Heterozygosity, Phased Genomes, and Personalized-omics

Michael Schatz

January 12, 2017
UMD CBCB

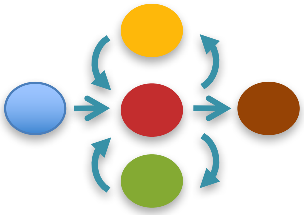
Outline

1. **Phased Genome Assembly**

   GenomeScope & FALCON-Unzip

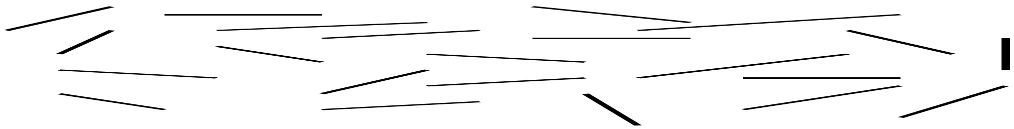
2. **Personalized-Omics**

   Complex SVs and oncogene amplifications in breast cancer



Sequence Assembly Problem

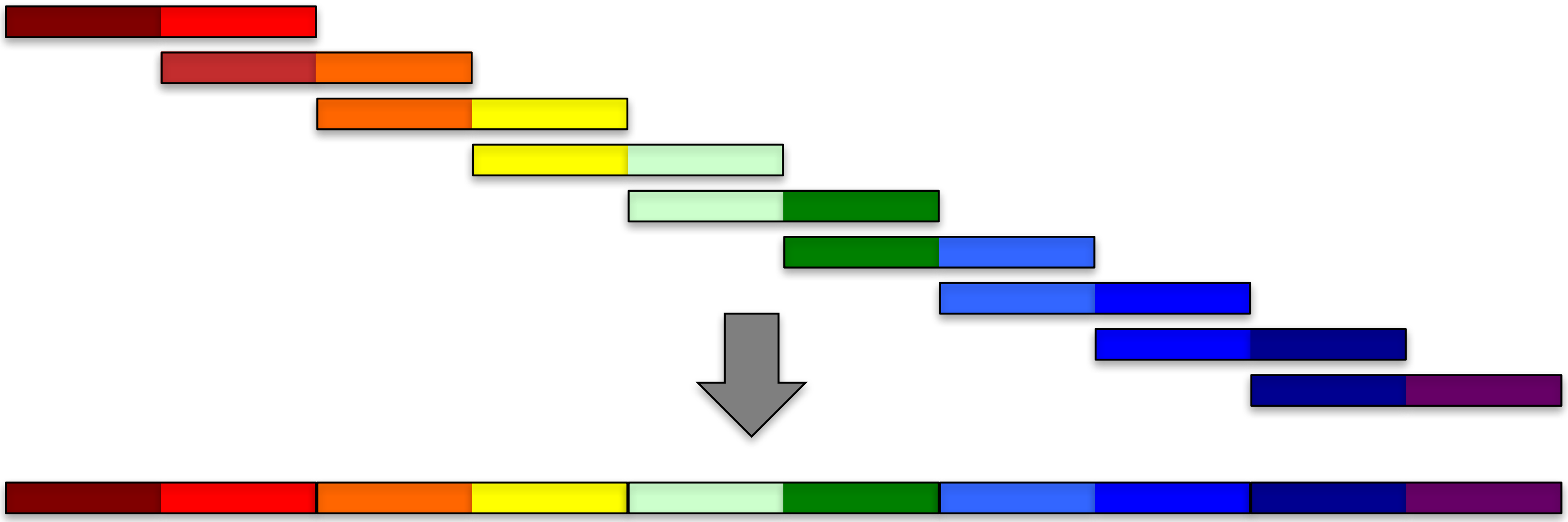
1. Shear & Sequence DNA



2. Construct assembly graph from overlapping reads

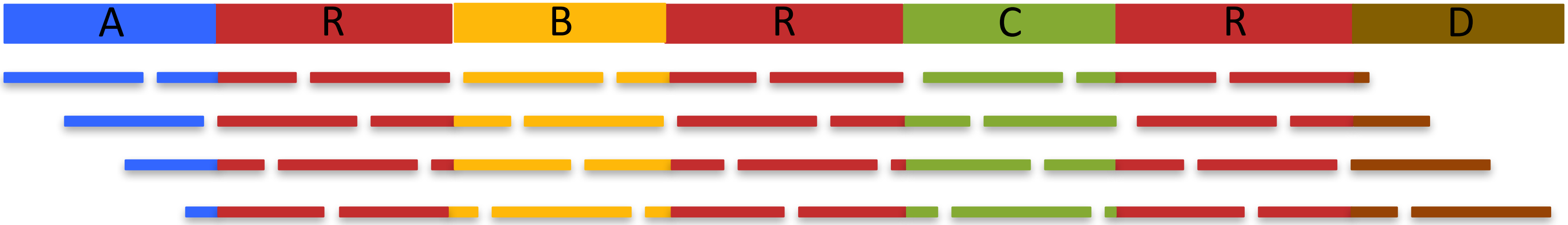
```
...AGCCTAGGGATGCGCGACACGT
          GGATGCGCGACACGTCGCATATCCGGTTTGGTCAACCTCGGACGGAC
                                          CAACCTCGGACGGACCTCAGCGAA...
```

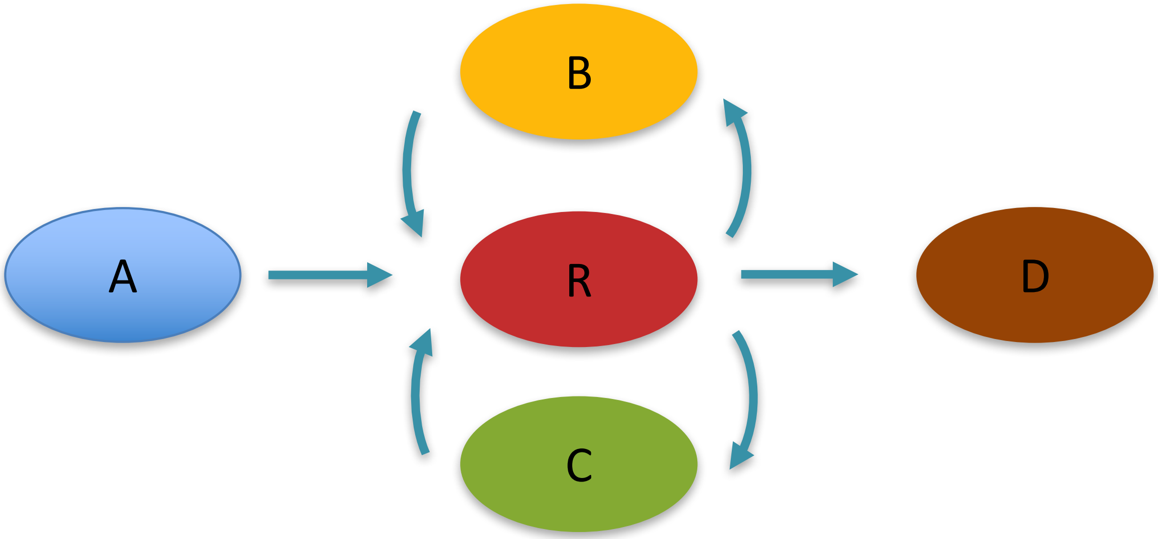
3. Simplify assembly graph



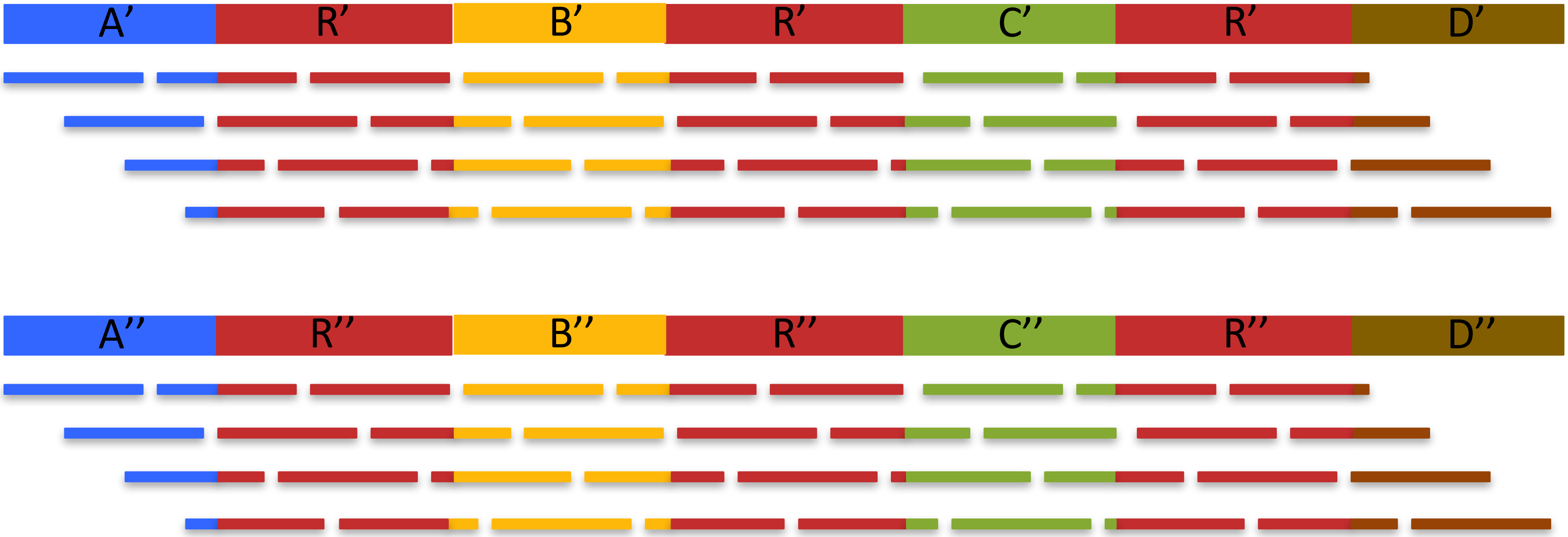
On Algorithmic Complexity of Biomolecular Sequence Assembly Problem
Narzisi, G, Mishra, B, Schatz, MC (2014) *Algorithms for Computational Biology.* Lecture Notes in Computer Science. *Vol. 8542*

Assembly Complexity

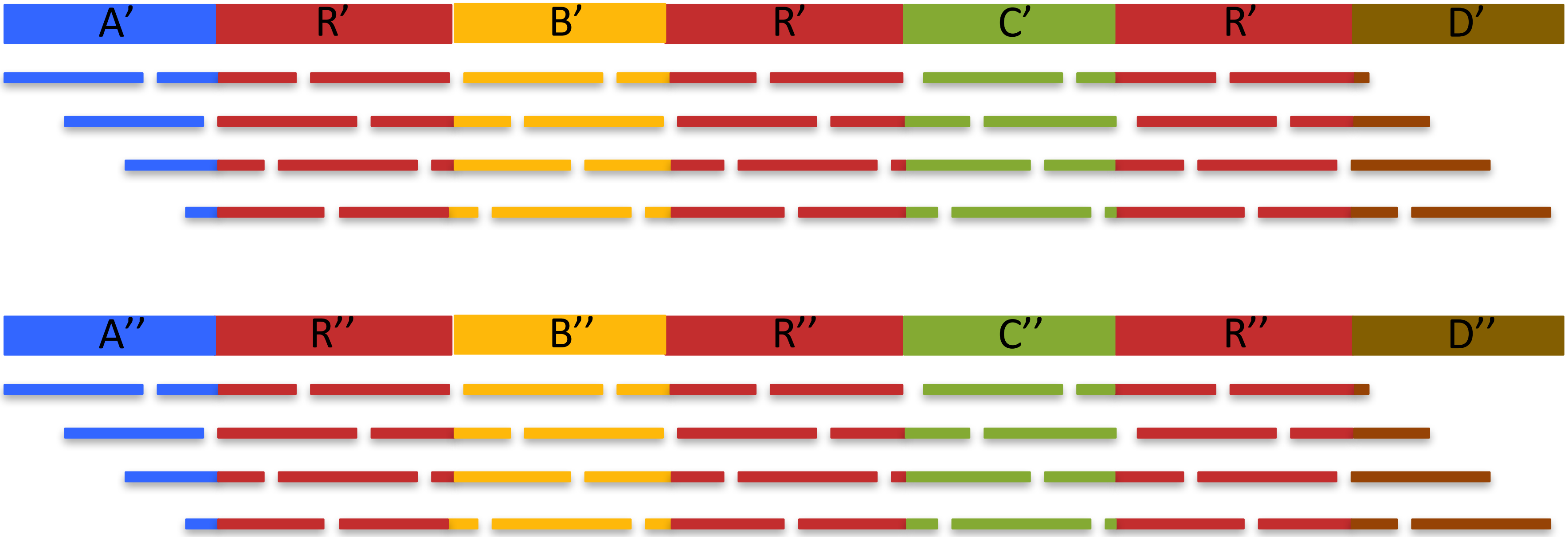


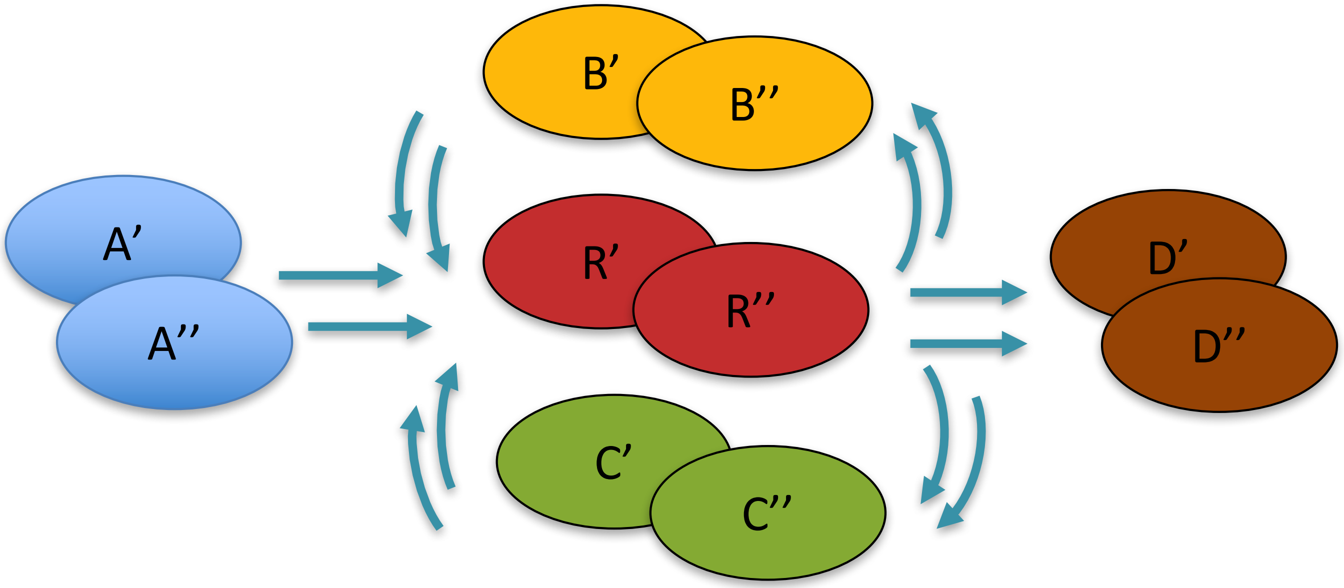


Diploid Assembly Complexity

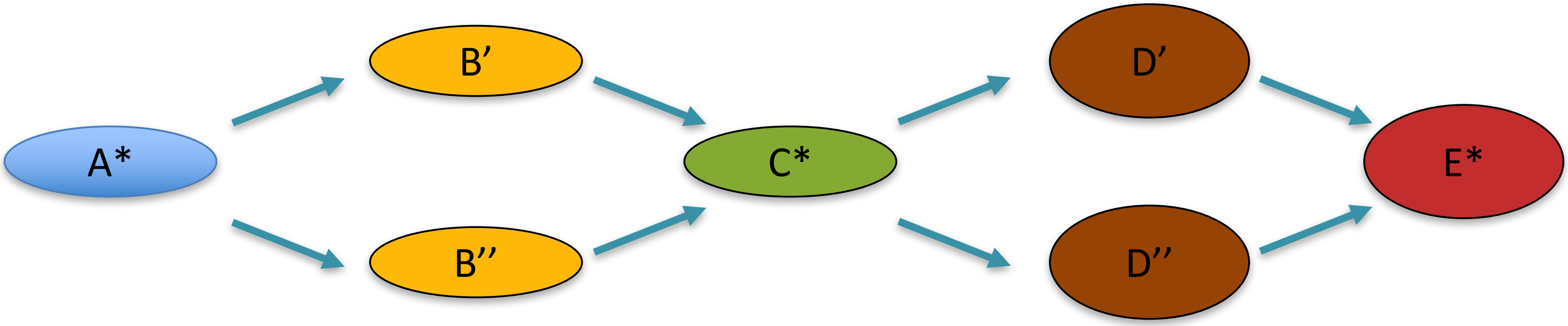


Diploid Assembly Complexity





Diploid Assembly Problems



Assembly becomes more fragmented

- *A. thaliana* inbred with short reads: ~100kbp contig N50
- *A. thaliana* outbred with short reads: ~10kbp contig N50

Assembly sequence & size will be distorted

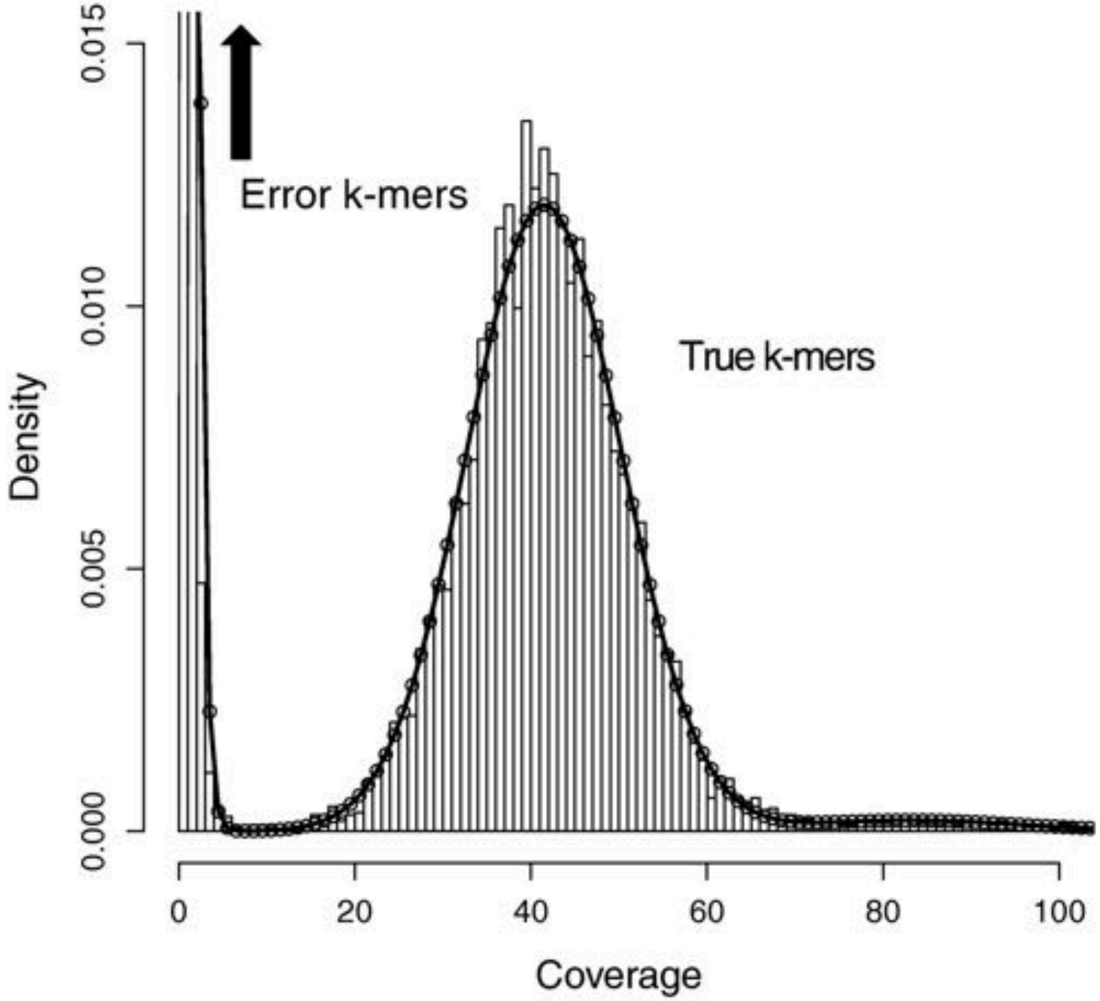
- Regions of low heterozygosity will be assembled together
  -> reduces assembly from true diploid size
- Regions of high heterozygosity will be split apart
  -> haplotypes may be next to each other in scaffolds or left out

"Mosaic" consensus sequences*

- Sequence will arbitrarily switch from maternal to paternal alleles
- May be "read incoherent" and not supported by any sequencing reads

Critical genes may be assembled into 0, 1, or 2 copies (or more)!

Quake: Quality-aware detection and correction of sequencing errors



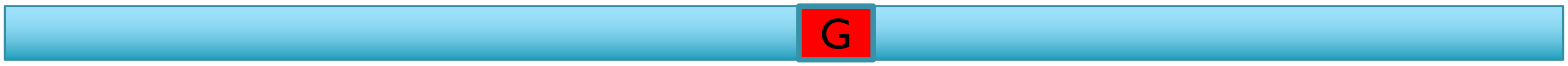
Reference-free approach for correcting sequencing errors

1. Scan reads, count #occurrences of all k-mers using Jellyfish
2. Analyze k-mer profile to find local minimum between error k-mers (occur < ~5 times) and trusted k-mers (occur > 5 times)
3. For each untrusted k-mer in a read, search for minimum # of substitutions to become trusted

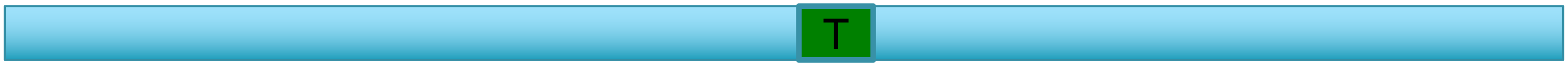
Quake: quality-aware detection and correction of sequencing errors
Kelley, DR, Schatz, MC, Salzberg, SL (2010) Genome Biology 11:R116

Heterozygous Kmer counting

Sequencing read from homologous chromosome 1A

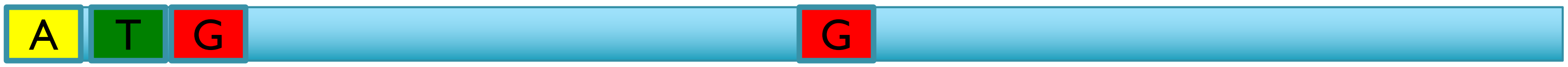


Sequencing read from homologous chromosome 1B

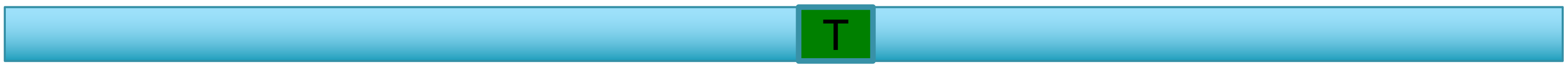


Heterozygous Kmer counting

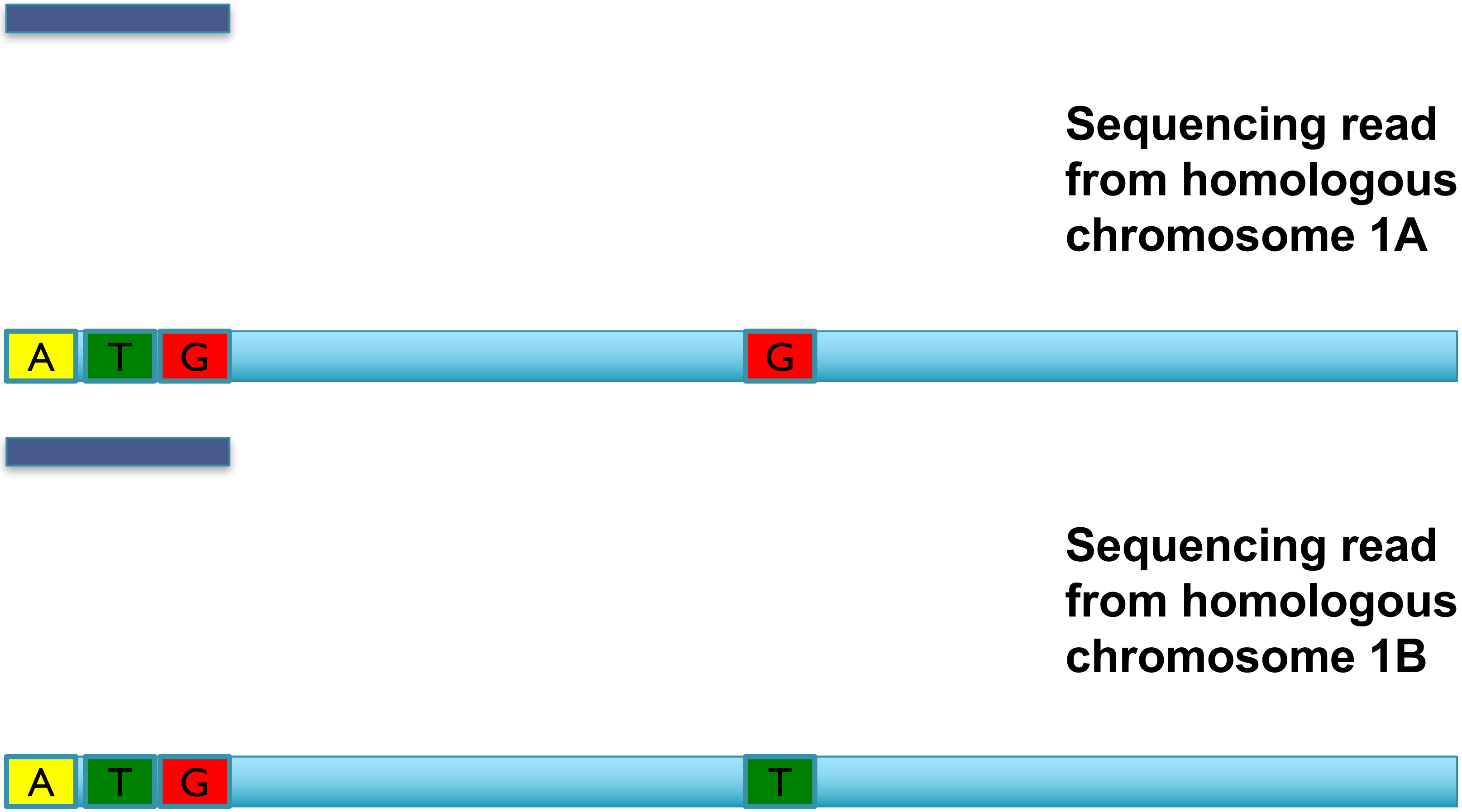
Sequencing read from homologous chromosome 1A



Sequencing read from homologous chromosome 1B



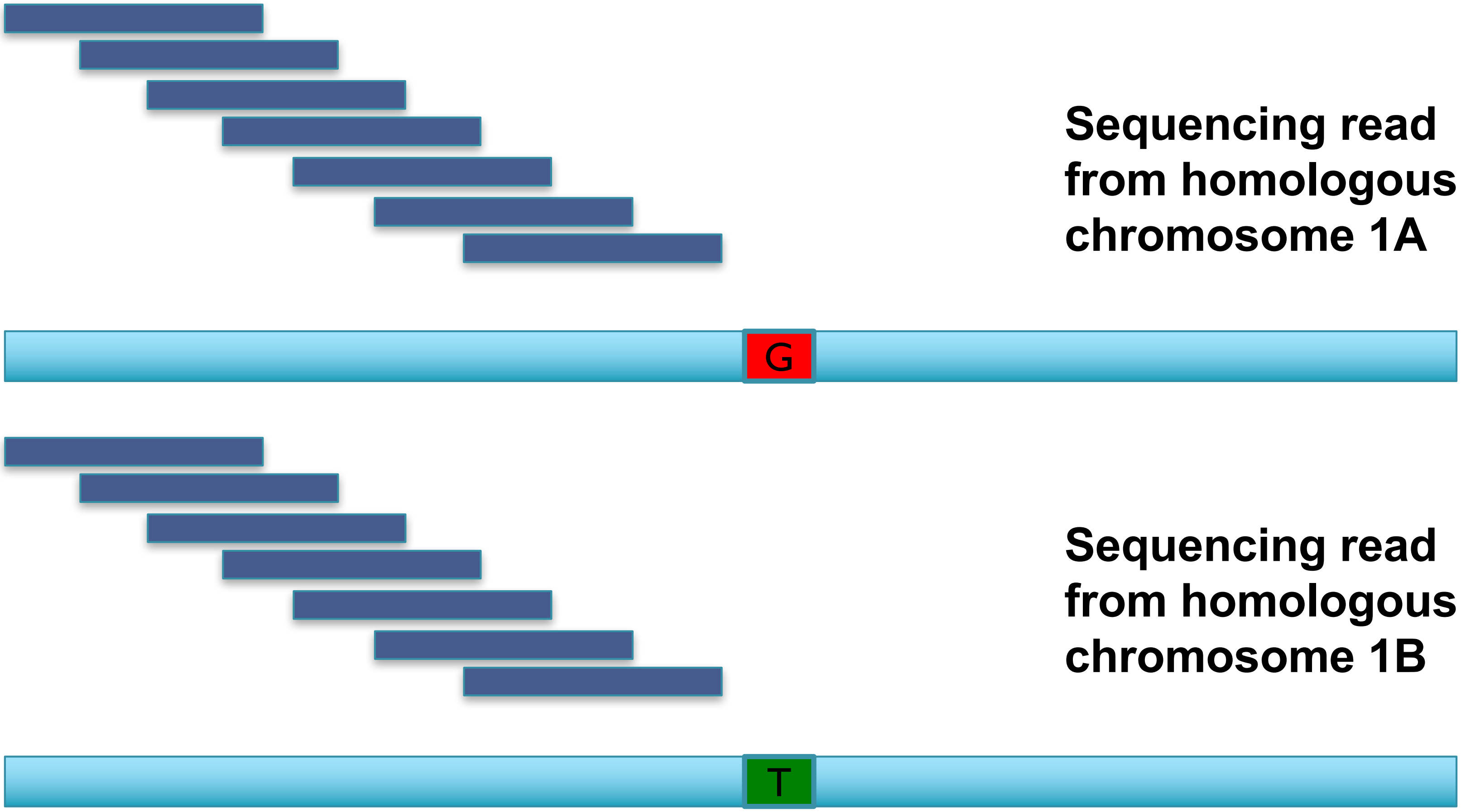
Heterozygous Kmer counting



Sequencing read
from homologous
chromosome 1A

Sequencing read
from homologous
chromosome 1B

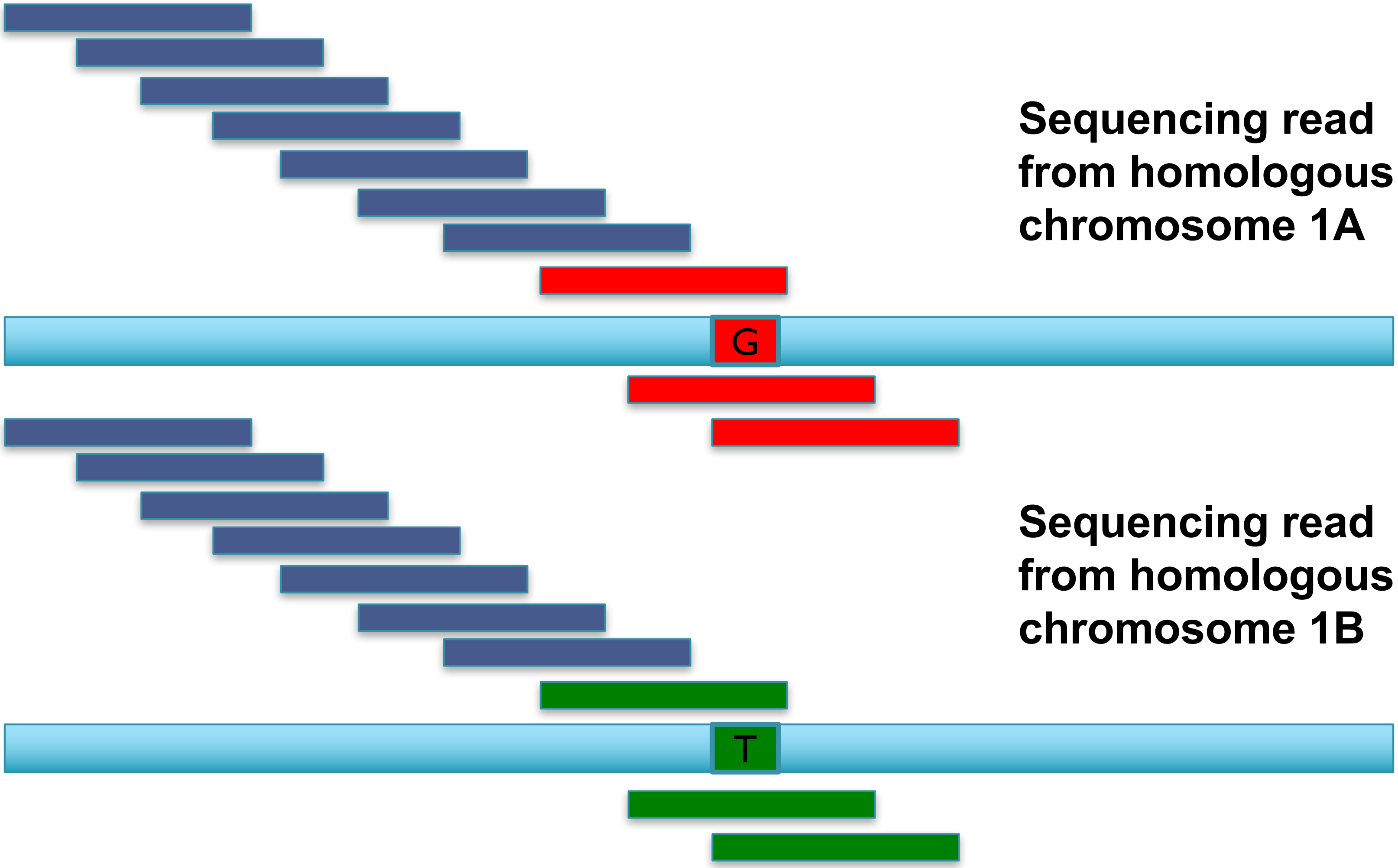
Heterozygous Kmer counting



Sequencing read
from homologous
chromosome 1A

Sequencing read
from homologous
chromosome 1B

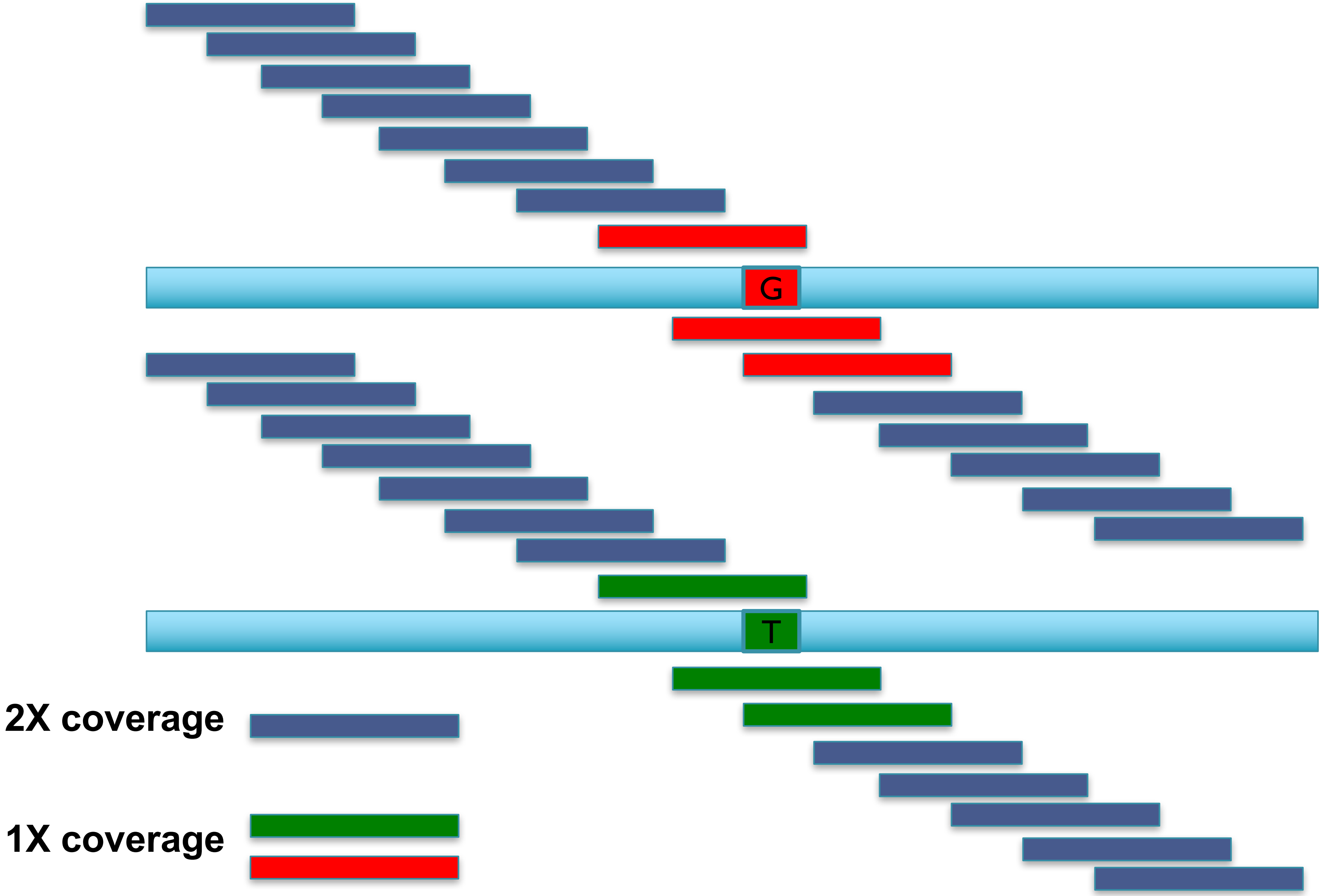
Heterozygous Kmer counting



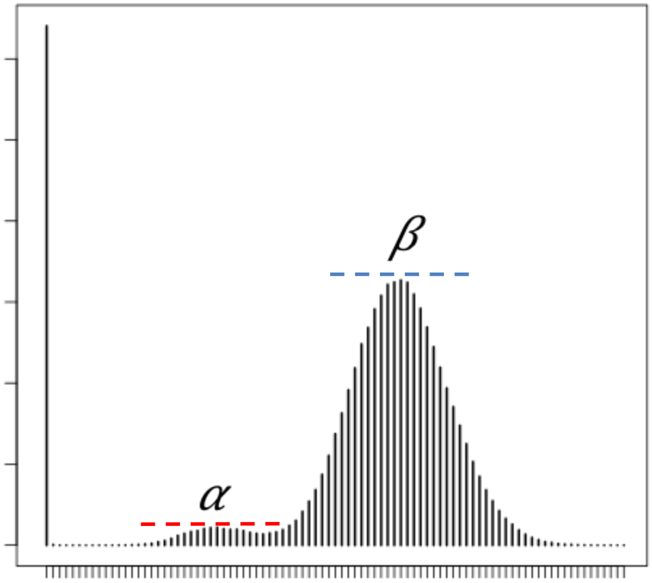
Sequencing read from homologous chromosome 1A

Sequencing read from homologous chromosome 1B

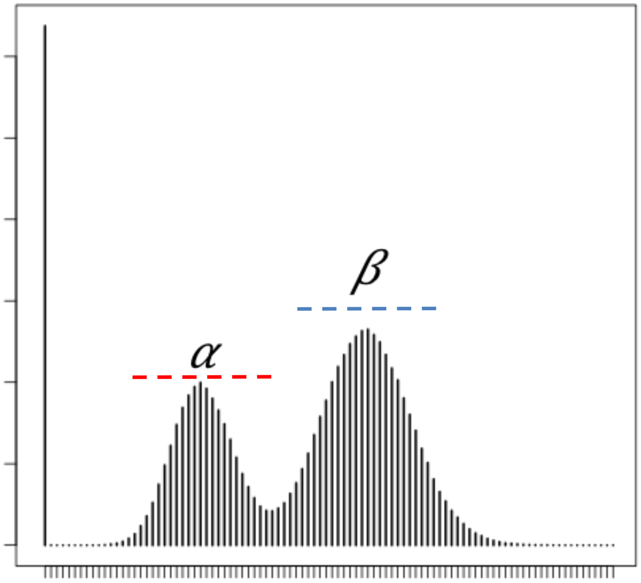
Heterozygous Kmer counting



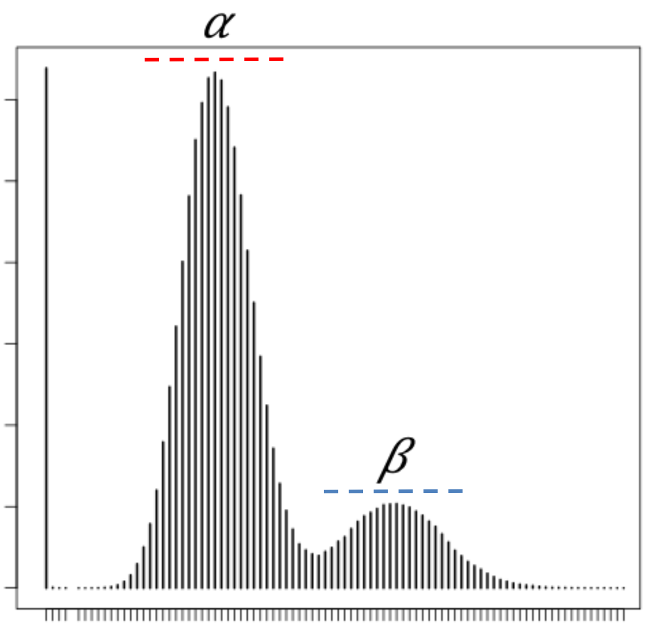
Heterozygous Kmer Profiles



0.1% heterozygosity



1% heterozygosity



5% heterozygosity

- ***Heterozygosity creates a characteristic "double-peak" in the Kmer profile***
  - Second peak at twice k-mer coverage as the first: heterozygous kmers average 50x coverage, homozygous kmers average 100x coverage

- ***Relative heights of the peaks is directly proportional to the heterozygosity rate***
  - The peaks are balanced at around 1.25% because each heterozygous SNP creates 2*k heterozygous kmers (typically k = 21)

GenomeScope Model

$$f(x) = G\Big\{\alpha NB(x,\lambda,\lambda/\rho) + \beta NB(x,2\lambda,2\lambda/\rho) + \gamma NB(x,3\lambda,3\lambda/\rho) + \delta NB(x,4\lambda,4\lambda/\rho)\Big\}$$

Analyze k-mer profiles using a mixture model of 4 negative binominal components

- Components centered at 1,2,3,4 * λ

- Four components capture heterozygous and homozygous unique (α,β) and 2 copy repeats (γ,δ). Higher order repeats do not contribute a significant number of kmers

- Negative binomial instead of Poisson to account for over dispersion observed in real data (especially PCR duplicates); variance modeled by ρ

$$\begin{aligned}
\alpha &= 2(1-d)(1-(1-r)^k) \\
\beta &= (1-2d)(1-r)^k + d(1-(1-r)^k)^2 \\
\gamma &= 2d(1-r)^k(1-(1-r)^k) \\
\delta &= d(1-r)^{2k}
\end{aligned}$$

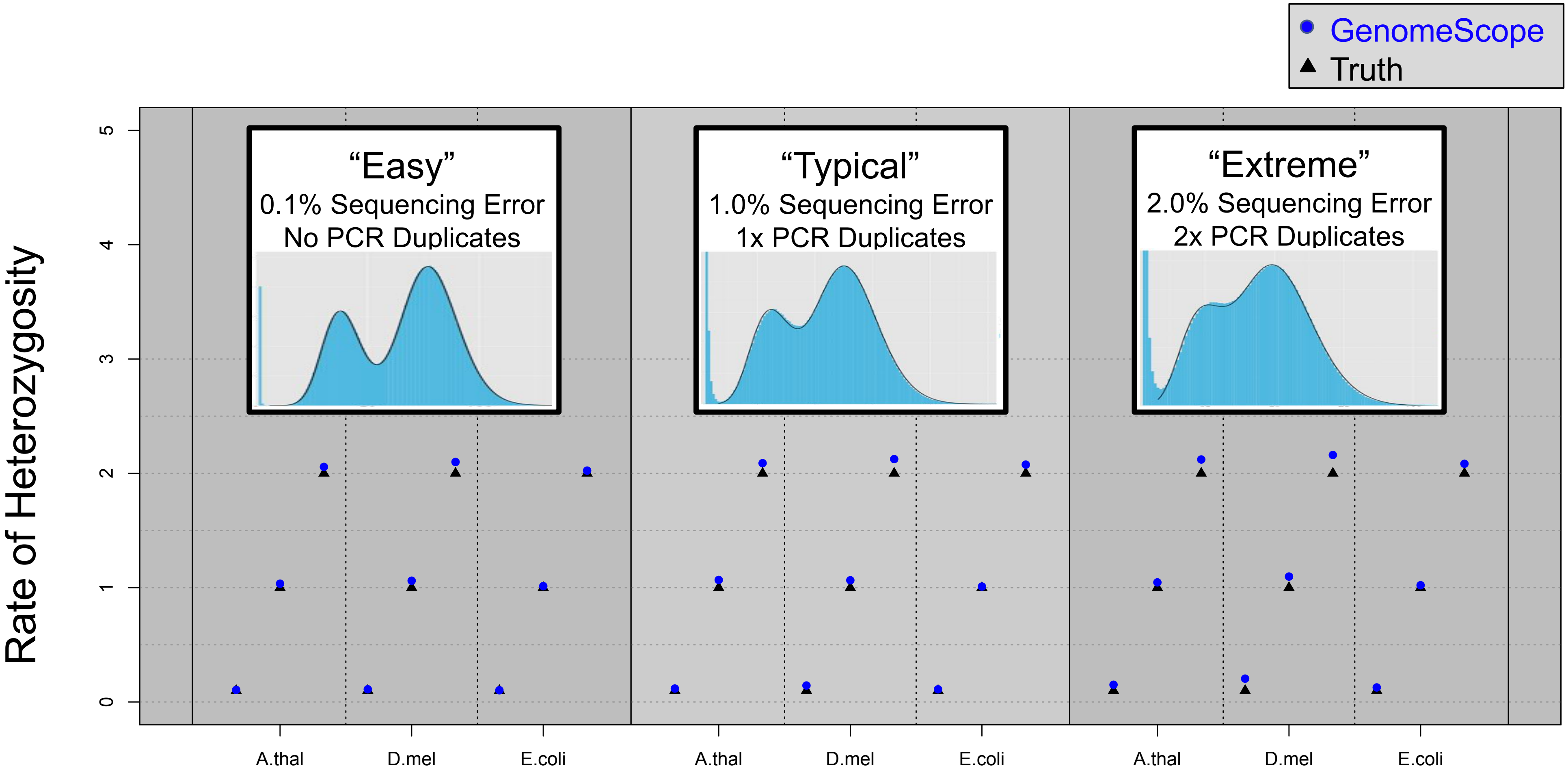
k is the *k-mer* length used when constructing the k-mer profile.

r is the rate of heterozygosity between sets of chromosomes

d represents the percentage of the genome that is two-copy repeat

Fit model with nls, infer rate of heterozygosity, genome size, unique/repetitive content, sequencing error rate, rate of PCR duplicates

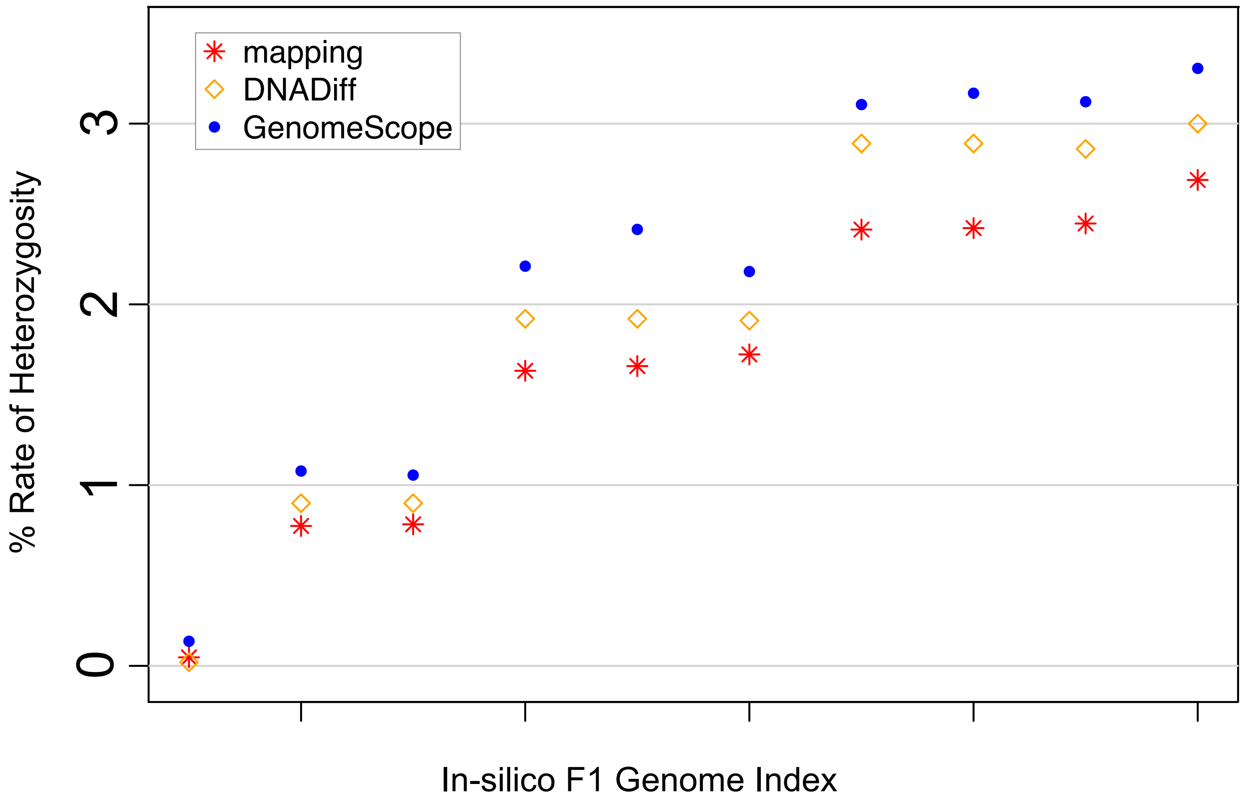
Simulated Results



Introduce SNPs into A. thaliana, D. melanogaster, or E. coli at known rates, simulate shotgun sequencing with specified rates of sequencing error and PCR duplications

In silico E. coli population sequencing

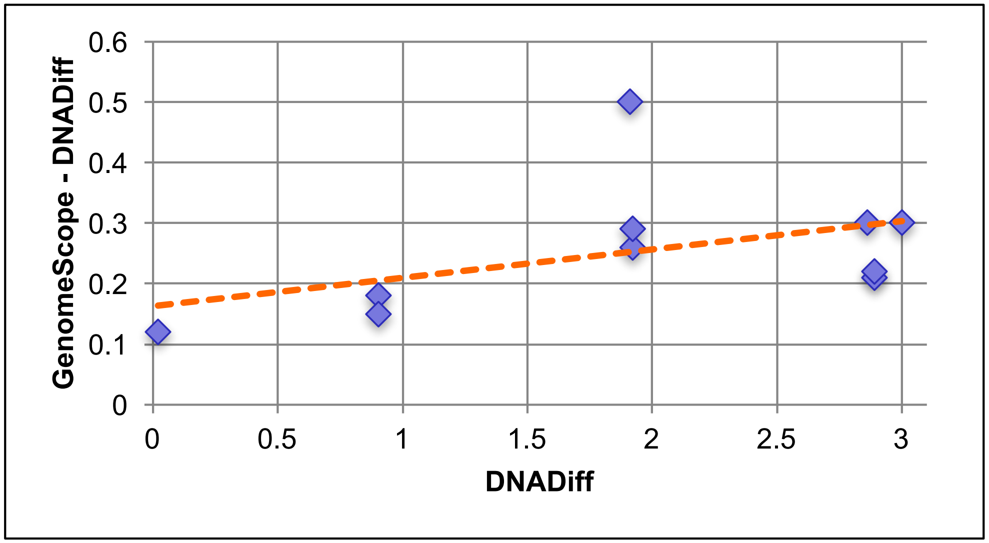
“Synthetic F1 Genome”

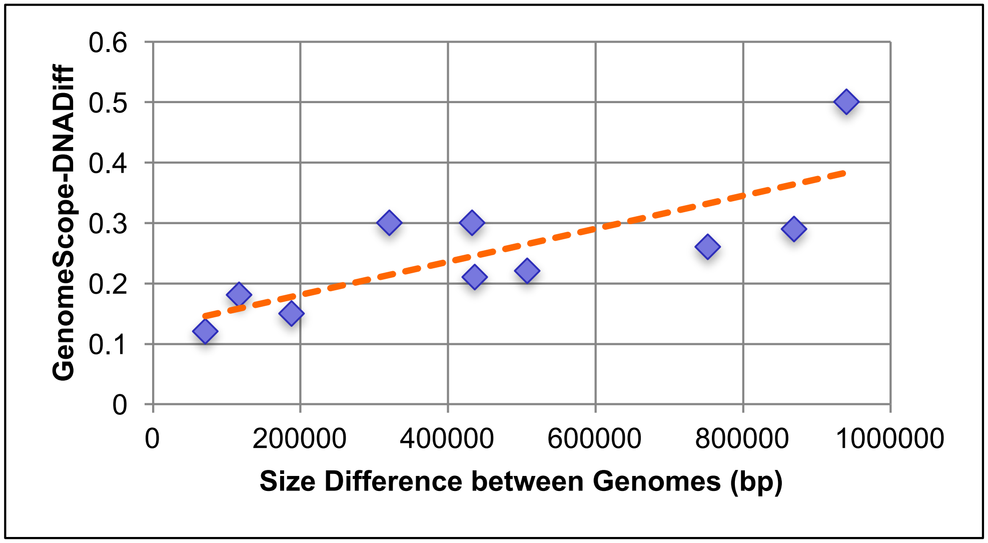


Mix equal numbers of real Illumina reads from pairs of 5 different E. coli isolates that have finished genomes with varying rates of similarity

Compare results to mapping pipeline (BWA+SAMTools) and MUMmer/DNADiff

Understanding DNADiff





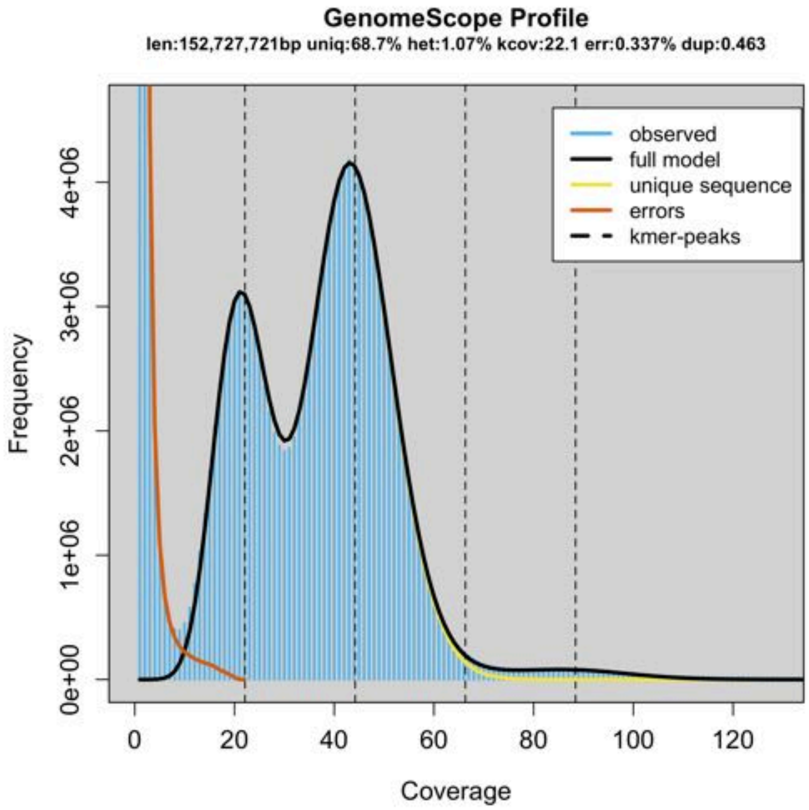
Observe that the difference between the rate of heterozygosity estimated by GenomeScope was generally higher than DNADiff, and that it was correlated with the rate of heterozygosity

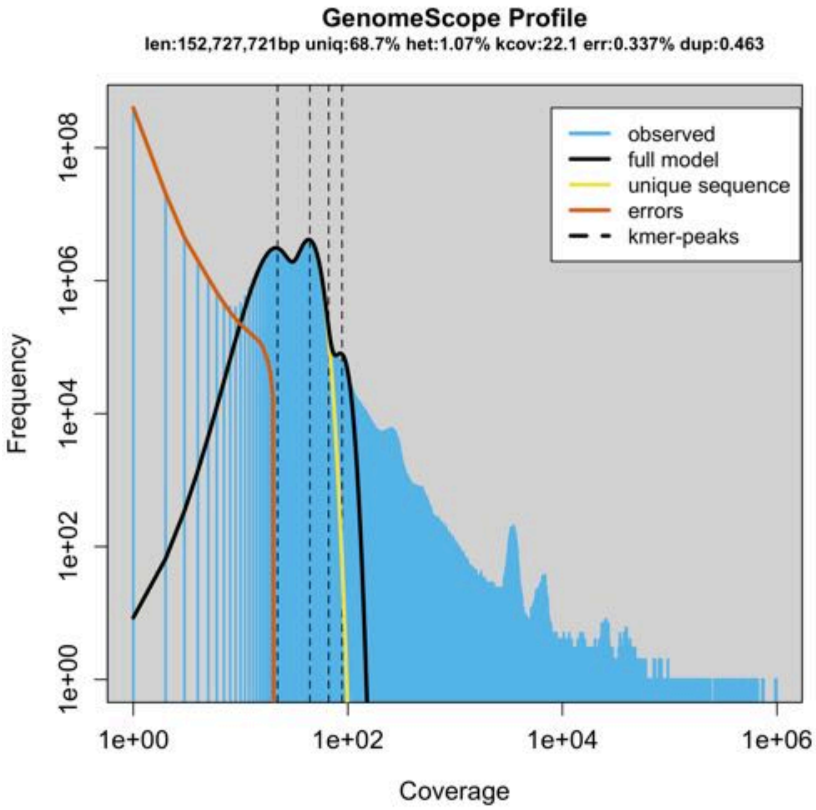
The difference was strongly correlated with the size difference between the genomes

Conclude that DNADiff is underestimating the true rate because it doesn't include bases in regions that don't align!

GenomeScope: Fast genome analysis from short reads

http://qb.cshl.edu/genomescope/



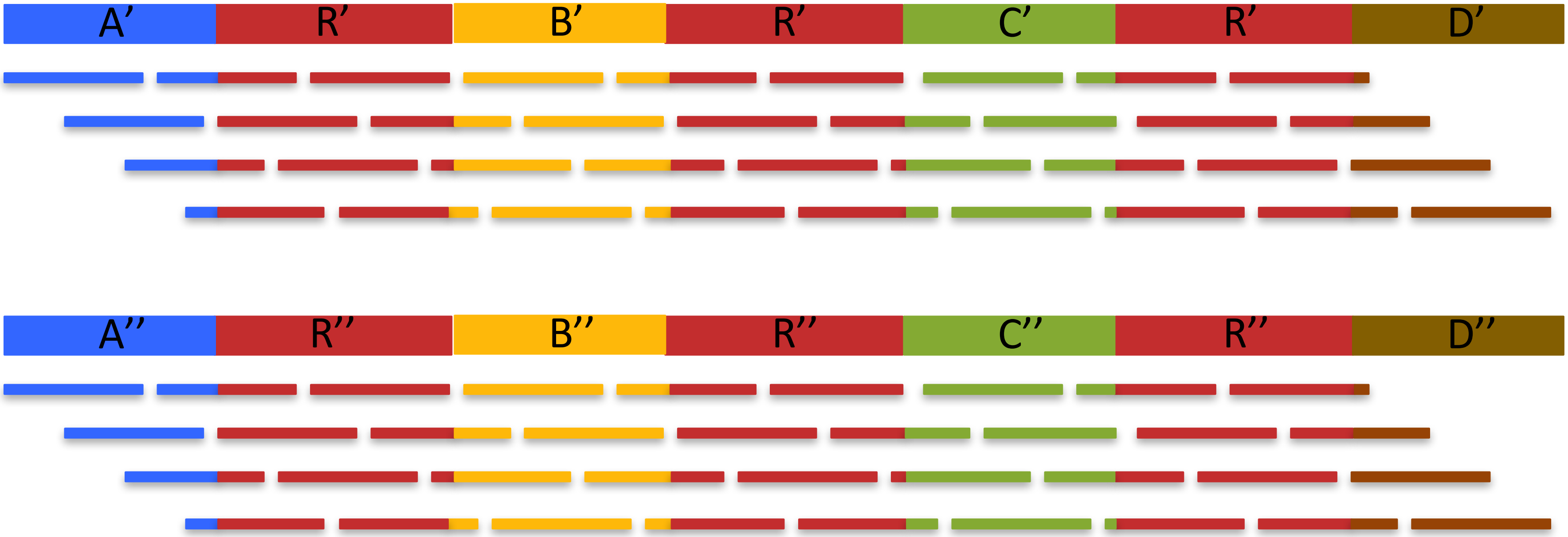


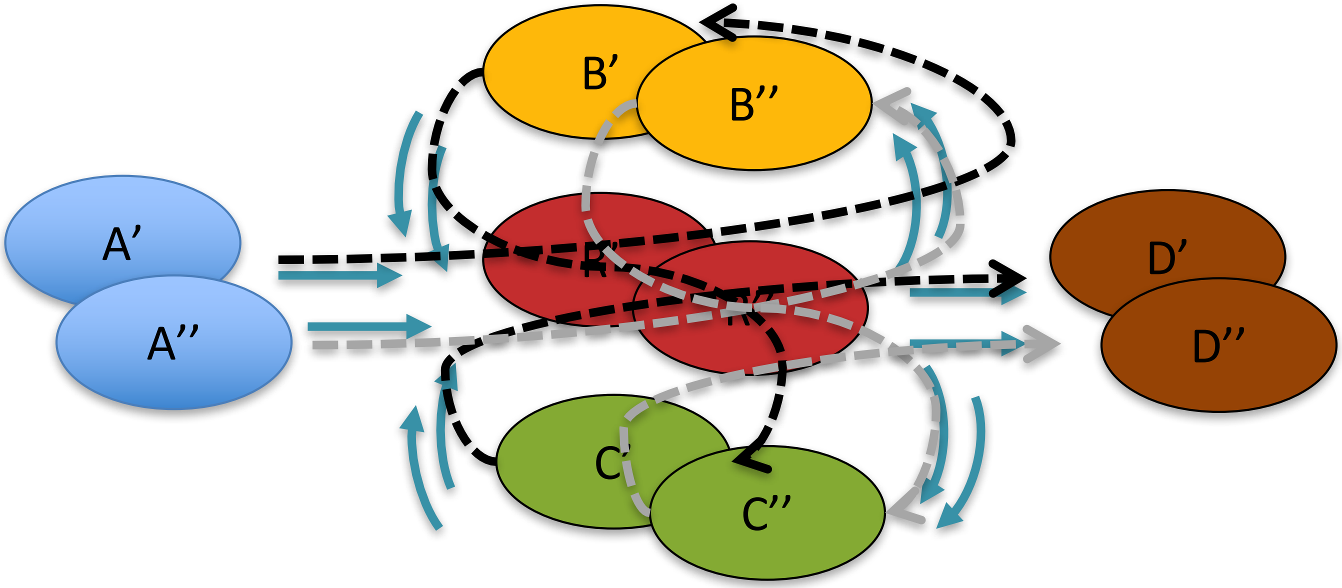
Evaluated on several genomes with published rates of heterozygosity:

- *L. calcarifer* (Asian seabass), *D. melanogaster* (fruit fly), *M. undulates* (budgerigar), *A. thaliana* Col-Cvi F1 (thale cress), *P. bretschneideri* (pear), C. gigas (Pacific oyster)
- Agrees well with published results:
  - Rate of heterozygosity is typically higher but likely correct.
  - Genome size of plants inflated by organelle sequences (exclude very high freq. kmers)

Vurture, GW*, Sedlazeck FJ*, Nattestad, M, Underwood, C, Fang, H, Gurtowski, J, Schatz, MC. *bioRxiv*

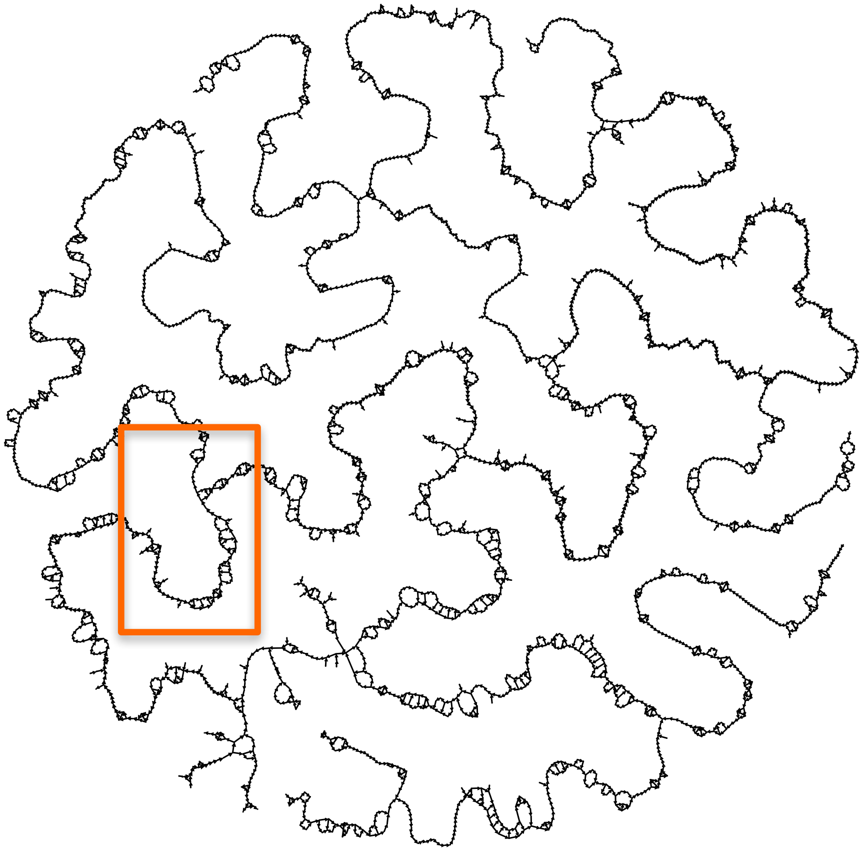
Assembly Complexity



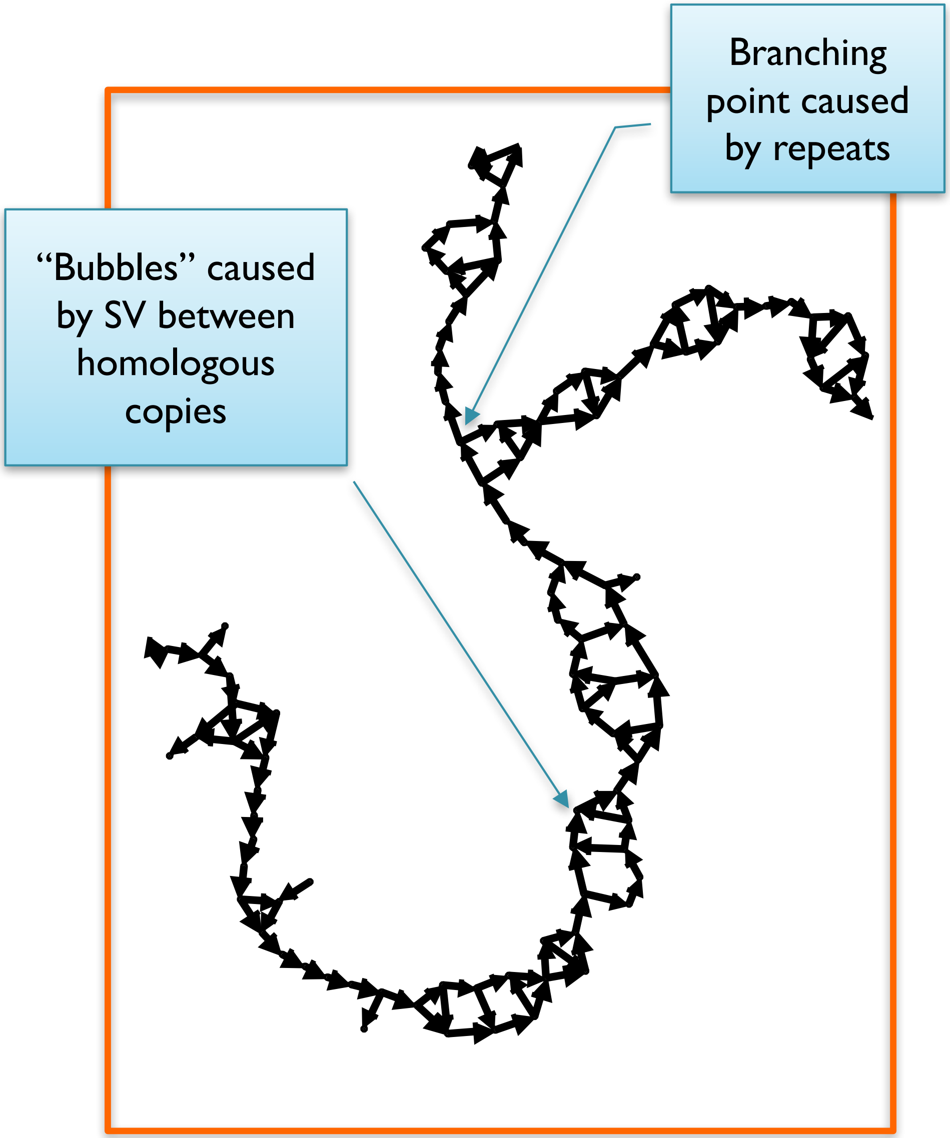


FALCON-unzip: Phased Diploid Genome Assembly

Assembly graph from A. thaliana Ler-0 + Col-0 data



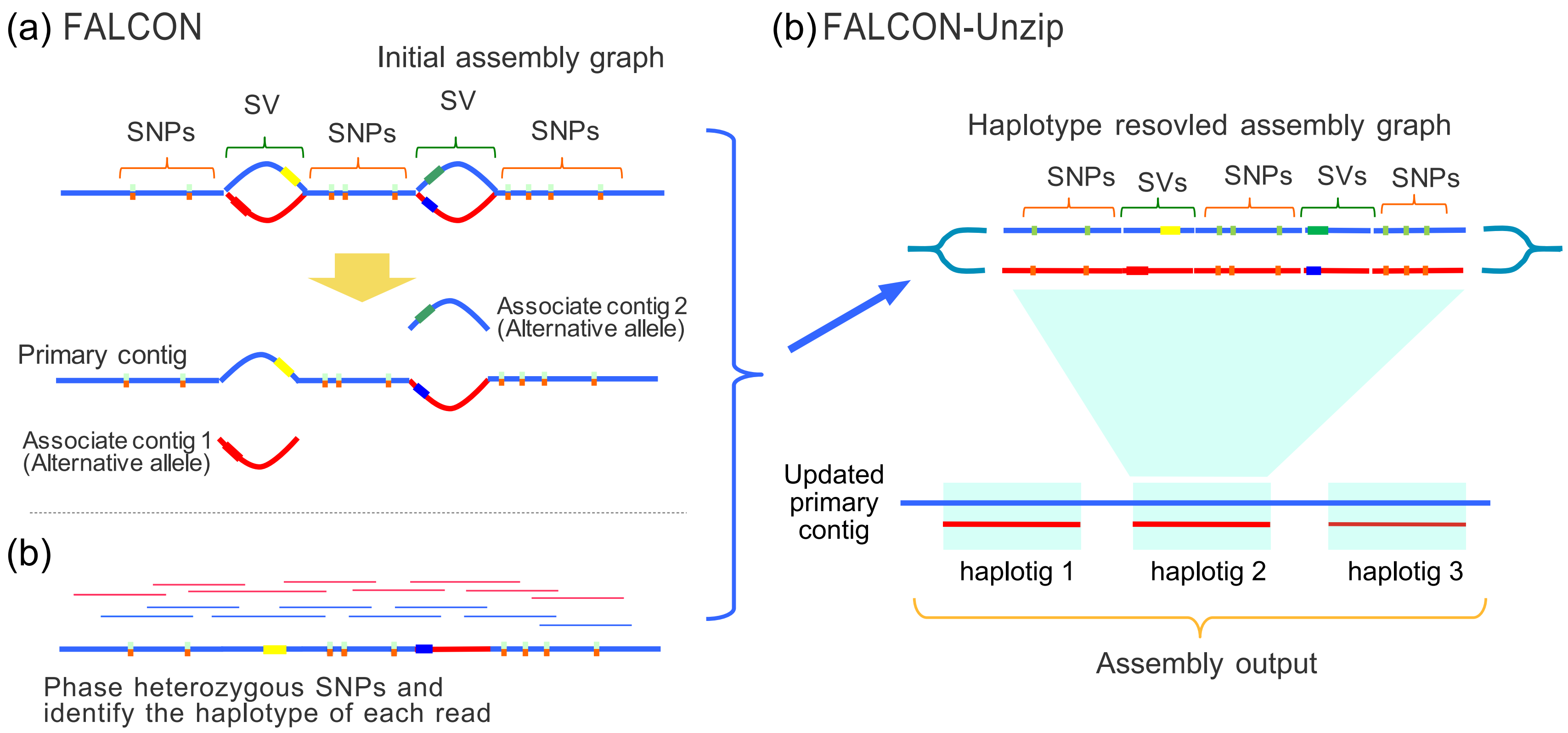
The graph "diameter" ~ 12 M bp
Mean edge size=17.4 k bp



Branching point caused by repeats

"Bubbles" caused by SV between homologous copies

Algorithm overview



1. Assemble Genome with FALCON

- Consensus is a mosaic of the two alleles, except large SVs that form bubbles

2. Use bubbles to seed phasing in flanking regions

- Greedy analysis of heterozygous SNPs flanking SV regions

3. Update Assembly graph with phased sequences: Phased Haplotigs

A. thaliana Assemblies

Two inbred lines, CVI-0 and Col-0, were sequenced separately about 1.5 years ago with P5C3 chemistry

- Compare Col-0 assembly to TAIR reference
- Establish very high quality reference for CVI

Characterize the variations between the two strains with the per-strain haploid assemblies:

- High SV density: big SV every 80 kb
- High SNP density: SNP every 100 to 300 bp

In silico diploid dataset:

- Mixture of the two datasets to emulate a diploid genome at about 80x coverage.
- Useful for testing and development

Genuine diploid dataset:

- Sequencing of an F1 progeny to 120x coverage

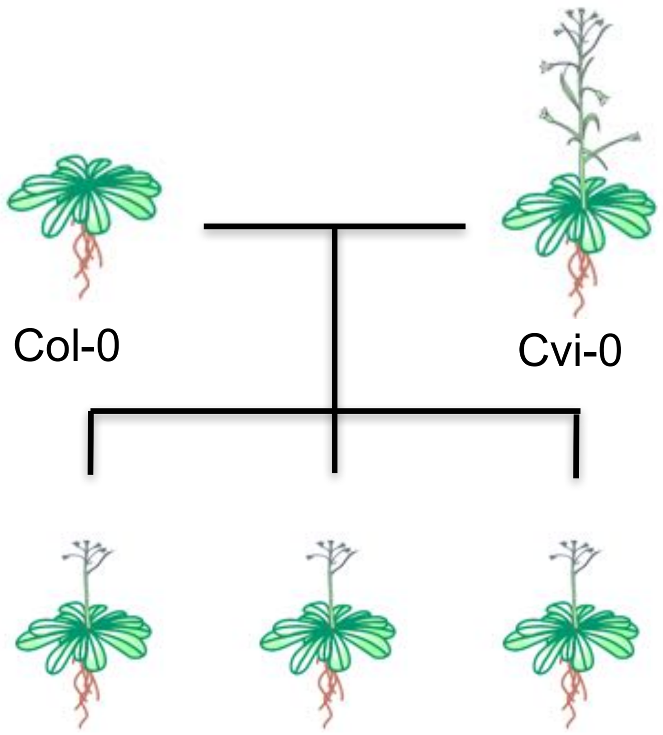
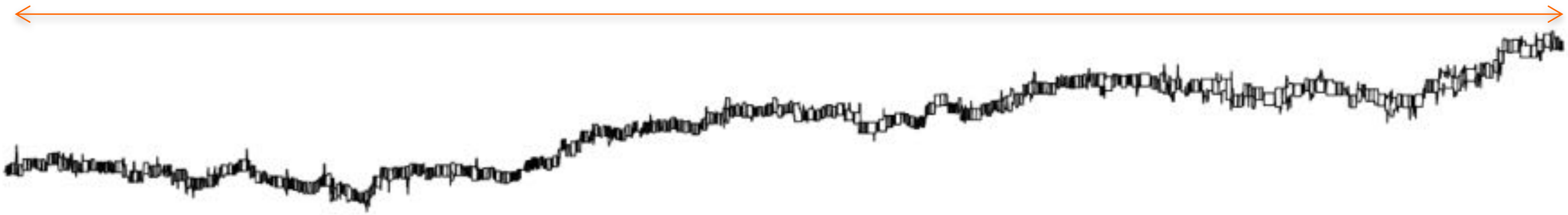
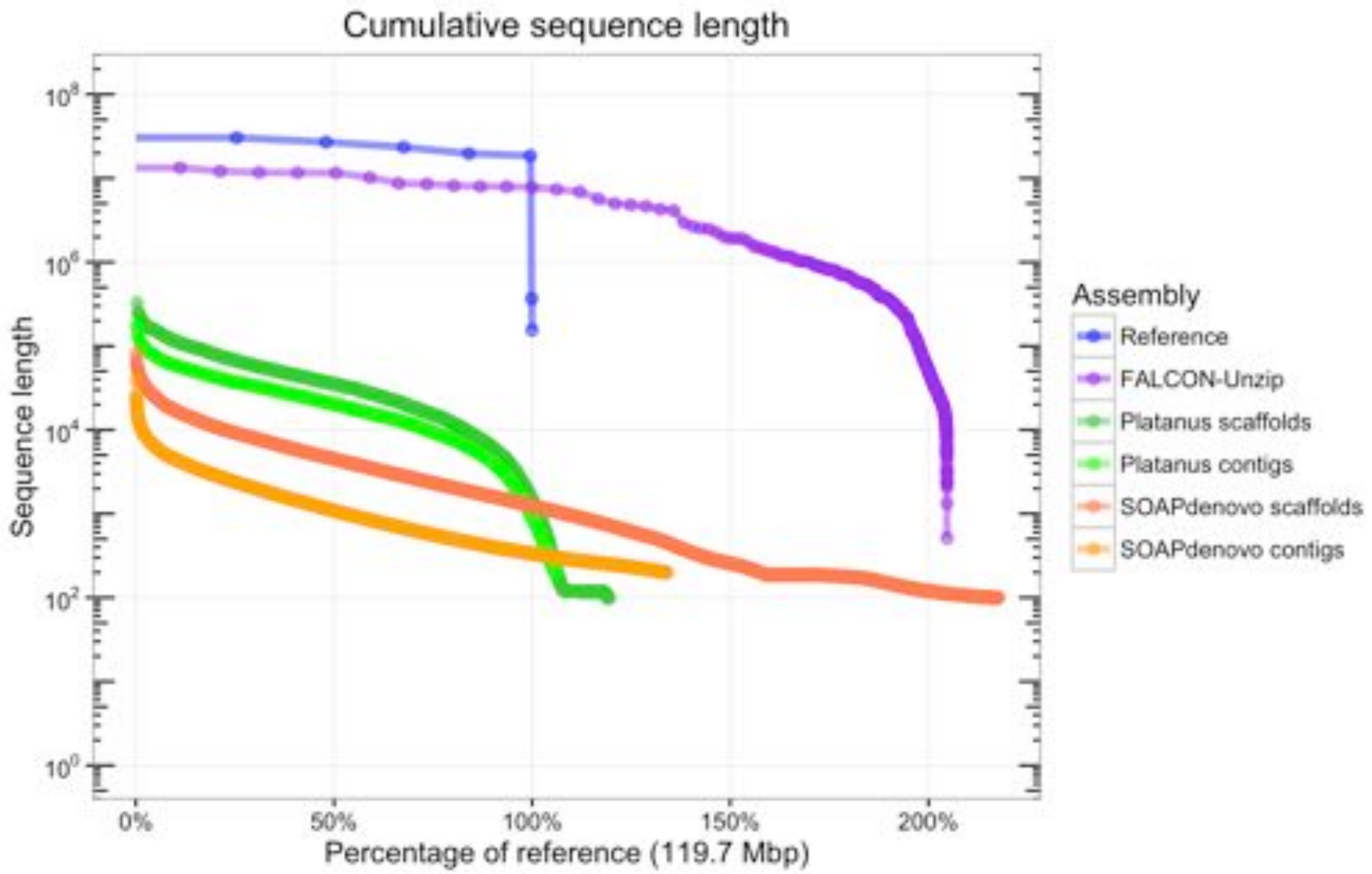


Image credits:
Pajoro, et al, Trends in plant science 21.1 (2016): 6-8.

9.49 Mb haplotype fused assembly graph



A. thaliana F1 Assembly Results



Cumulative sequence length of three *Arabidopsis* F1 assemblies created by FALCON-Unzip, Platanus, and SOAPdenovo compared to the TAIR10 reference.

FALCON-unzip:
Phased Diploid Genome Assembly with PacBio Long Reads

	C. pyxidata (Coral fungus)	Cabernet Sauvignon	*T. guttata* (Zebra finch)$*	Human*
Haploid Genome Size:	~ 44 Mb	~ 500 Mb	~1.2 Gb	~ 3 Gb
Sequencing Coverage	4.1 Gb / 95x	73.7 Gb / 147x	50 Gb / 42x	255 Gb / 85x
Primary contig size	41.9 Mb	591.0 Mb	1.07 Gb	2.76 Gb
Primary contig N50	1.5 Mb	2.2 Mb	3.23 Mb	22.9 Mb
Haplotig size	25.5 Mb	372.2 Mb	0.84 Gb	2.0 Gb
Haplotig N50	872 kb	767 kb	910 kb	330 kb

$ Thanks to Erich Jarvis for permission to use preliminary data

* Preliminary results. Fast file system and efficient computational infrastructure are currently needed for large genomes.

Phased Diploid Genome Assembly with Single Molecule Real-Time Sequencing
Chin, CS, Peluso, P, Sedlazeck, FJ, Nattestad, M, Concepcion, GT, Clum, A, Dunn, C, O'Malley, R, Figueroa-Balderas, R, Morales-Cruz, A, Cramer, GR, Delledonne, M, Luo, C, Ecker, JR, Cantu, D, Rank, DR., Schatz, MC
(2016) Nature Methods doi:10.1038/nmeth.4035

Outline

1. **Phased Genome Assembly**

   GenomeScope & FALCON-Unzip

2. **Personalized-Omics**

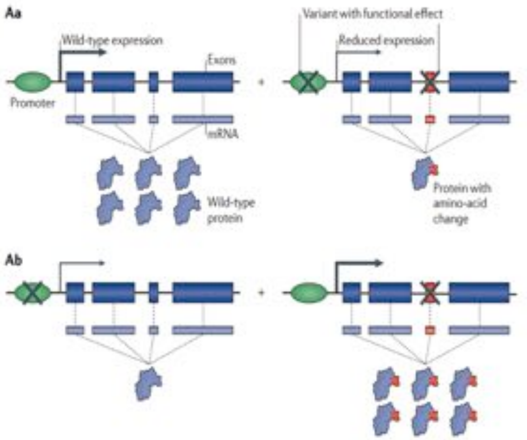
   Complex SVs and oncogene amplifications in breast cancer

Importance of Personal Genomes

Functional data analysis often performed relative to a standard reference genome, but there are many reasons to analyze relative to a phased personal genome

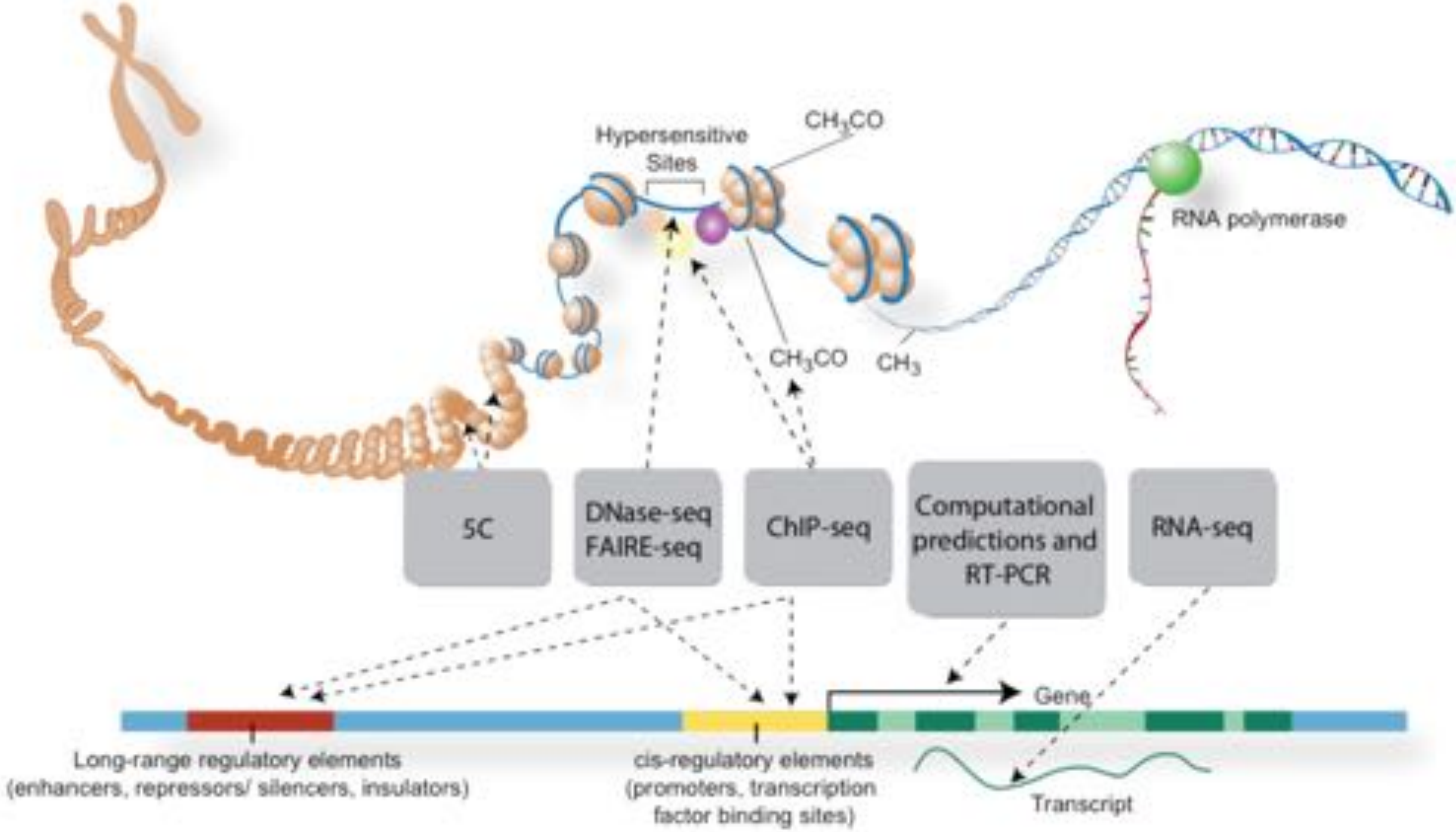
- ***More accurate read mapping***: especially reads spanning significant structural variations
- ***Genomic insights into the expression program***: mutations of splice sites or regulatory elements, CNVs modulate expression levels, gene fusions
- ***Relate regulatory variants to expression of genes***: *cis* versus *trans* effects, *allele-specific expression, allele-specific binding*
- ***Detailed analysis of inheritance and haplotypes***

- …



Tewhey et al (2011)
Nat. Rev. Gen.

Personal Genome Projects


Hypersensitive Sites
CH$_3$CO
RNA polymerase
CH$_3$CO
CH$_3$
5C
DNase-seq FAIRE-seq
ChIP-seq
Computational predictions and RT-PCR
RNA-seq
Gene
Long-range regulatory elements (enhancers, repressors/ silencers, insulators)
cis-regulatory elements (promoters, transcription factor binding sites)
Transcript

ENCODE

Genomic: Illumina + PacBio + 10X
Functional: RNA-seq, ChipSeq, DNAase-seq
4 individuals: 2 male + 2 female

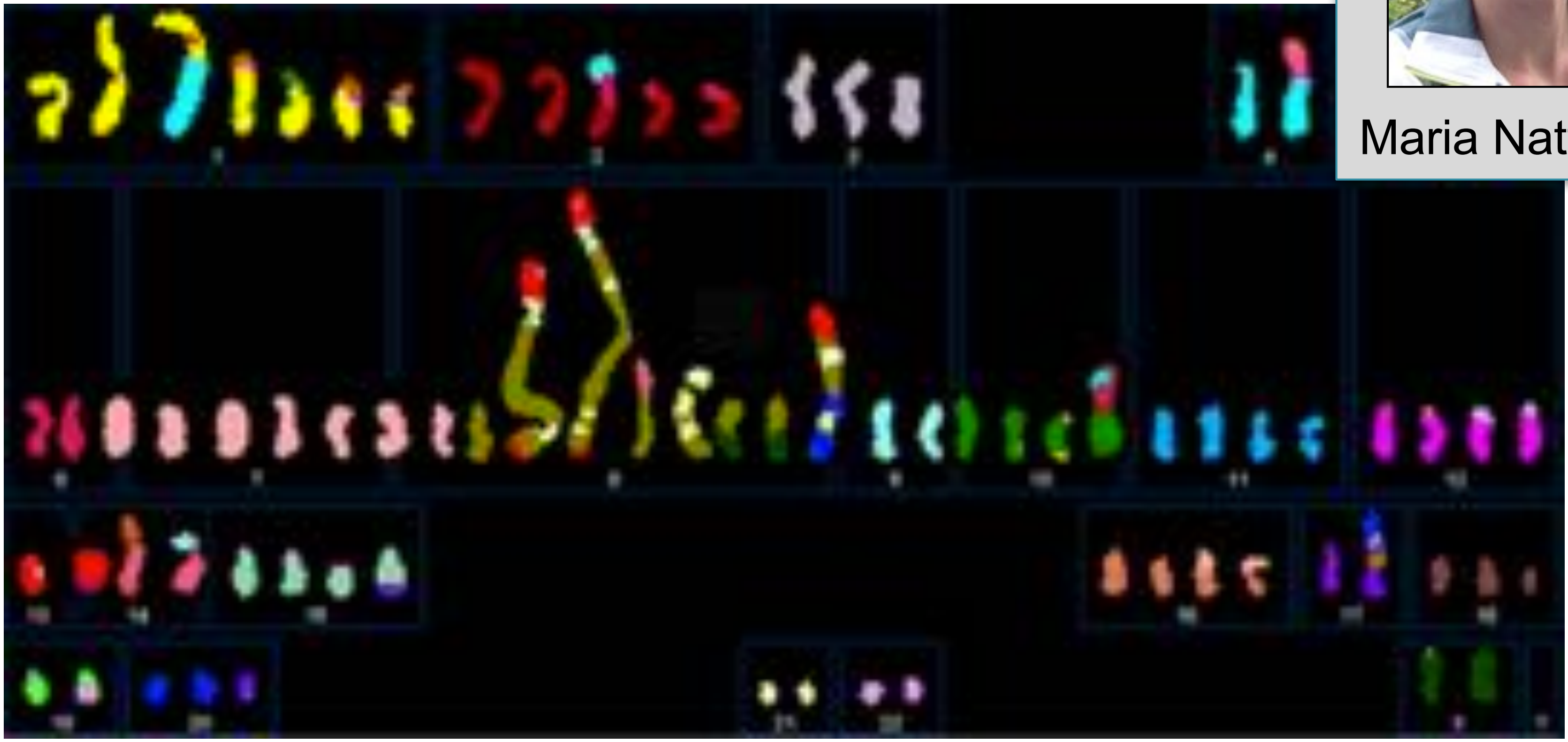
MaizeCode

Genomic: Illumina + PacBio + 10X
Functional: RNA-seq, ChipSeq, MNase-seq
4 accessions: 2 maize + 2 teosinte

SK-BR-3

Most commonly used Her2-amplified breast cancer

Maria Nattestad

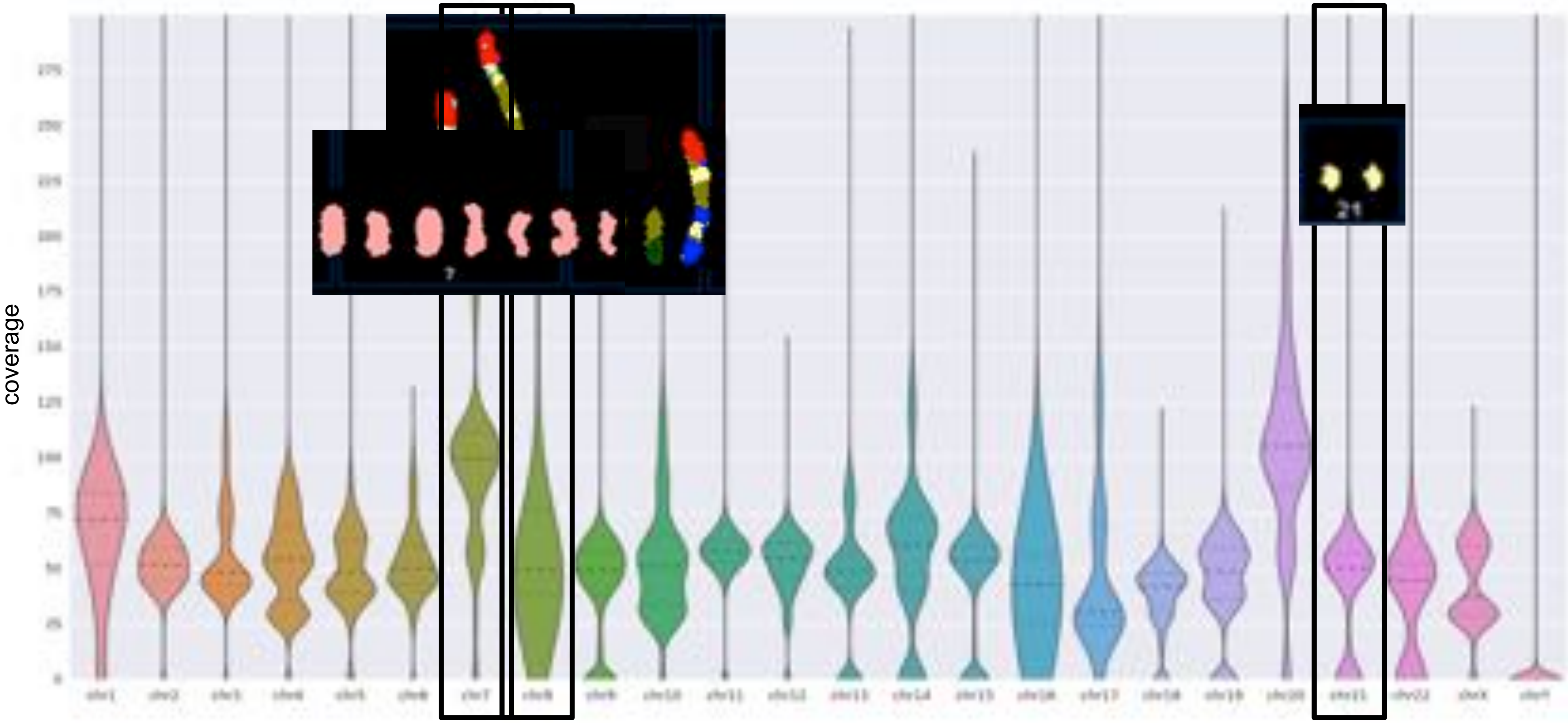


(Davidson et al, 2000)

Can we resolve the complex structural variations, especially around Her2?

Ongoing collaboration between JHU, CSHL and OICR to *de novo* assemble the complete cell line genome with PacBio long reads

Genome Wide Coverage Analysis

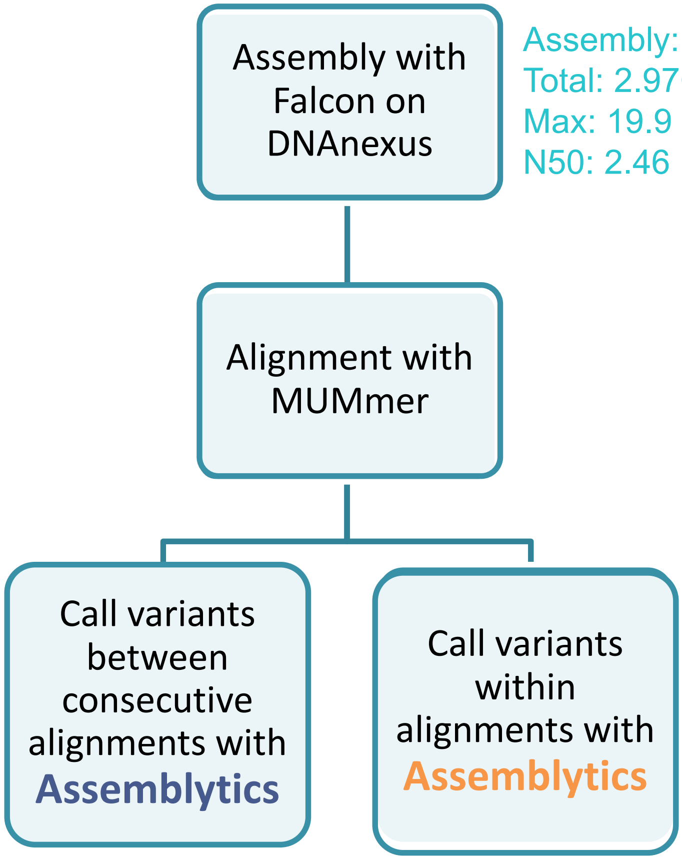


Genome-wide coverage averages around 54X

Coverage per chromosome varies greatly as expected from previous karyotyping results

Structural Variation Analysis

Assembly-based



Assembly with Falcon on DNAnexus

Assembly:
Total: 2.97Gb
Max: 19.9 Mb
N50: 2.46 Mb

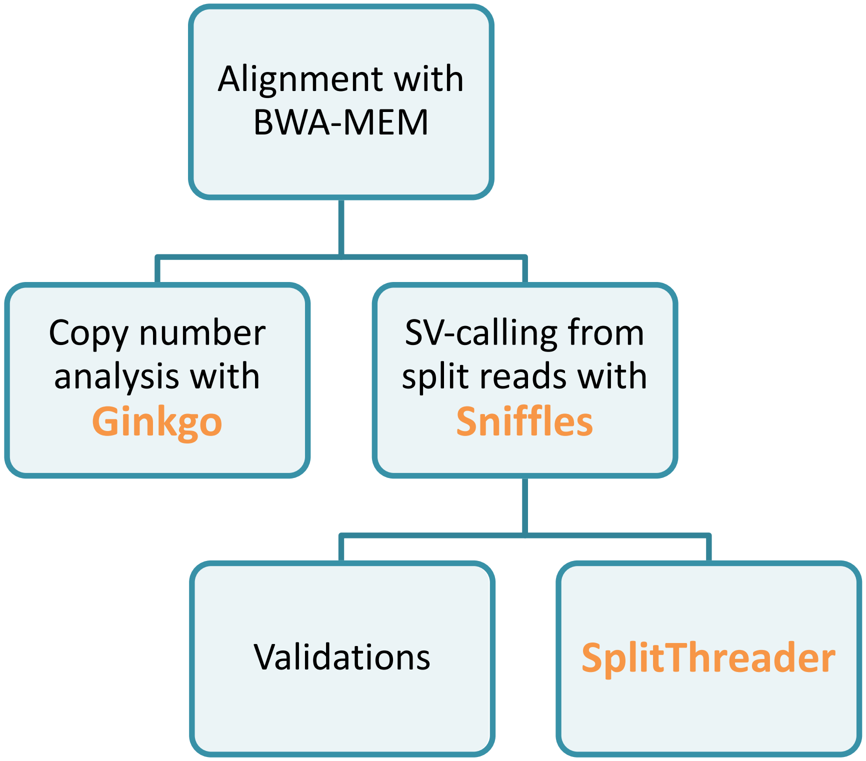
Alignment with MUMmer

Call variants between consecutive alignments with **Assemblytics**

Call variants within alignments with **Assemblytics**

~ 11,000 local variants
50 bp to 10 kbp

Split-Read based



Alignment with BWA-MEM

Copy number analysis with **Ginkgo**

SV-calling from split reads with **Sniffles**

Validations

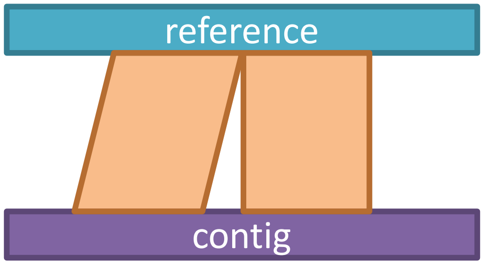
SplitThreader

~ 750 long-range variants
>10kb distance

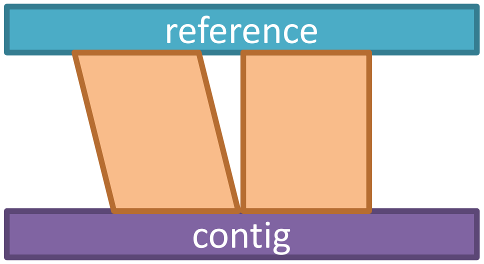
Assemblytics: Assembly-Based Variant-Caller

Defined point

Insertion

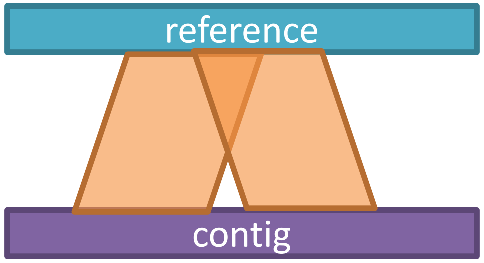


Deletion

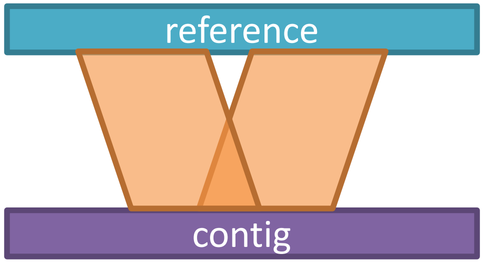


Overlapping alignments suggest tandem repeat

Tandem Expansions

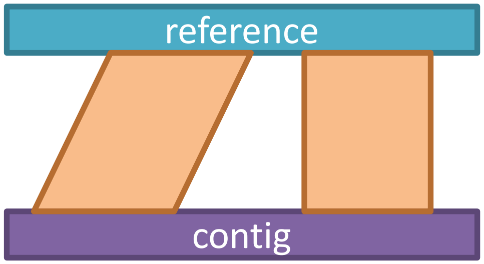


Tandem Contractions

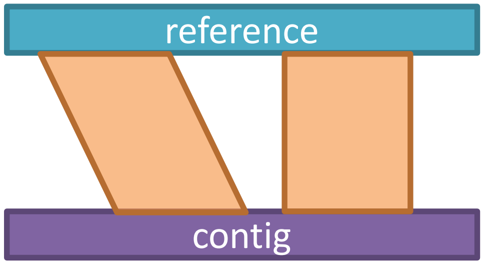


Gap where sequences do not align uniquely suggests a repeat

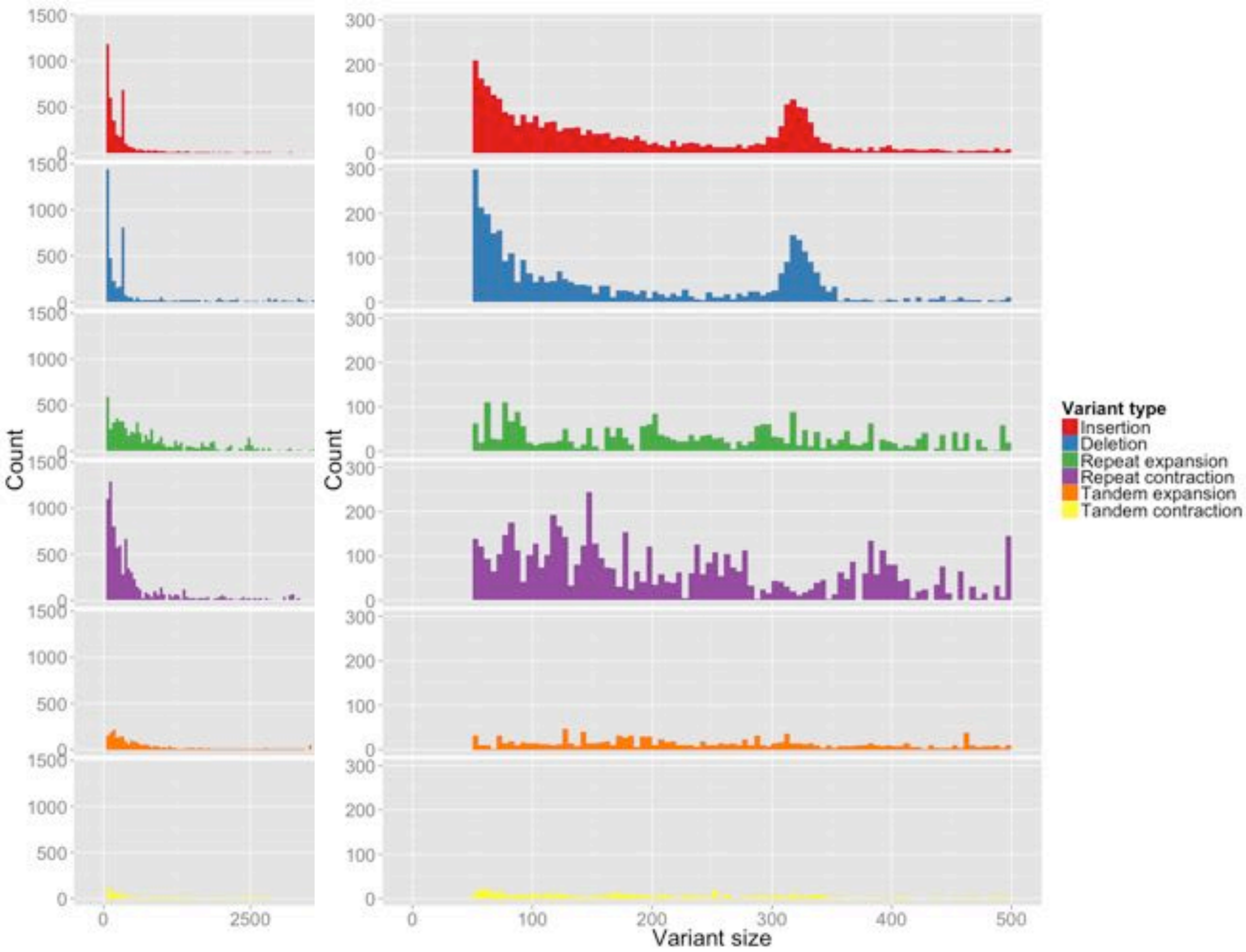
Repeat Expansions

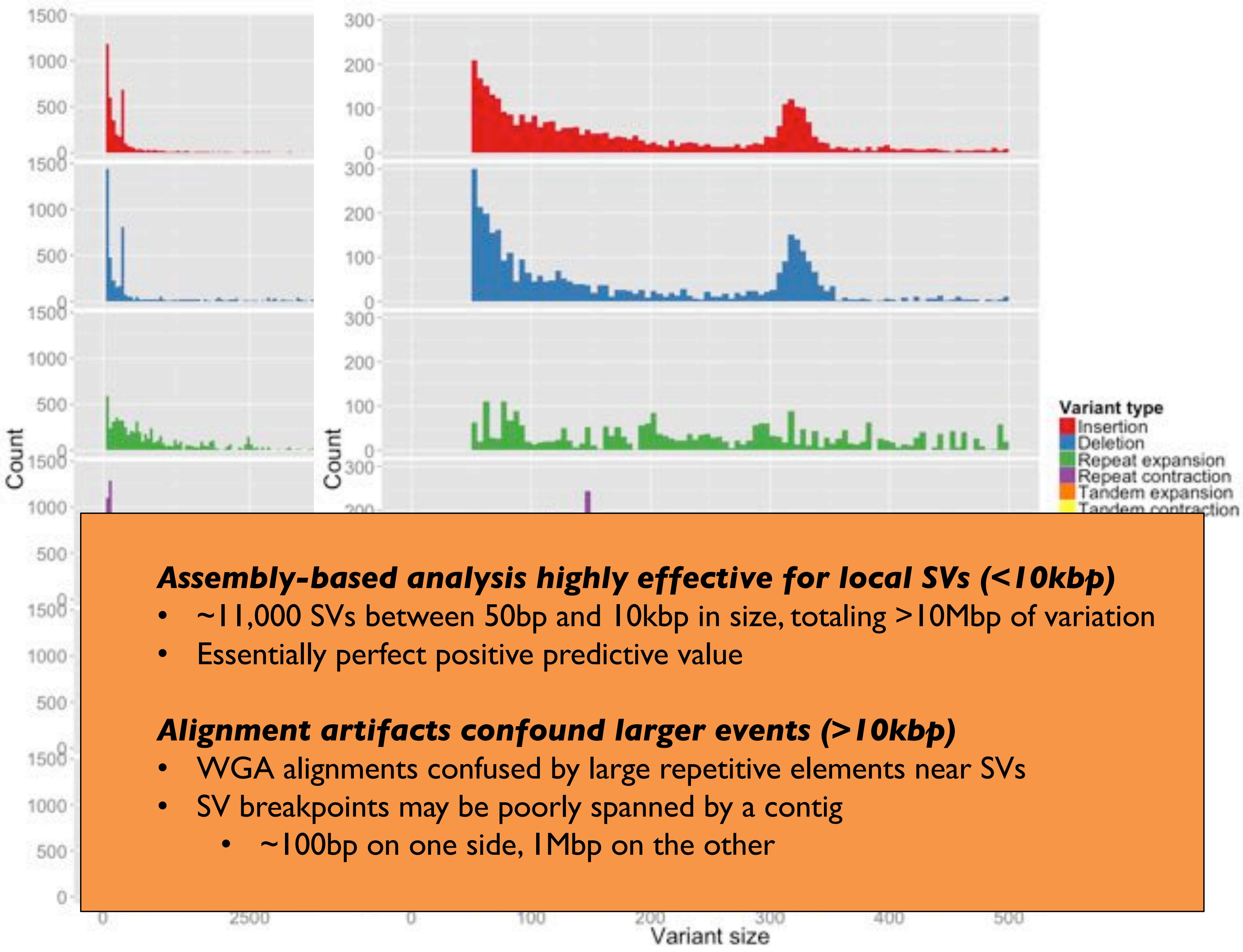


Repeat Contractions



Assemblytics: a web analytics tool for the detection of variants from an assembly
Nattestad, M, Schatz, MC (2016) Bioinformatics doi: 10.1093/bioinformatics/btw369


1500
1000
500
0
300
200
100
0
2500
Count
Variant size
Variant type
Insertion
Deletion
Repeat expansion
Repeat contraction
Tandem expansion
Tandem contraction
400
500



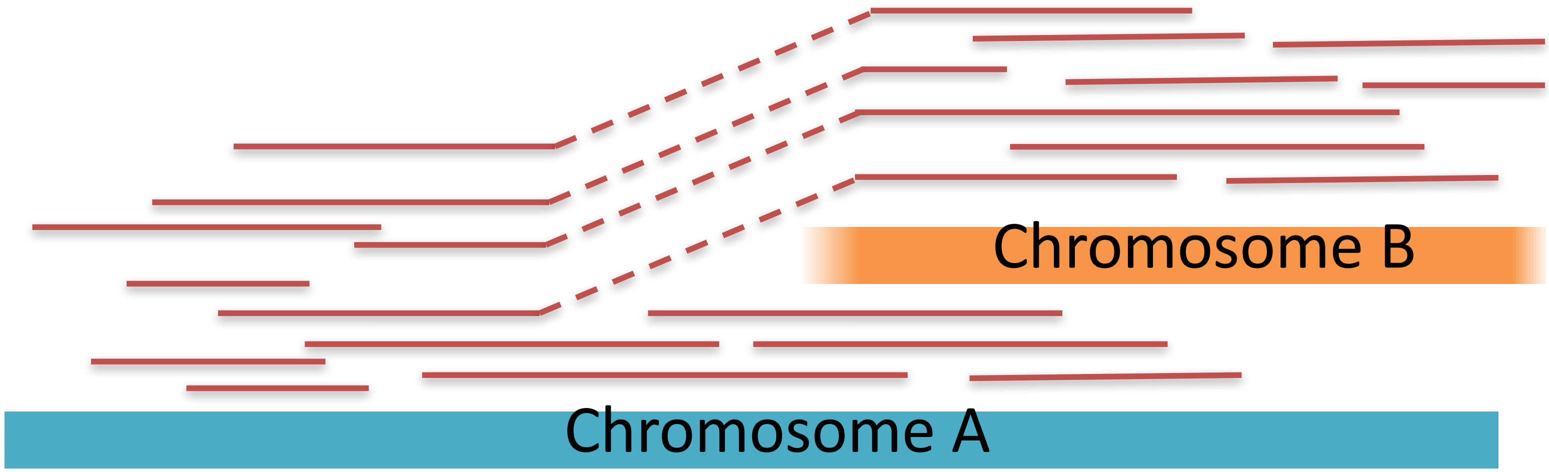
Assembly-based analysis highly effective for local SVs (<10kbp)

- ~11,000 SVs between 50bp and 10kbp in size, totaling >10Mbp of variation
- Essentially perfect positive predictive value

Alignment artifacts confound larger events (>10kbp)

- WGA alignments confused by large repetitive elements near SVs
- SV breakpoints may be poorly spanned by a contig
  - ~100bp on one side, 1Mbp on the other

Alignment-Based Structural Variation Analysis



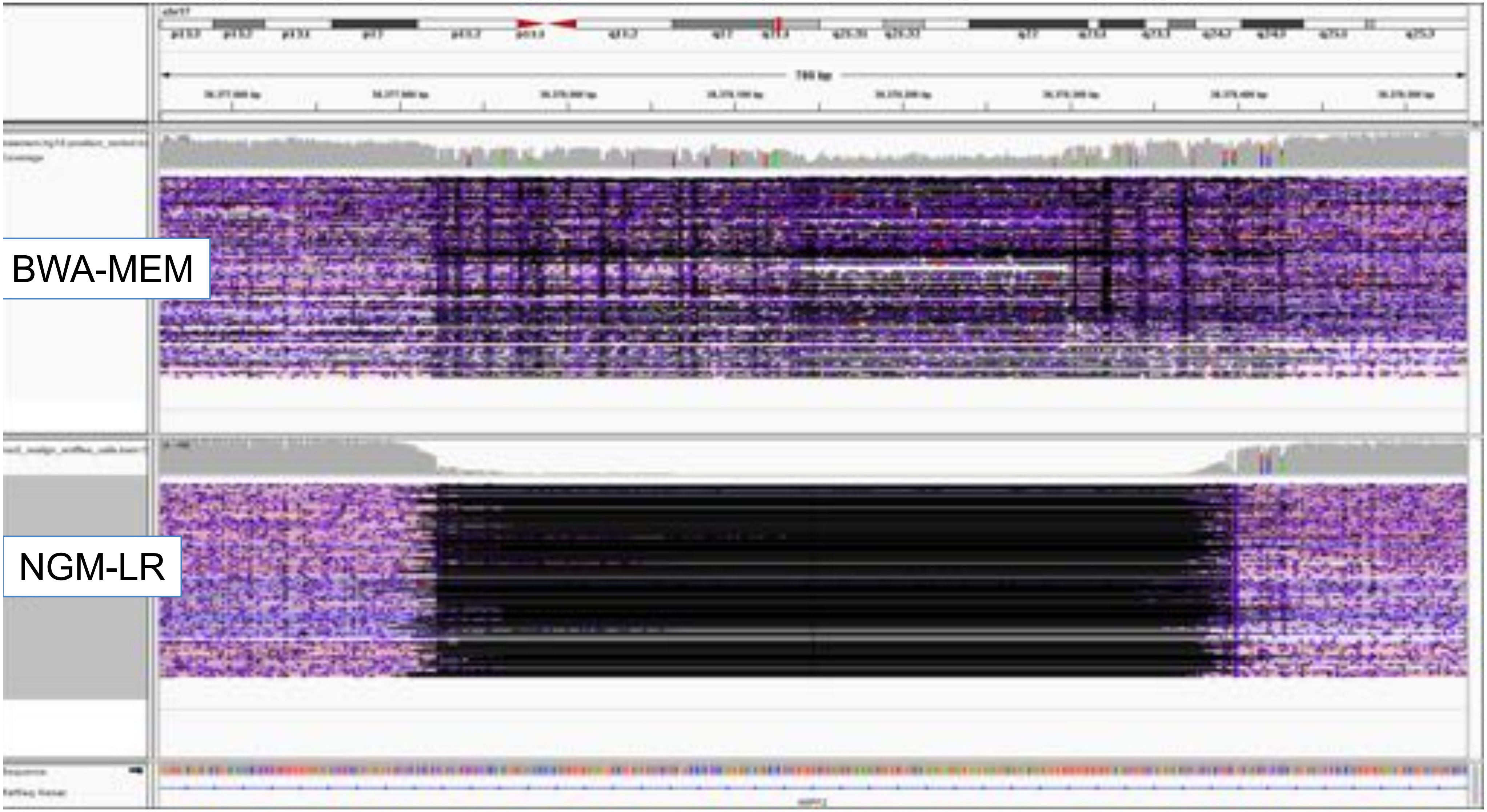
Alignment based analysis greatly improved by long reads

- More confident mappings, Improved chances of spanning events
- However, many SVs lost due to poor alignments and poor PacBio support
  - LUMPY fails on reads that span more than 1 breakpoint, poor localization

New methods in development: NGM-LR + Sniffles

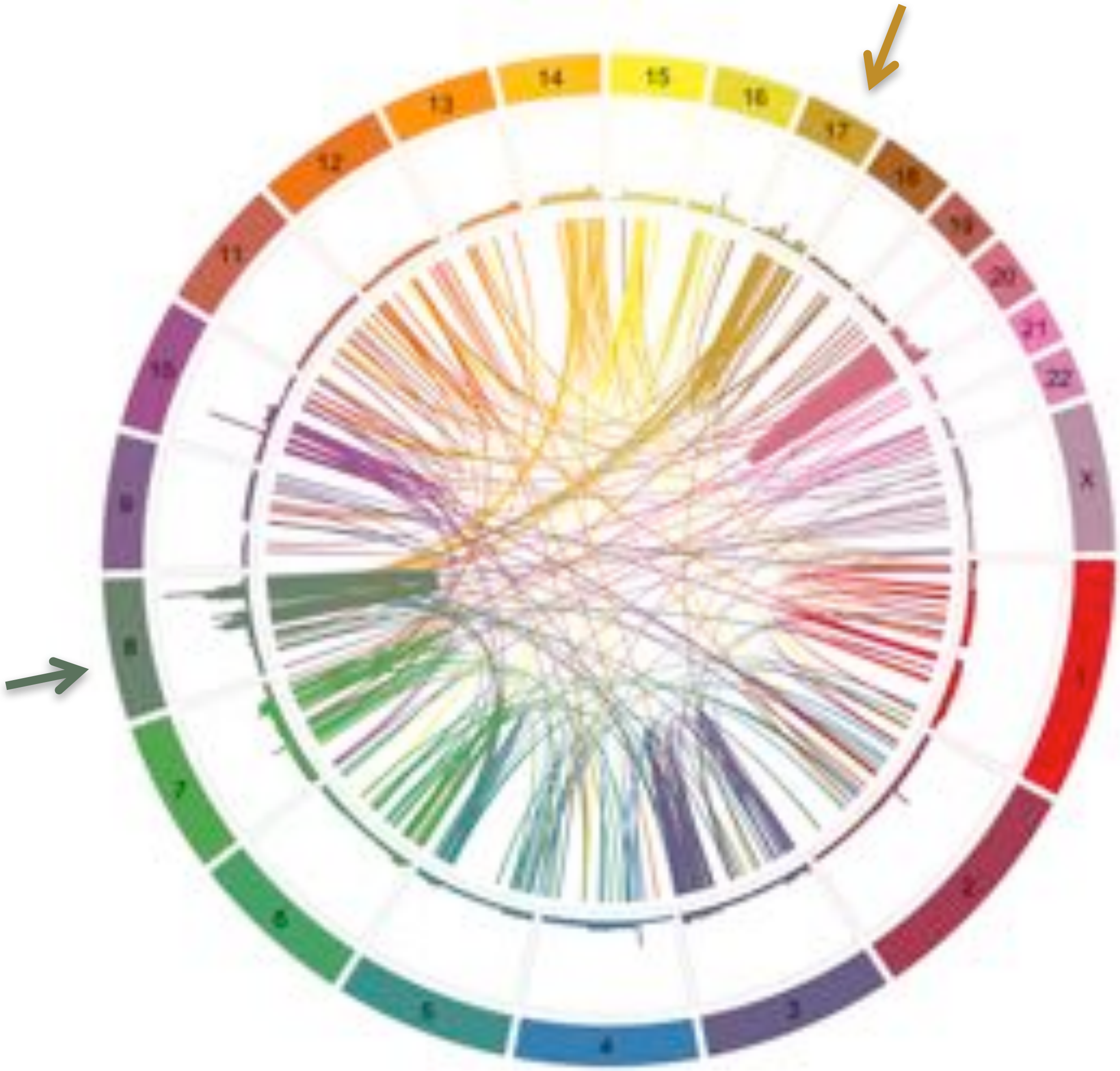
1. ***NGM-LR***: Improve mapping of noisy long reads
2. ***Sniffles***: Integrates SV evidence from split-read alignments, alignment fidelity (CIGAR strings and MD tags)

Mapping a ~500bp deletion



Similar issues for insertions, inversions; or Nanopore sequencing
Improved seeding, improved gap scoring: convex instead of affine

Long Range Variations in SK-BR-3

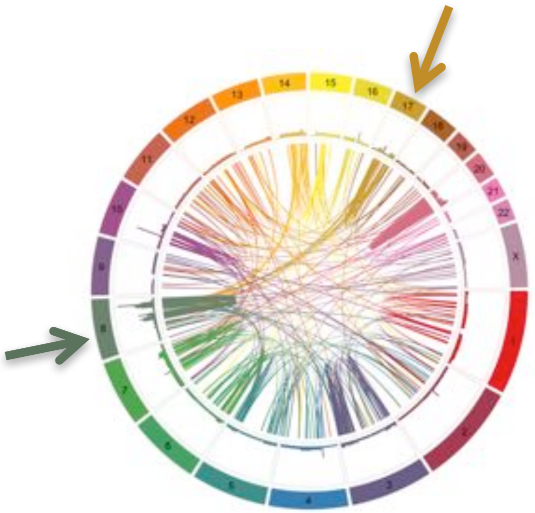


Fritz Sedazeck

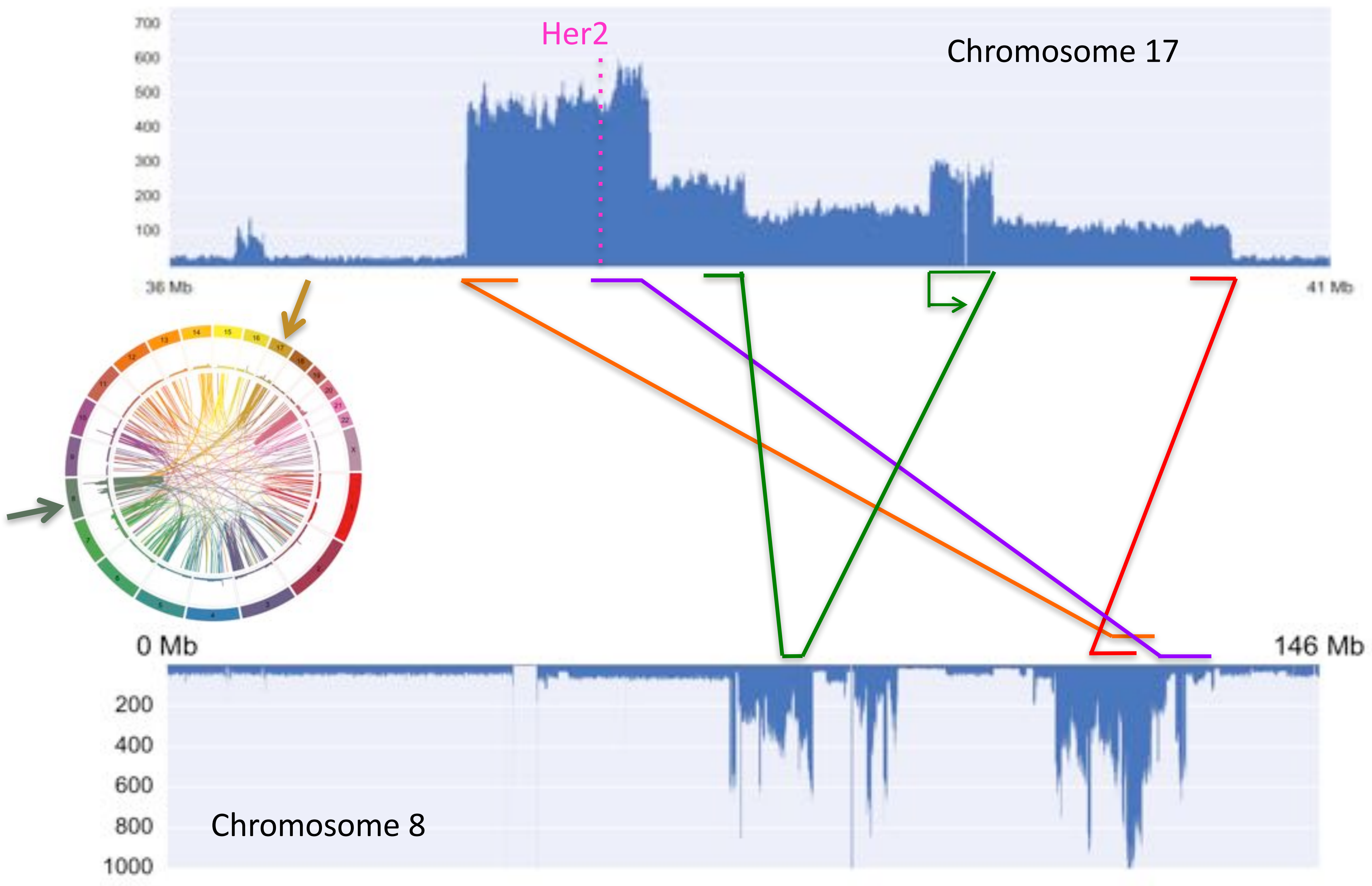
Analysis by Sniffles

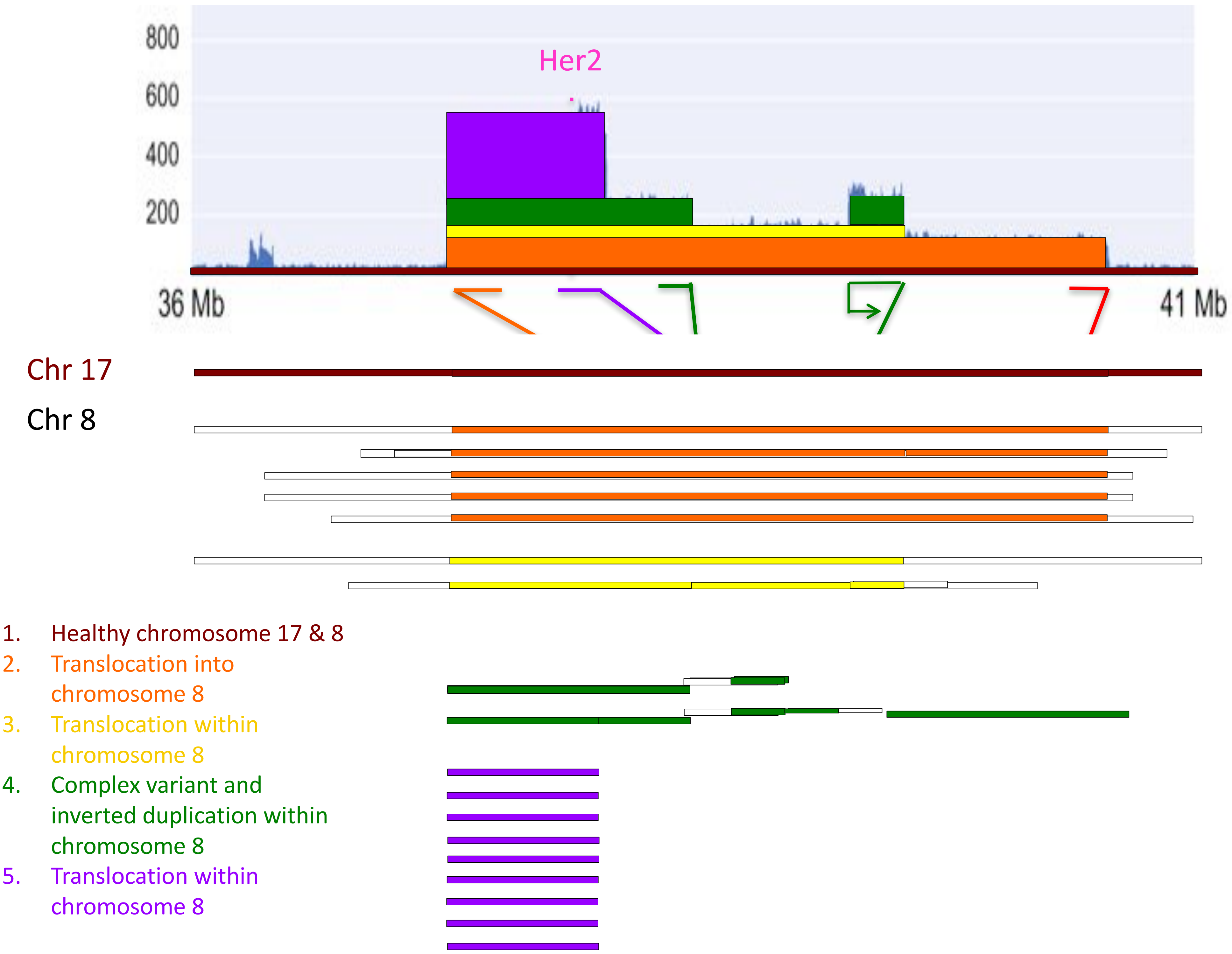
- ~750 variants >= 10kbp
- ~200 balanced translocations
- Requires 10 split reads broken within a 200 bp interval on both sides

Long-range structural variants found by Sniffles

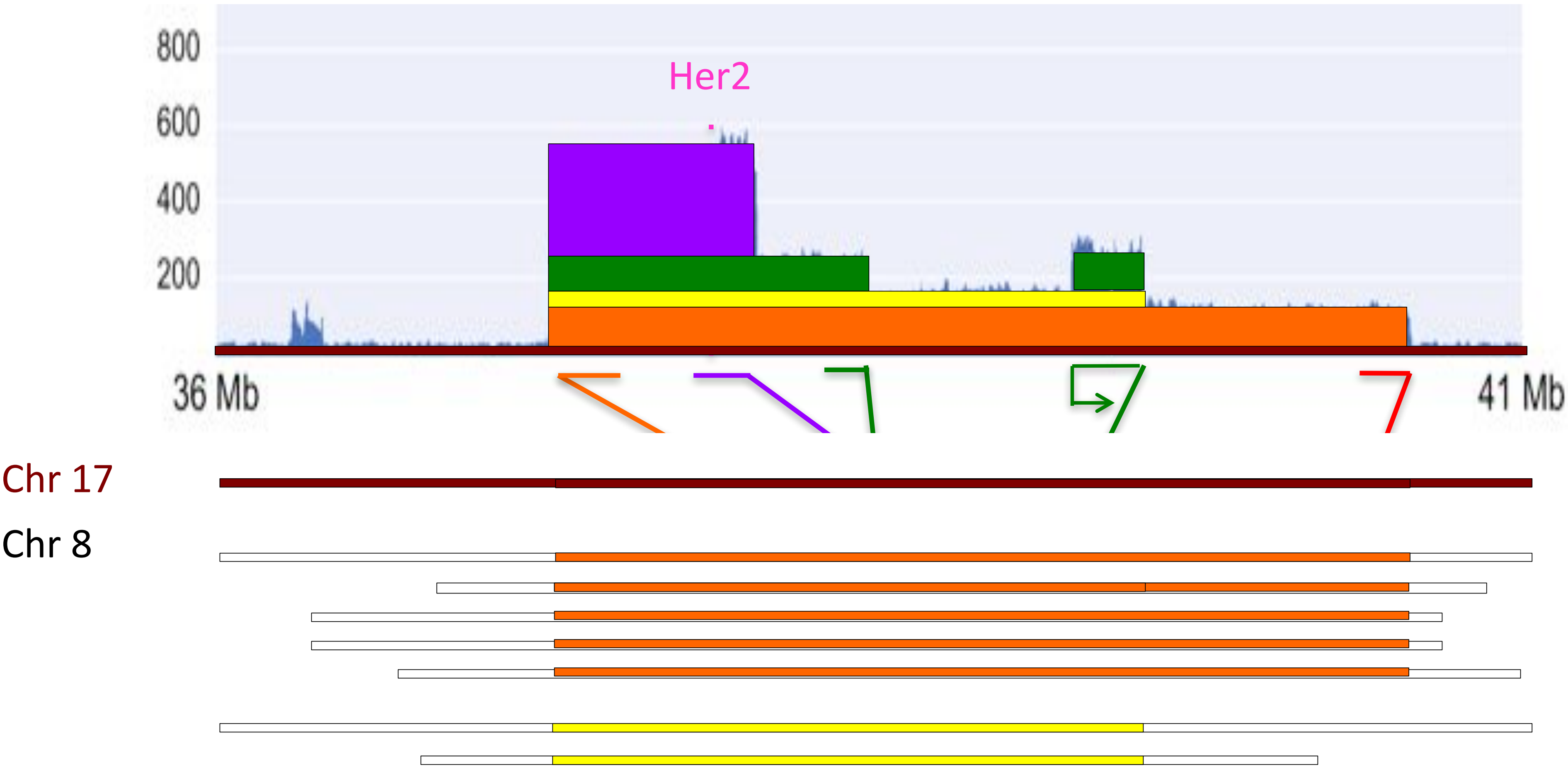


Long-range structural variants found by Sniffles

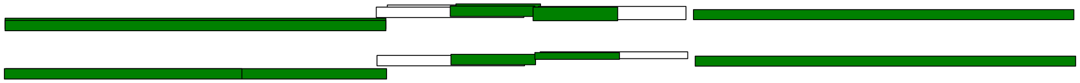


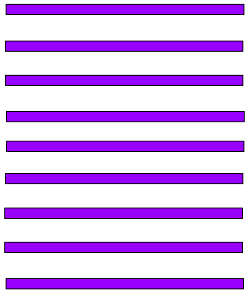


1. Healthy chromosome 17 & 8
2. Translocation into chromosome 8
3. Translocation within chromosome 8
4. Complex variant and inverted duplication within chromosome 8
5. Translocation within chromosome 8



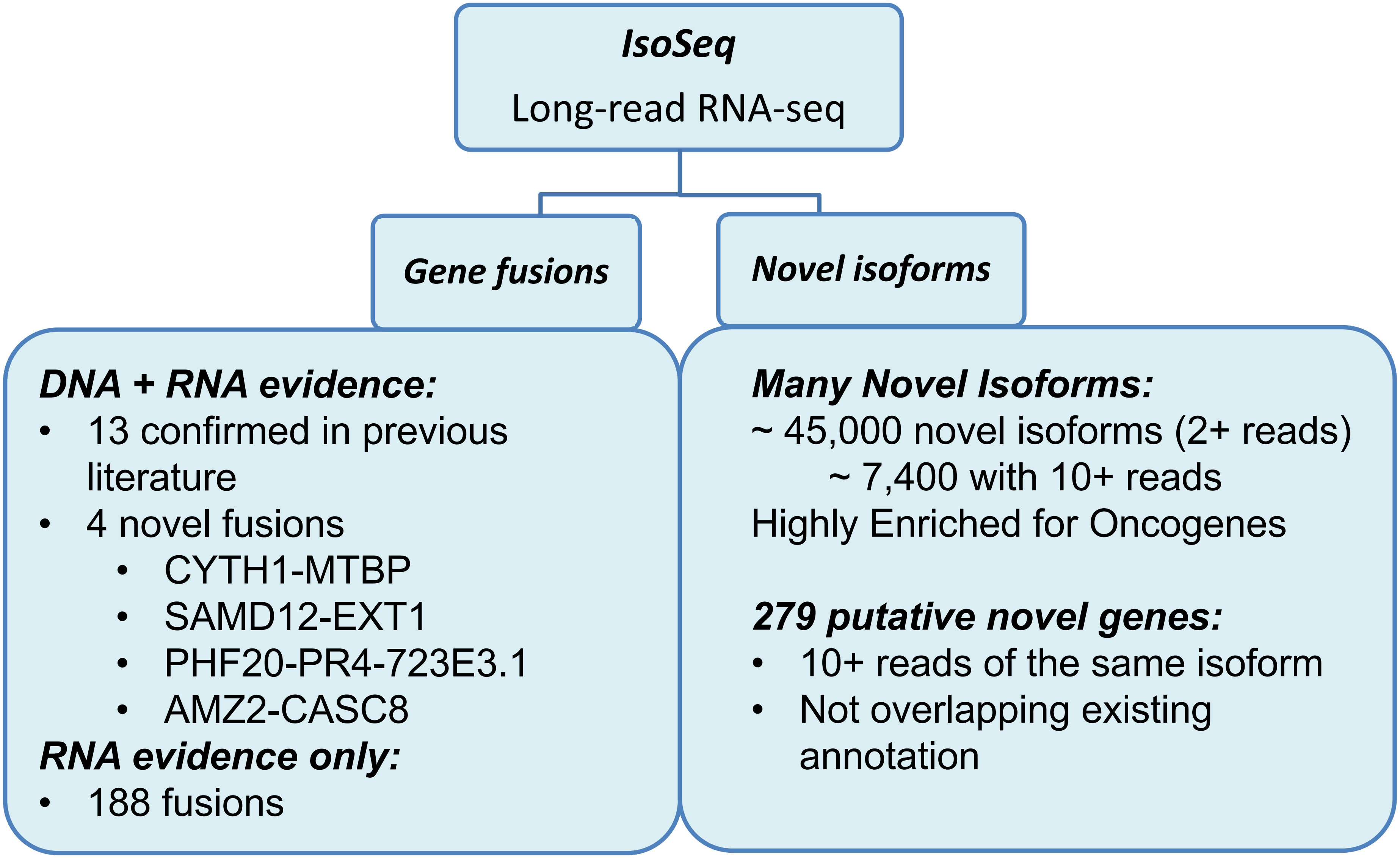
1. Healthy chromosome 17 & 8
2. Translocation into chromosome 8
3. Translocation within chromosome 8
4. Complex variant and inverted duplication within chromosome 8
5. Translocation within chromosome 8





Inferring the evolution of genome structure

Transcriptome analysis with IsoSeq



IsoSeq

Long-read RNA-seq

Gene fusions

DNA + RNA evidence:

- 13 confirmed in previous literature
- 4 novel fusions
  - CYTH1-MTBP
  - SAMD12-EXT1
  - PHF20-PR4-723E3.1
  - AMZ2-CASC8

RNA evidence only:

- 188 fusions

Novel isoforms

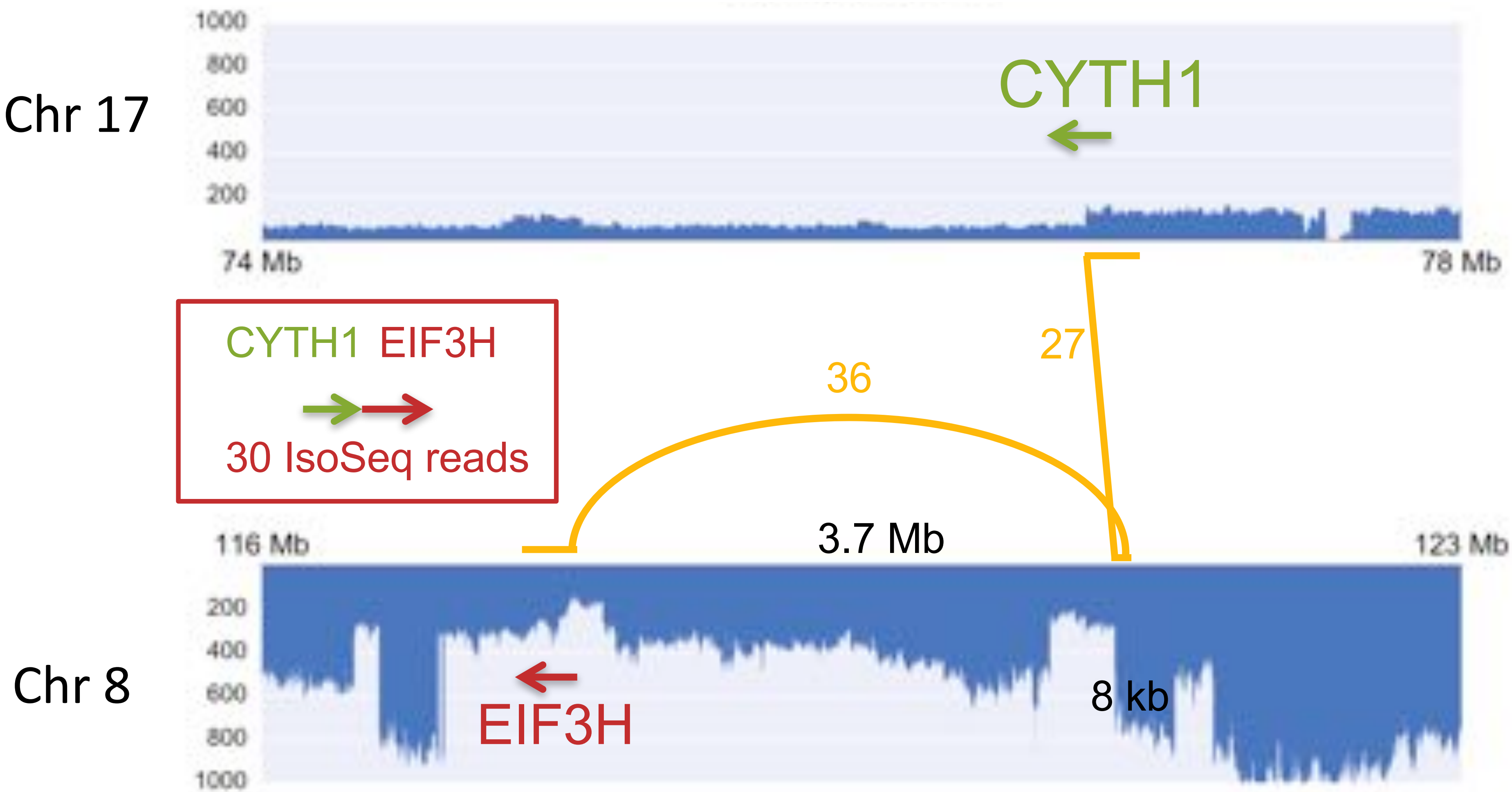
Many Novel Isoforms:

~ 45,000 novel isoforms (2+ reads)
~ 7,400 with 10+ reads
Highly Enriched for Oncogenes

279 putative novel genes:

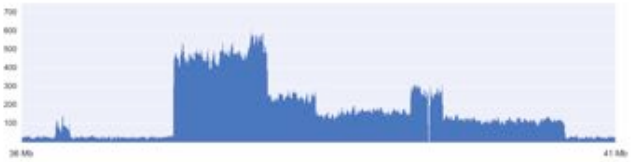
- 10+ reads of the same isoform
- Not overlapping existing annotation

CYTH1-EIF3H gene fusion

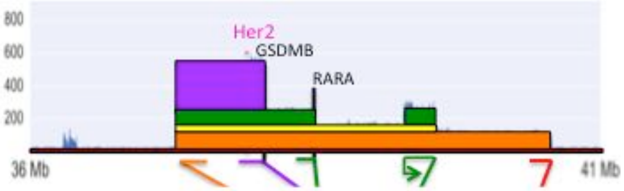


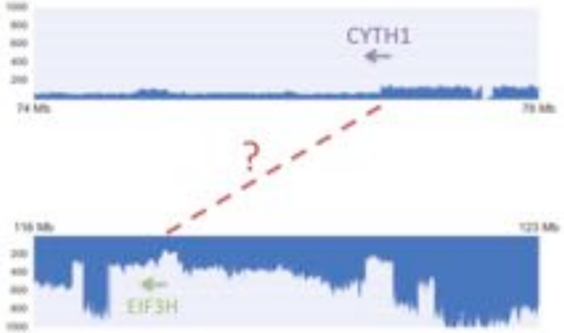
SplitThreader: Exploration and analysis of rearrangements in cancer genomes
Nattestad, M, Alford, MC, Sedlazeck, FJ, Schatz, MC (2016) bioRxiv. doi: https://doi.org/10.1101/087981

The genome informs the transcriptome

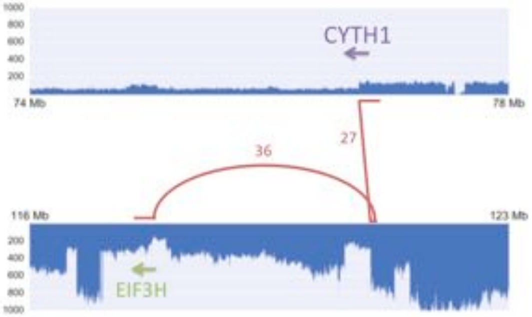


Explain amplifications



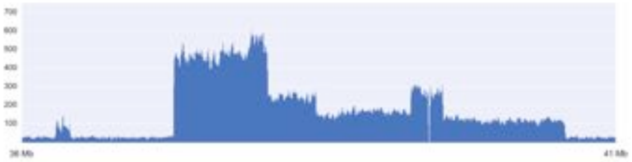


Trace gene fusions

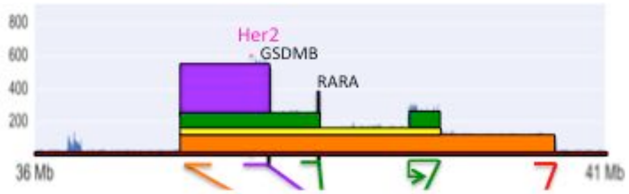


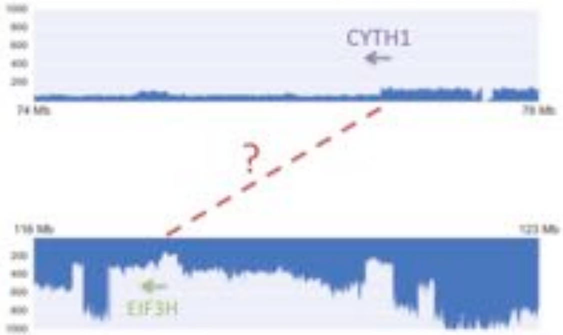
Data and additional results: http://schatzlab.cshl.edu/data/skbr3/

The genome informs the transcriptome … and informs the prognosis

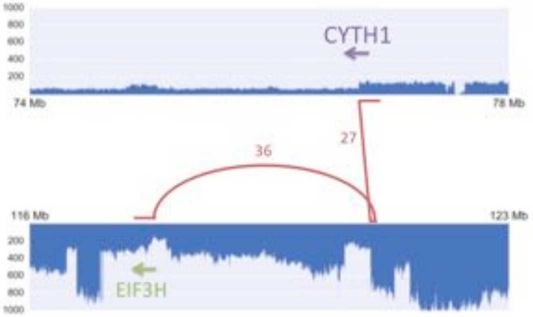


Explain amplifications





Trace gene fusions



Data and additional results: http://schatzlab.cshl.edu/data/skbr3/

PacBio Roadmap

PacBio RS II

$750k instrument cost
1895 lbs

~$75k / human @ 50x

SMRTcell

150k Zero Mode Waveguides
~10kb average read length
~1 GB / SMRTcell
~$500 / SMRTcell

PacBio Roadmap

PacBio Sequel

$350k instrument cost
841 lbs

~$30k / human @ 50x

SMRTcell v2

1M Zero Mode Waveguides
~15kb average read length
~10 GB / SMRTcell
~$1000 / SMRTcell

Oxford Nanopore

MinION

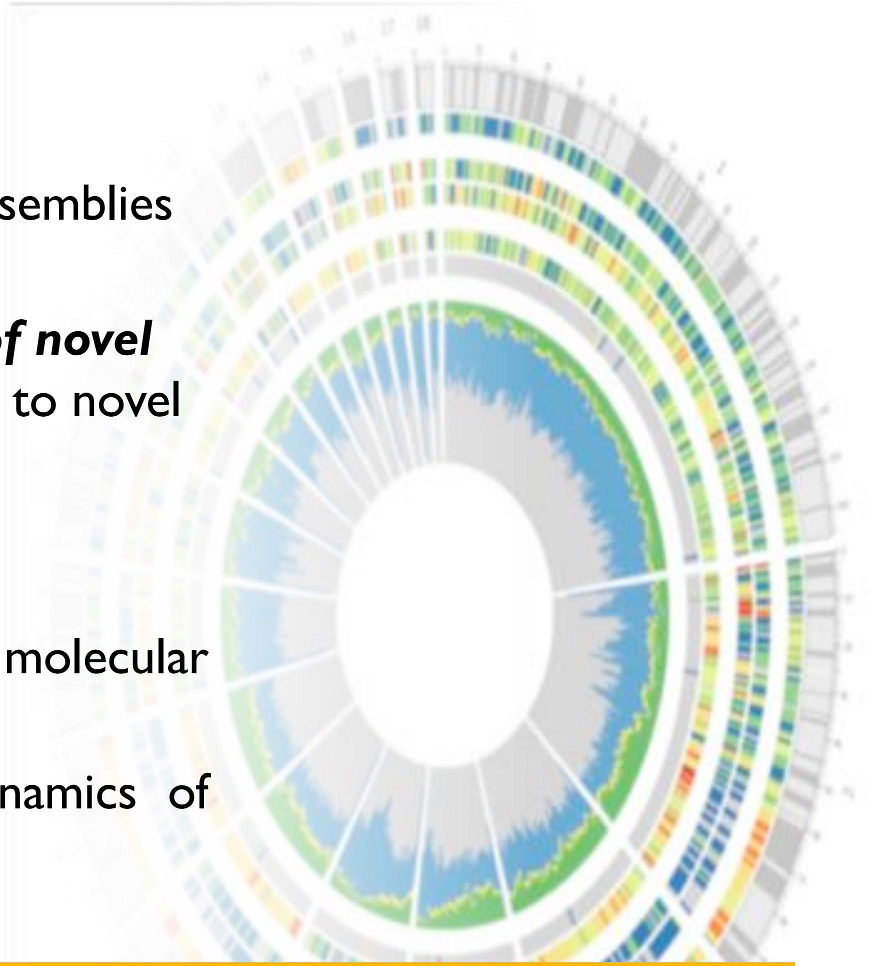
$2k / instrument
1-2 GB / day
~$300k / human @ 50x

PromethION

$75k / instrument
>>100GB / day
??? / human @ 50x

Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome
Goodwin, S, Gurtowski, J, Ethe-Sayers, S, Deshpande, P, Schatz MC* McCombie, WR* (2015) Genome Research doi: 10.1101/gr.191395.115

Understanding Genome Structure & Function



Single Molecule Sequencing

- Now have the ability create ***reference quality*** assemblies of many microbes, fungi, plants, and animals
- Using this technology to find ***10s of thousands of novel structural variations*** per human genome leading to novel gene structures and regulatory contexts

Single Cell Sequencing

- Exciting technologies to probe the genetic and molecular ***composition of complex environments***
- We have only begun to explore the rich dynamics of genomes, transcriptomes, and epigenomics

These advances give us incredible power to study how genomes mutate and evolve

With several new biotechnologies in hand, we are now largely limited only by our quantitative power to make comparisons and find patterns

Acknowledgements

Schatz Lab
Rahul Amin
Rhyker Benavidez
Han Fang
Tyler Gavin
James Gurtowski
Hayan Lee
Zak Lemmon
Giuseppe Narzisi
Maria Nattestad
Aspyn Palatnick
Srividya Ramakrishnan
Fritz Sedlazeck
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

SBU
Skiena Lab
Patro Lab

JHU
Cormack Lab
Langmead Lab
Salzberg Lab
Timp Lab
Wheelan Lab

Cornell
Susan McCouch
Lyza Maron
Mark Wright

OICR
John McPherson
Karen Ng
Timothy Beck
Yogi Sundaravadanam

NYU
Jane Carlton
Elodie Ghedin

Oxford NANOPORE Technologies

PACIFIC BIOSCIENCES®

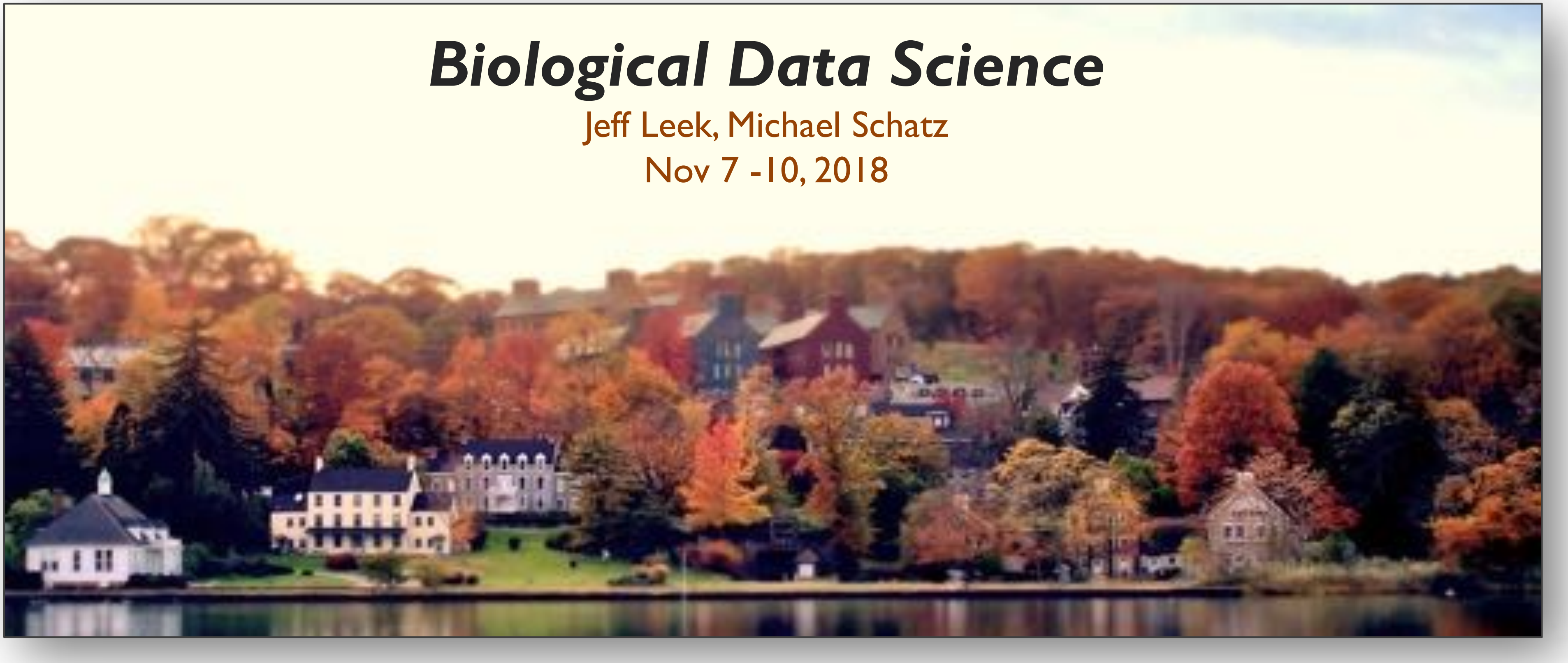
NSF

National Human Genome Research Institute

U.S. DEPARTMENT OF ENERGY

SFARI
SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

ALFRED P. SLOAN FOUNDATION



Biological Data Science

Jeff Leek, Michael Schatz

Nov 7 -10, 2018

Thank you

http://www.cs.jhu.edu/~mschatz

@mike_schatz