indel LGDs in autistic children overlapped with the 842 FMRP-associated genes
  - Trends further confirmed in larger study over the entire collection that is currently under review

**Accurate de novo and transmitted indel detection in exome-capture data using microassembly.**  
Narzisi et al. (2014) *Nature Methods* doi:10.1038/nmeth.3069

**The burden of de novo coding mutations in autism spectrum disorders.**  
Iossifov et al (2014) *Under review.*

# Understanding Genome Structure & Function

## **Biotechnology**

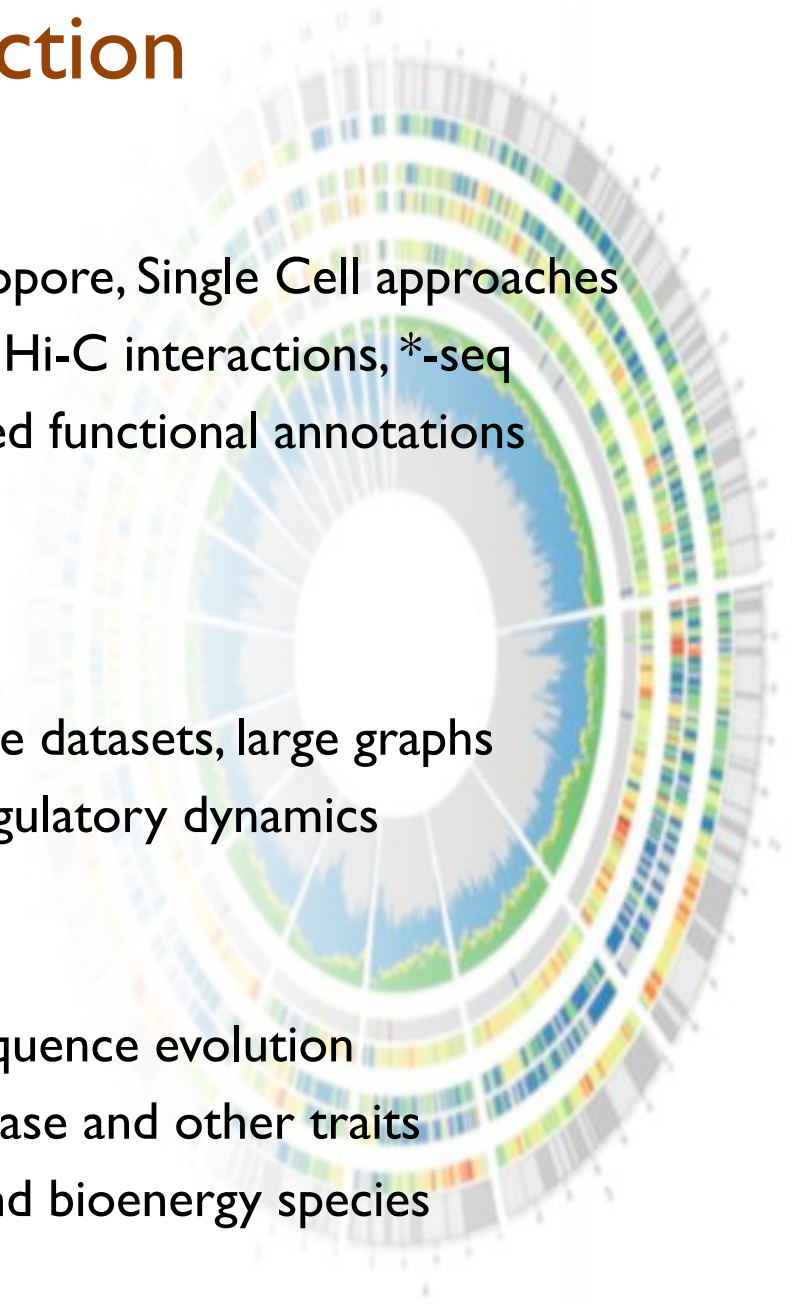
- Sequencing: Illumina, PacBio, Oxford Nanopore, Single Cell approaches
- Biochemical assays: RNA-seq, Methyl-seq, Hi-C interactions, \*-seq
- More accurate sequencing & more detailed functional annotations

## **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**

Rahul Amin  
Tyler Gavin  
James Gurtowski  
Han Fang  
Hayan Lee  
Maria Nattestad  
Aspyn Palatnick  
Srividya  
Ramakrishnan  
  
Eric Biggers  
Ke Jiang  
Shoshana Marcus  
Giuseppe Narzisi  
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  
  
Pacific Biosciences  
Oxford Nanopore



National Human  
Genome Research  
Institute



U.S. DEPARTMENT OF  
**ENERGY**

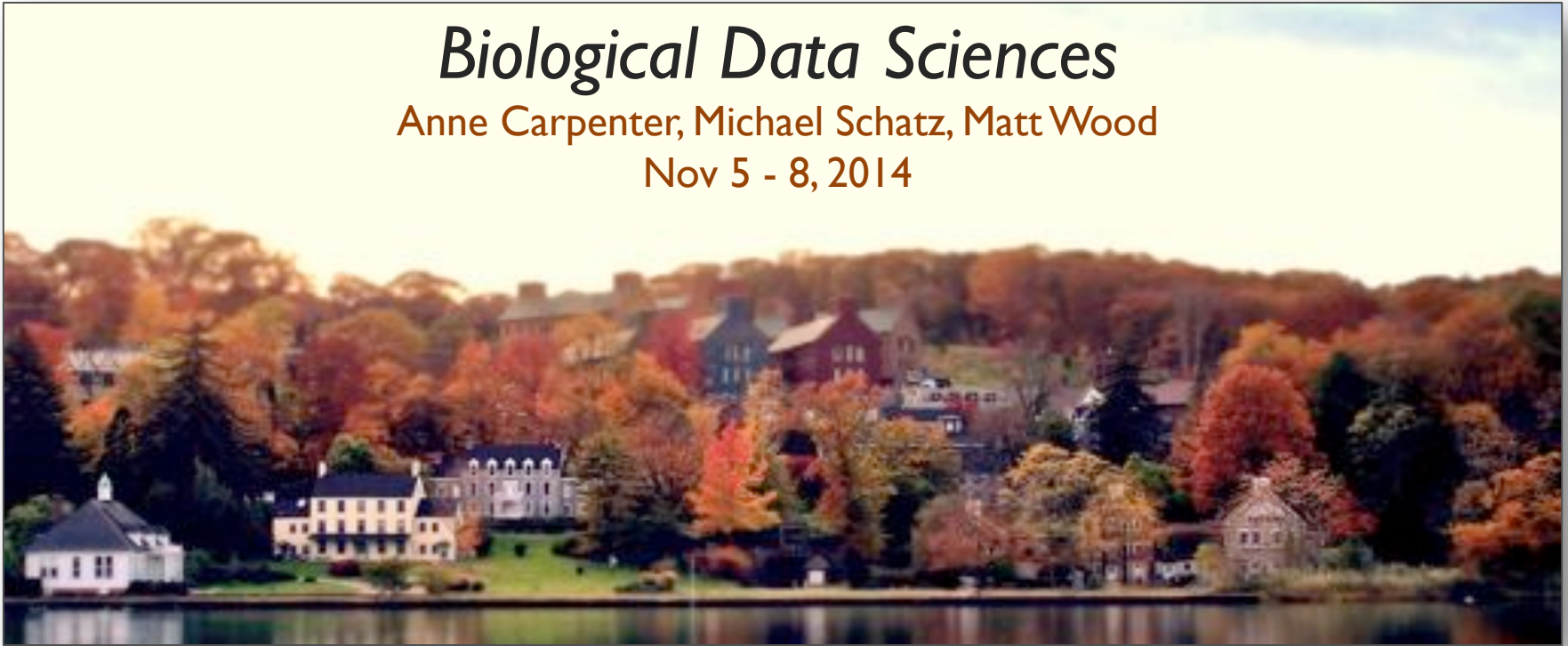
**SFARI**

SIMONS FOUNDATION  
AUTISM RESEARCH INITIATIVE

# *Biological Data Sciences*

Anne Carpenter, Michael Schatz, Matt Wood

Nov 5 - 8, 2014



# Thank you

<http://schatzlab.cshl.edu>

@mike\_schatz