

Analysis of Structural Variants using 3rd generation Sequencing

Michael Schatz

January 12, 2016

Bioinformatics / PAG XXIV



[@mike_schatz](#) / [#PAGXXIV](#)

Analysis of Structural Variants using 3rd generation Sequencing

Michael Schatz

January 12, 2016

Bioinformatics / PAG XXIV



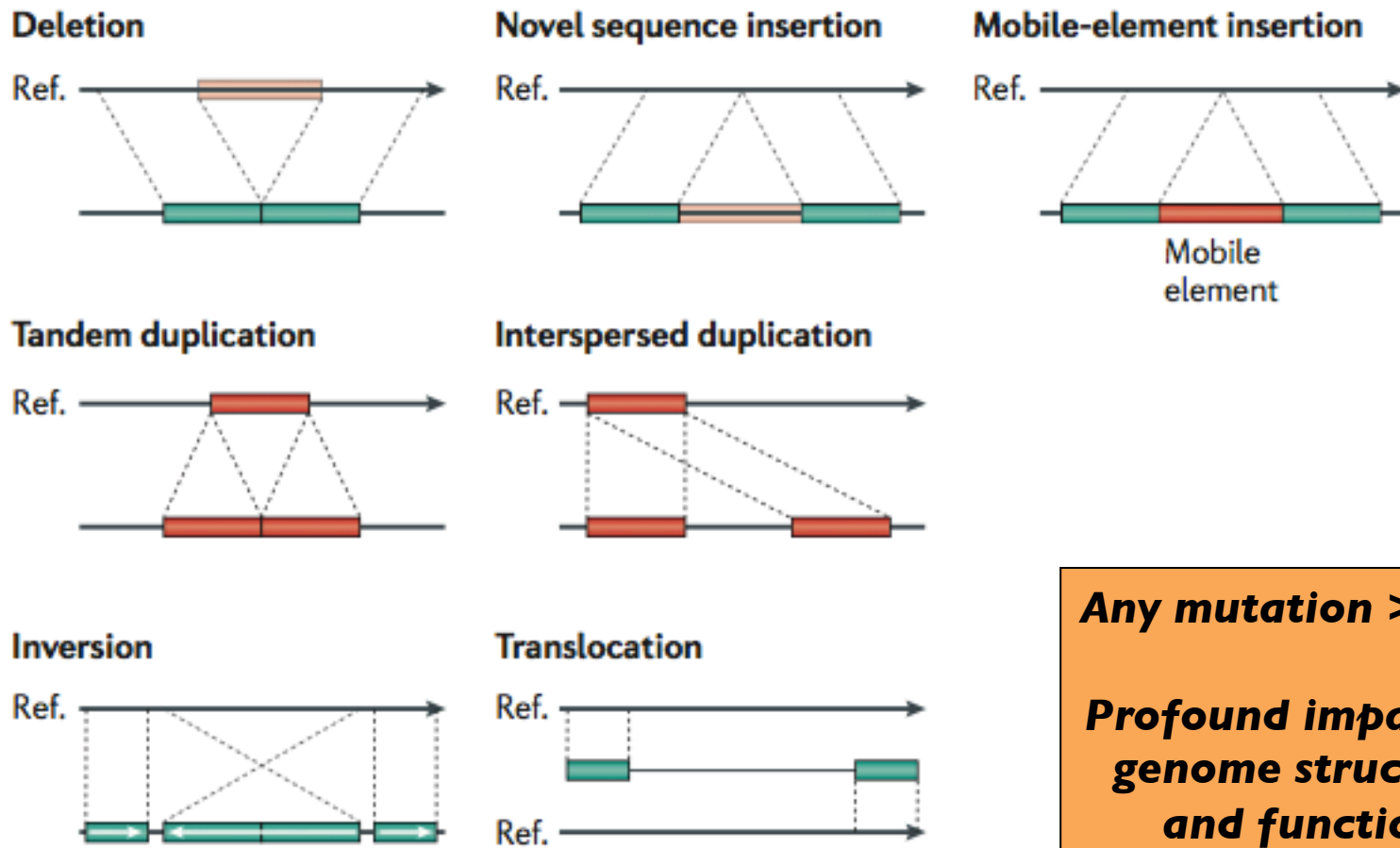
@mike_schatz / #PAGXXIV

The Resurgence of Reference Quality Genomes

Michael Schatz & Daniel Rokhsar
Tuesday, January 12, 2016 @ 4pm – 6pm
Town & Country - Pacific Salon I

4:00pm	<i>The Resurgence of Reference Quality Genomes</i> Michael Schatz, CSHL + JHU
4:20pm	<i>High Quality, Highly Contiguous Genome Assemblies Now</i> Richard Green, Dovetail Genomics
4:40pm	<i>Scalable Parallel Algorithms for de novo Assembly of Complex Genomes</i> Aydin Buluc, Lawrence Berkeley National Laboratory
5:00pm	<i>Using PacBio Long Reads to Generate a High Quality Reference for the Allotetraploid Coffea arabica and its Maternal Diploid Ancestor Coffea eugeniodes</i> Marcela Yepes, Cornell University
5:20pm	<i>MaSuRCA Mega-Reads Assembly Technique for Haplotype Resolved Genome Assembly of Hybrid PacBio and Illumina Data</i> Aleksy Zimin, University of Maryland
5:40pm	<i>How to Compare and Cluster Every Known Genome in about an Hour</i> Sergey Koren, NHGRI

Structural Variations





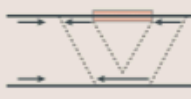
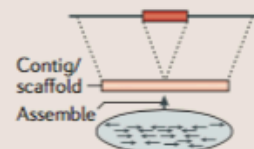
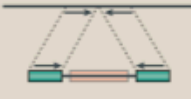

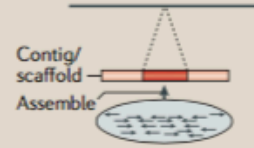

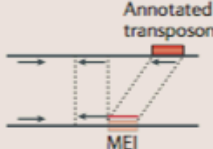
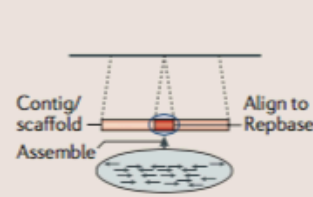

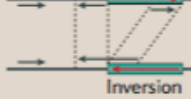

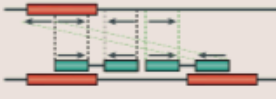

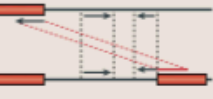
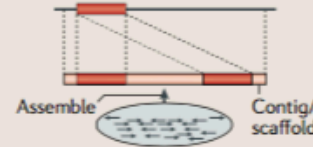



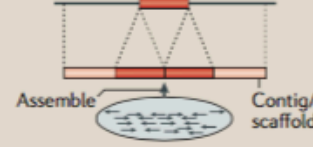
Any mutation >50bp

**Profound impact on
genome structure
and function**



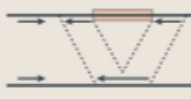
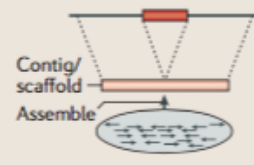
Genome structural variation discovery and genotyping

Alkan, C, Coe, BP, Eichler, EE (2011) *Nature Reviews Genetics*. May;12(5):363-76. doi: 10.1038/nrg2958.

Structural Variation Sequence Signatures

SV classes	Read pair	Read depth	Split read	Assembly
Deletion				
Novel sequence insertion		Not applicable		
Mobile-element insertion		Not applicable		
Inversion		Not applicable		
Interspersed duplication				
Tandem duplication				

Structural Variation Sequence Signatures

SV classes	Read pair	Read depth	Split read	Assembly
Deletion				



PacBio Sequel

>10kbp Mean Read Lengths
~\$15k / Mammalian-sized genome

Single Molecule Sequencing

- No amplification artifacts
- More uniform coverage
- Essentially no GC biases

Long read lengths

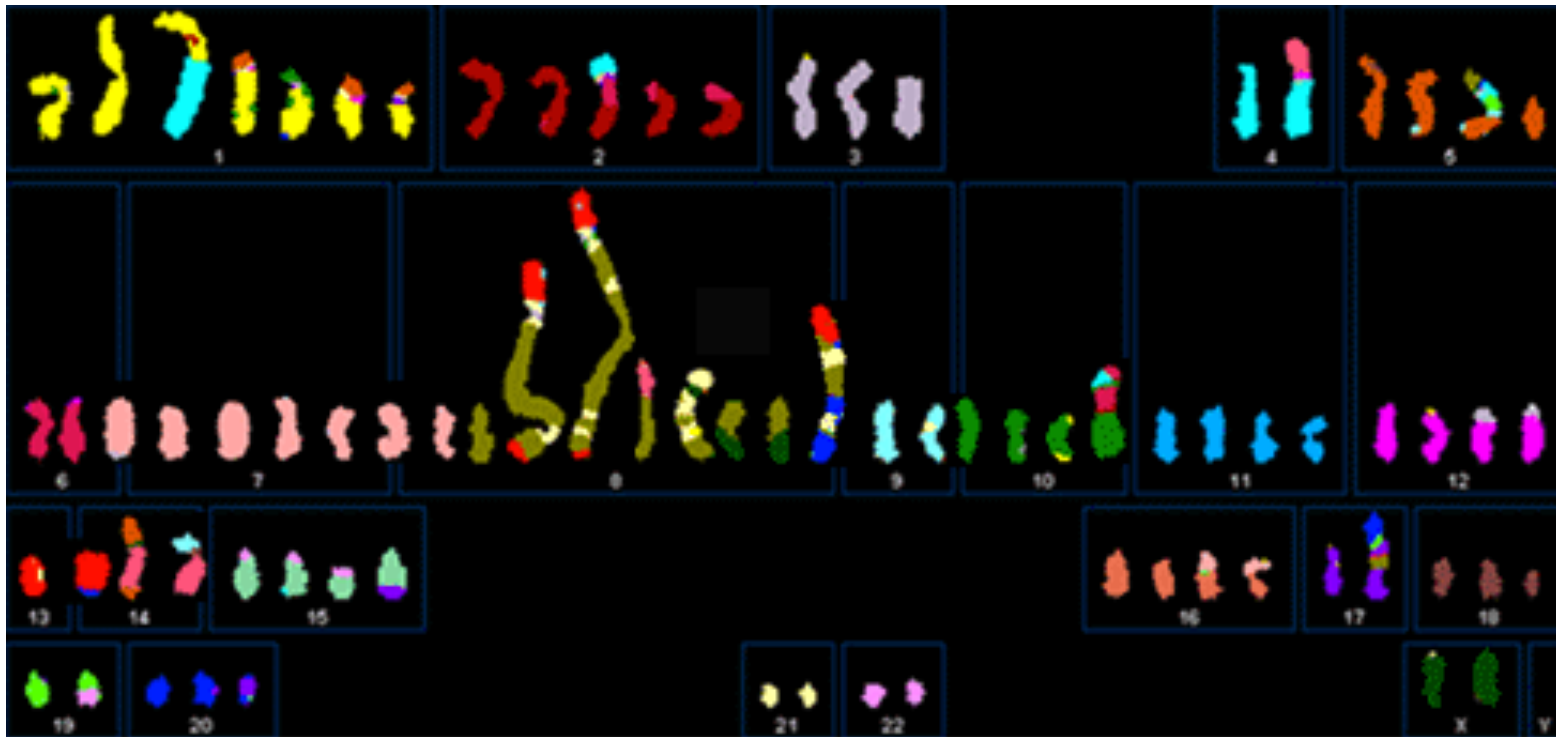
- Improved mappability
- More likely to span breakpoints
- More robust split read analysis
- More robust assemblies

**Basepair resolution for
50bp through 50Mbp events**

SK-BR-3



Most commonly used Her2-amplified breast cancer cell line



(Davidson et al, 2000)

Highly-rearranged Mammalian genome

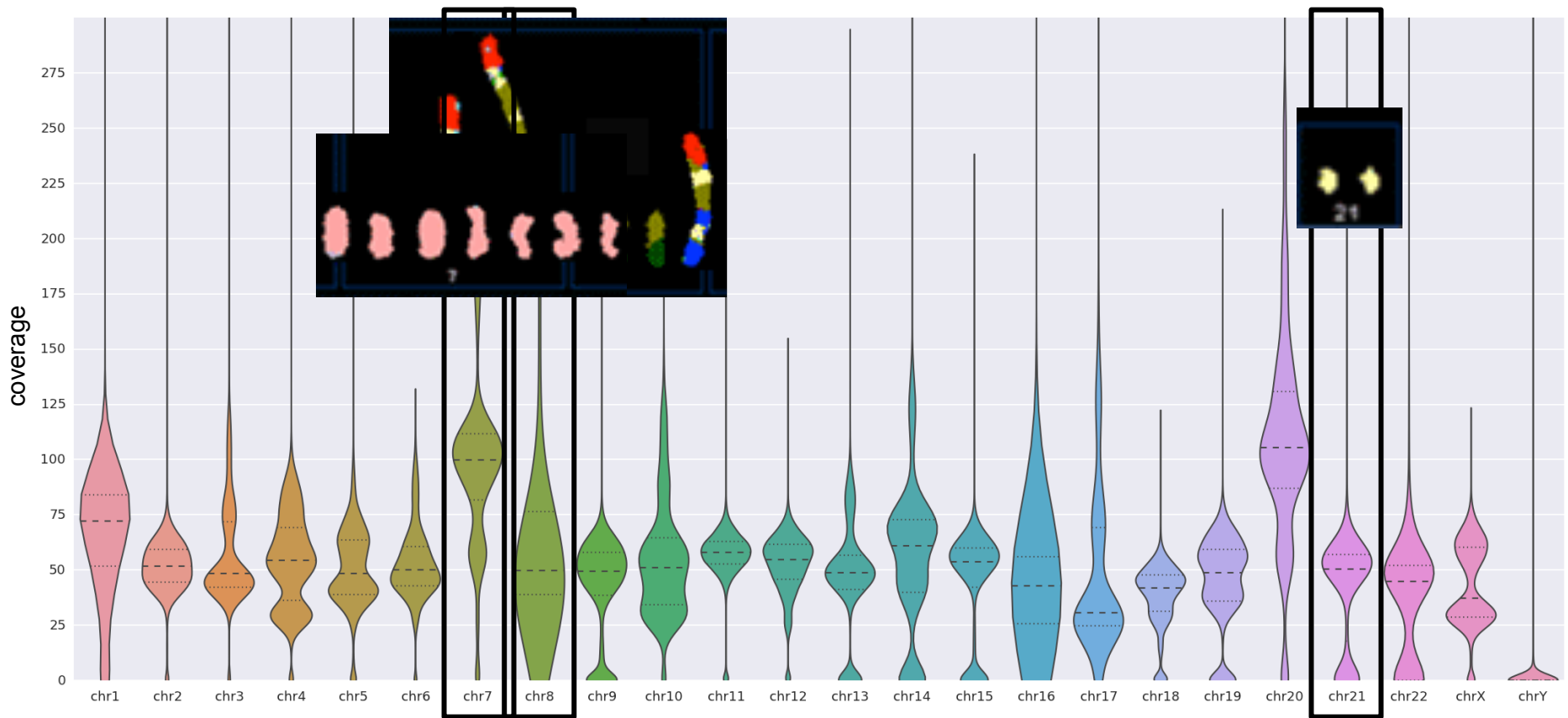
80 chromosomes instead of 46

Numerous chromosome fusions, rearrangements, other SVs

PacBio Long-Read Sequencing

mean read length: 9 kb
max read length: 71 kb

72X overall coverage

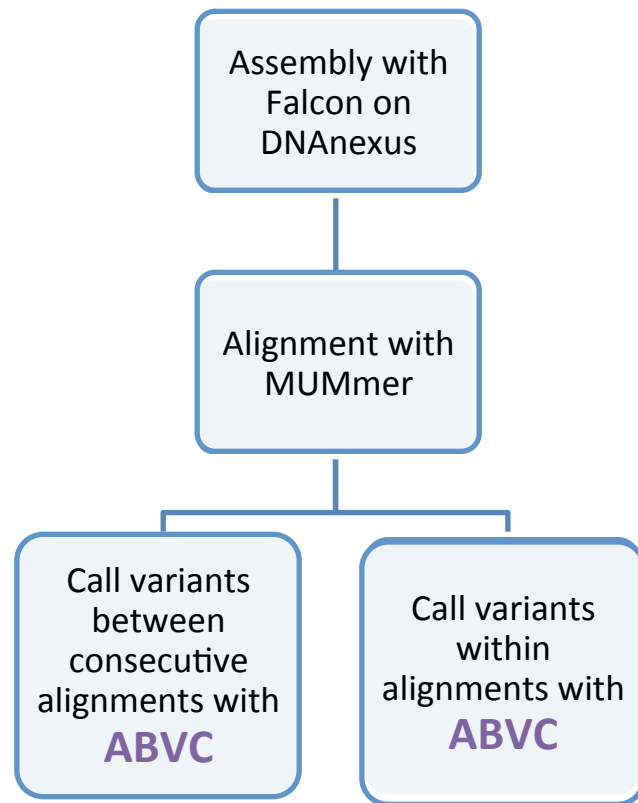


Genome-wide coverage averages around 54X

Coverage per chromosome varies greatly as expected from previous karyotyping results

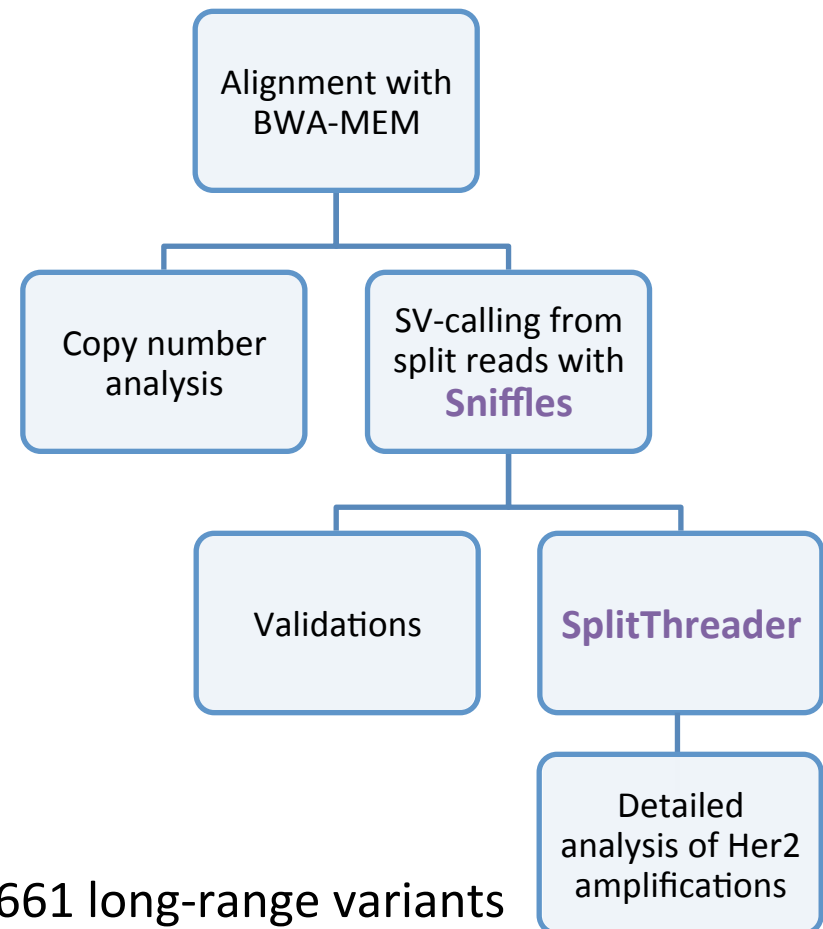
Genome structural analysis

Assembly-based



~ 11,000 local variants
50 bp < size < 10 kbp

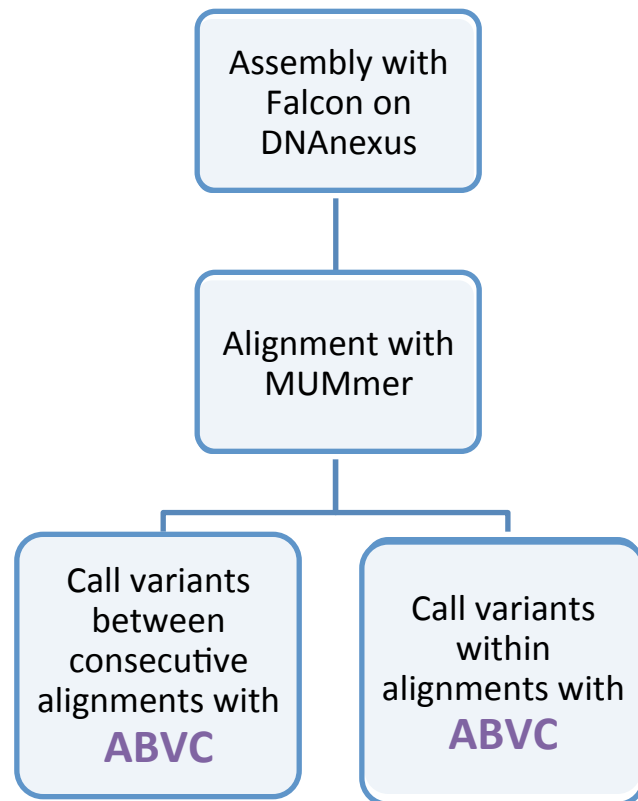
Alignment-based



661 long-range variants
(>10kb distance)

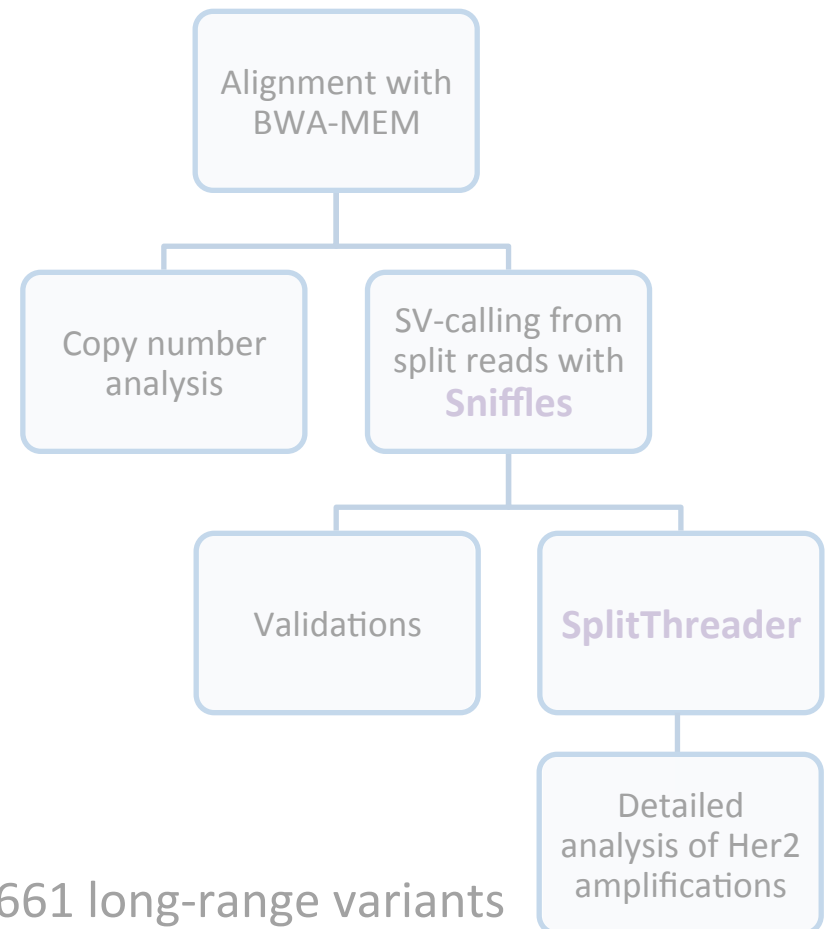
Genome structural analysis

Assembly-based



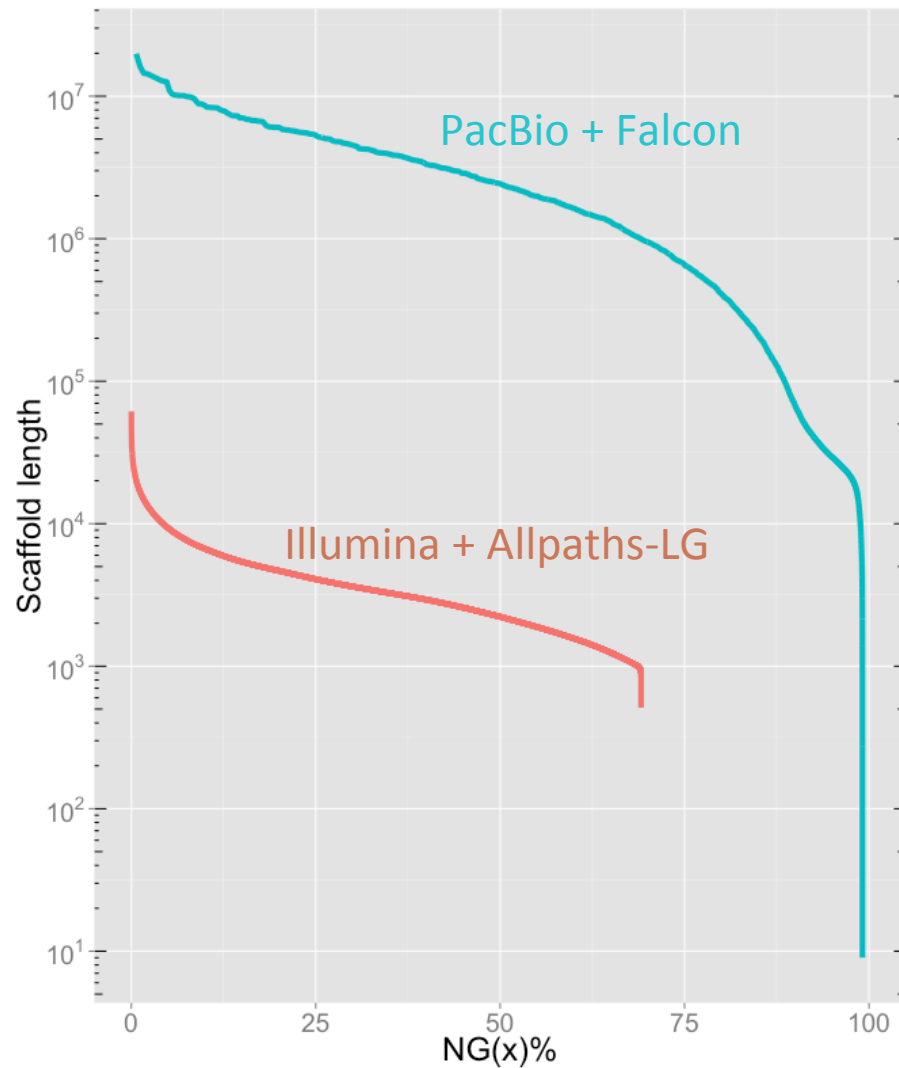
~ 11,000 local variants
50 bp < size < 10 kbp

Alignment-based



661 long-range variants
(>10kb distance)

Assembly using PacBio yields far better contiguity

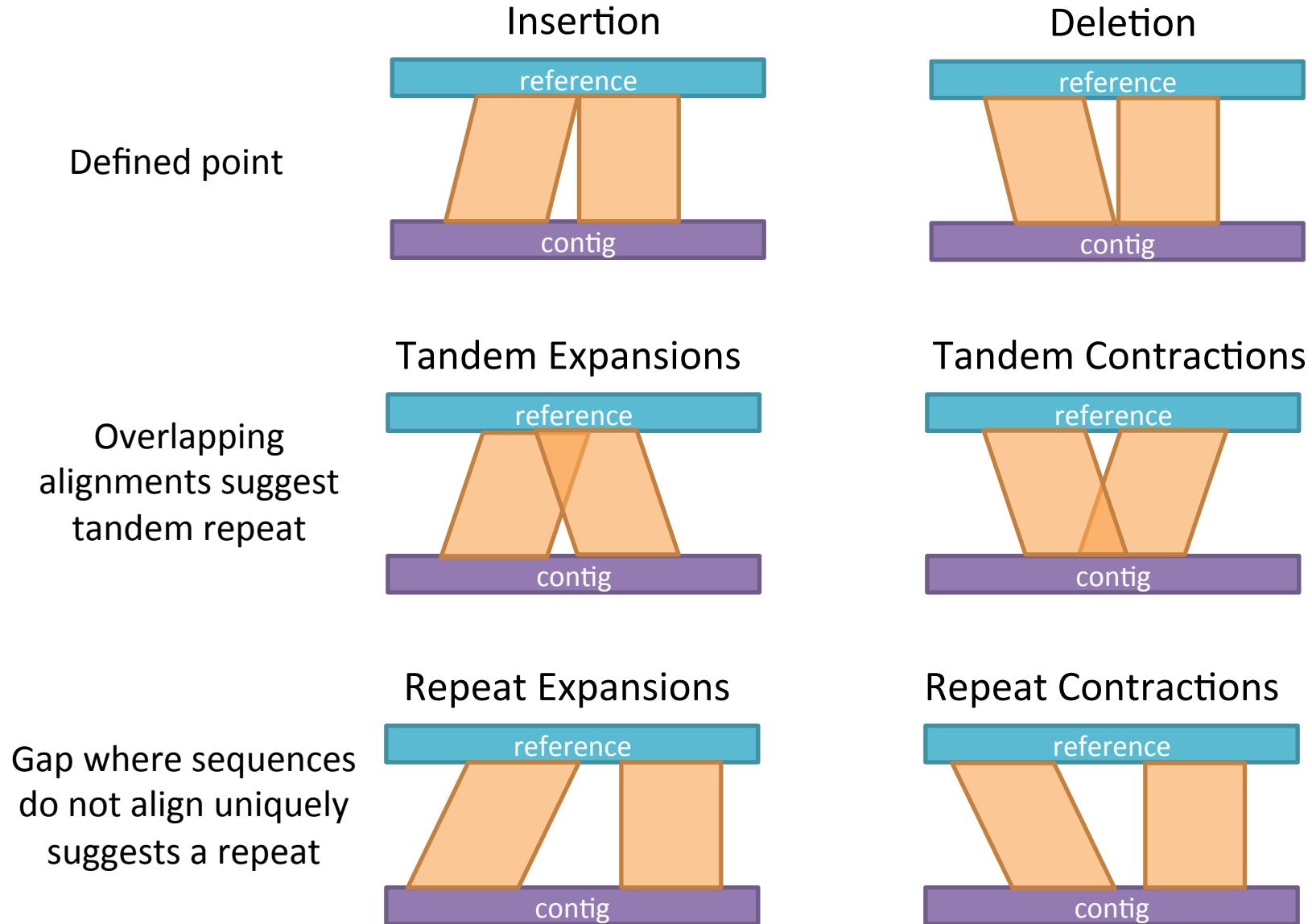


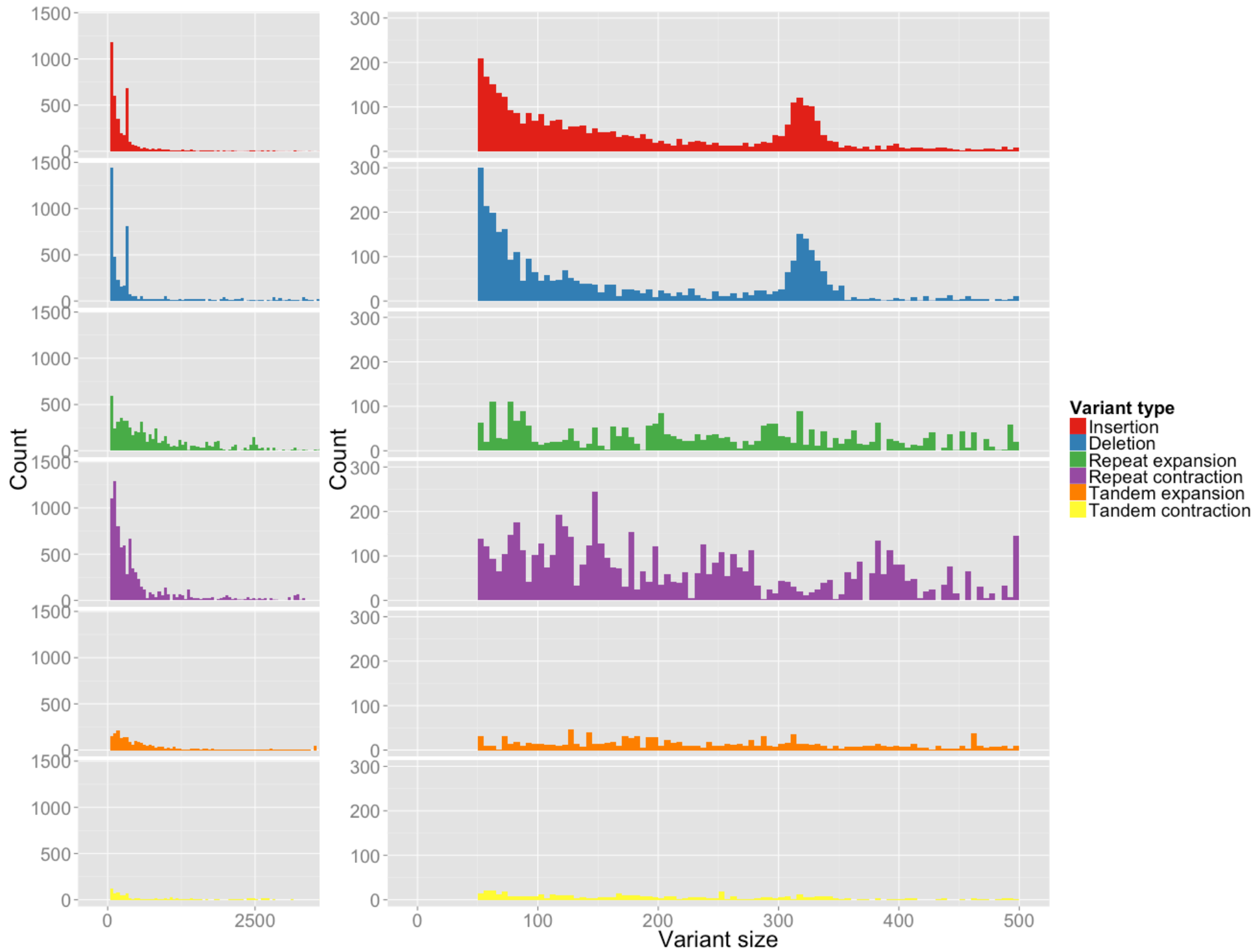
Number of sequences: 13,532
Total sequence length: 2.97Gb
Mean: 266 kb
Max: 19.9 Mb
N50: 2.46 Mb

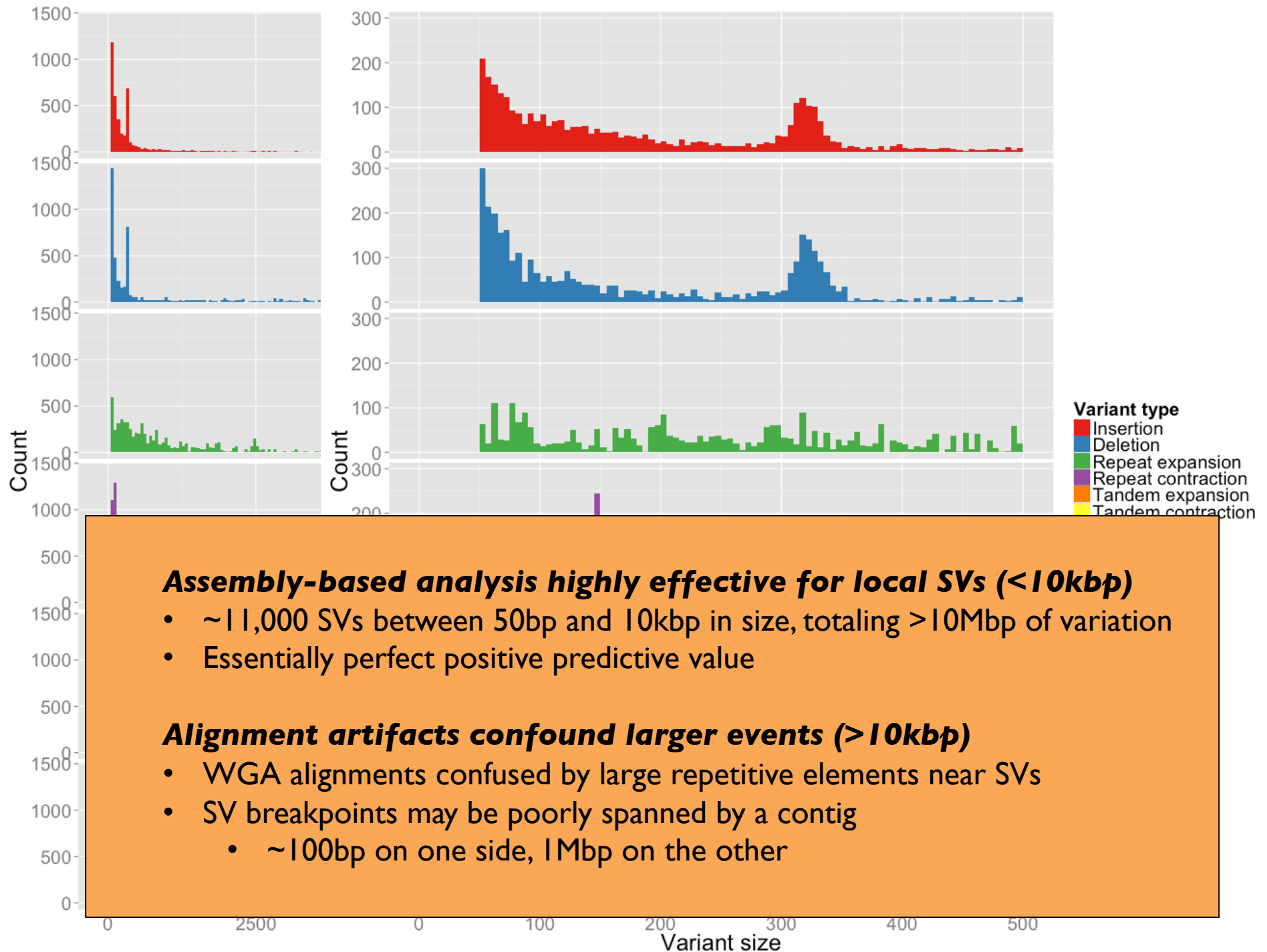
Relative to a genome size of 3 Gb

Number of sequences: 748,955
Total sequence length: 2.07 Gb
Mean: 2.8 kb
Max: 61 kb
N50: 3.3 kb

ABVC: Assembly-Based Variant-Caller

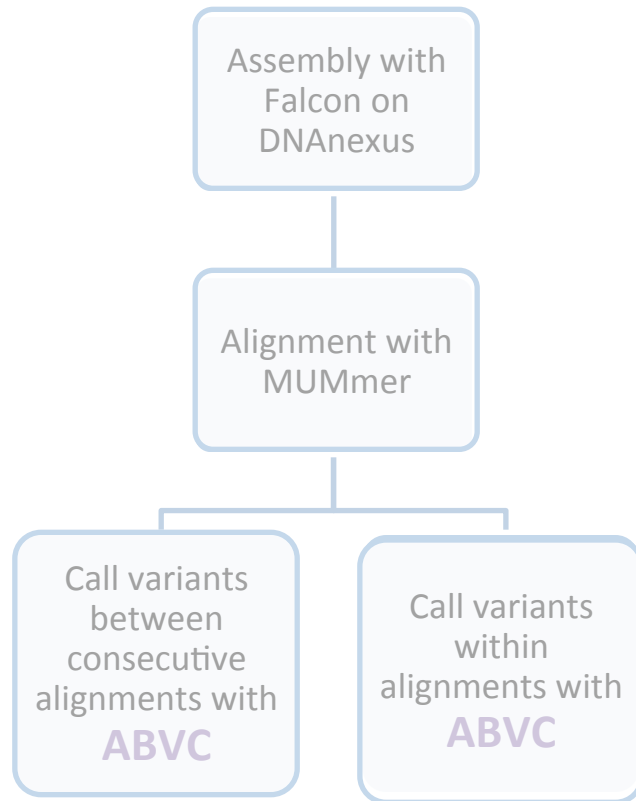






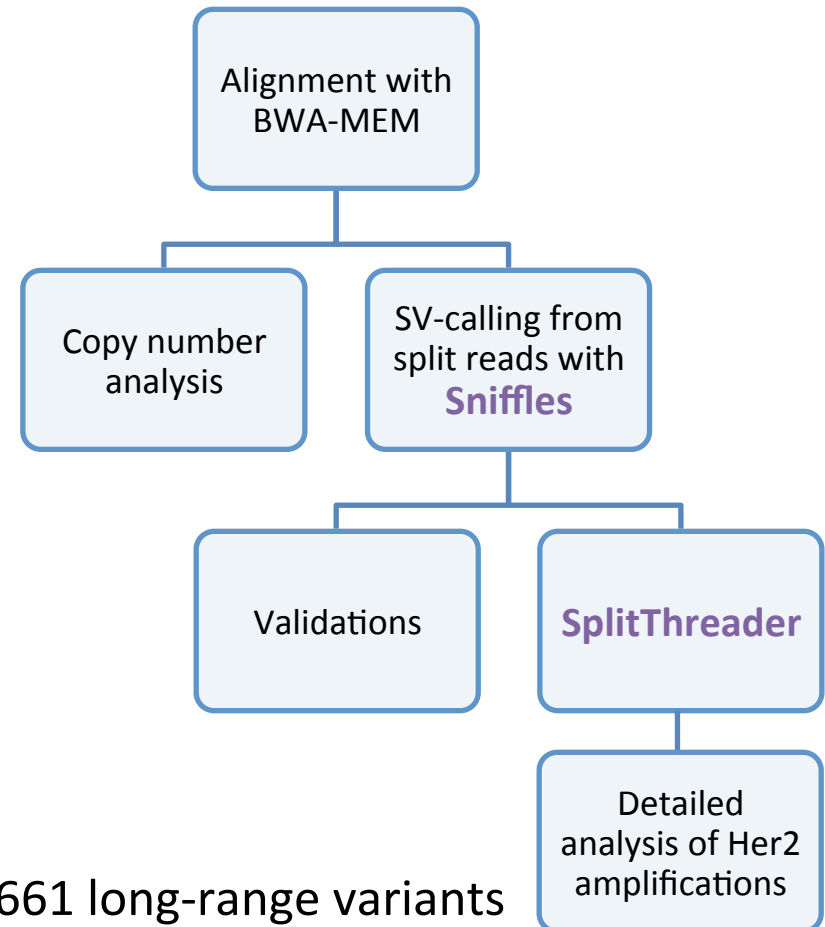
Genome structural analysis

Assembly-based



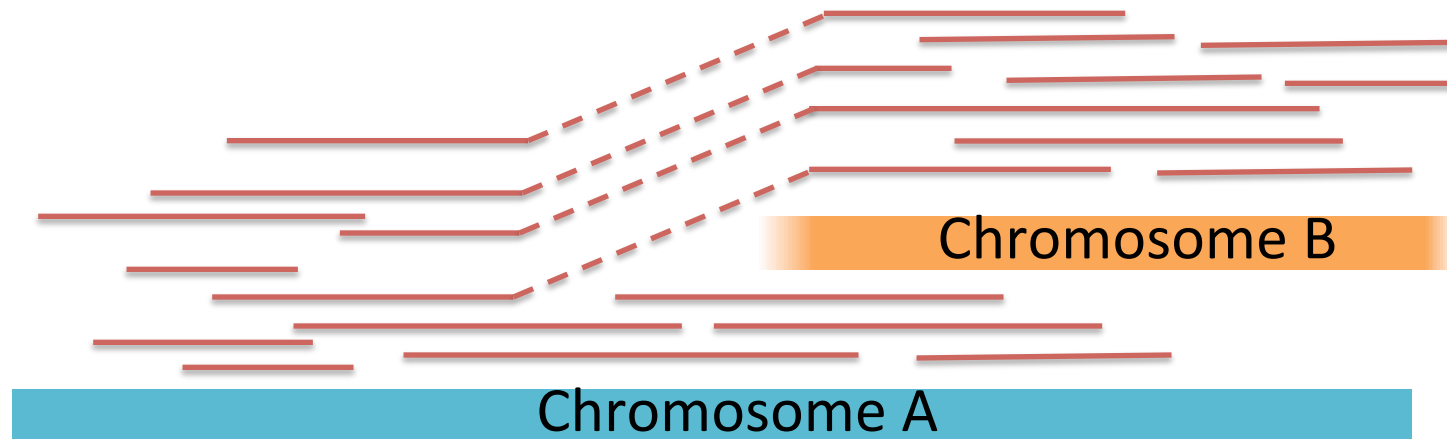
~ 11,000 local variants
50 bp < size < 10 kbp

Alignment-based



661 long-range variants
(>10kb distance)

Long Read Structural Variation Analysis



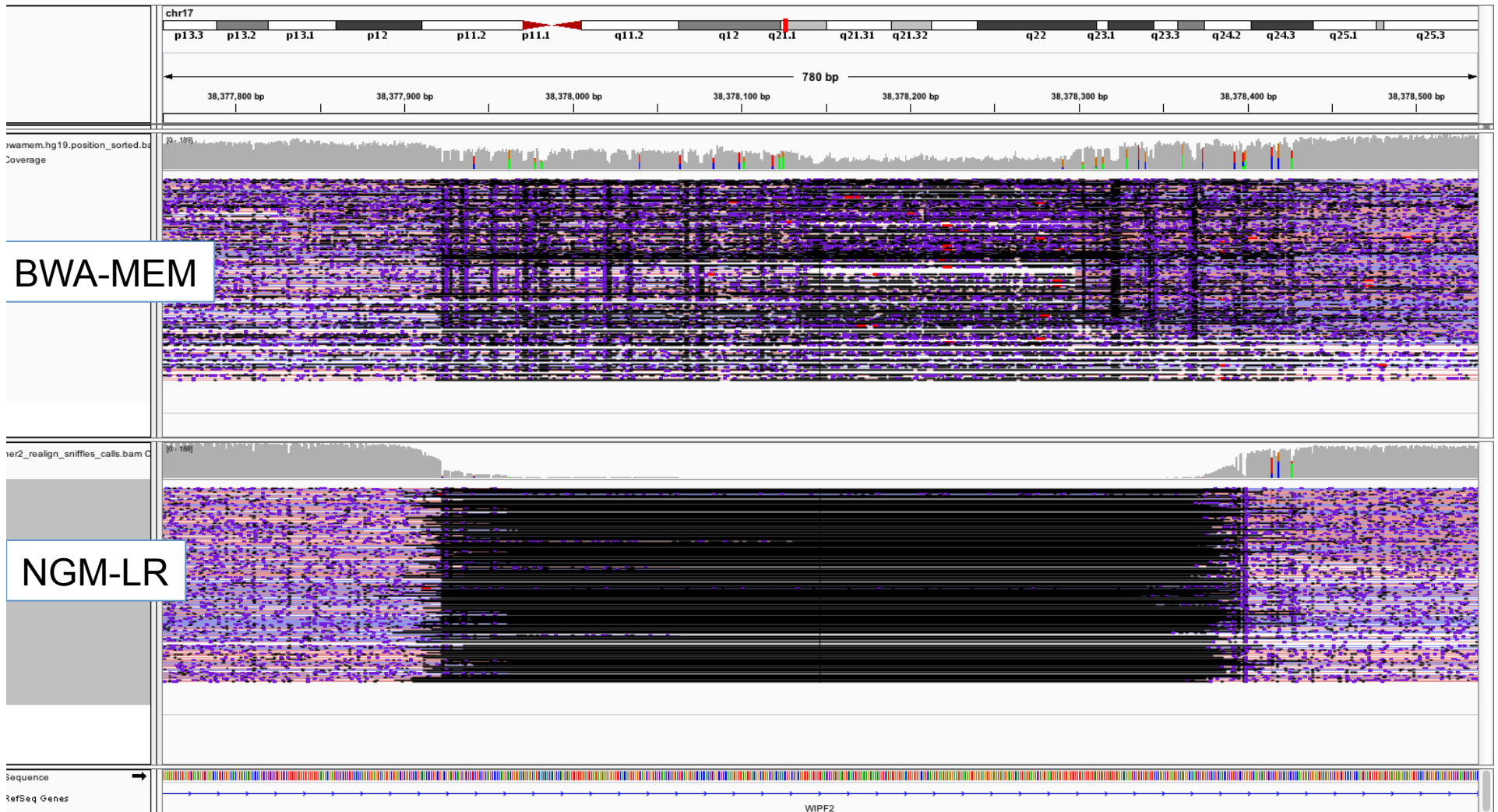
Split-read analysis greatly improved by long reads

- Improved chances of spanning event, including nested 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

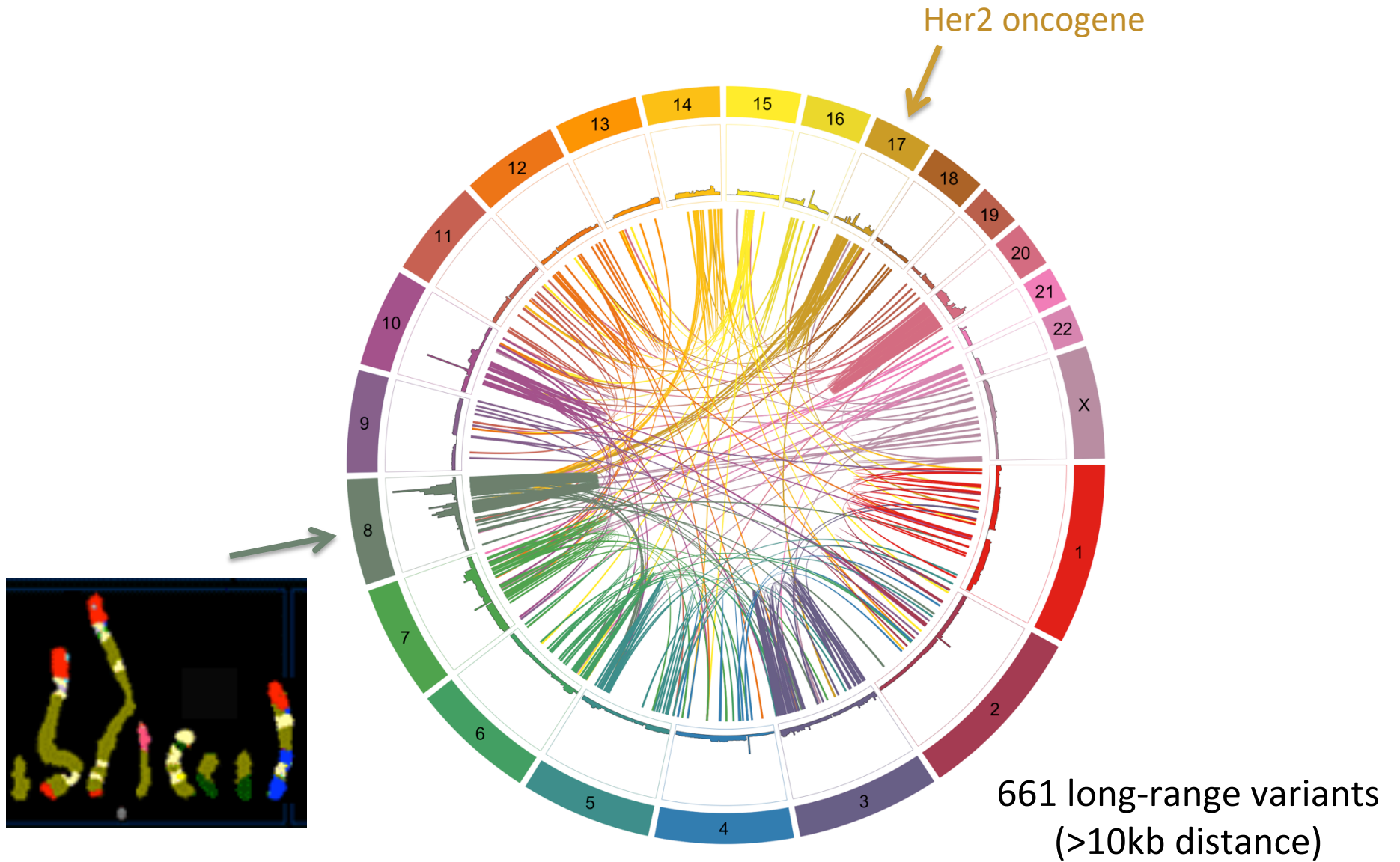
- **NGM-LR**: Improve mapping of noisy long reads
- **Sniffles**: Integrates SV evidence from split alignments, alignment fidelity (CIGAR string and MD tag)

Mapping a ~500bp deletion

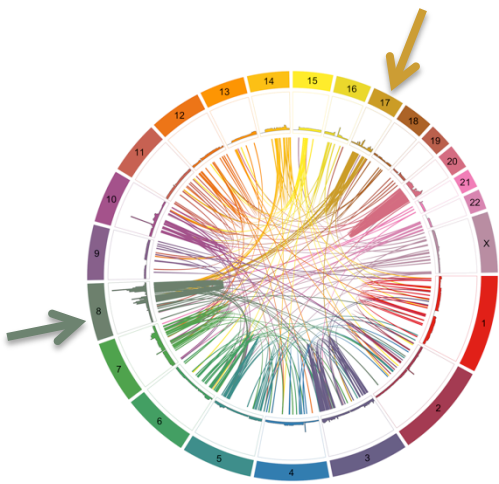


Similar issues for insertions, inversions; or Nanopore sequencing
Improved seeding, improved gap scoring: convex instead of affine

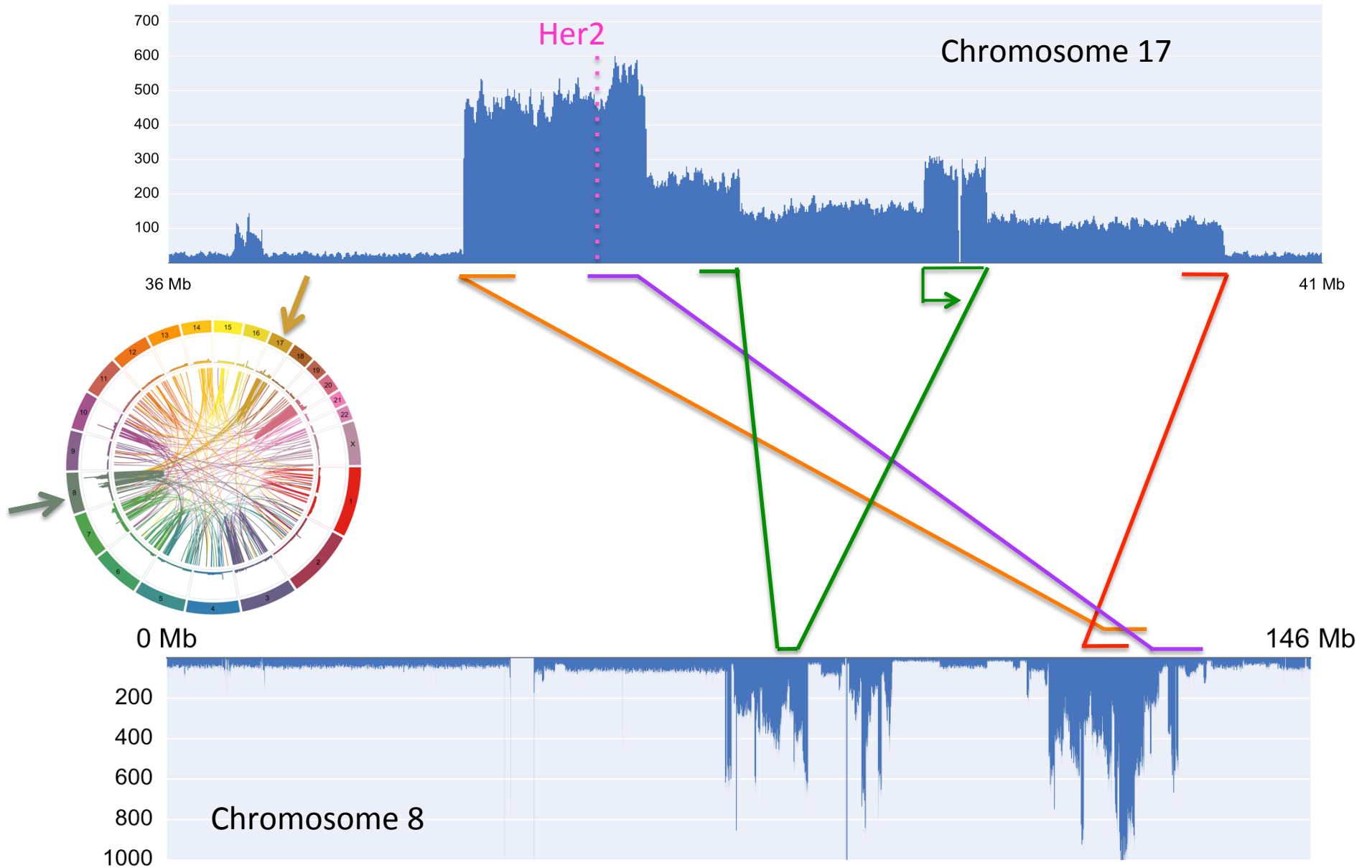
Long-range structural variants found by Sniffles



Long-range structural variants found by Sniffles

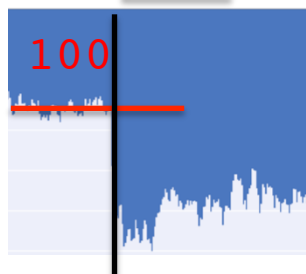
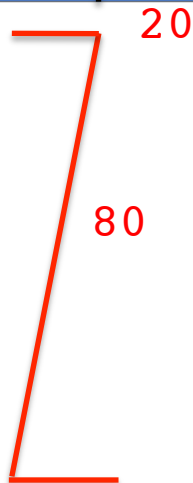
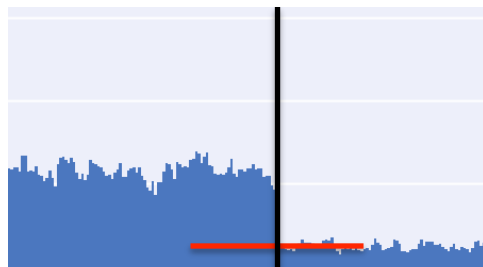


Long-range structural variants found by Sniffles



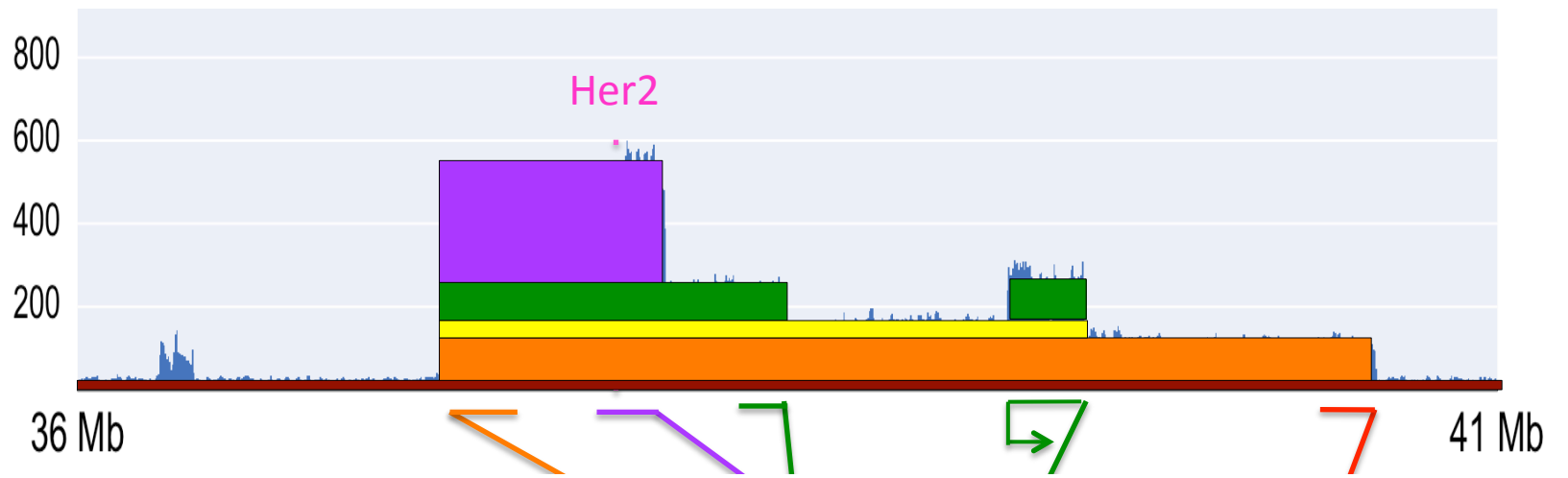
SplitThreader

Threading SV breakpoints to
Infer the history of rearrangements in complex genomes



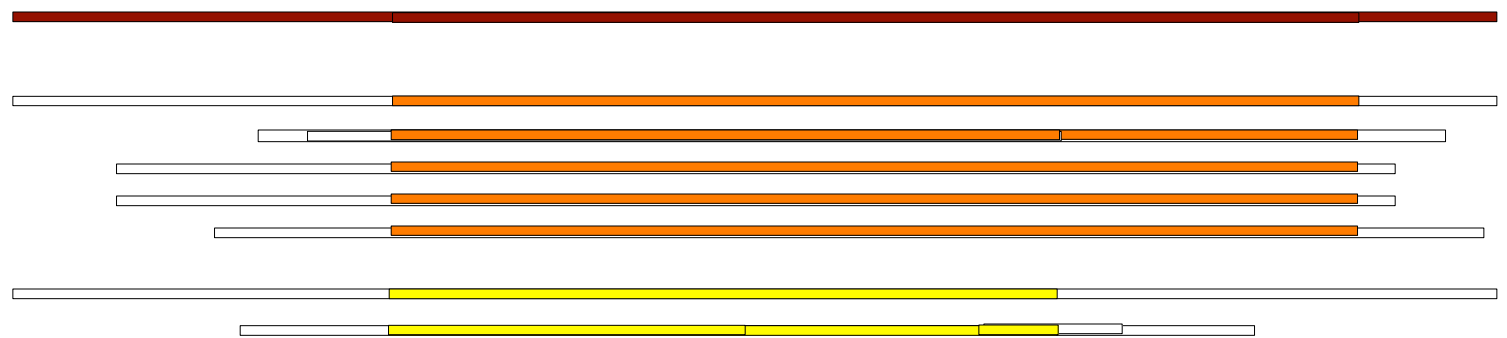
80

100

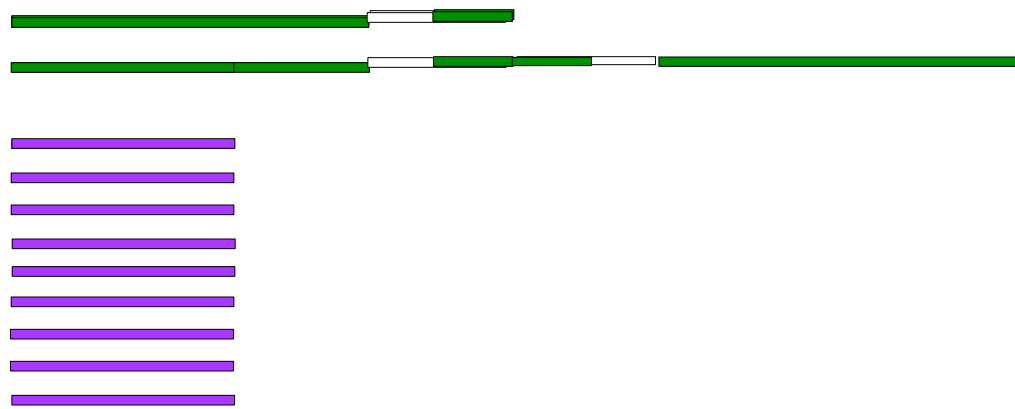


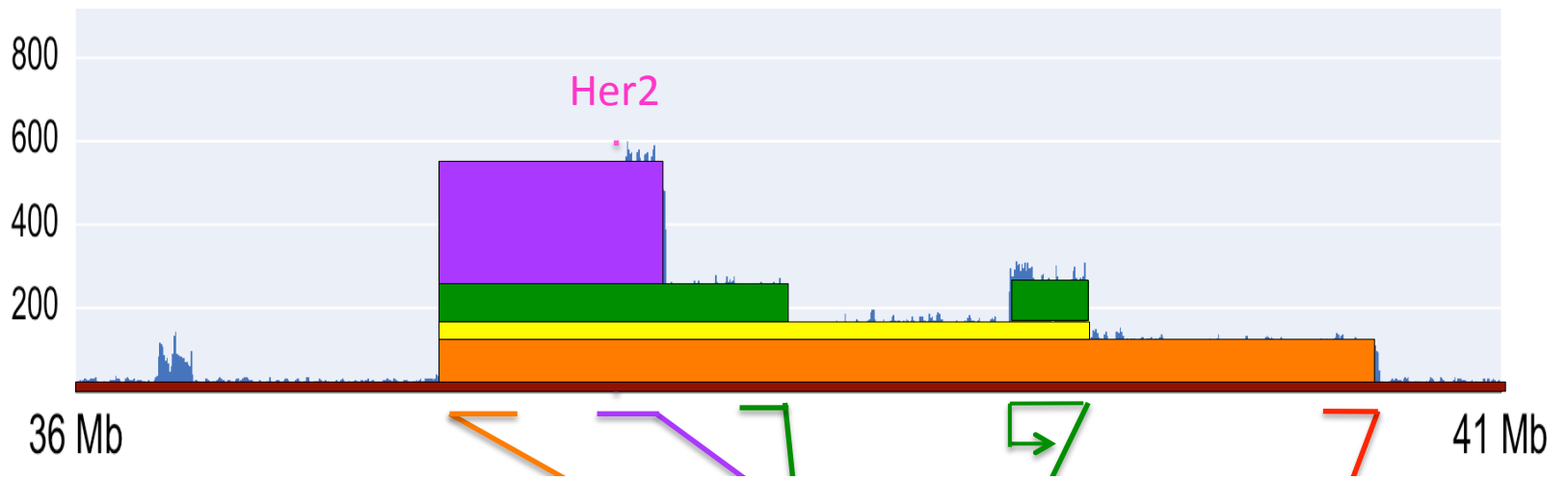
Chr 17

Chr 8



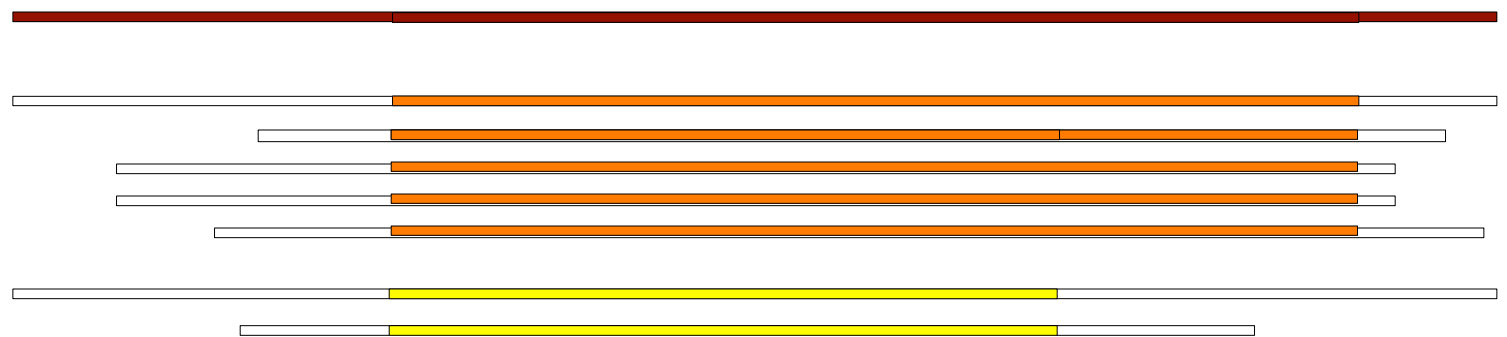
1. Healthy chromosome 17 & 8
2. Translocation into chromosome 8
3. Translocation within chromosome 8
4. Complex variant and inverted duplication within chromosome 8
5. Translocation within chromosome 8



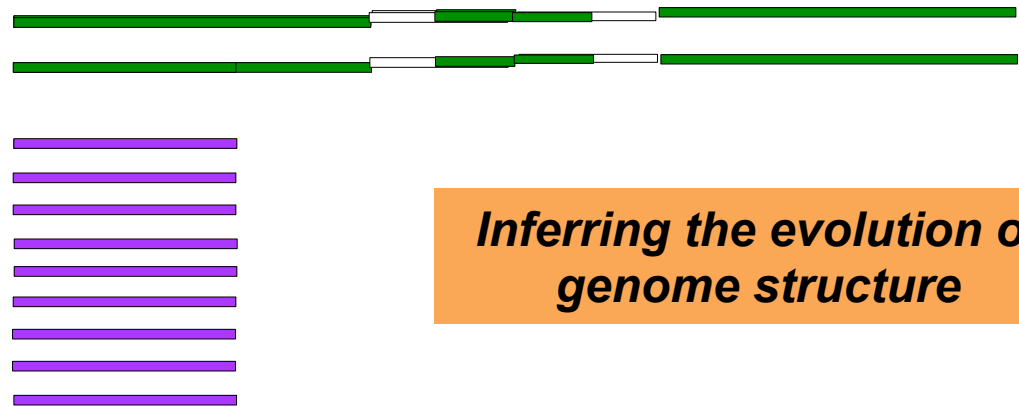


Chr 17

Chr 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

Summary & Acknowledgements



ABVC + SplitThreader by Maria Nattestad

- Assembly-based variant analysis is efficient and accurate
 - 10s of thousands variants present in mammalian-sized genomes
- SplitThreader infers the evolution to genome structure
 - Additional context as genes are moved next to new promoters and other regulatory elements



NGM-LR + Sniffles by Fritz Sedazeck

- Correct long read mapping is essential for SV analysis
 - Design the mapping strategy for the error model of the data
- Integrate all available information for robust SV calling
 - Currently extending to other long-range mapping technologies: Oxford Nanopore, BioNano, 10X Genomics



***Special thanks to Dick McCombie (CSHL), John McPherson (OICR), PacBio
Funding by NSF, NIH, DOE, Sloan Foundation***



Your new office?

<http://schatzlab.cshl.edu/apply/>

Thank you!

The Resurgence of Reference Quality Genomes

Michael Schatz & Daniel Rokhsar

Tuesday January 12, 2016 @ 4pm – 6pm

Town & Country - Pacific Salon I