Algorithms for the analysis of complex genomes

Michael Schatz

Oct 18, 2013
CSHL In House

Introductions

Srividya “Sri” Ramakrishnan

DOE Systems Biology Knowledgebase

Worlds fastest genomics pipelines

Tyler Garvin

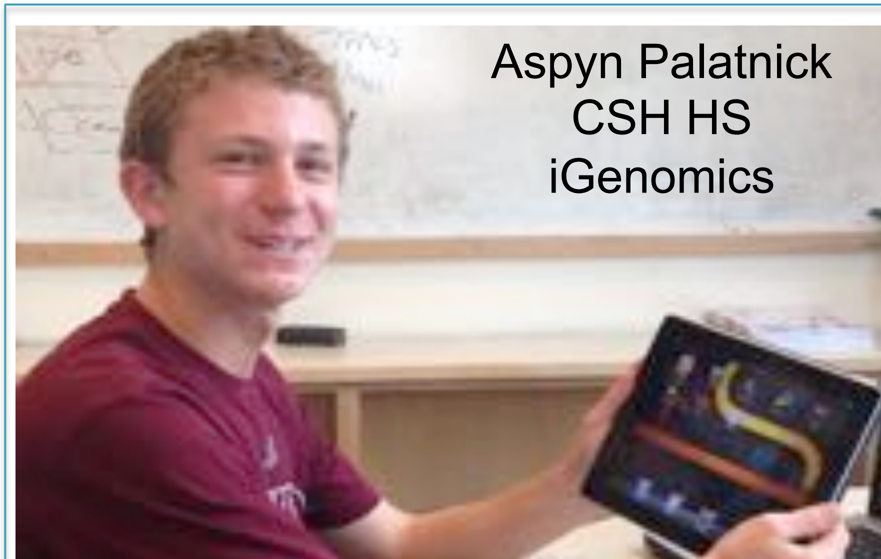
WSBS

Interactive CNV and QC of single cell sequencing

Greg Vurture

CSHL URP / NYU

Mathematics of genomic architecture and heterozygosity



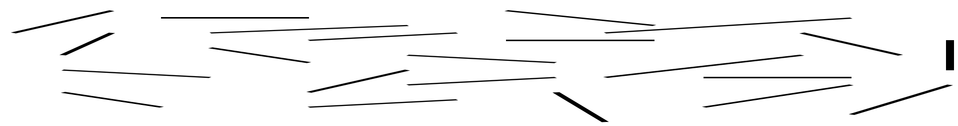
Aspyn Palatnick
CSH HS
iGenomics

Outline

1. Read length & assembly complexity
2. Single molecule assembly of rice
3. De novo indel mutations in autism

Assembling a Genome

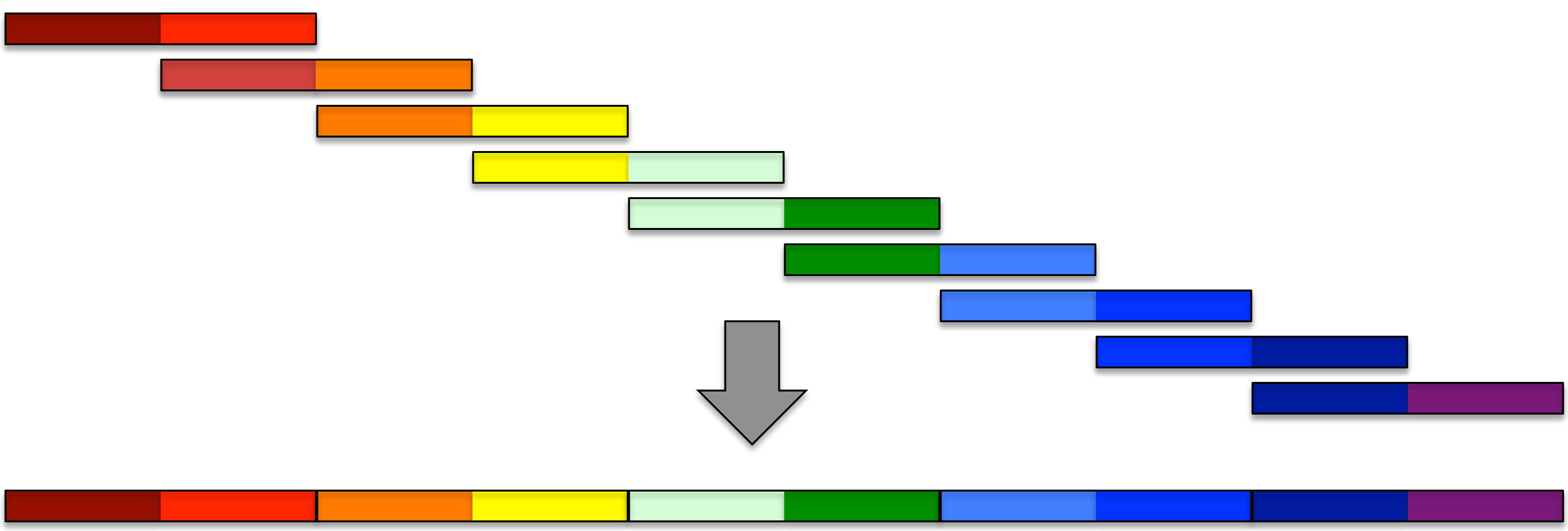
1. Shear & Sequence DNA



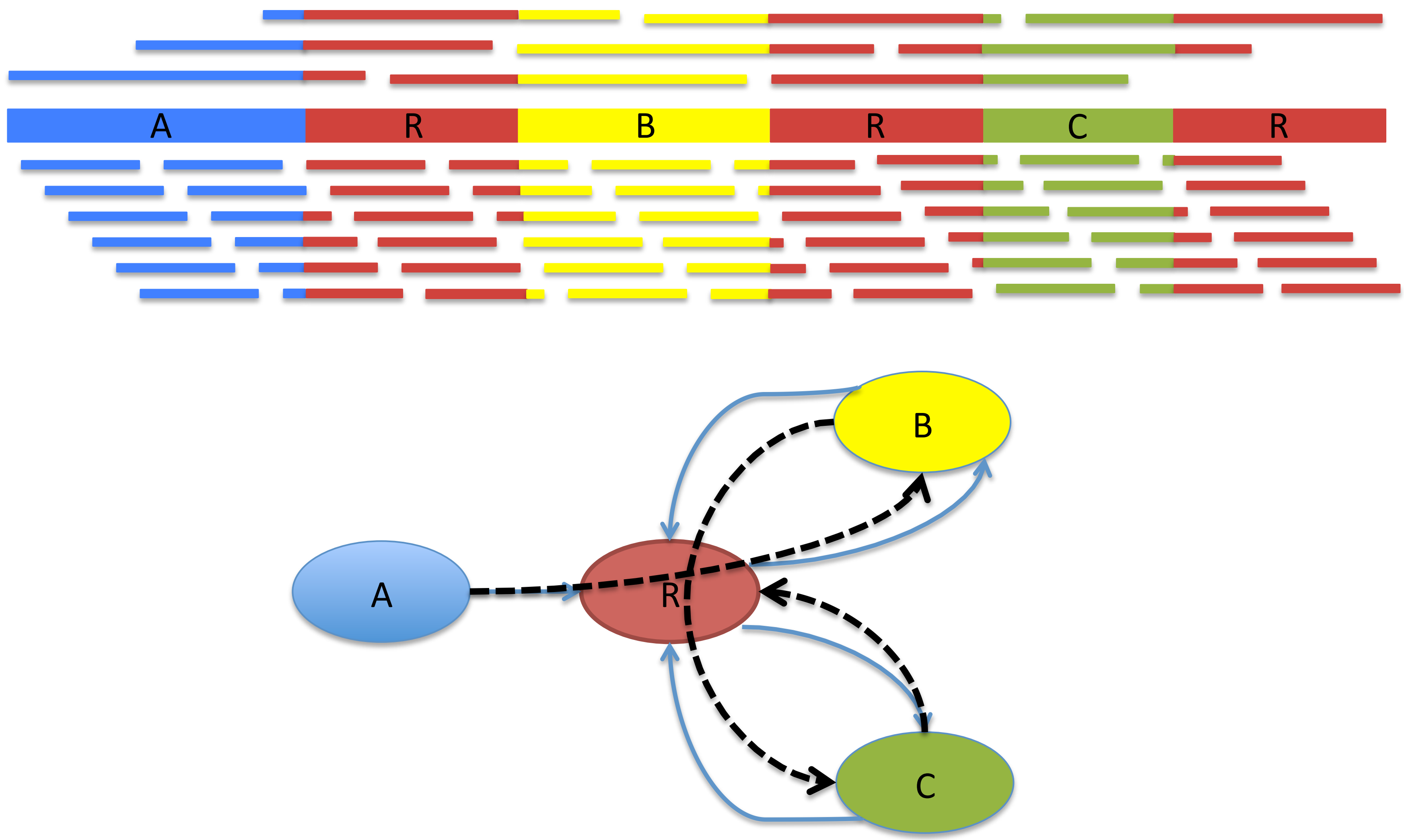
2. Construct assembly graph from overlapping reads

...AGCCTAGGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGTTTGGTCAACCTCGGACGGAC
CAACCTCGGACGGACCTCAGCGAA...

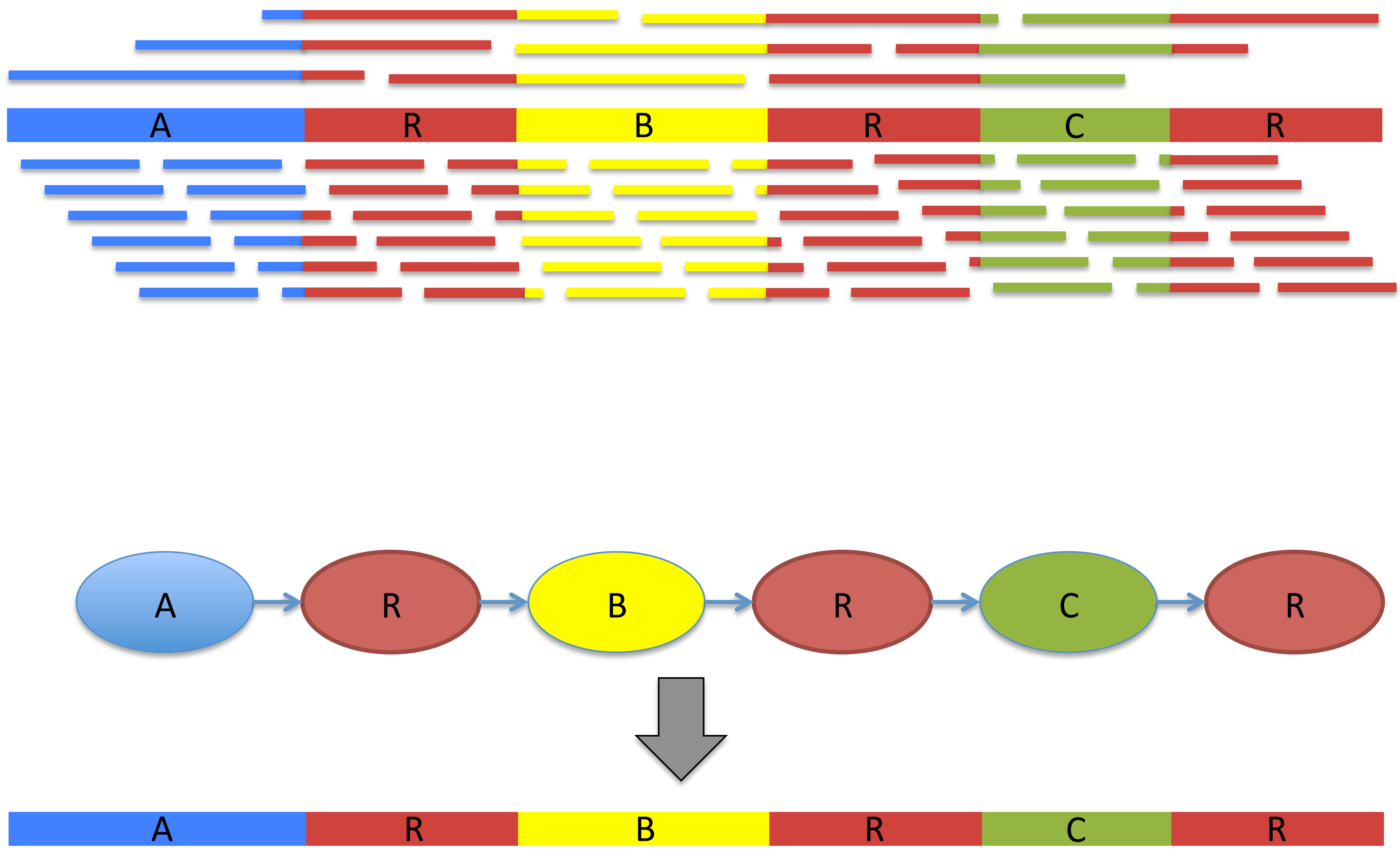
3. Simplify assembly graph



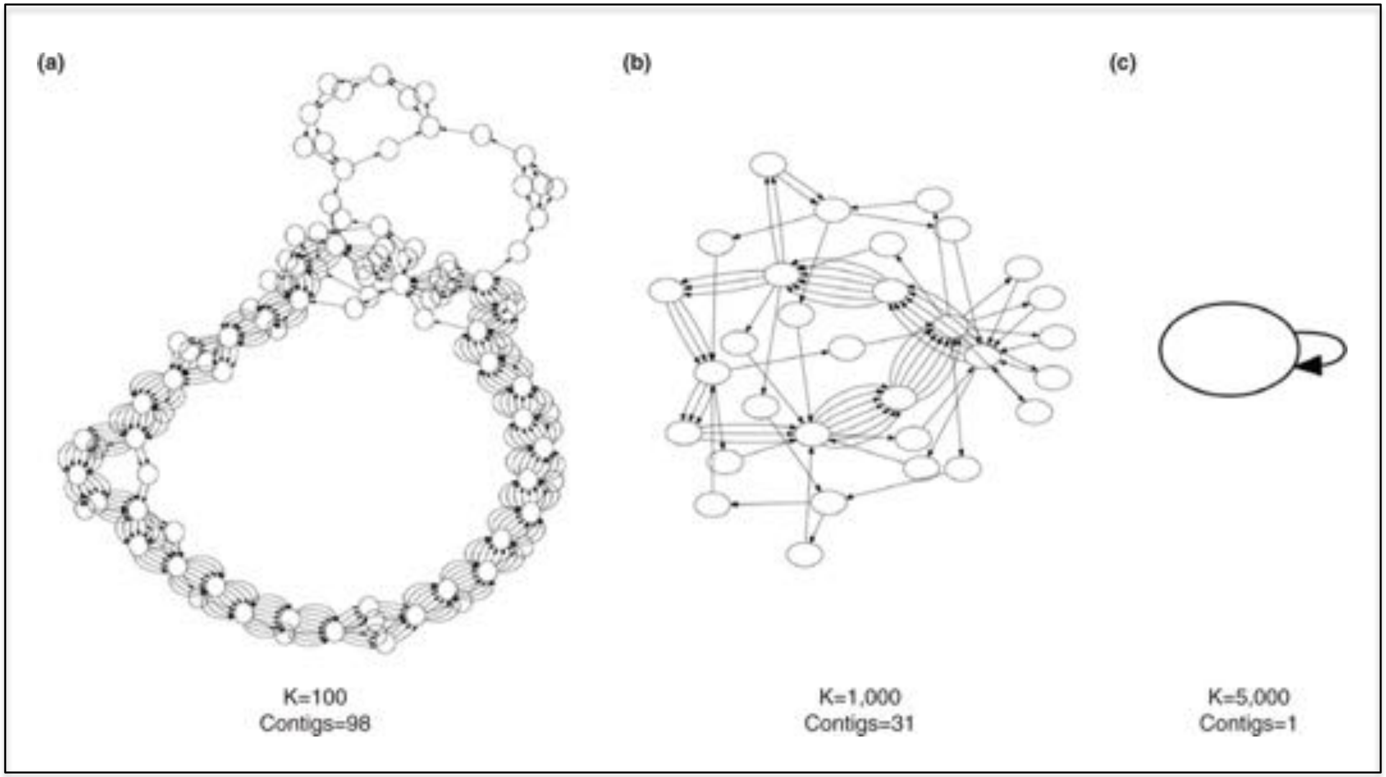
Assembly Complexity



Assembly Complexity



Reducing Complexity



Longer reads span more repeats, simplifying the assembly problem

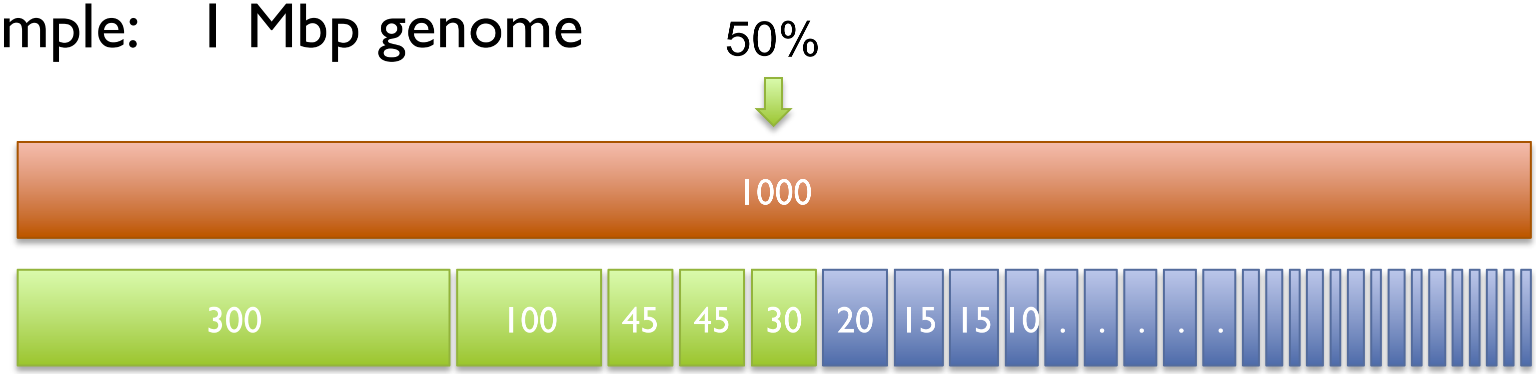
- Idealized assembly of *B. anthracis* reduces to a single contig with 5kb reads
- Exact improvement depends on the specific genome

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 >= 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.

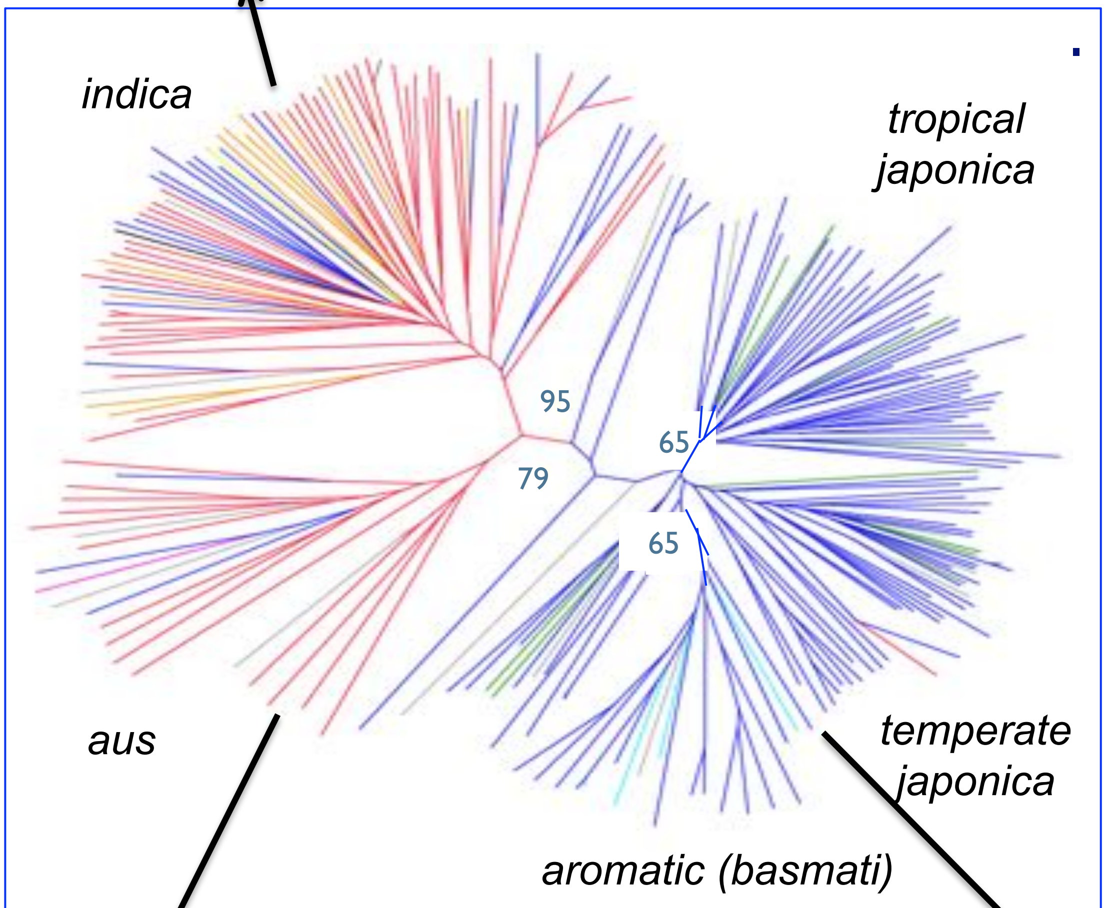
Outline

1. Read length & assembly complexity
2. Single molecule assembly of rice
3. De novo indel mutations in autism

Population structure of *Oryza sativa*

3 varieties selected for *de novo* sequencing

IR64

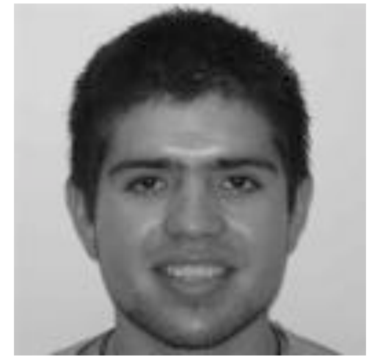


DJ123

Garris et al. (2005)
Genetics 169: 1631–1638

Nipponbare

Assembly and Annotation



Indica

Total Span: 344.3 Mbp
Scaffold N50: 293kbp
Contig N50: 22.2kbp
Unique genes: 598

Aus

Total Span: 344.9Mbp
Scaffold N50: 323kbp
Contig N50: 25.5kbp
Unique genes: 502

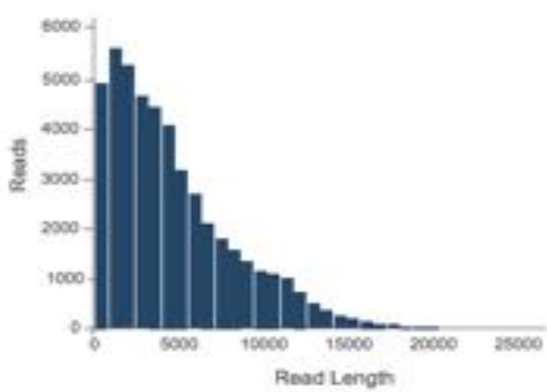
Nipponbare

Total Span: 354.9Mbp
Scaffold N50: 213kbp
Contig N50: 21.9kbp
Unique genes: 1093

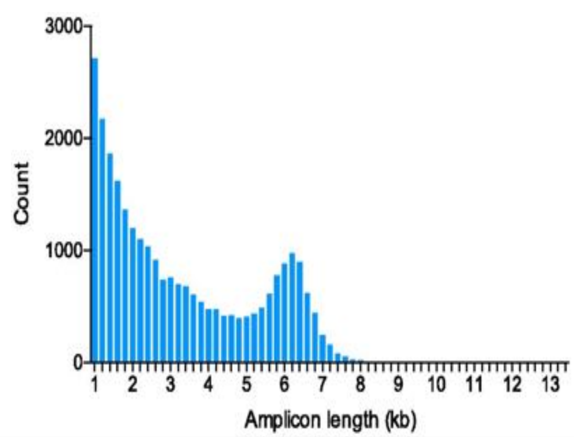
New whole genome de novo assemblies of three divergent strains of rice documents novel gene space of *Aus* and *Indica* subpopulations
Schatz, MC, McCombie, WR, Ware, DW, McCouch, S, *et al* (2013) *In preparation*

Single Molecule Sequencing Technology

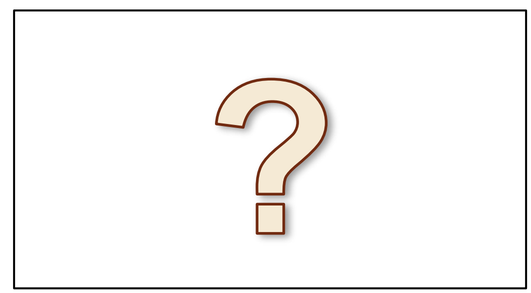
PacBio RS II



Moleculo

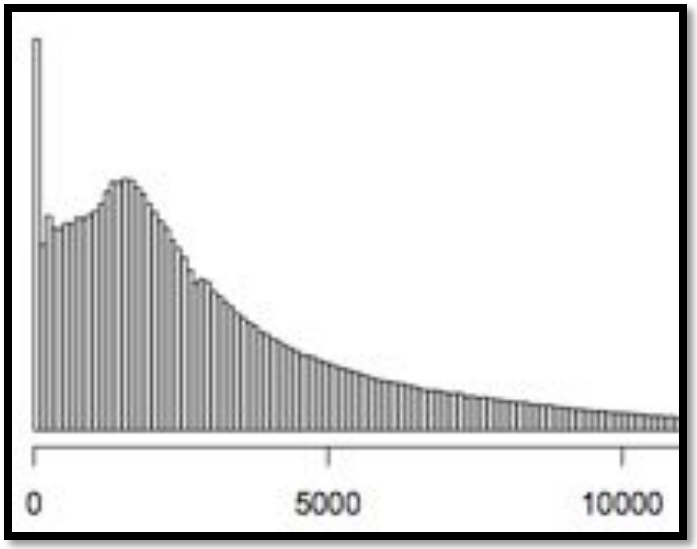


Oxford Nanopore



Clive G. Brown @Clive_G_Brown 9 Oct
I've reluctantly rejoined twitter purely so that I can make one tweet - when the appropriate time arises ...
Expand

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

PacBio Error Correction: HGAP

- With 50-100x of Pacbio coverage, virtually all of the errors can be eliminated
  - Works well for Microbial genomes: single contig per chromosome routinely achieved
  - Difficult to scale up for use with eukaryotic genomes

Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data
Chin, CS *et al.* (2013) *Nature Methods*. 10: 563-569

Hybrid Sequencing

Illumina
Sequencing by Synthesis

High throughput (60Gbp/day)
High accuracy (~99%)
Short reads (~100bp)

Pacific Biosciences
SMRT Sequencing

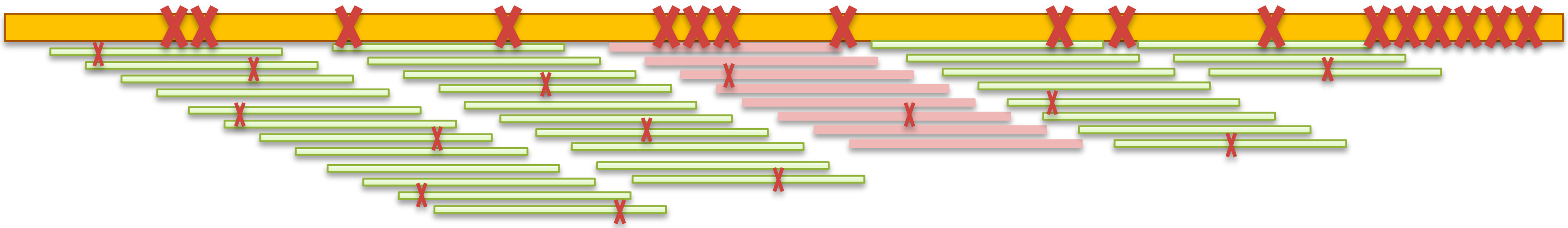
Lower throughput (1Gbp/day)
Lower accuracy (~85%)
Long reads (5kbp+)

Hybrid Error Correction: PacBioToCA

http://wgs-assembler.sf.net

1. Correction Pipeline
   1. Map short reads to long reads
   2. Trim long reads at coverage gaps
   3. Compute consensus for each long read

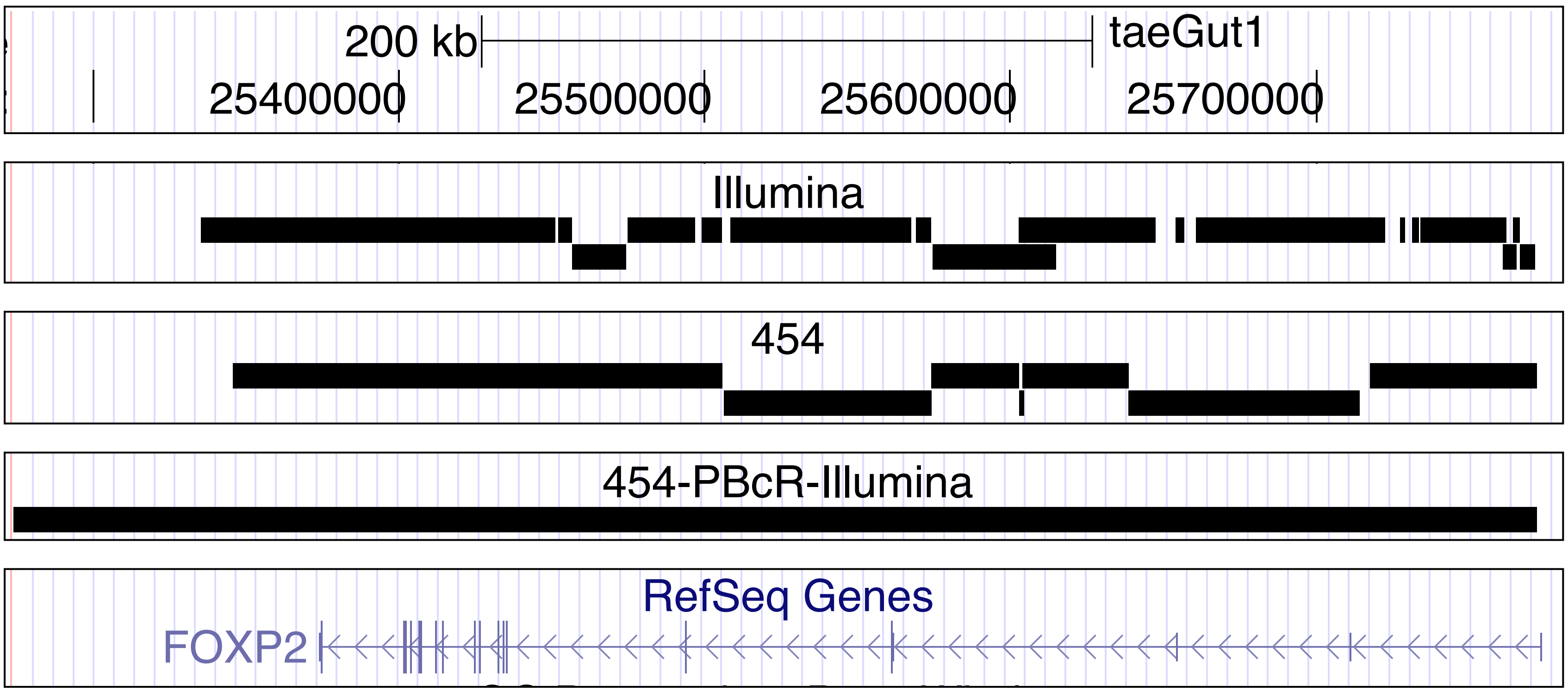
2. Error corrected reads can be easily assembled, aligned



Hybrid error correction and de novo assembly of single-molecule sequencing reads.
Koren, S, Schatz, MC, *et al.* (2012) *Nature Biotechnology*. doi:10.1038/nbt.2280

Improved Gene Reconstruction

FOXP2 assembled in a single contig in the PacBio parrot assembly

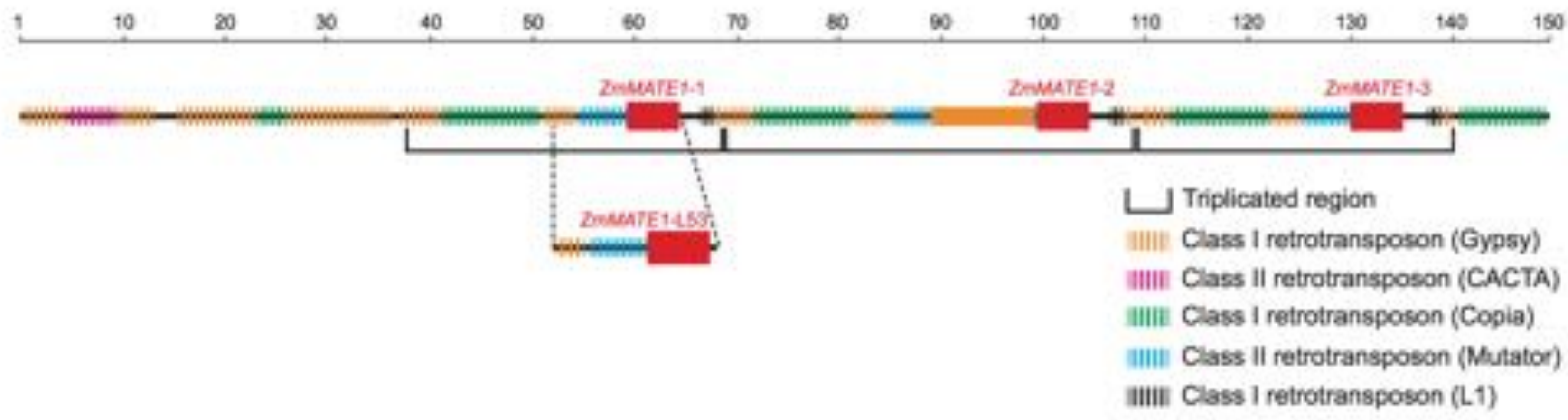


Hybrid error correction and de novo assembly of single-molecule sequencing reads.
Koren, S, Schatz, MC, *et al.* (2012) *Nature Biotechnology*. doi:10.1038/nbt.2280

Long Read CNV Analysis

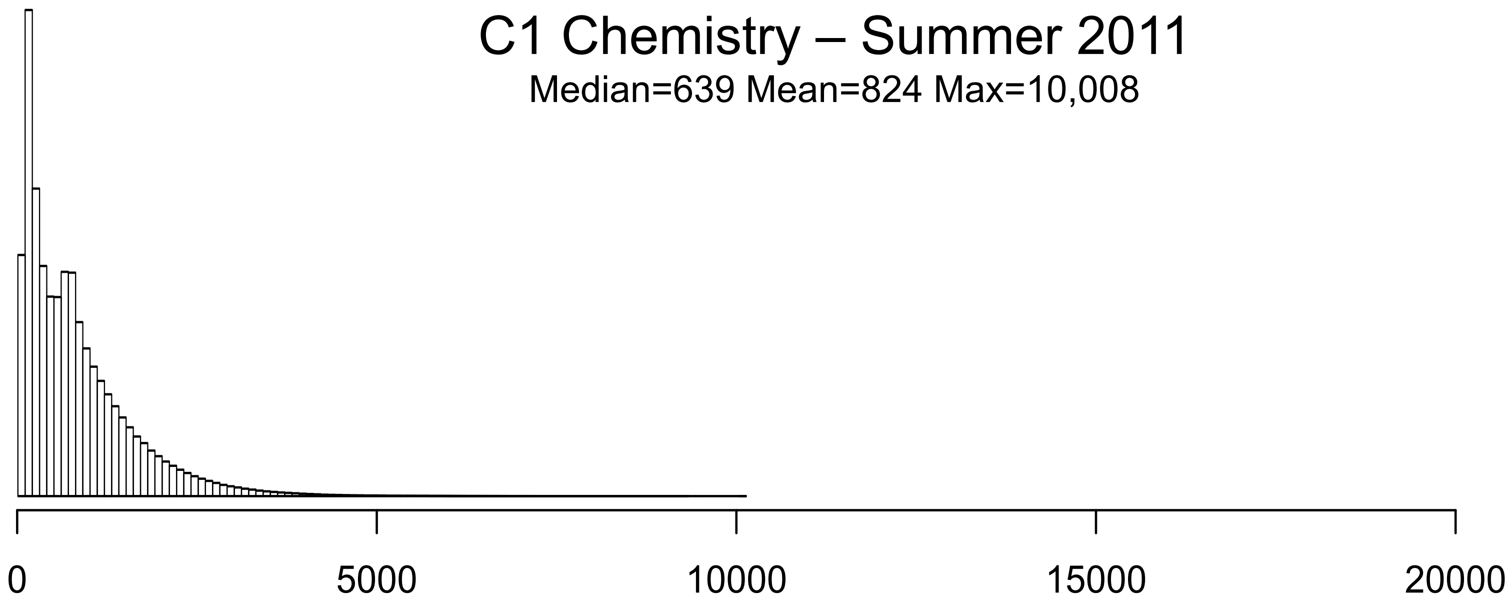
Aluminum tolerance in maize is important for drought resistance and protecting against nutrient deficiencies

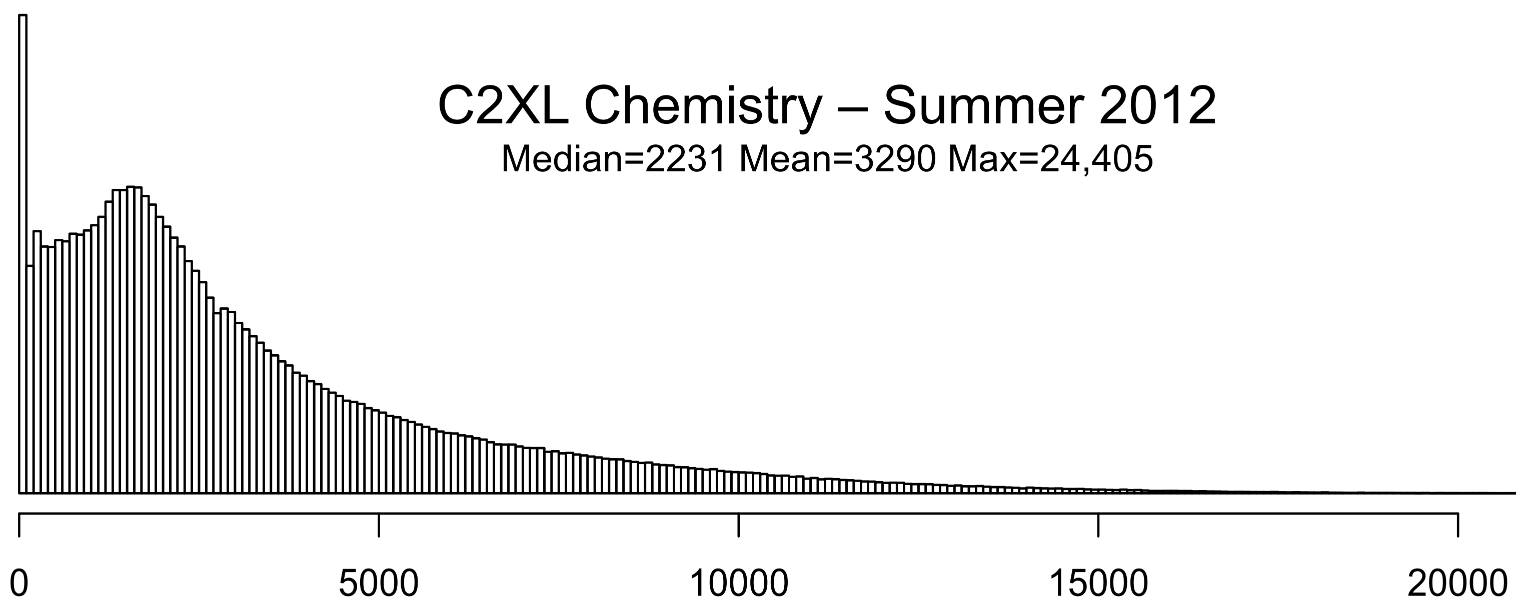
- Segregating population localized a QTL on a BAC, but unable to genotype with Illumina sequencing because of high repeat content and GC skew
- Long read PacBio sequencing corrected by CCS reads revealed a triplication of the ZnMATE1 membrane transporter



A rare gene copy-number variant that contributes to maize aluminum tolerance and adaptation to acid soils
Maron, LG *et al.* (2013) PNAS doi: 10.1073/pnas.1220766110

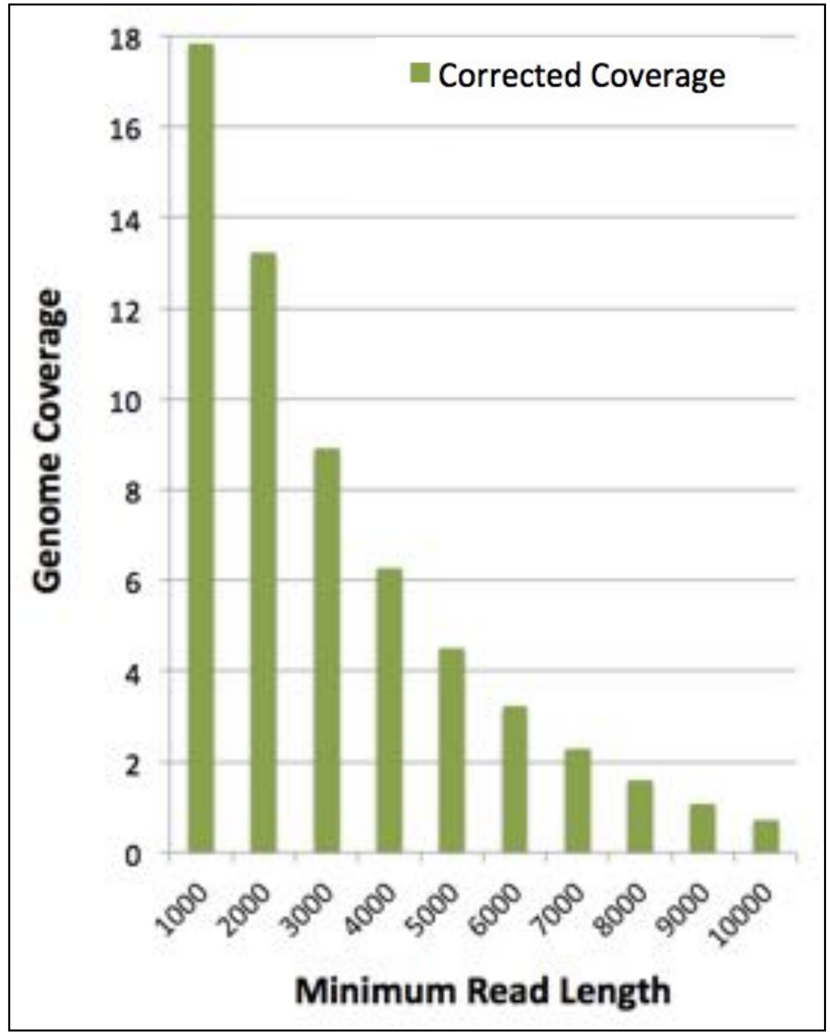
PacBio Long Read Rice Sequencing





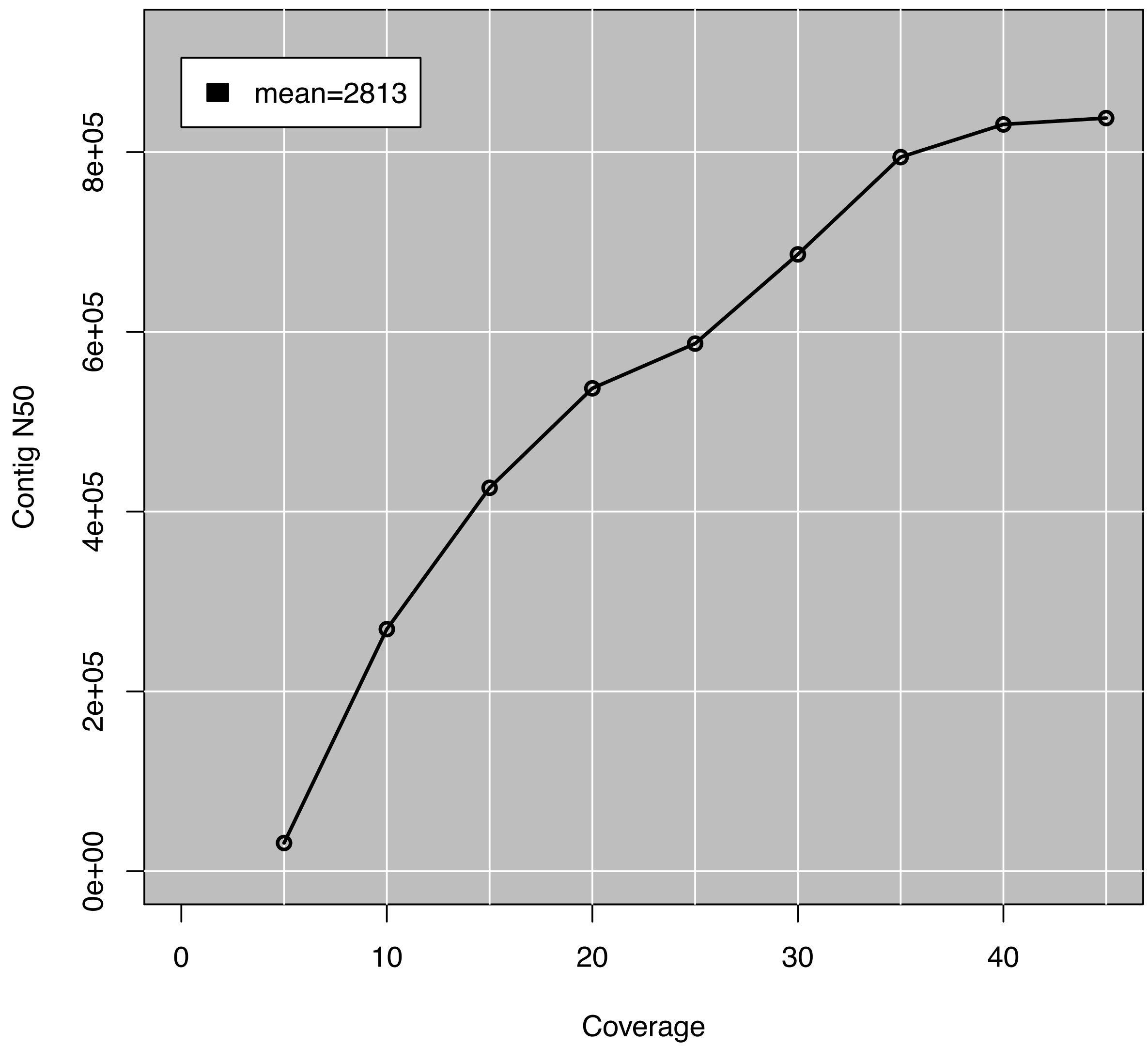
Preliminary Rice Assemblies

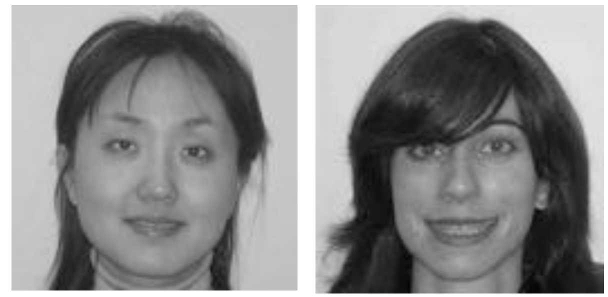
Assembly	Contig NG50
HiSeq Fragments 50x 2x100bp @ 180	3,925
MiSeq Fragments 23x 459bp 8x 2x251bp @ 450	6,332
"ALLPATHS-recipe" 50x 2x100bp @ 180 36x 2x50bp @ 2100 51x 2x50bp @ 4800	18,248



In collaboration with McCombie & Ware labs @ CSHL

Assembly Coverage Model





Simulate PacBio-like reads to predict how the assembly will improve as we add additional coverage

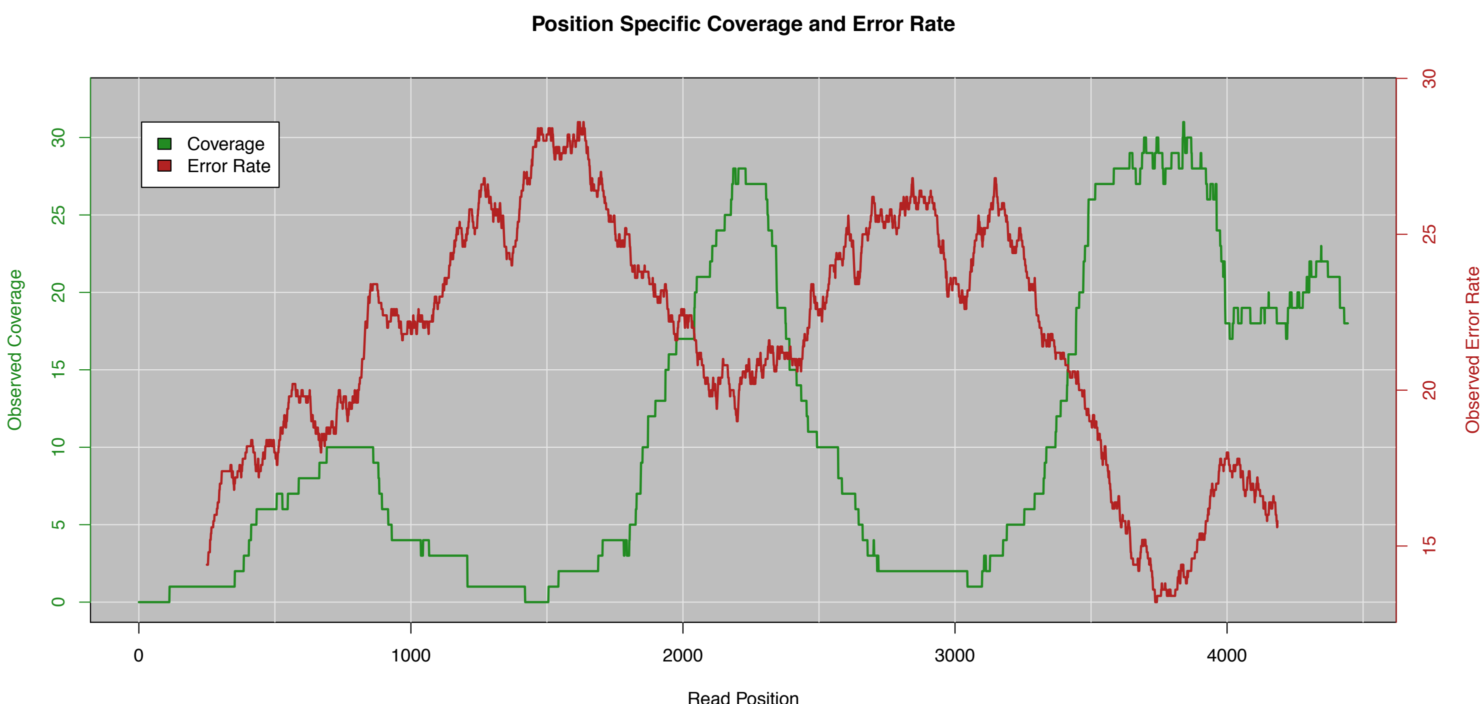
Only 8x coverage is needed to sequence every base in the genome, but 40x improves the chances repeats will be spanned by the longest reads

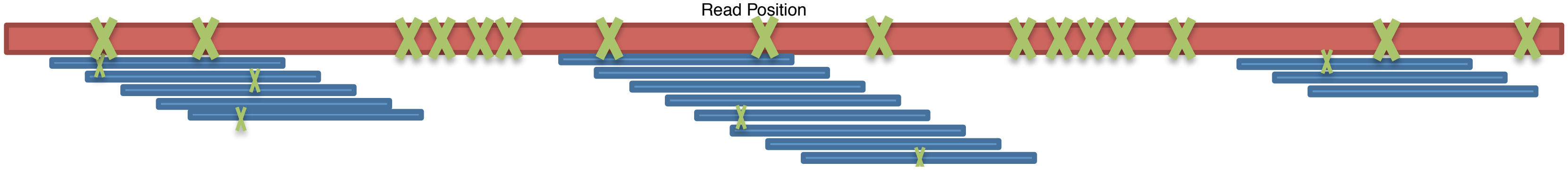
Assembly complexity of long read sequencing
Lee, H*, Gurtowski, J*, Yoo, S, Marcus, S, McCombie, WR, Schatz MC *et al.* (2013) *In preparation*

Enhanced PacBio Error Correction

PacBioToCA fails in complex regions

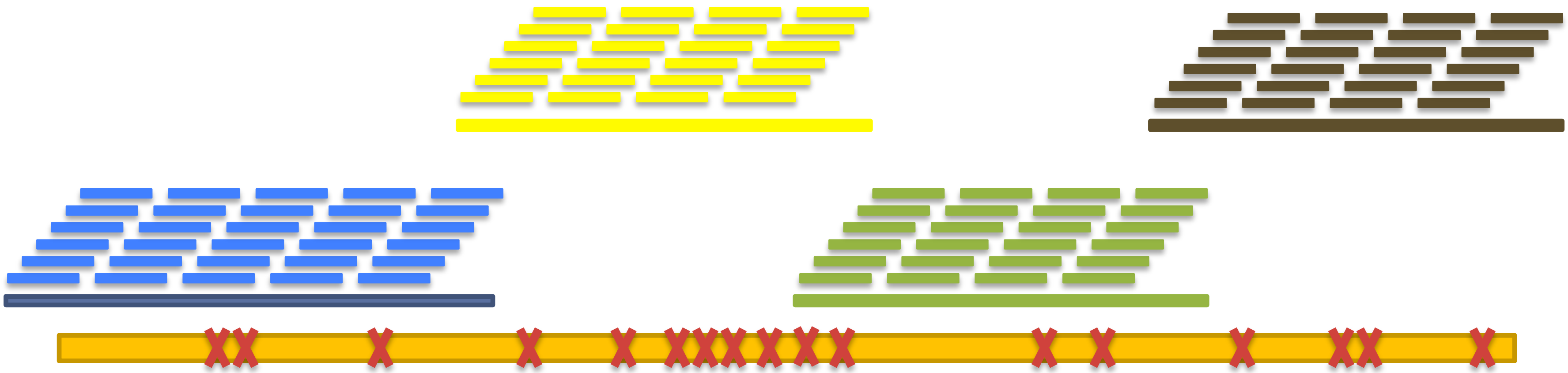
1. Simple Repeats – Kmer Frequency Too High to Seed Overlaps
2. Error Dense Regions – Difficult to compute overlaps with many errors
3. Extreme GC – Lacks Illumina Coverage





Error Correction with pre-assembled Illumina reads

https://github.com/jgurtowski/pbtools



Short Reads -> Assemble Unitigs -> Align & Select - > Error Correct

Unitigs:

High quality contigs formed from unambiguous, unique overlaps of reads

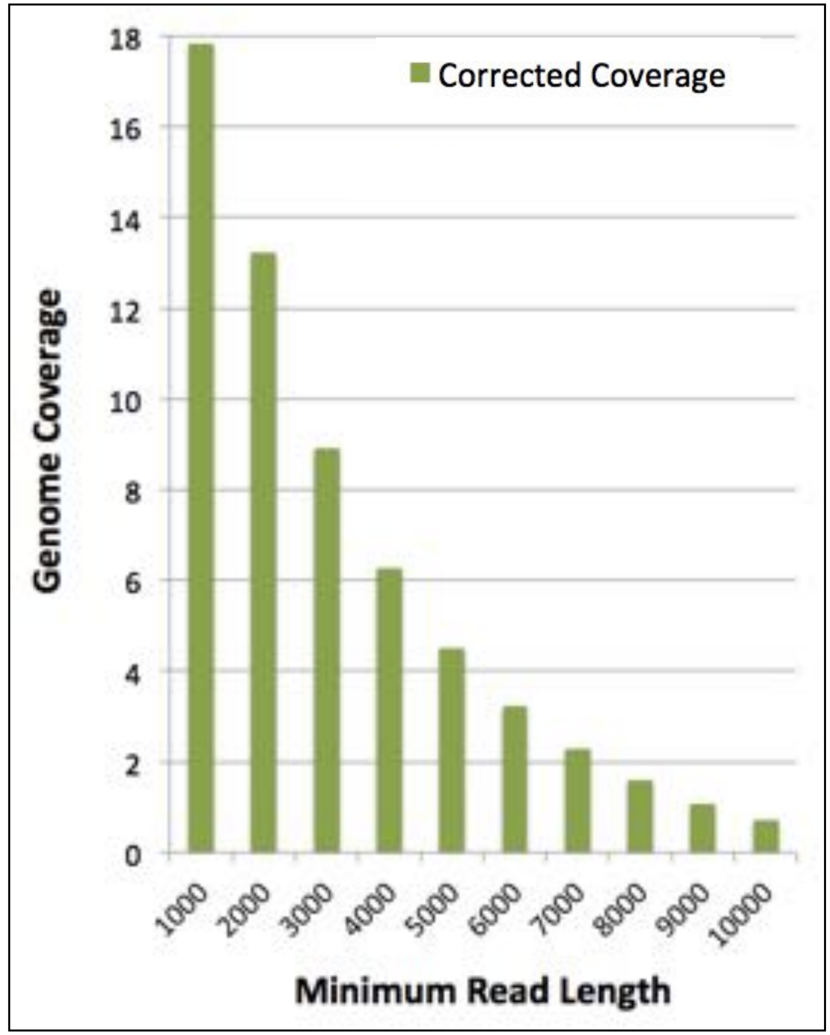
Each read is placed into a single unitig

Can Help us overcome:

1. **Simple Repeats – Kmer Frequency Too High to Seed Overlaps**
2. **Error Dense Regions – Difficult to compute overlaps with many errors**

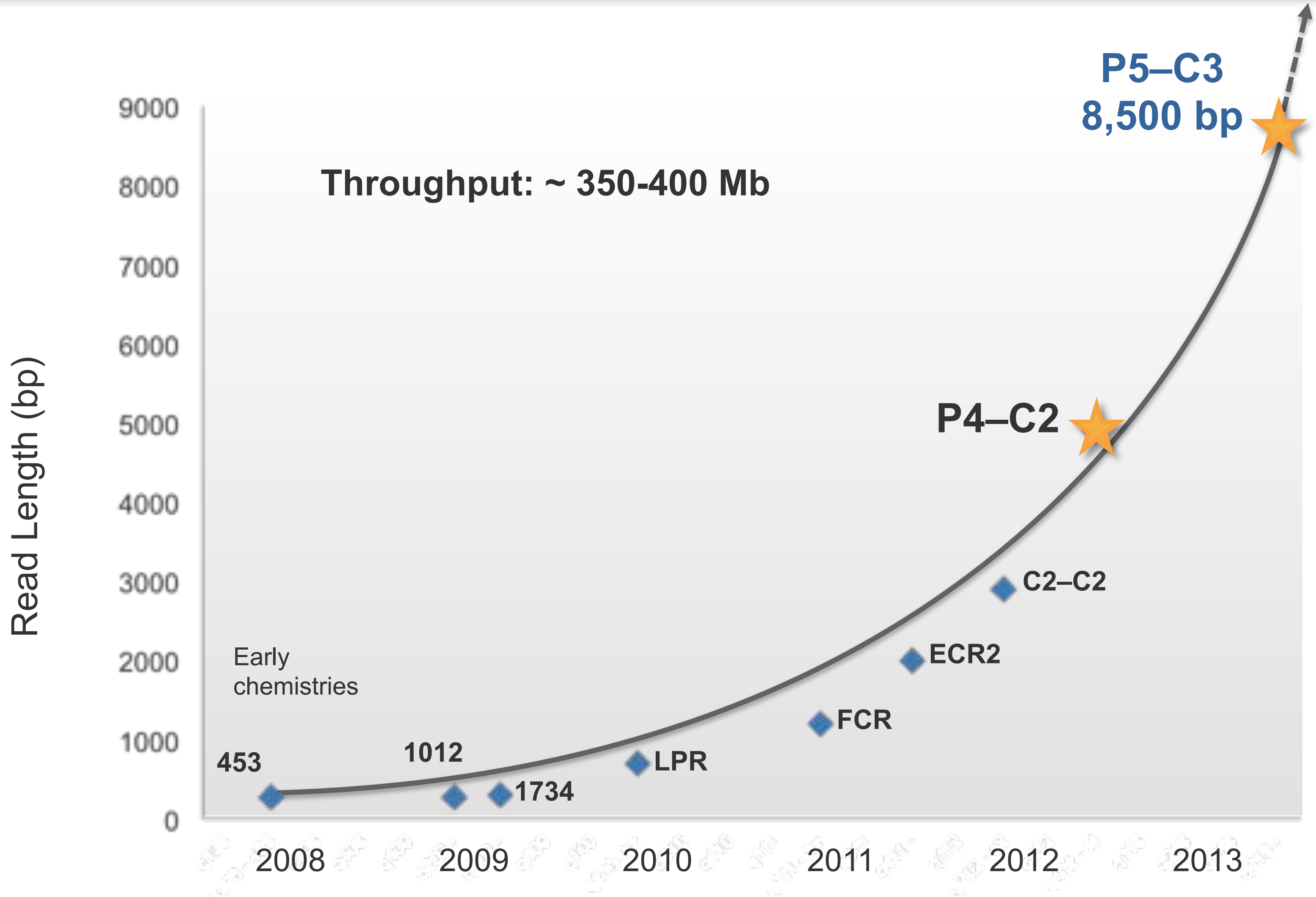
Preliminary Rice Assemblies

Assembly	Contig NG50
HiSeq Fragments 50x 2x100bp @ 180	3,925
MiSeq Fragments 23x 459bp 8x 2x251bp @ 450	6,332
"ALLPATHS-recipe" 50x 2x100bp @ 180 36x 2x50bp @ 2100 51x 2x50bp @ 4800	18,248
PBeCR Reads 19x @ 3500 ** MiSeq for correction	50,995



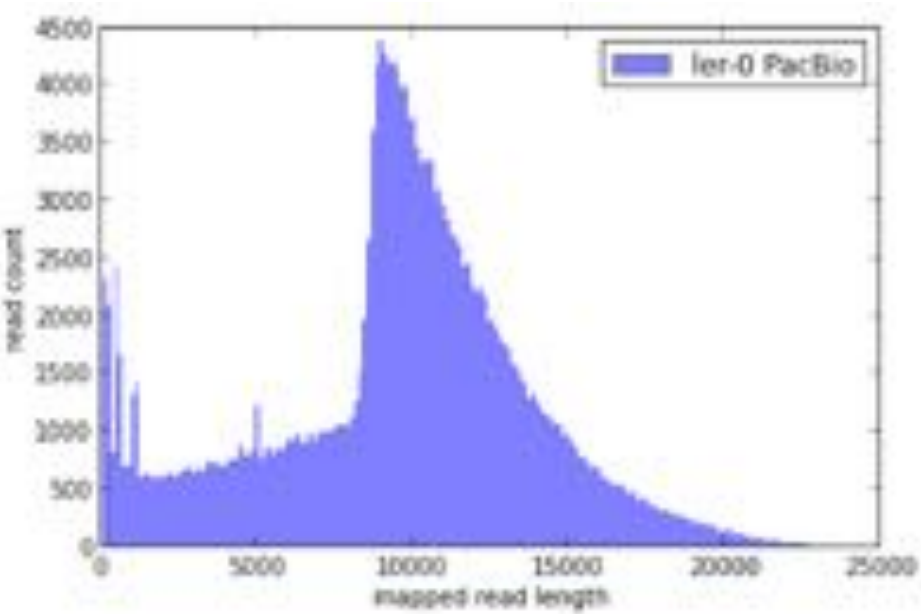
In collaboration with McCombie & Ware labs @ CSHL

P5-C3 Chemistry Read Lengths



De novo assembly of Arabidopsis

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

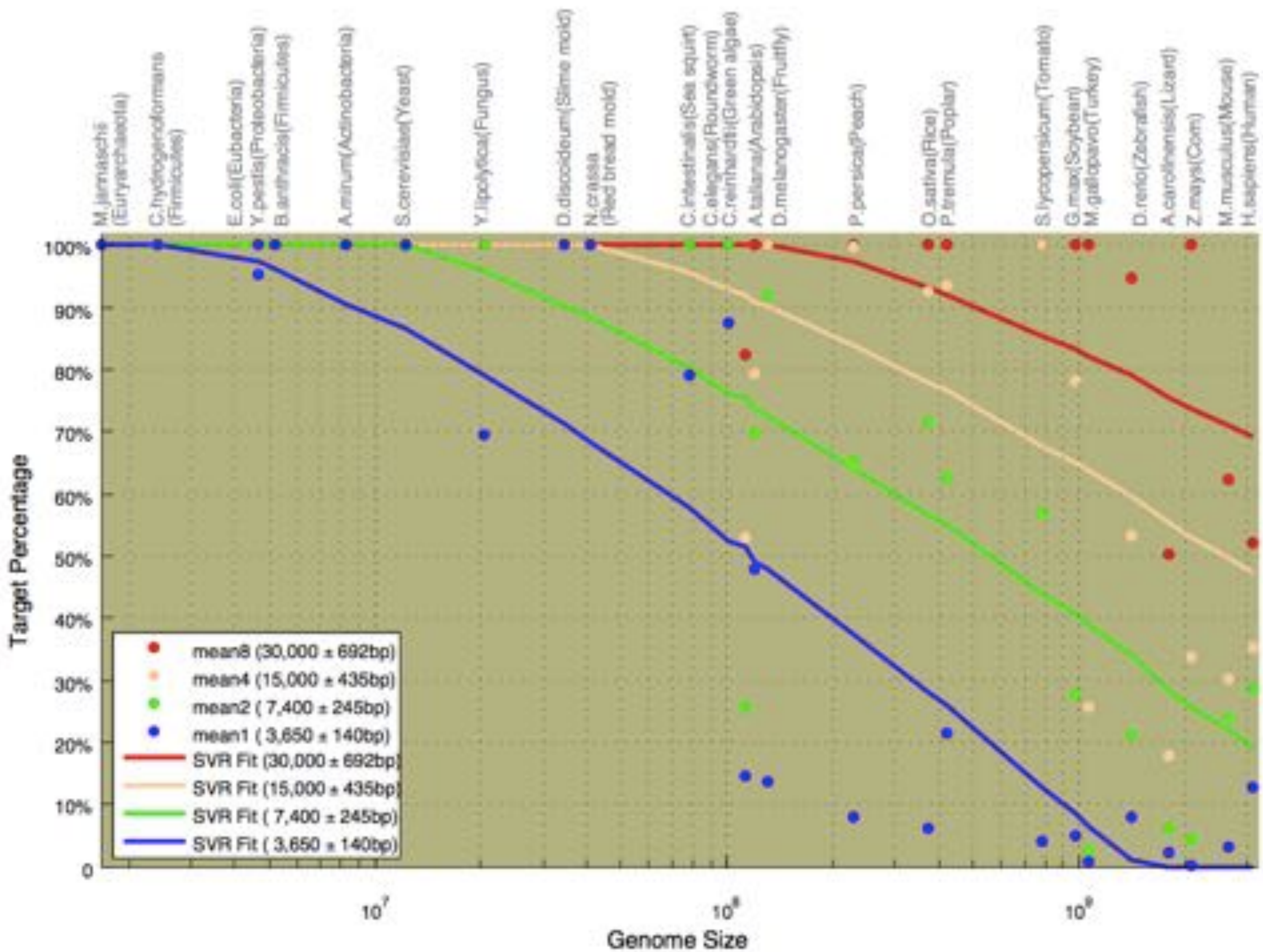


A. thaliana Ler-0 sequenced at PacBio

- Sequenced using the latest 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 >100x

Genome size:	124.6 Mb	Sum of Contig Lengths:	149.5Mb
GC content:	33.92%	Number of Contigs:	1788
Raw data:	11 Gb	Max Contig Length:	12.4 Mb
Assembly coverage:	15x over 9kbp	N50 Contig Length:	8.4 Mb

Assembly Complexity of Long Reads



Outline

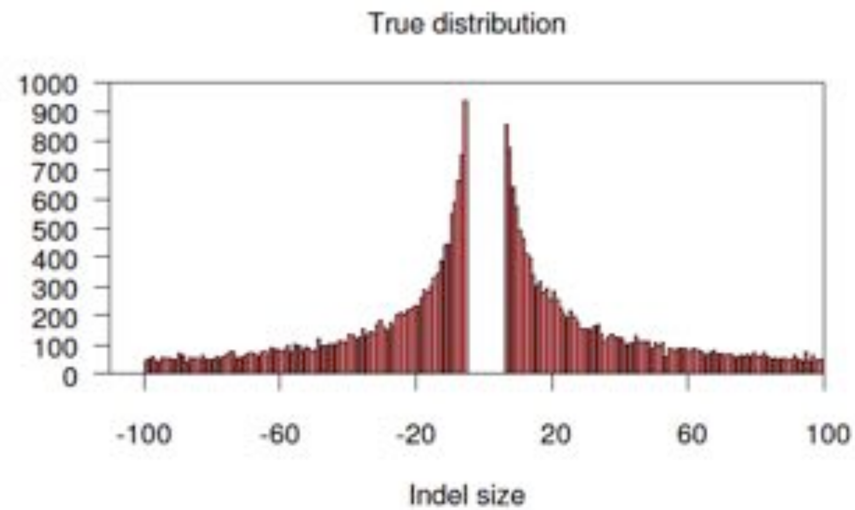
1. Read length & assembly complexity
2. Single molecule assembly of rice
3. De novo indel mutations in autism

Variation Detection Complexity

SNPs + Short Indels

High precision and sensitivity

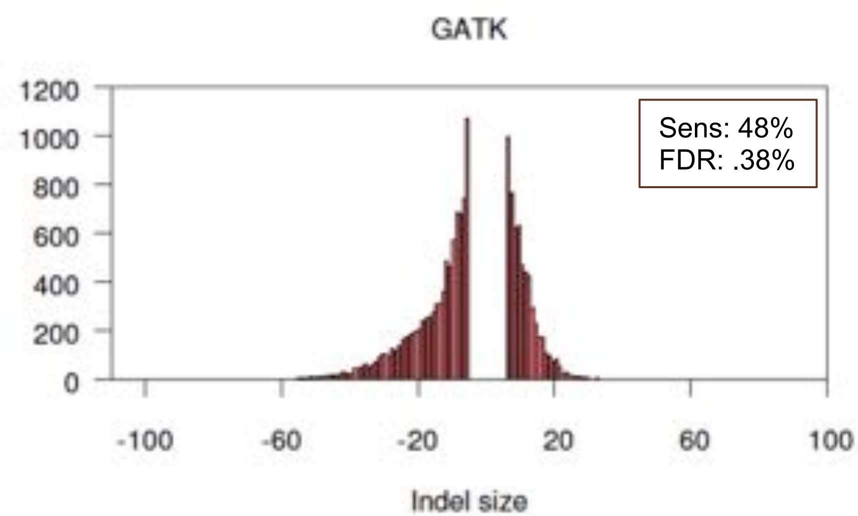
```
..TTTAGAATAG-CGAGTGC...
     ||||||| ||||
     AGAATAGGCGAG
```



"Long" Indels (>5bp)

Reduced precision and sensitivity

```
..TTTAG--------AGTGC...
  |||||        |||||
  TTTAGAATAGGC |||||
        ATAGGCGAGTGC
```



Analysis confounded by sequencing errors, localized repeats, allele biases, and mismapped reads

Scalpel: Haplotype Microassembly

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

Features

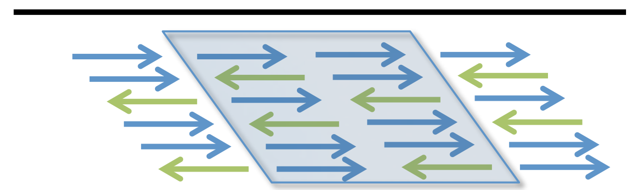
1. Combine mapping and assembly
2. Exhaustive search of haplotypes
3. De novo mutations

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

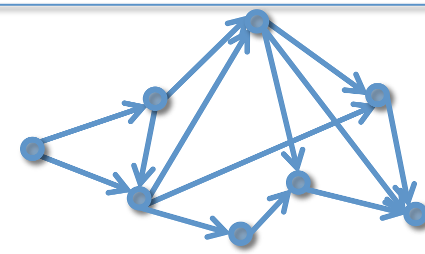
SCALPEL: Micro-assembly approach to accurately detect de novo and transmitted indel mutations within exome-Capture data
Narzisi, G, O'Rawe, J, Iossifov, I, Lee, Y, Wang, Z, Wu, Y, Lyon, G, Wigler, M, Schatz, MC (2013) *In preparation*

Scalpel Pipeline

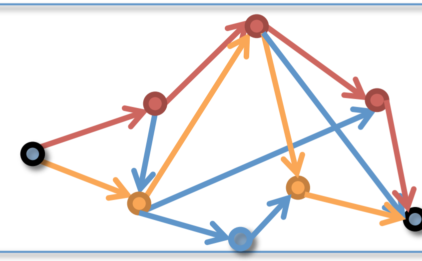
Extract reads mapping within the exon including (1) well-mapped reads, (2) soft-clipped reads, and (3) anchored pairs



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



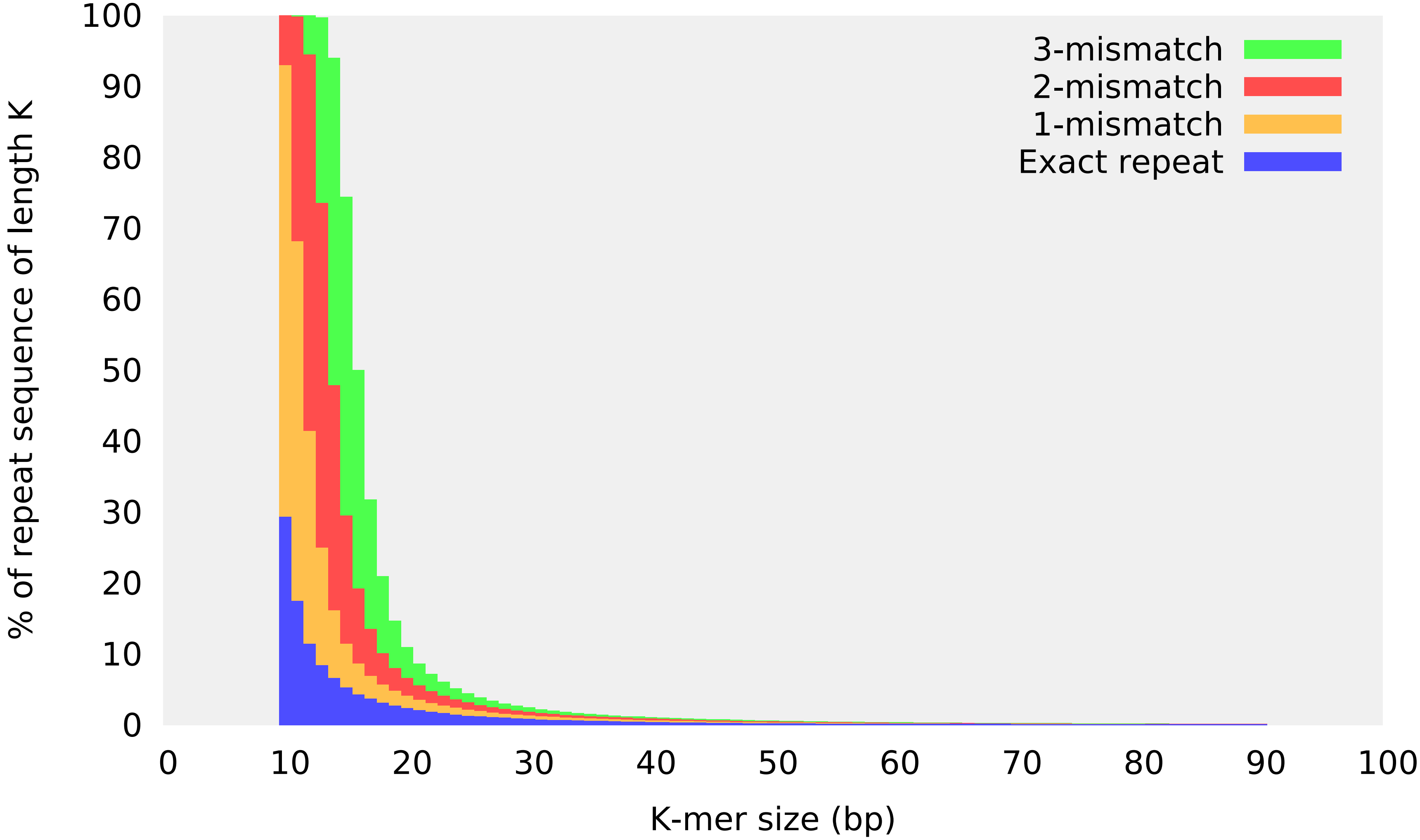
Find end-to-end haplotype paths spanning the region



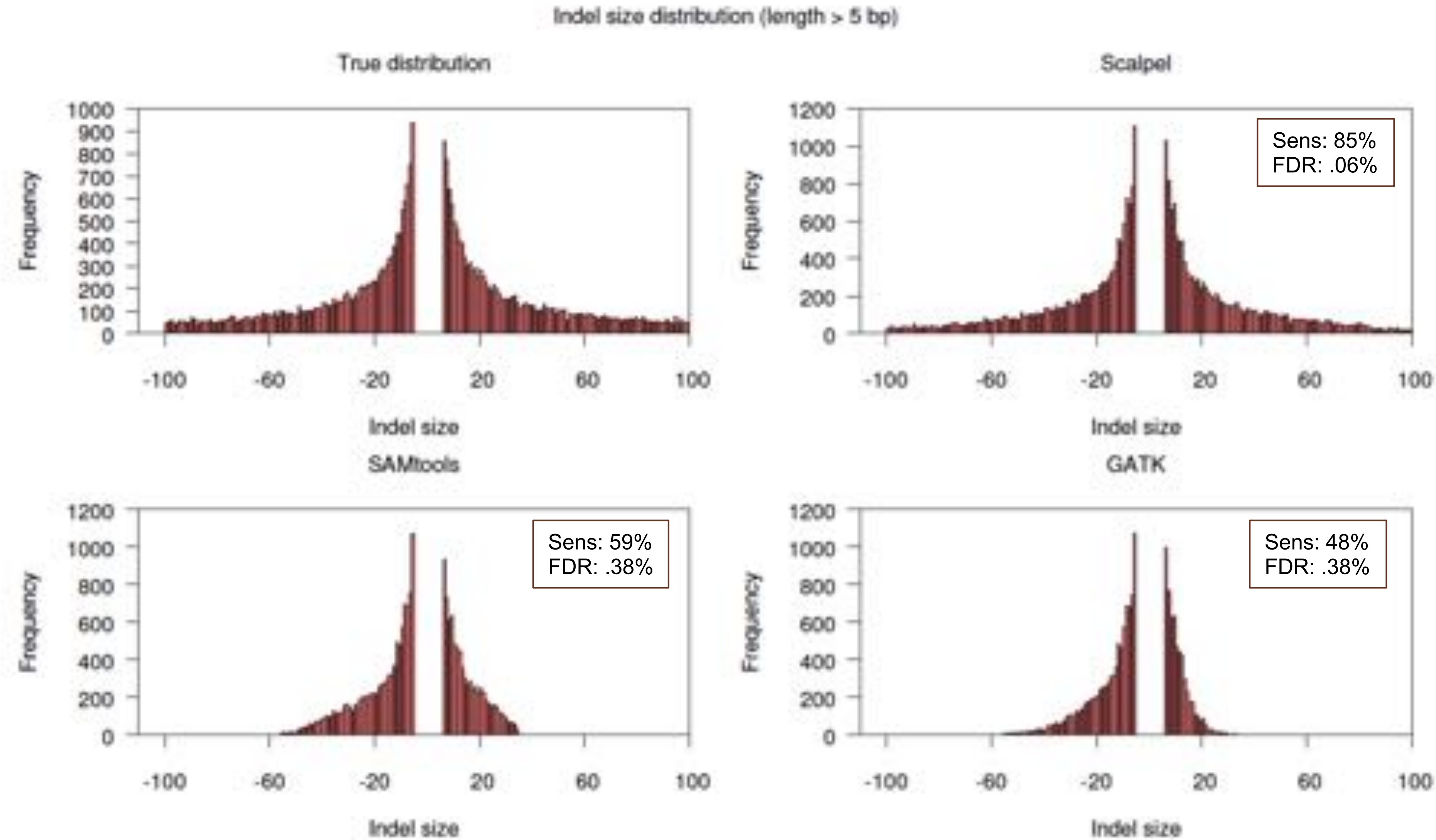
Align assembled sequences to reference to detect mutations

Repeats in the Exome

Specificity Challenge: 30% of exons have a perfect 10bp or larger repeat
Compute an on-the-fly analysis of repeat composition



Simulation Analysis



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

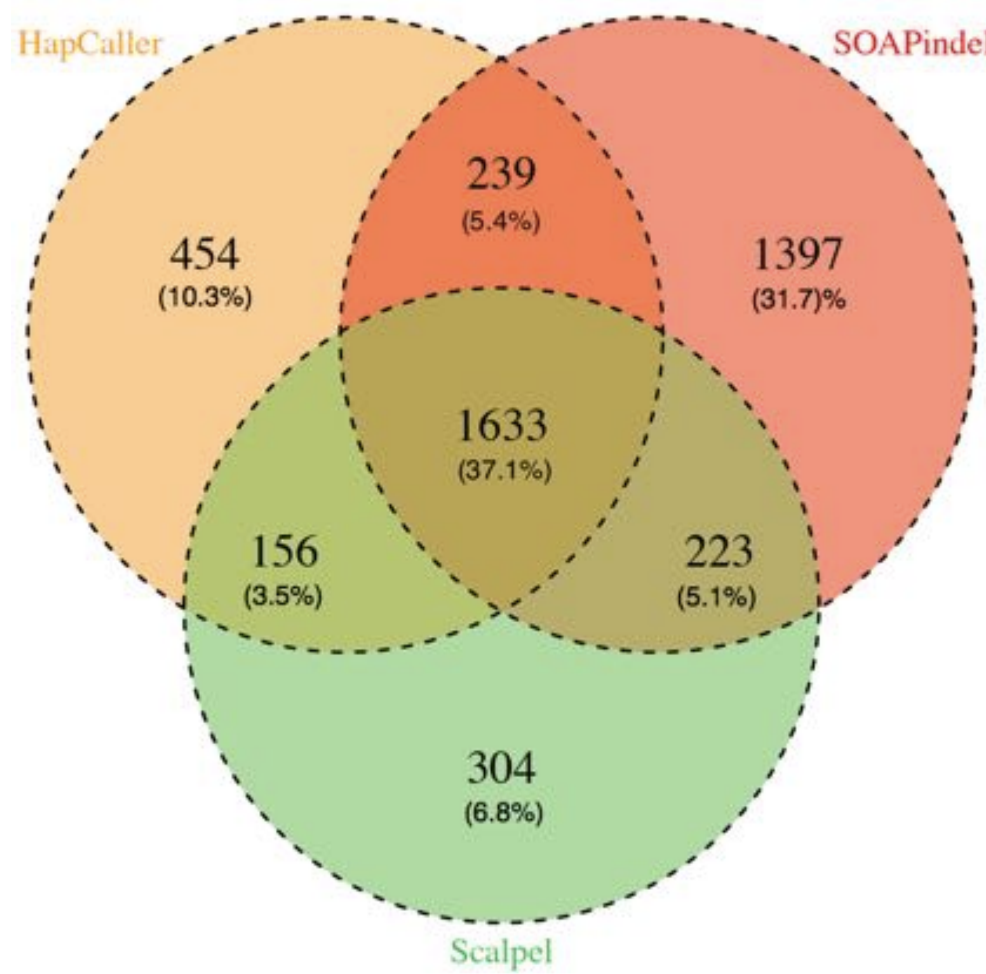
Experimental Analysis & Validation

Selected one deep coverage exome for deep analysis

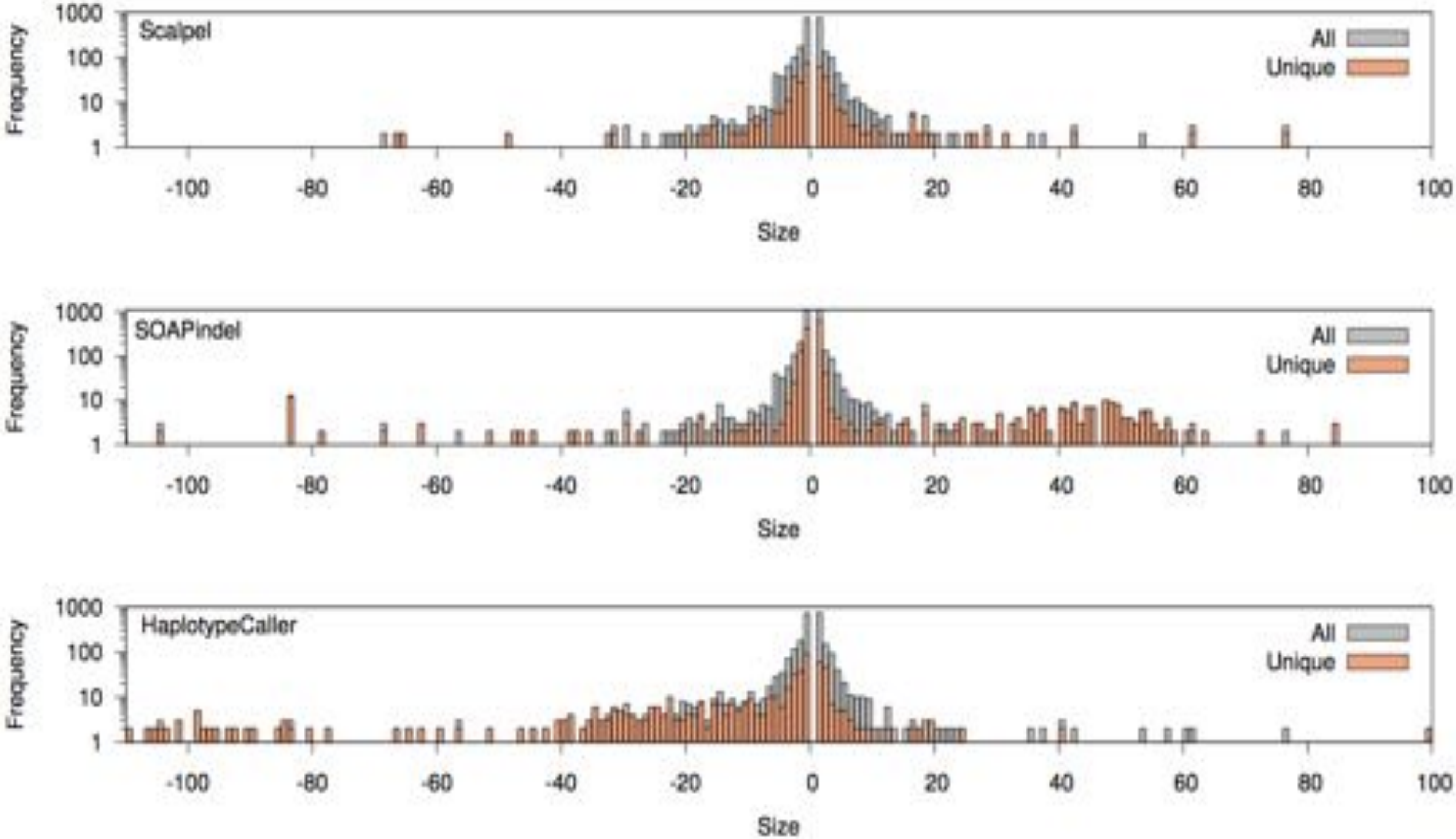
- Individual was diagnosed with ADHD (See Gholson for details)
- 80% of the target at >20x coverage
- Evaluated with Scalpel, SOAPindel, and GATK Haplotype Caller

1000 indels selected for validation

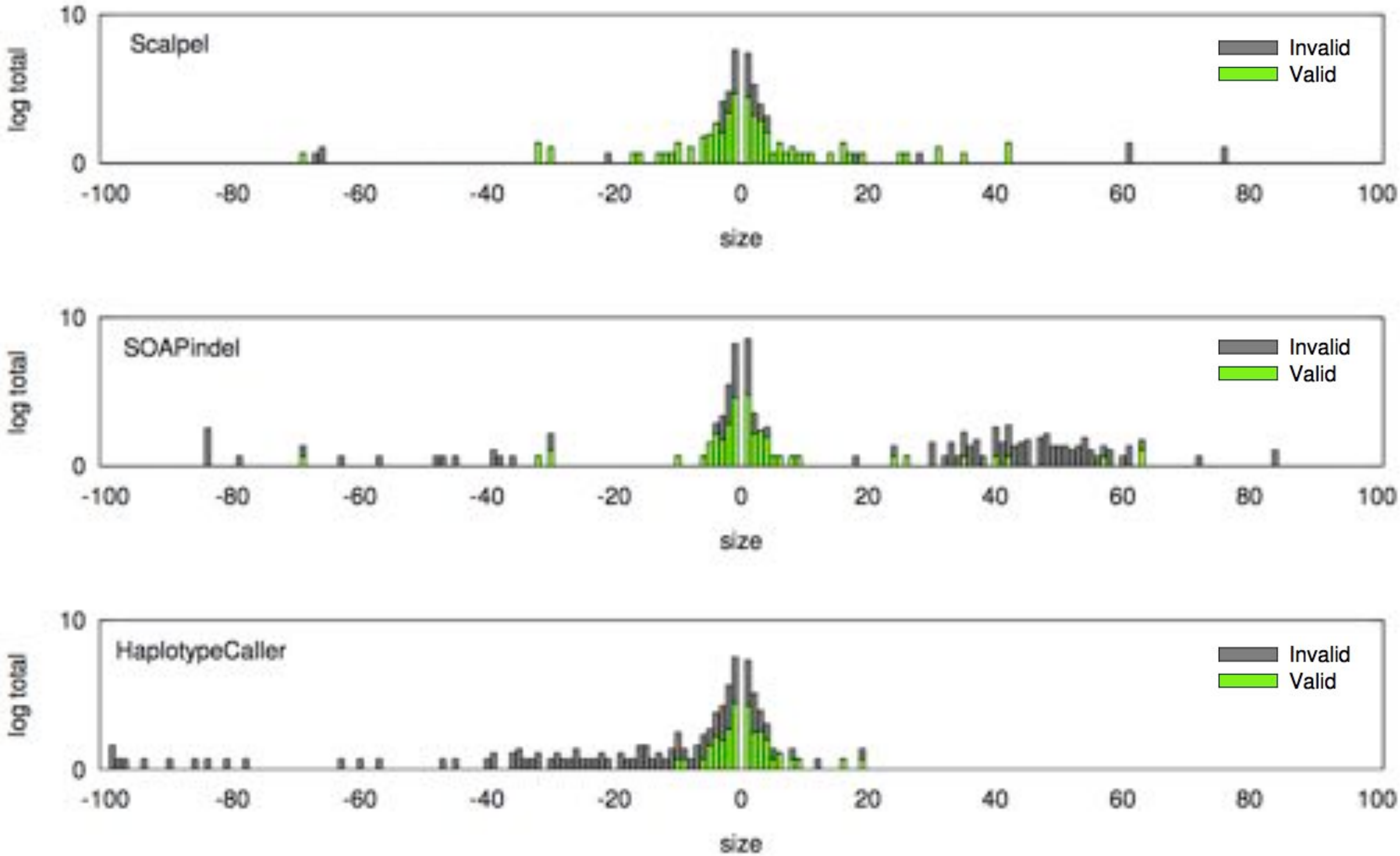
- 200 Scalpel
- 200 GATK Haplotype Caller
- 200 SOAPindel
- 200 within the intersection
- 200 long indels (>30bp)



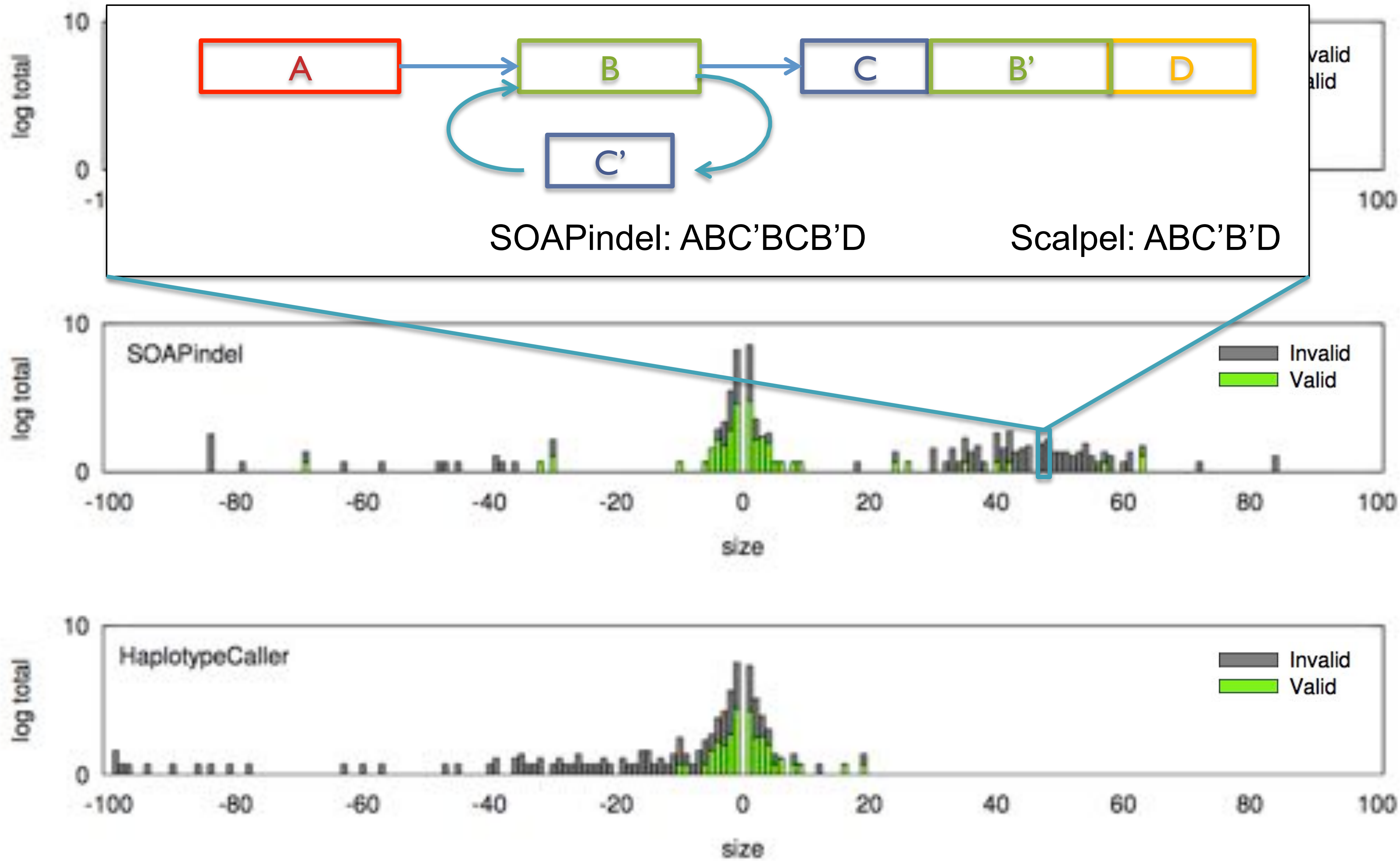
Scalpel Indel Discovery



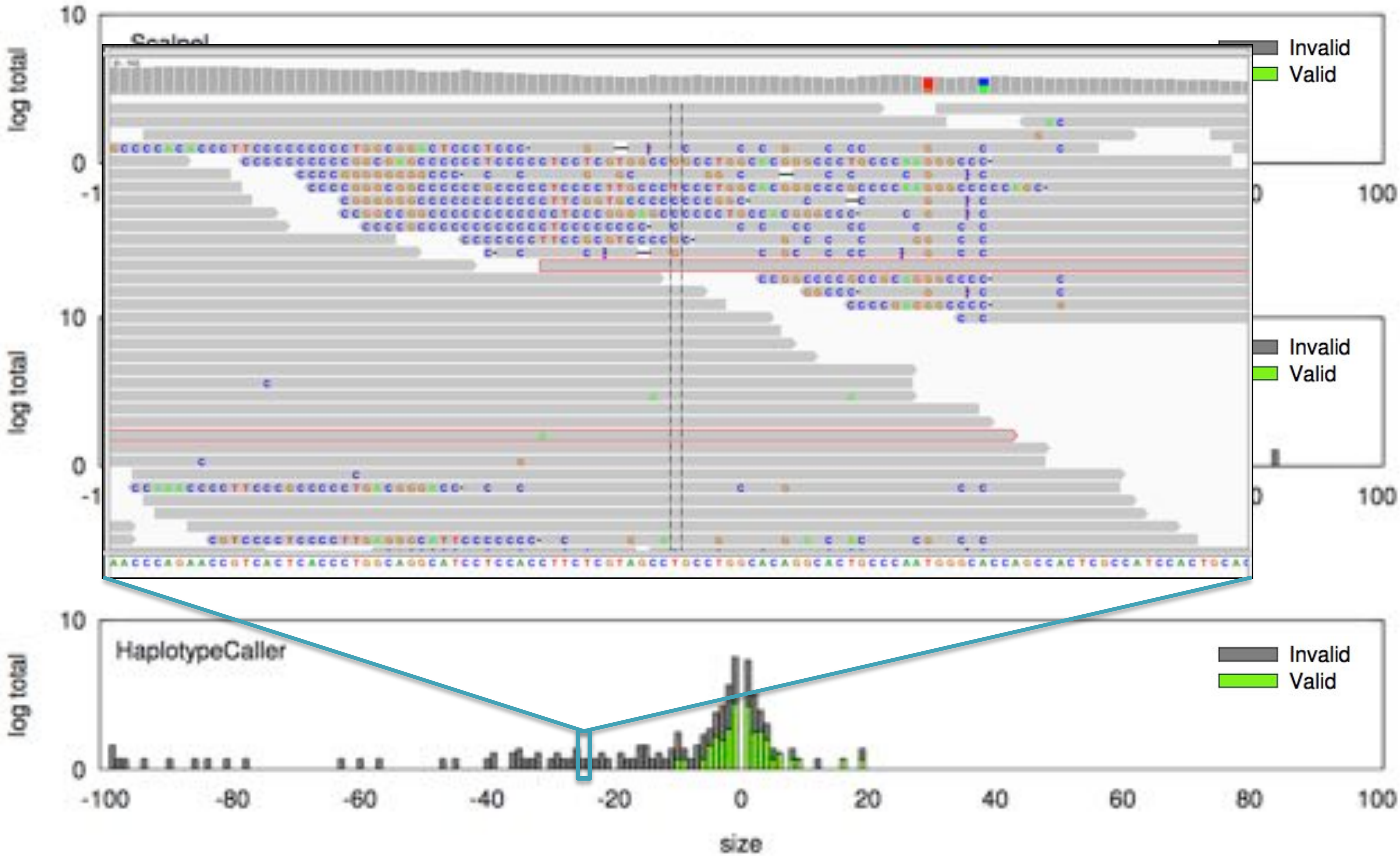
Scalpel Indel Discovery



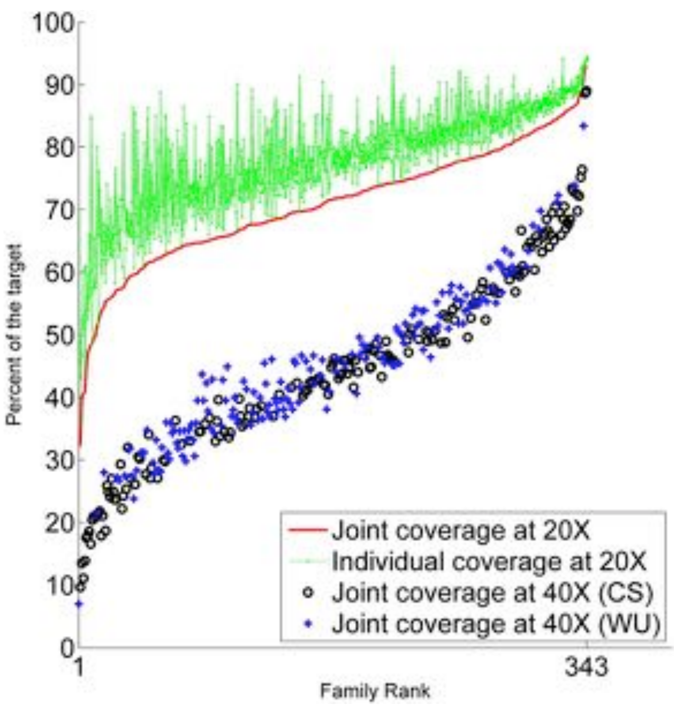
Scalpel Indel Discovery



Scalpel Indel Discovery



Exome sequencing of the SSC



Last year saw 3 reports of >593 families from the Simons Simplex Collection

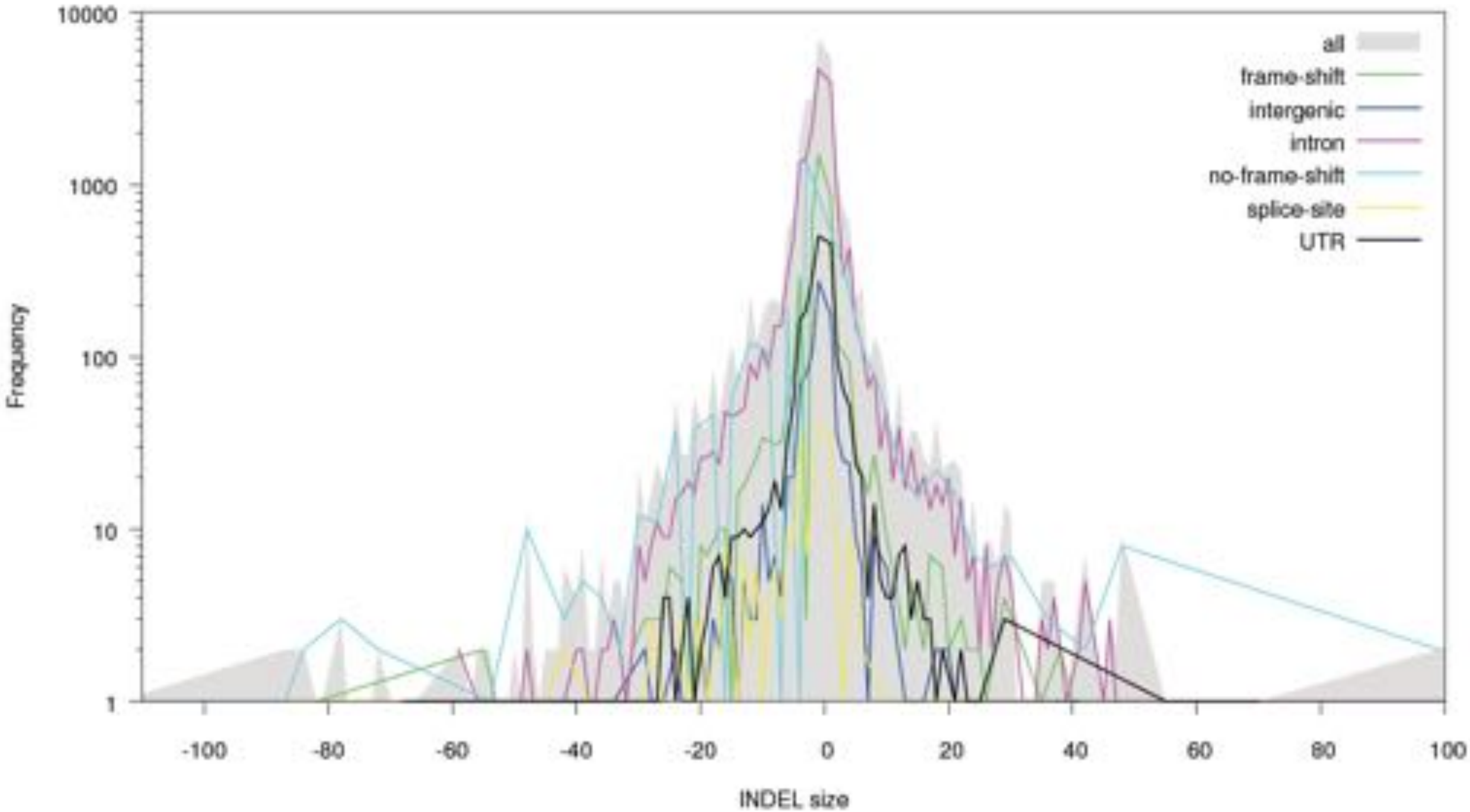
- Parents plus one child with autism and one non-autistic sibling
- All reported strong enrichment for de novo gene killing mutations (nonsense, frameshift, splice site mutations)
- Iossifov (343) and O'Roak (50) used GATK, Sanders (200) didn't attempt to identify indels

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.

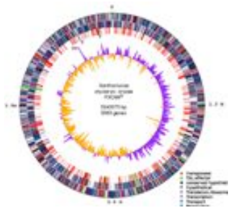
Revised Analysis of the SSC



Constructed database of >1M transmitted and de novo indels

Strengthened enrichment for de novo frameshift mutations (35:16)

Many new gene candidates identified, population analysis underway



Summary

- Hybrid assembly let us combine the best characteristics of 2^{nd} and 3^{rd} gen sequencing
  - Long reads and good coverage are the keys to a good de novo assembly
  - Single contig de novo assemblies of entire microbial chromosomes are now routine; Single contig de novo assemblies of entire plant and animal chromosomes on the horizon

- Assembly is the missing link towards high accuracy indel mutation discovery
  - Allows the algorithm to break free from the expectations of the reference
  - Pinpointing de novo mutations require both high sensitivity and specificity

- We are starting to apply these technologies to discover significant biology that is otherwise impossible to measure

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
Ware Lab
Wigler Lab

IT Department

SFARI
SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

National Human Genome Research Institute

U.S. DEPARTMENT OF ENERGY

NSF

See you at

Genome Informatics

Oct 30 – Nov 2

http://schatzlab.cshl.edu
@mike_schatz