Comparative genomics

Bioinformatics Applications (PLPTH813)

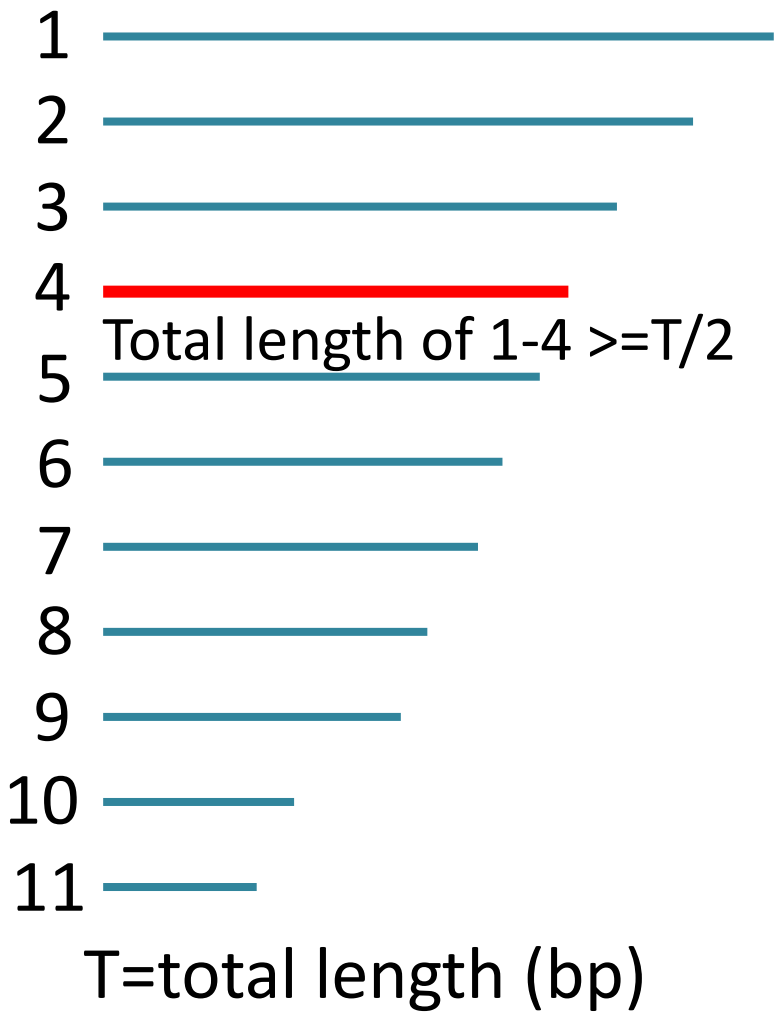
Sanzhen Liu

4/11/2019

Assembly statistics – N50

- N50: A statistic used for assessing the contiguity of a genome assembly.

- The contigs in an assembly are **sorted** by size and added, starting with the largest. The contig N50 is **the size of the contig that makes the total greater than or equal to 50% of the total contig size.**



Long-read assemblies

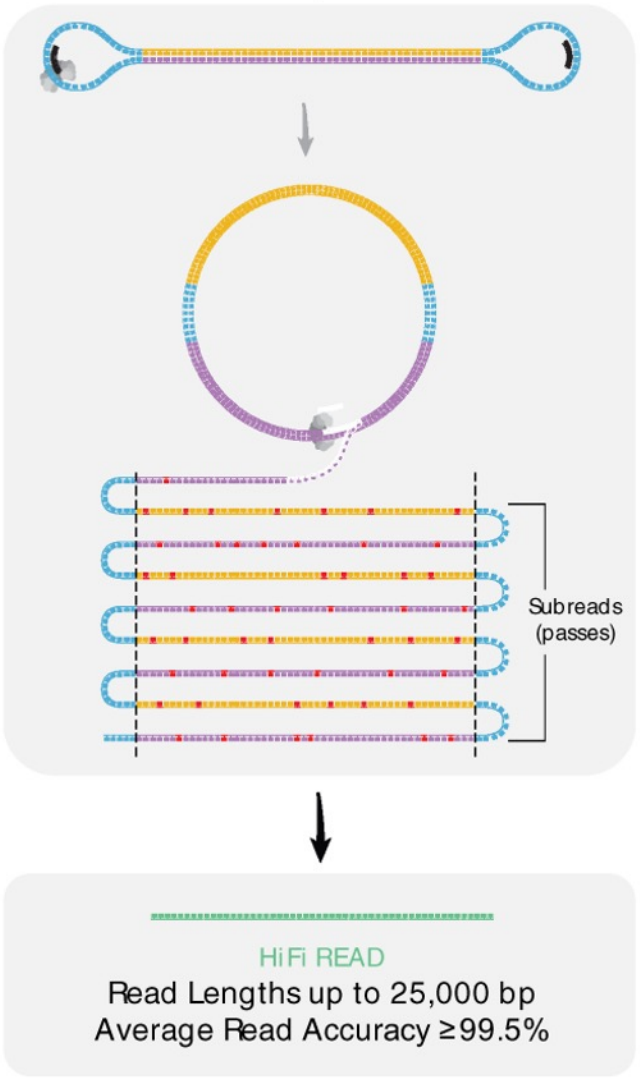
Software	Description
Canu	K-mer-based overlap computation
FALCON	Assembles phased diploid genomes
Flye	Uses A-Bruijn graph

Contig polishing

Software	Description
Arrow	Assembly error correction with PacBio long-read alignments
Nanopolish	Assembly error correction using voltage (raw) Nanopore data

Pilon	Assembly error correction using Illumina data

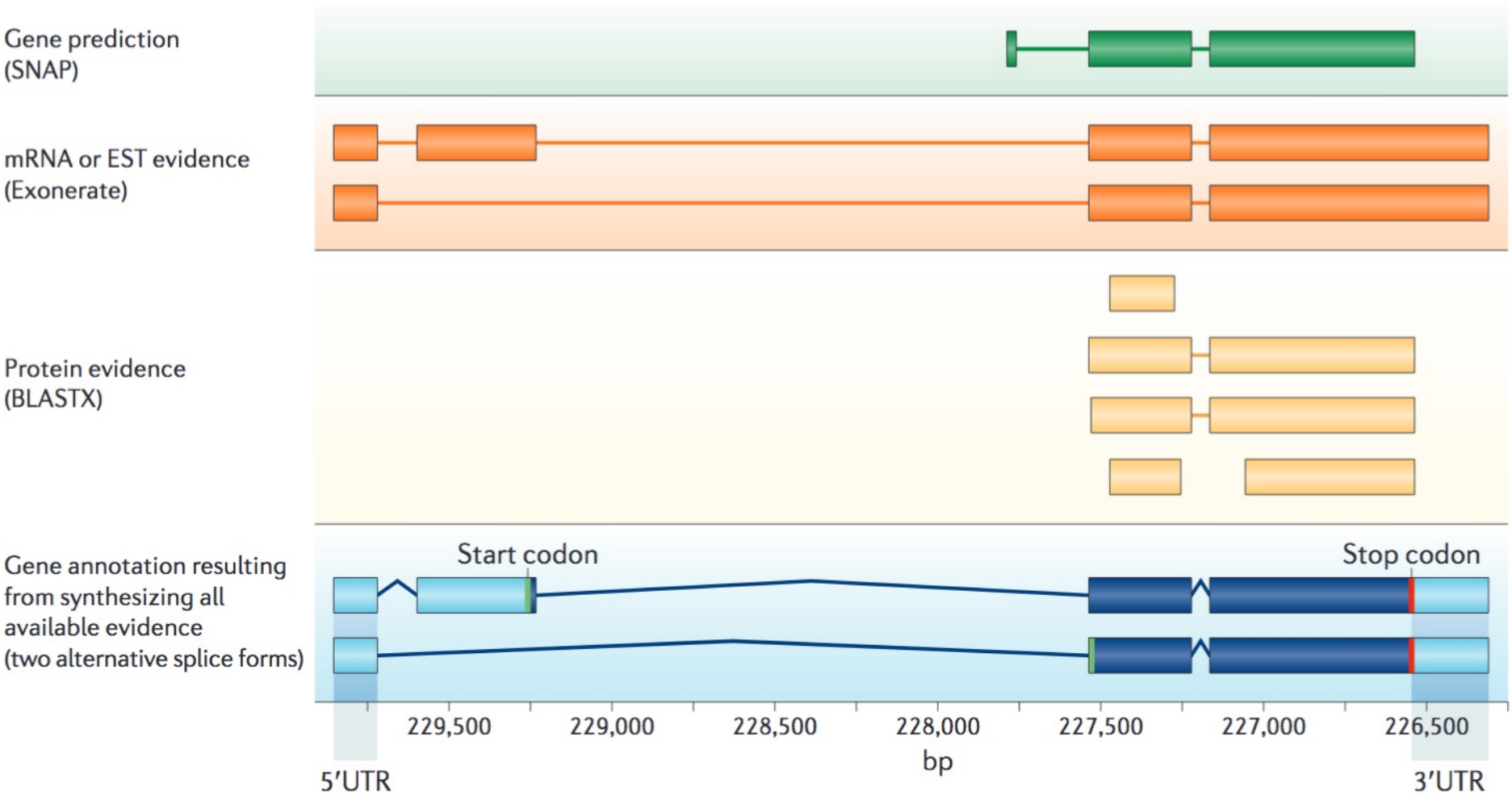
Assembly meets long HiFi PacBio reads



hifiasm:
High-contiguity assembly
fast

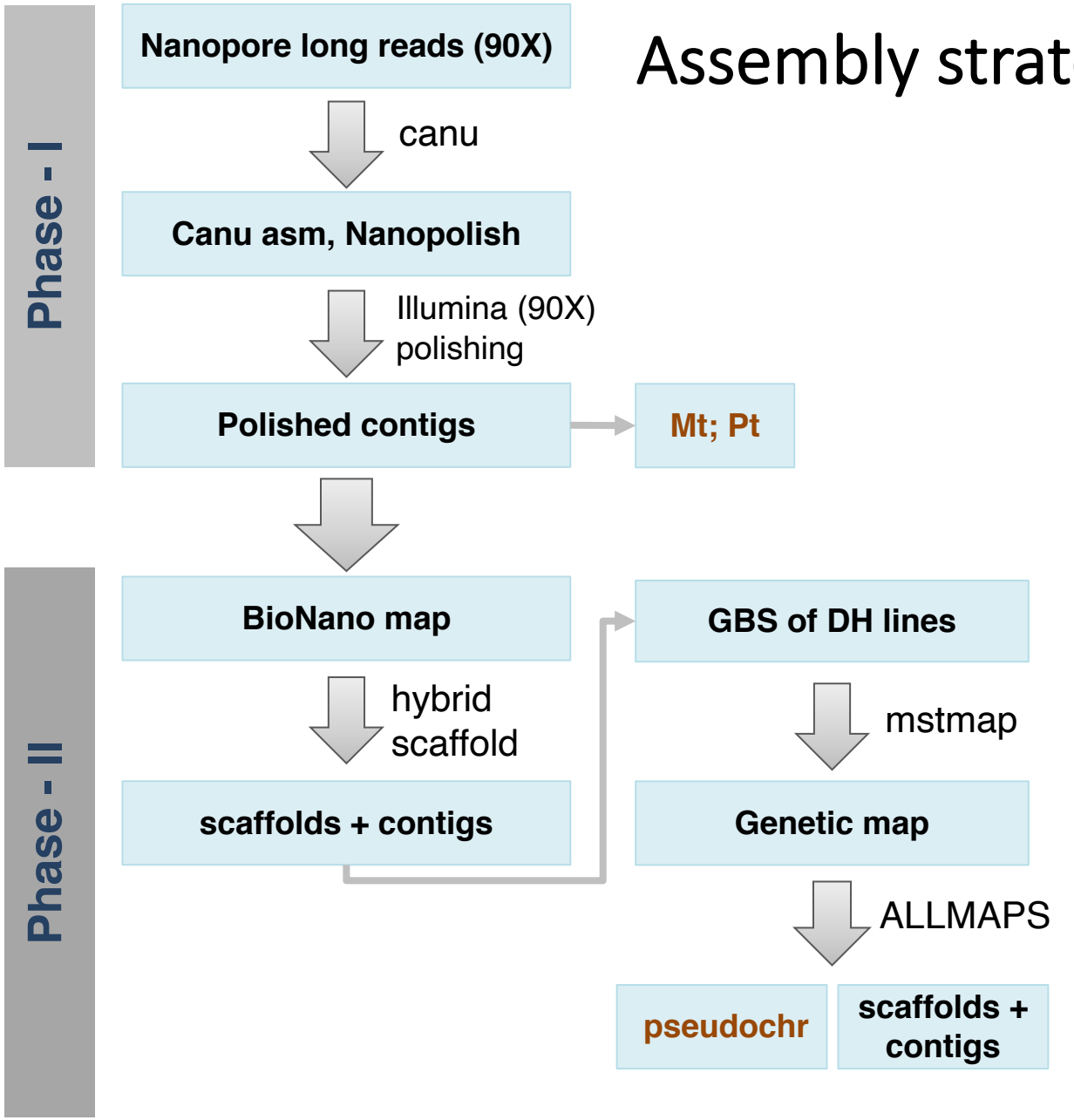
https://github.com/chhylp123/hifiasm

Gene annotation: *ab initio* prediction + evidence



Nature Reviews Genetics 13:329–342 (2012)

Assembly strategy



Outline

- Introduction of comparative genomics and structural variation
- Approaches
  1. Comparative genome hybridization
  2. Paired-end reads
  3. Read depth
  4. Whole genome assembly
- Pangenomics
- Case study: CGRD

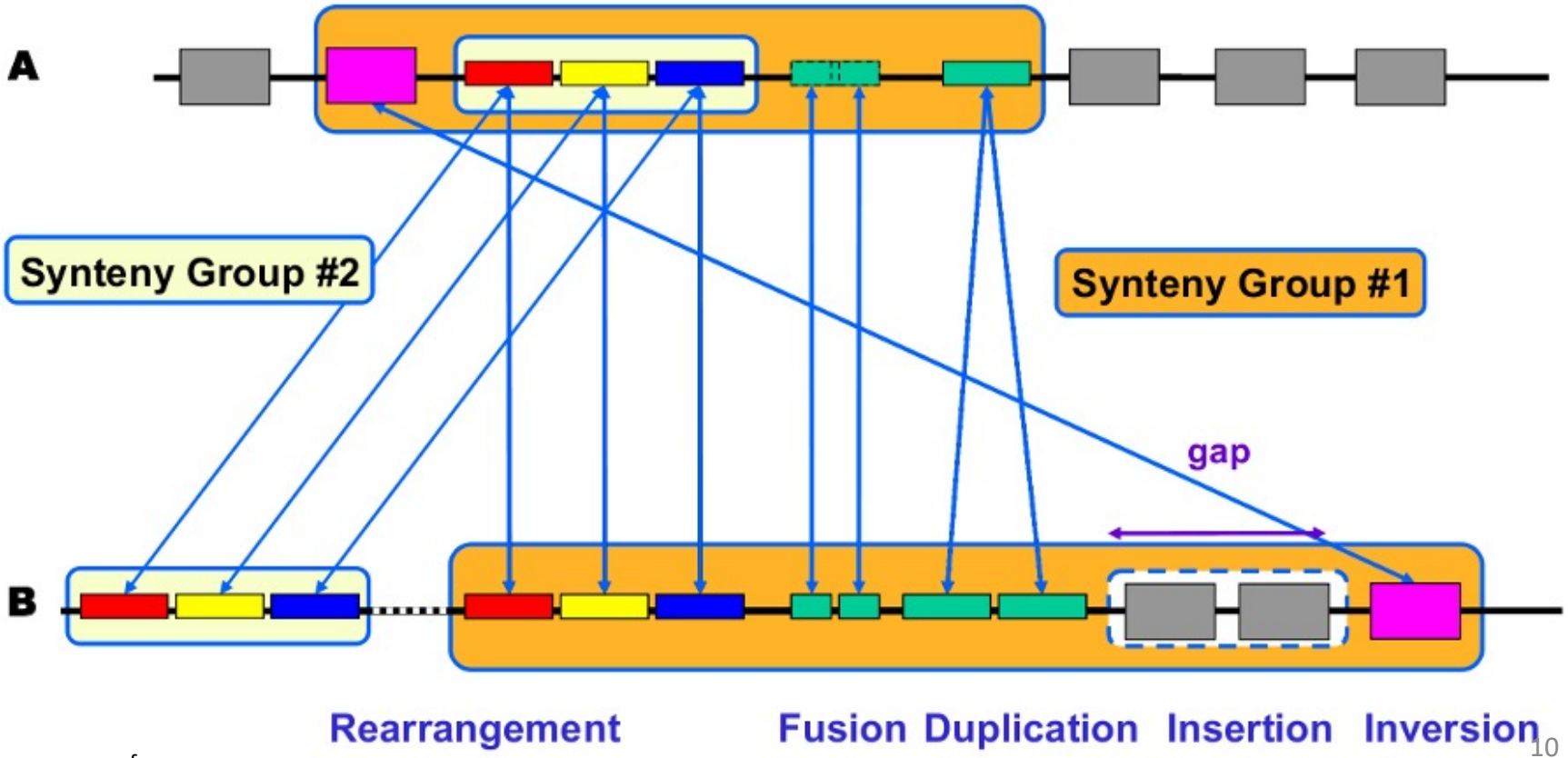
Comparative genomics among genomes

Conservation and variability

- Homologs: Genomic regions derived from a common ancestral gene.
- Orthologs: Homologs from the divergence of lineages.
- Paralogs: Homologs derived from their duplication within a lineage.
- Homeologs: The subset of paralogs created by WGD. (synonyms: ohnolog; syntenic paralog)

Synteny

Synteny is usually referred to as the conservation of blocks of order within two sets of chromosomes that are being compared with each other. Syntenic regions are evidence by homologous genes arranged in a collinear order.



Intra-species genome rearrangements and structural variation (SV)

Variation of *hundreds of bp* to several Mb in size

Balanced variation

- Inversion

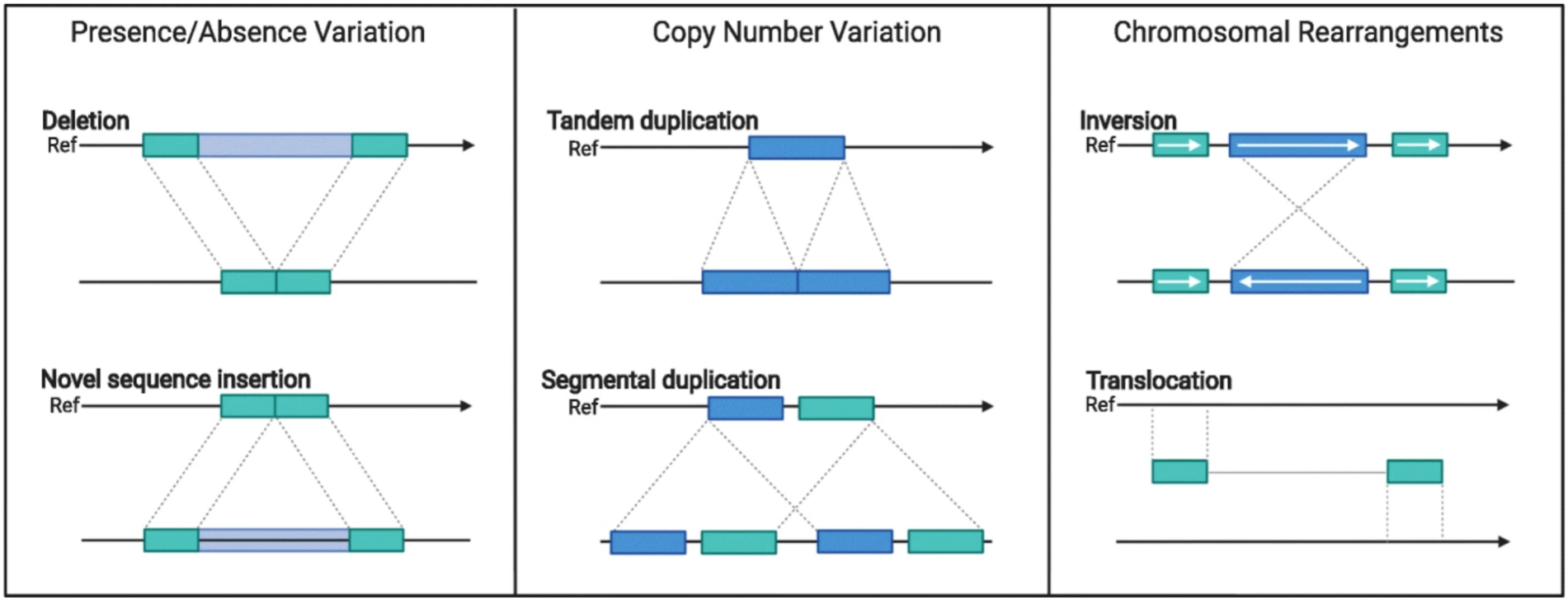
- Translocation

Unbalanced variation:

- Copy number variation (CNV)

(Presence/Absence variation, PAV)

Structure variation types within species



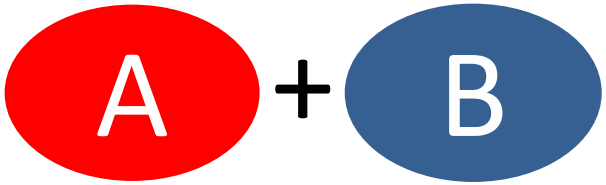
Coletta et al., Genome Biol, 2021

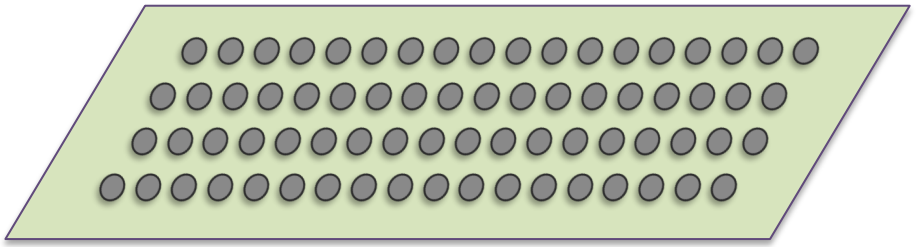
Outline

- Introduction of comparative genomics and structural variation
- Approaches
  1. Comparative genome hybridization
  2. Paired-end reads
  3. Read depth
  4. Whole genome assembly
- Pangenomics
- Case study: CGRD

array Comparative Genomic Hybridization (aCGH)

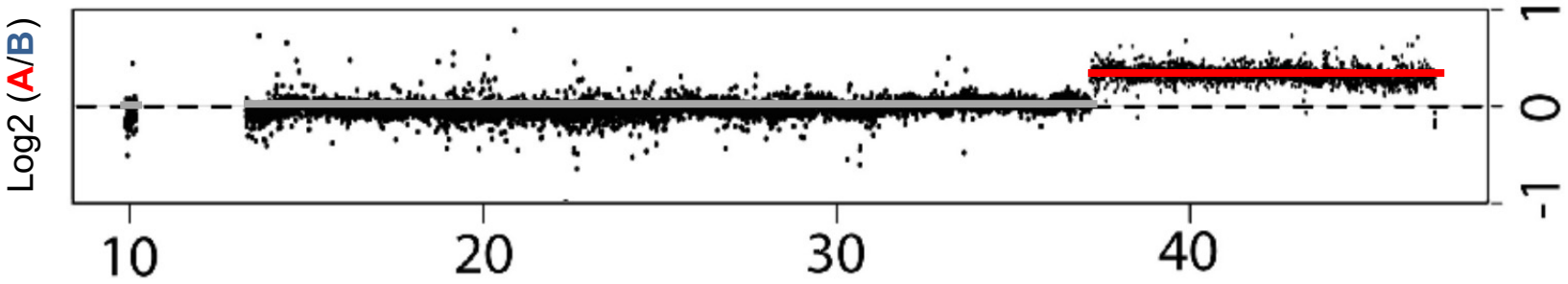
array





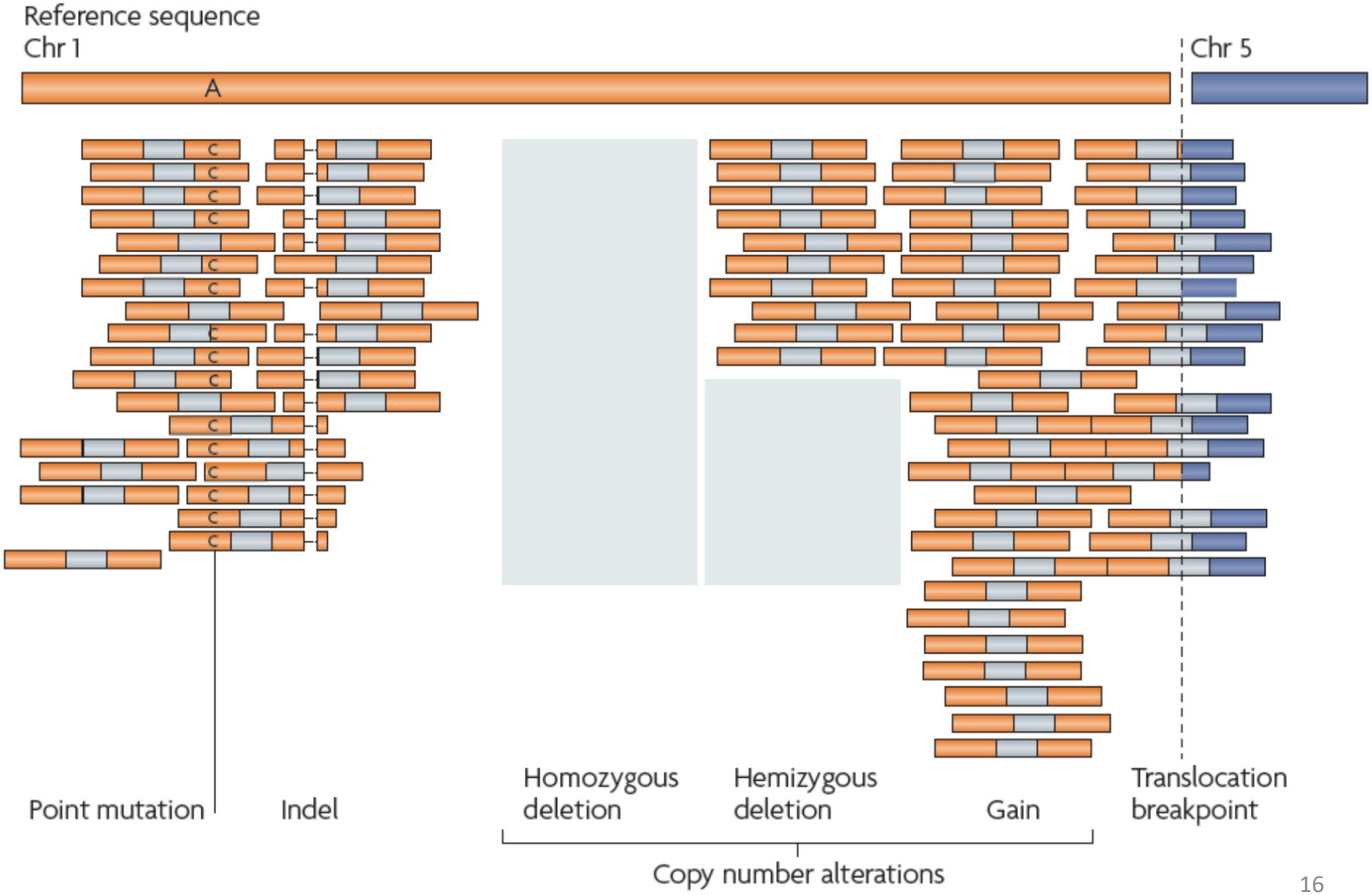
Red signal (A)

Blue signal (B)

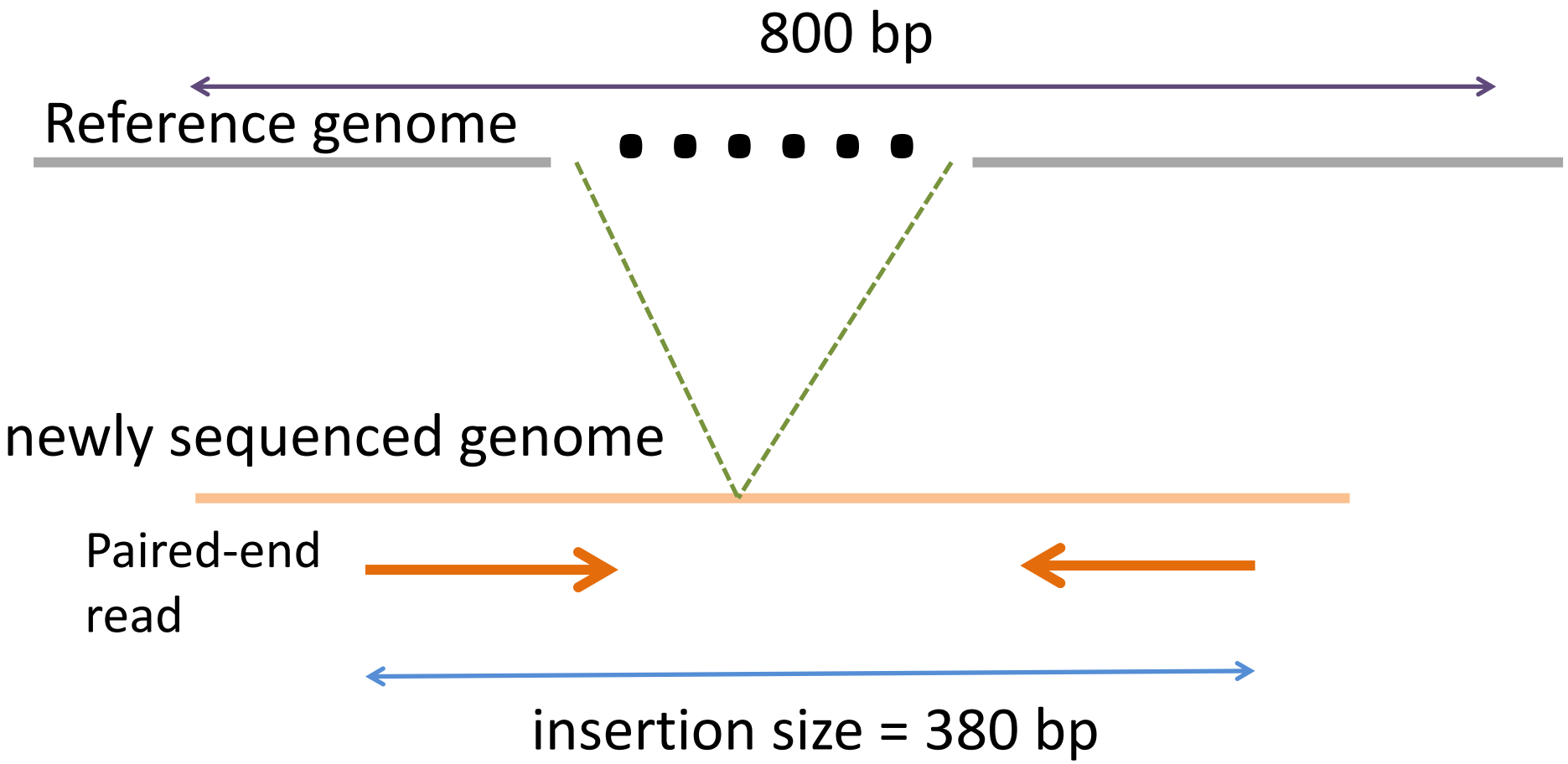


NGS provides information for the discovery of variants, including structural variation

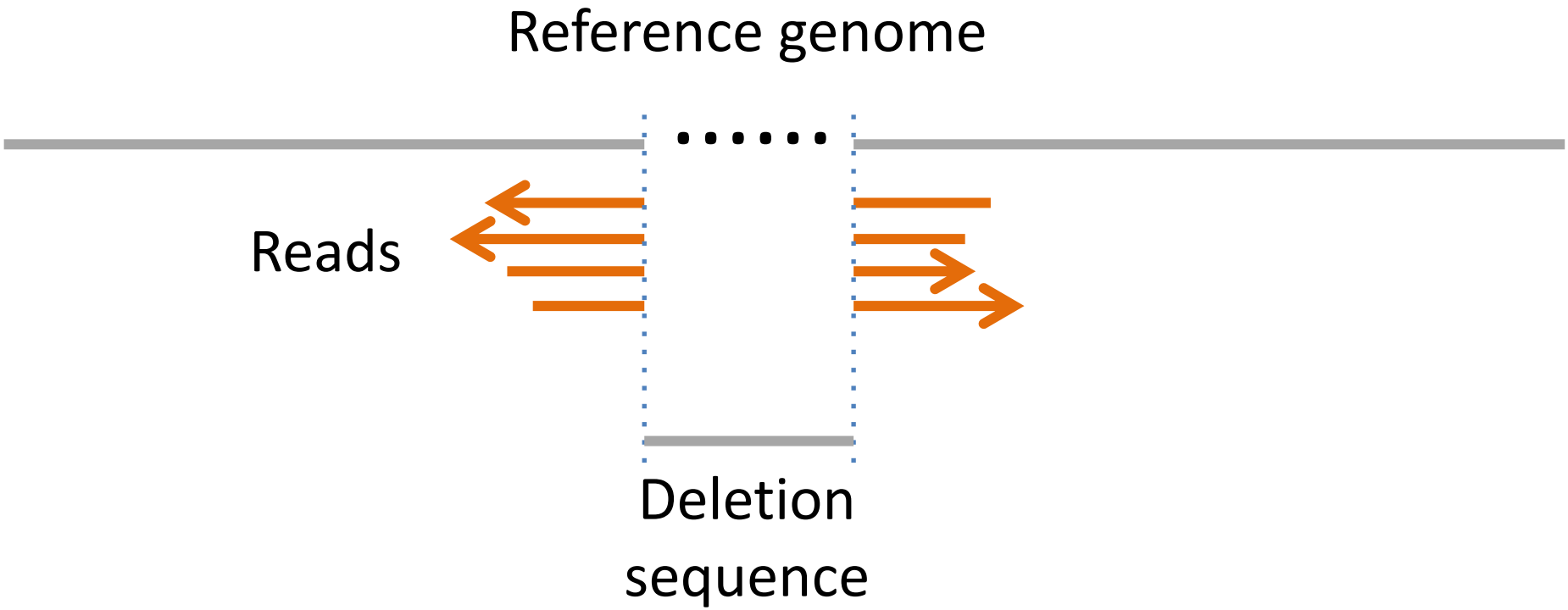
Variants in sequencing reads



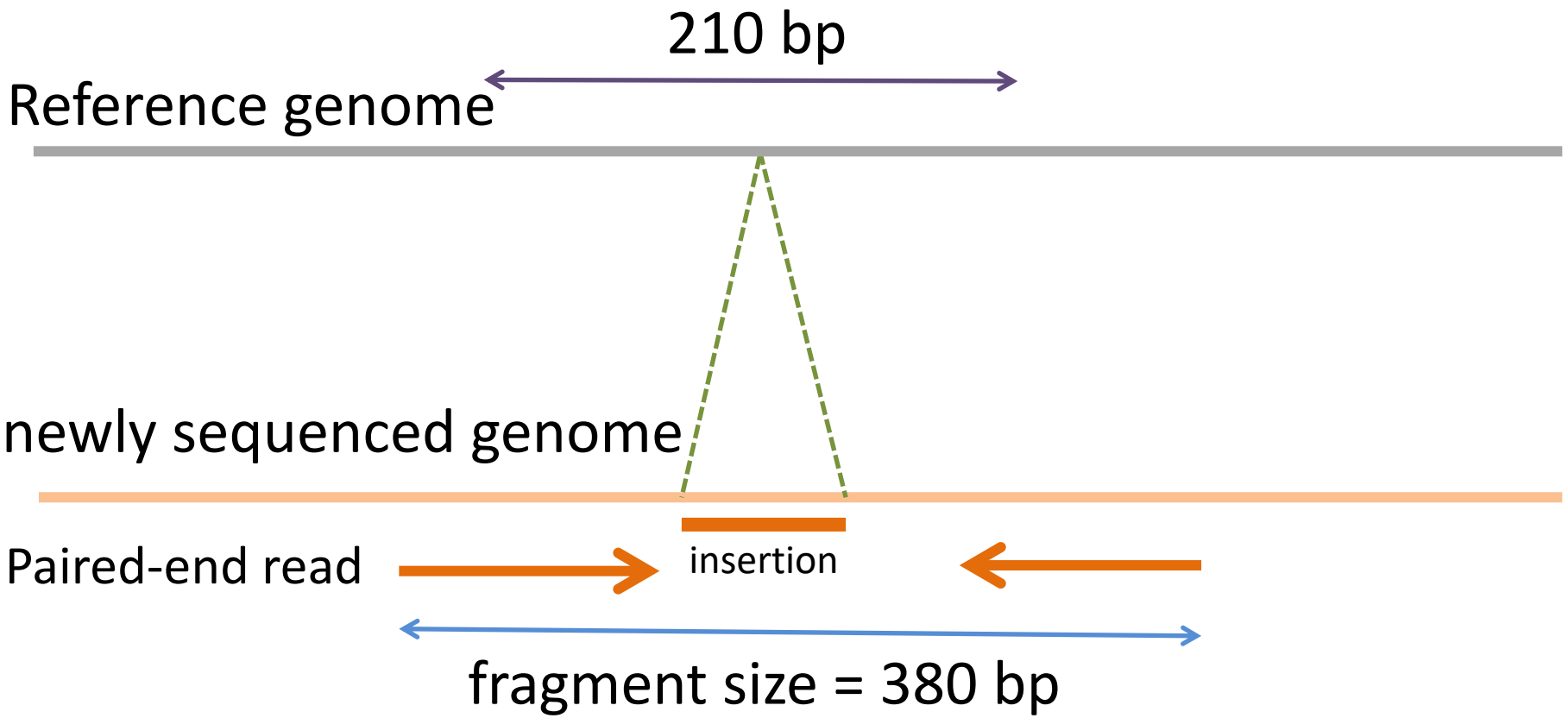
Paired-end reads to find "deletion" relative to the reference



Split reads to find "exact deletion sequence"

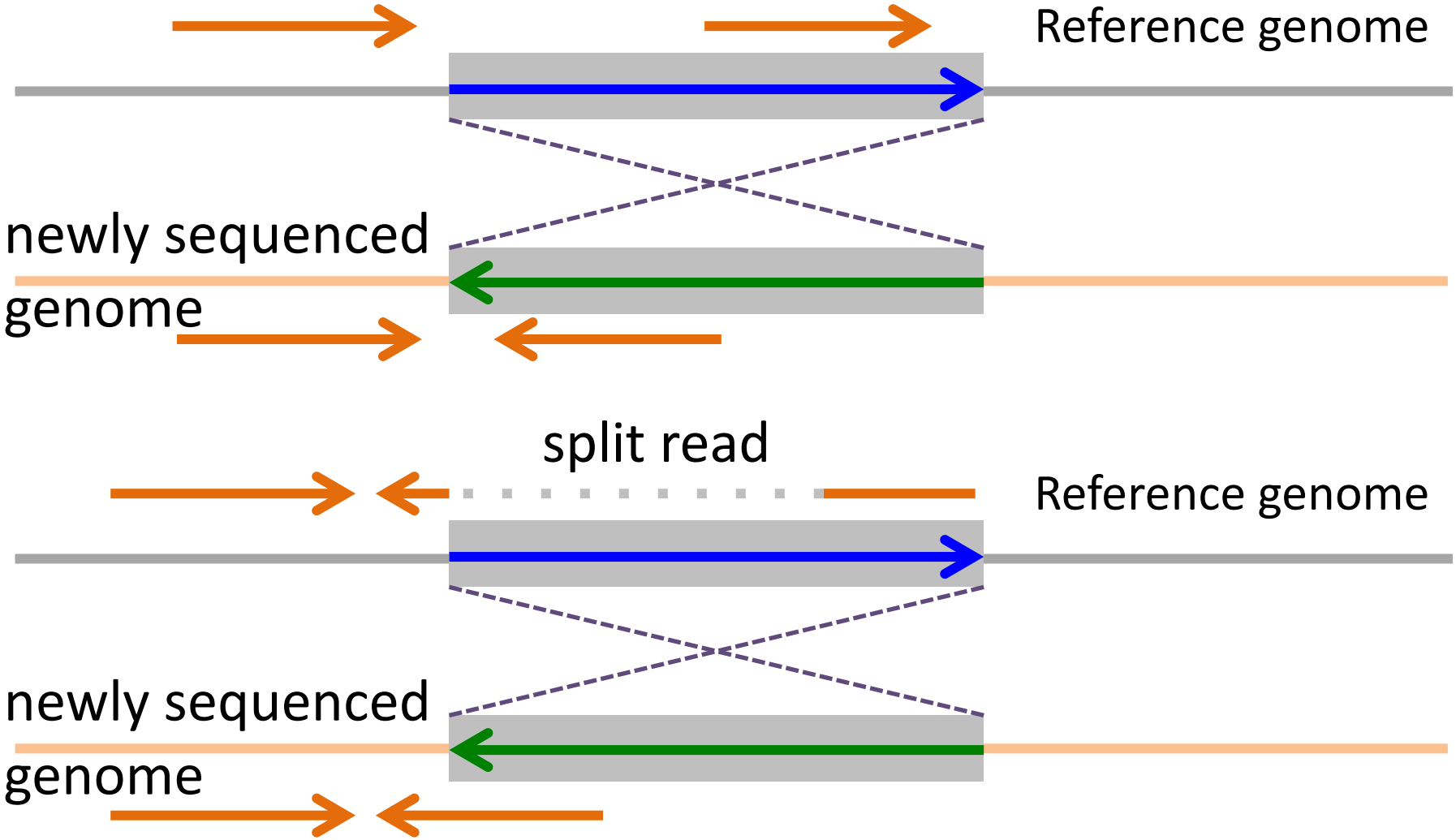


Paired-end reads to find "insertion" relative to the reference

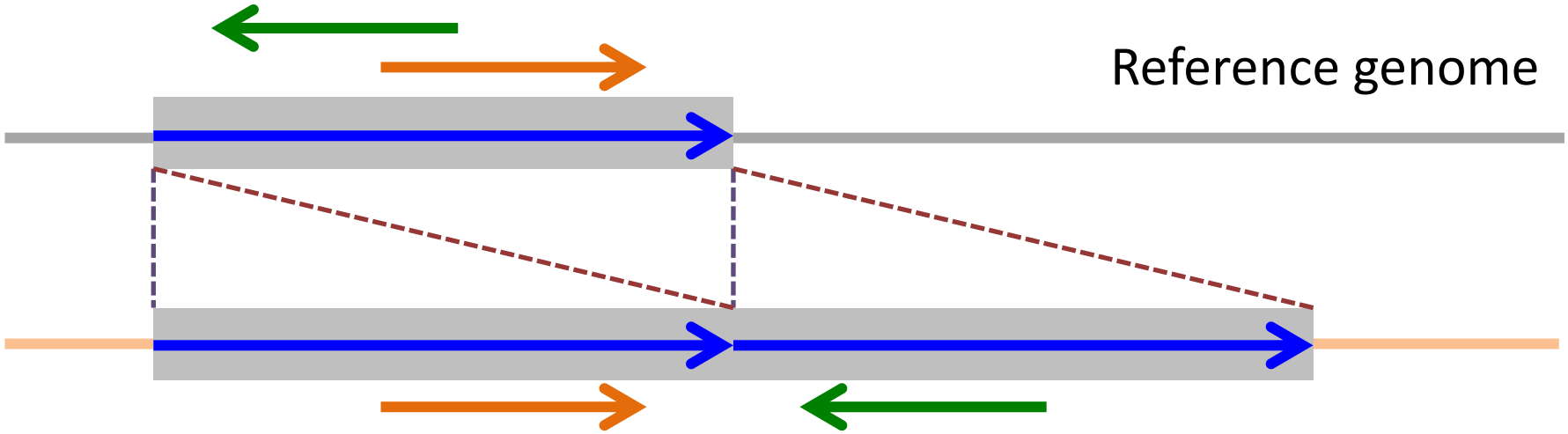


The size of insertions that can be identified by PE reads is determined by fragment sizes and read lengths.

inversion

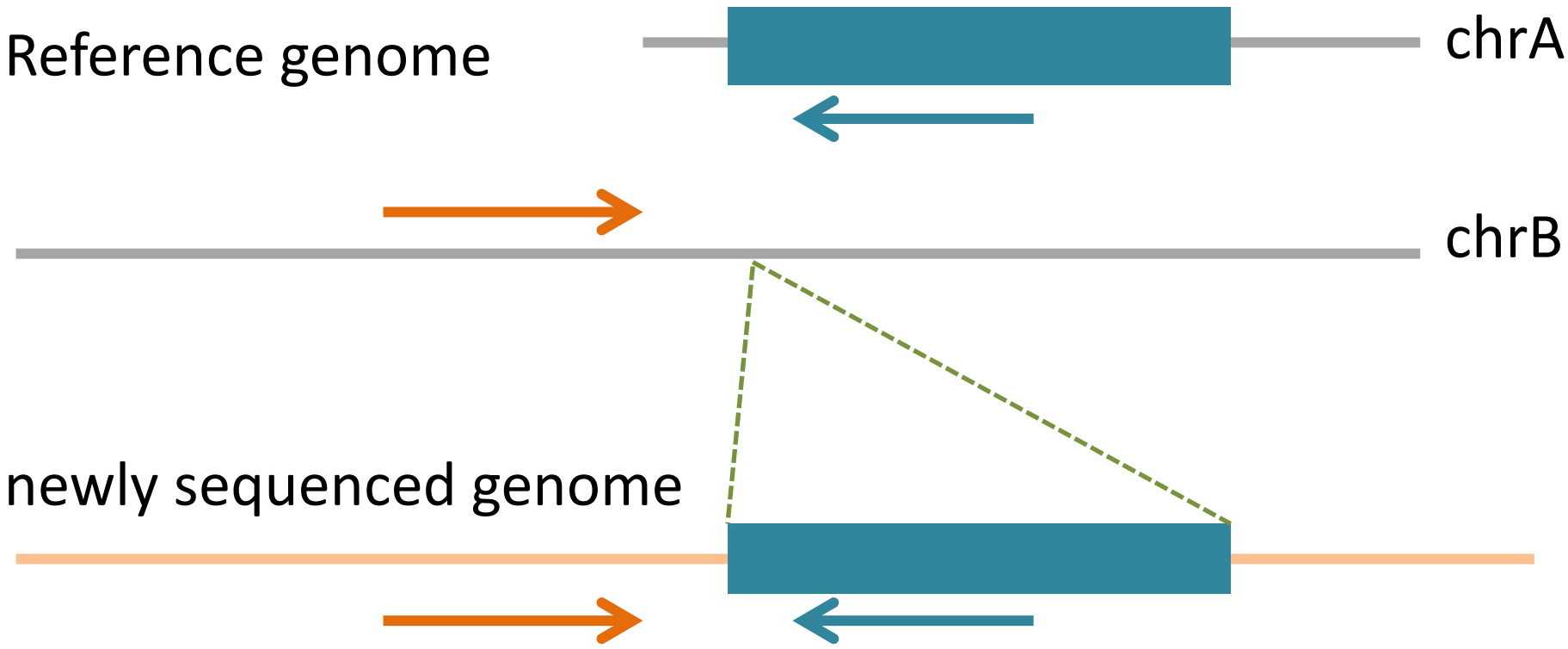

Reference genome
newly sequenced genome
split read
Reference genome
newly sequenced genome

Tandem duplication

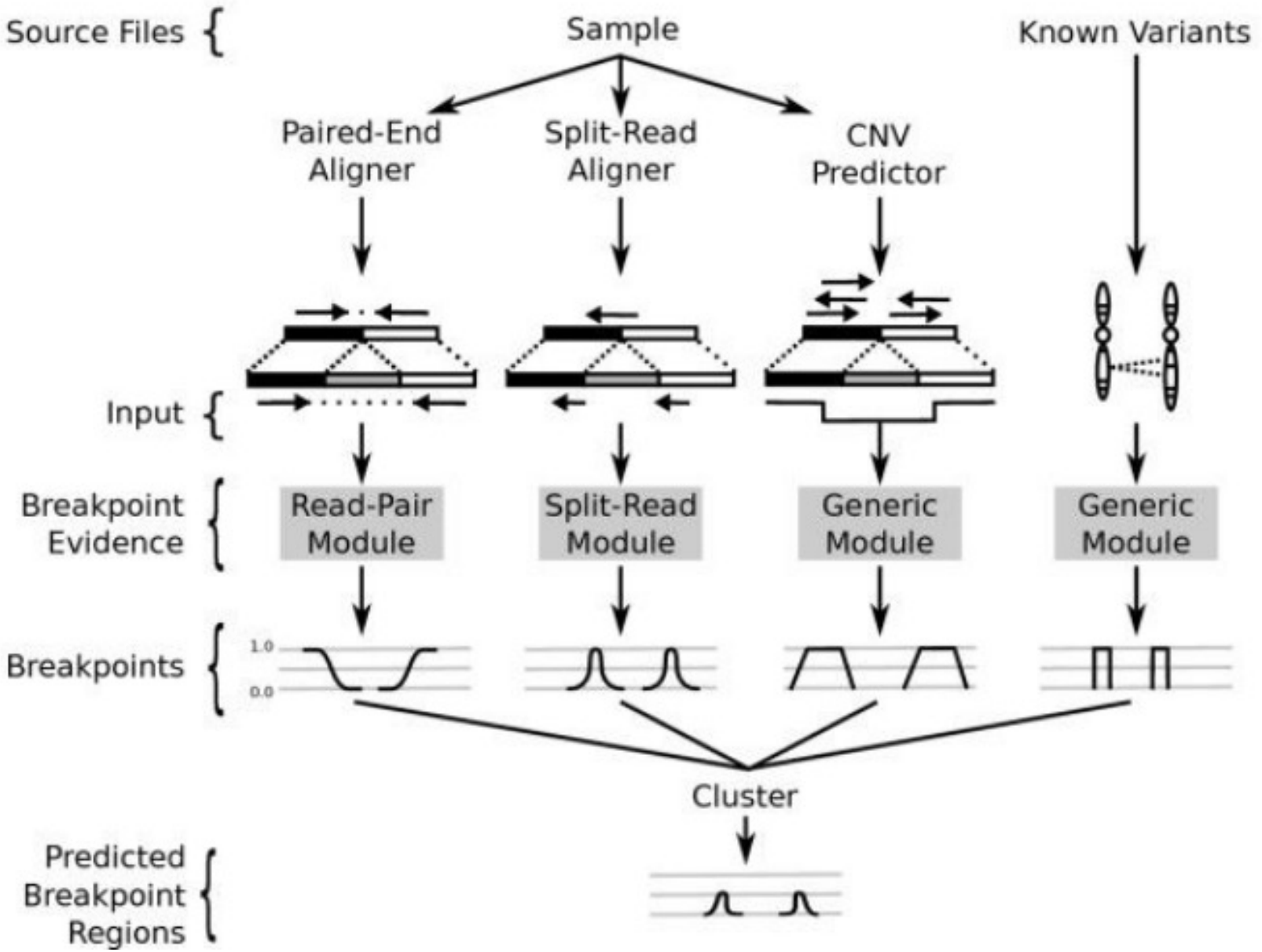


newly sequenced genome

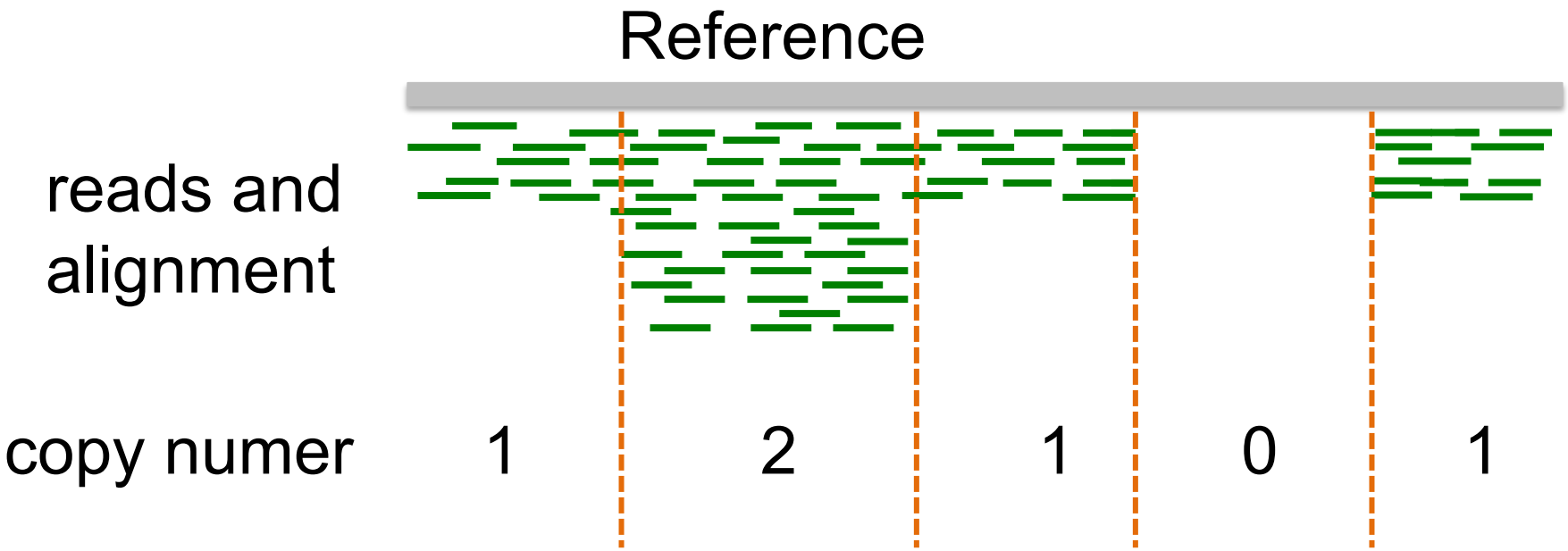
Translocation



LUMPY: an integrative framework for SV discovery

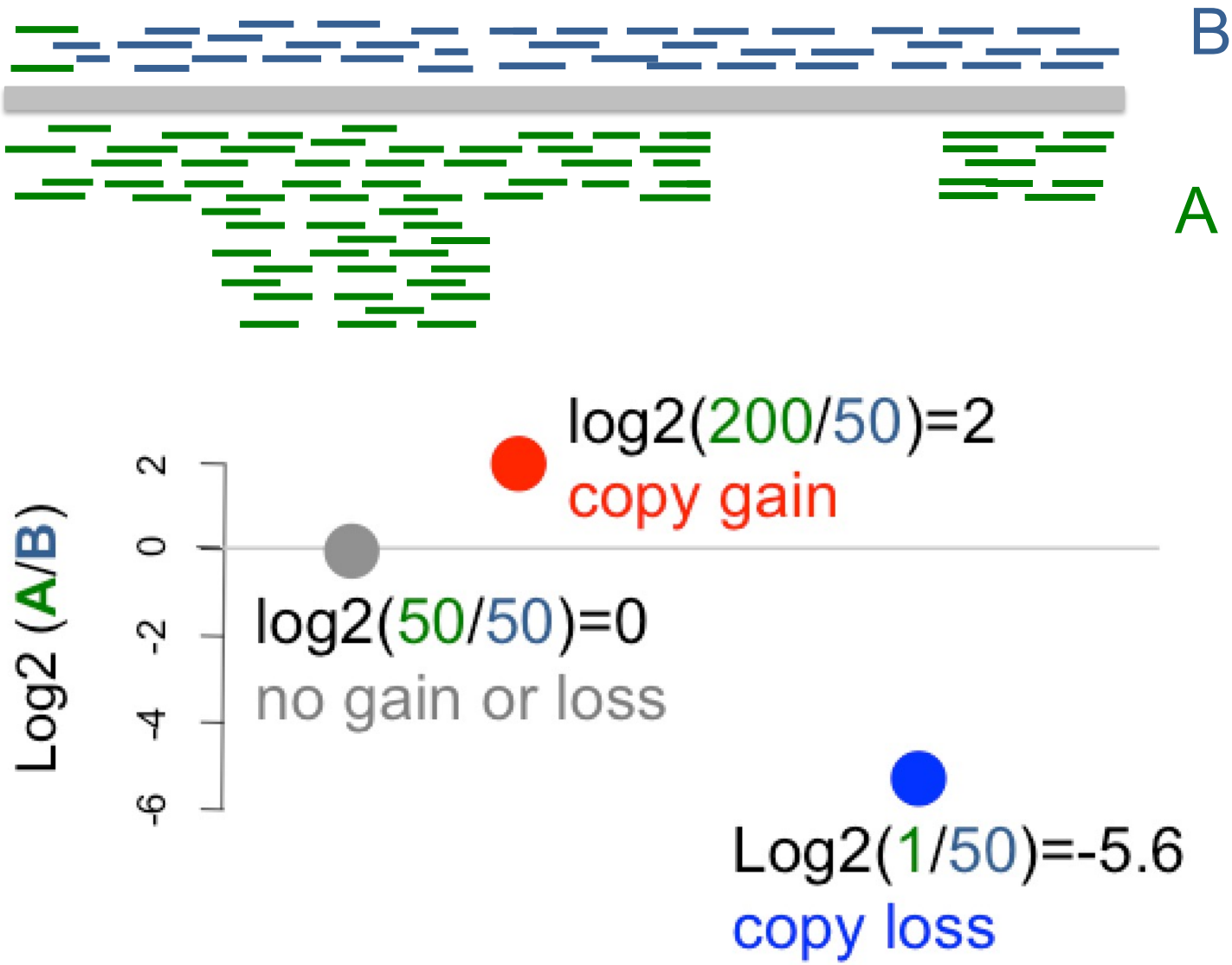


Read depths



1. PCR bias (GC content, complicated structure)
2. Alignment bias (e.g., highly repetitive regions)
3. Nucleotide polymorphisms

Read depths

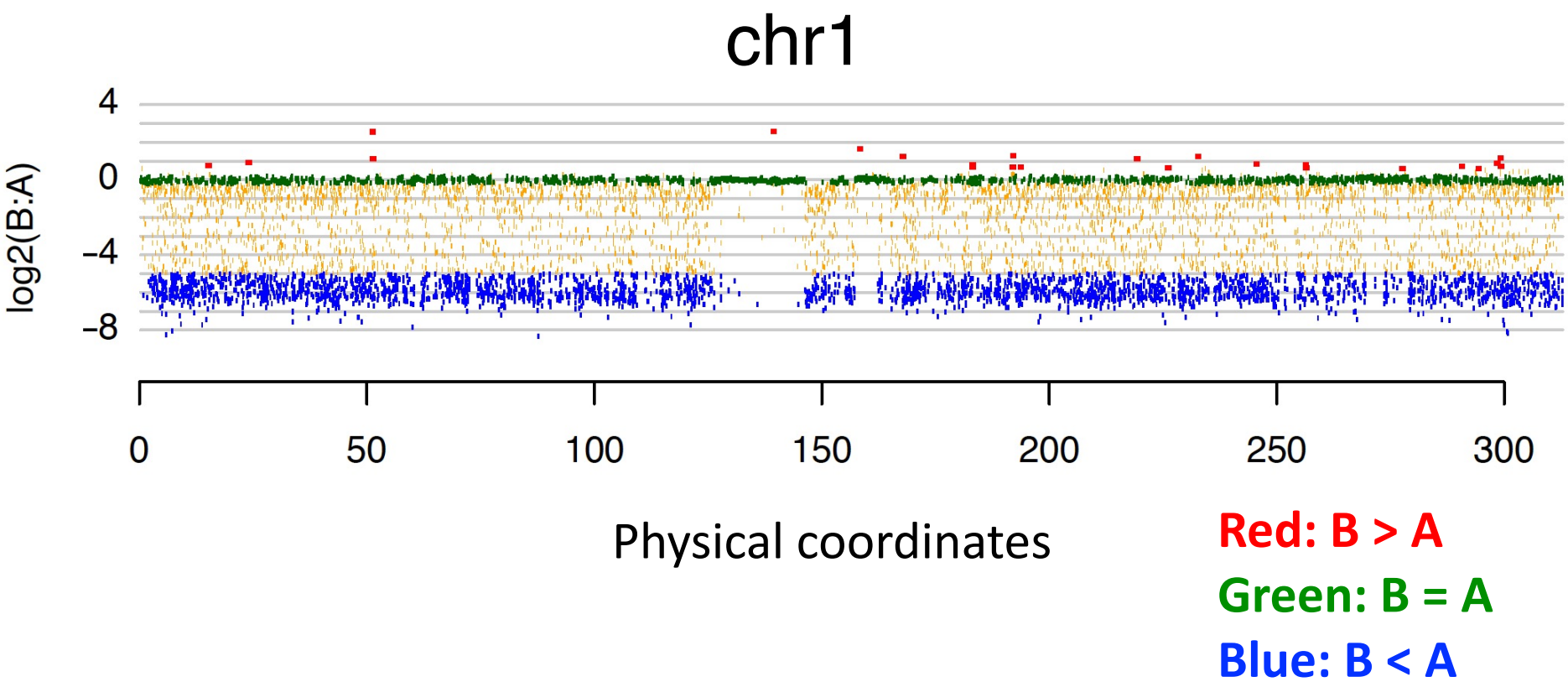

B
A
log2(200/50)=2
copy gain
log2(50/50)=0
no gain or loss
Log2(1/50)=-5.6
copy loss
Log2 (A/B)
2
0
-2
-4
-6

Comparative Genomic Read Depth (CGRD)

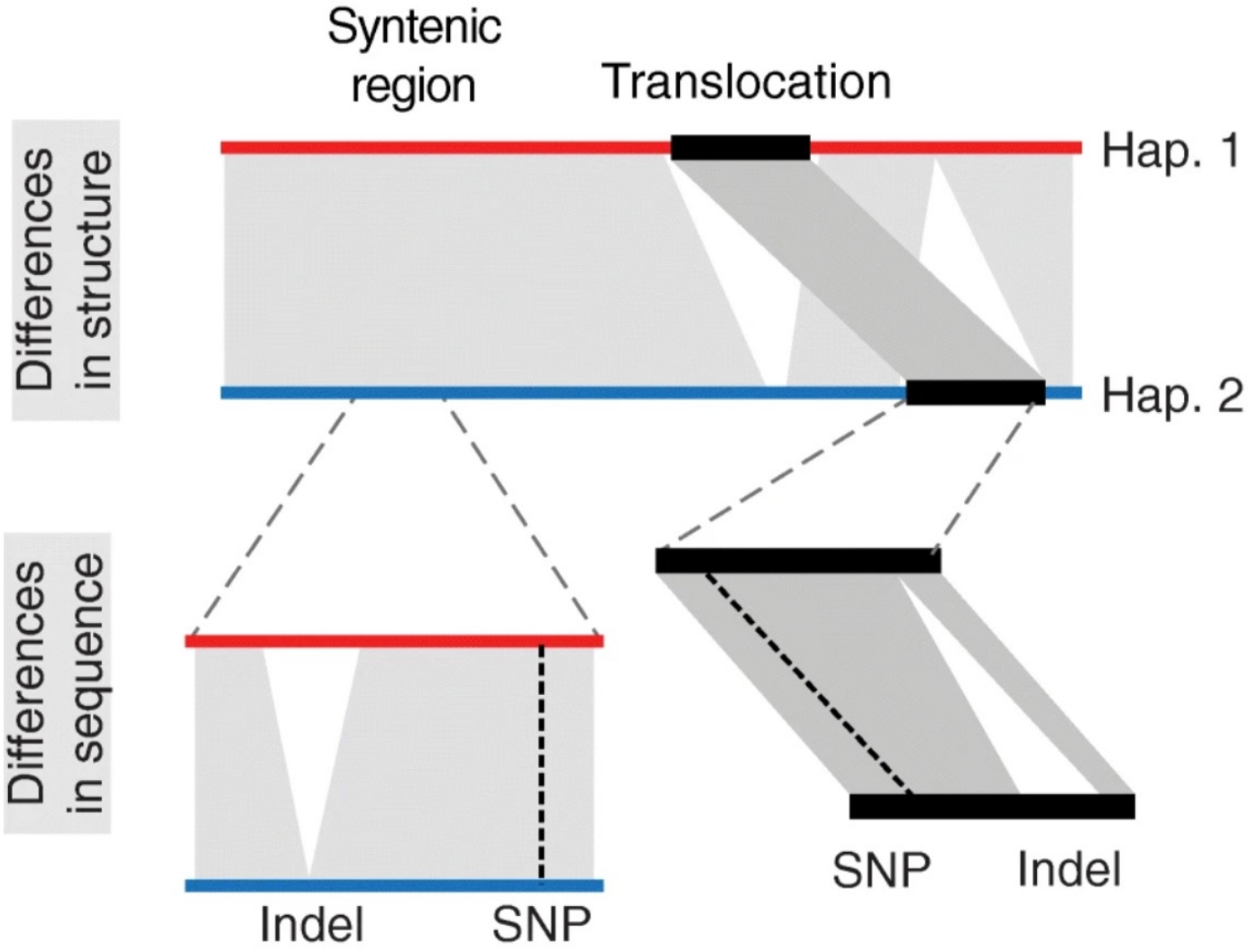
CGRD is a pipeline to compare sequencing read depths from two samples along a reference genome.

1. define effective genomic bins each of which harbors certain non-repetitive sequences
2. align reads and count read depths per bin for both samples
3. combine neighbor bins with similar fold changes in read depth between the two samples (segmentation)

Comparative Genomic Read Depths (CGRD)

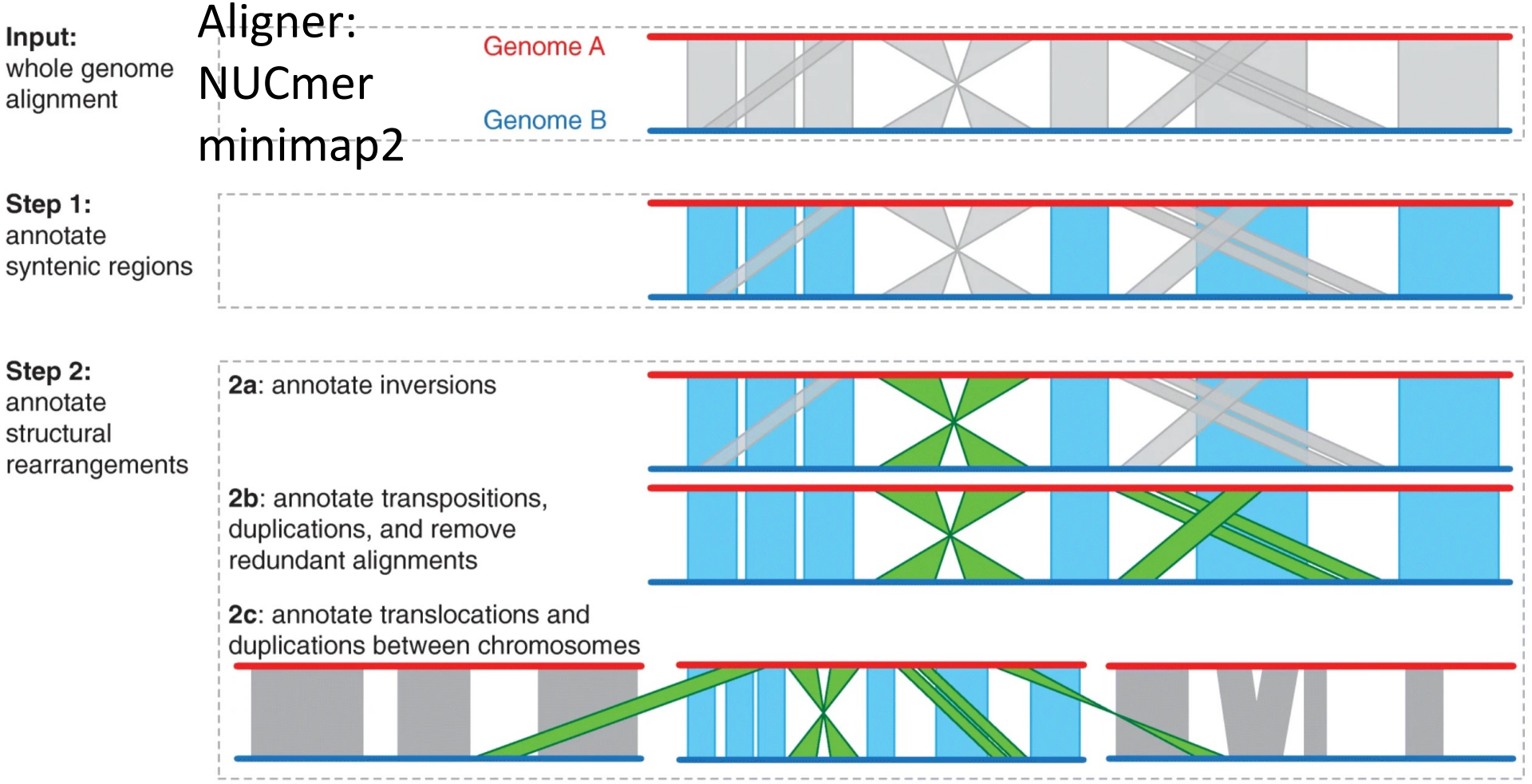


SyRI: finding genomic variation using whole-genome assemblies



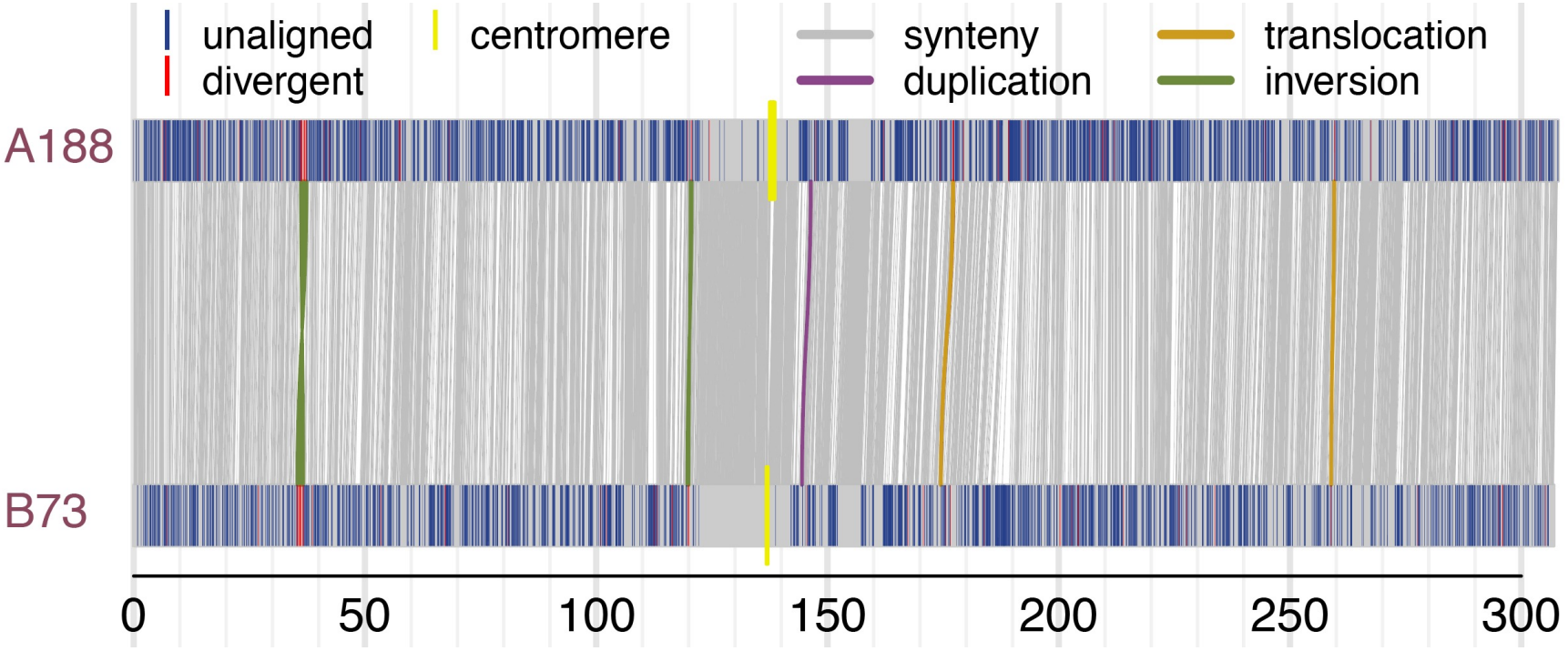
Goel et al, 2019, Genome Biology, 20:277

SyRI - algorithm



Structure variation with SyRI

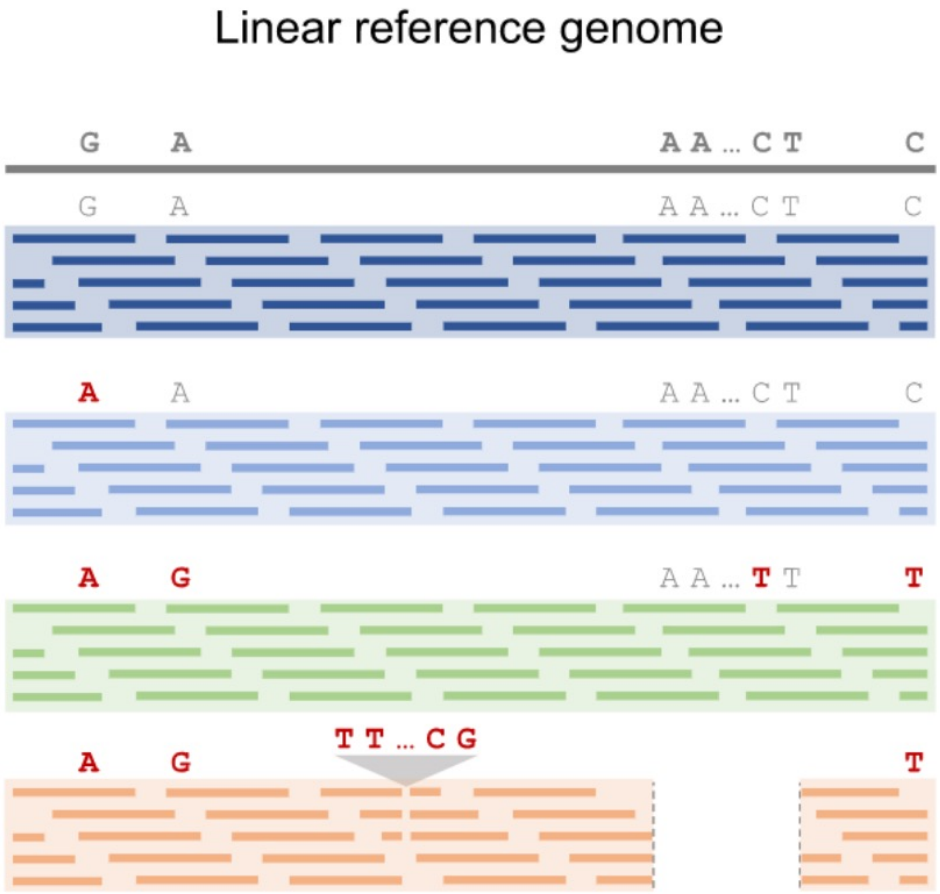
- SyRI is a tool using the **genome assemblies** to discover genomic rearrangement and local sequence differences (Goel et al., 2019)

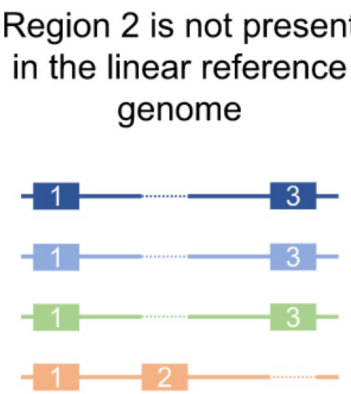


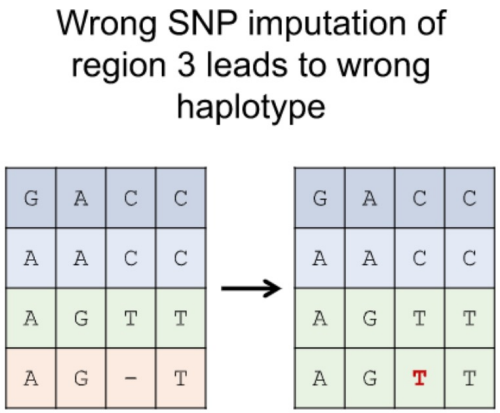
Outline

- Introduction of comparative genomics and structural variation
- Approaches
  1. Comparative genome hybridization
  2. Paired-end reads
  3. Read depth
  4. Whole genome assembly
- Pangenomics
- Case study: CGRD

Issues associated with single reference genomes

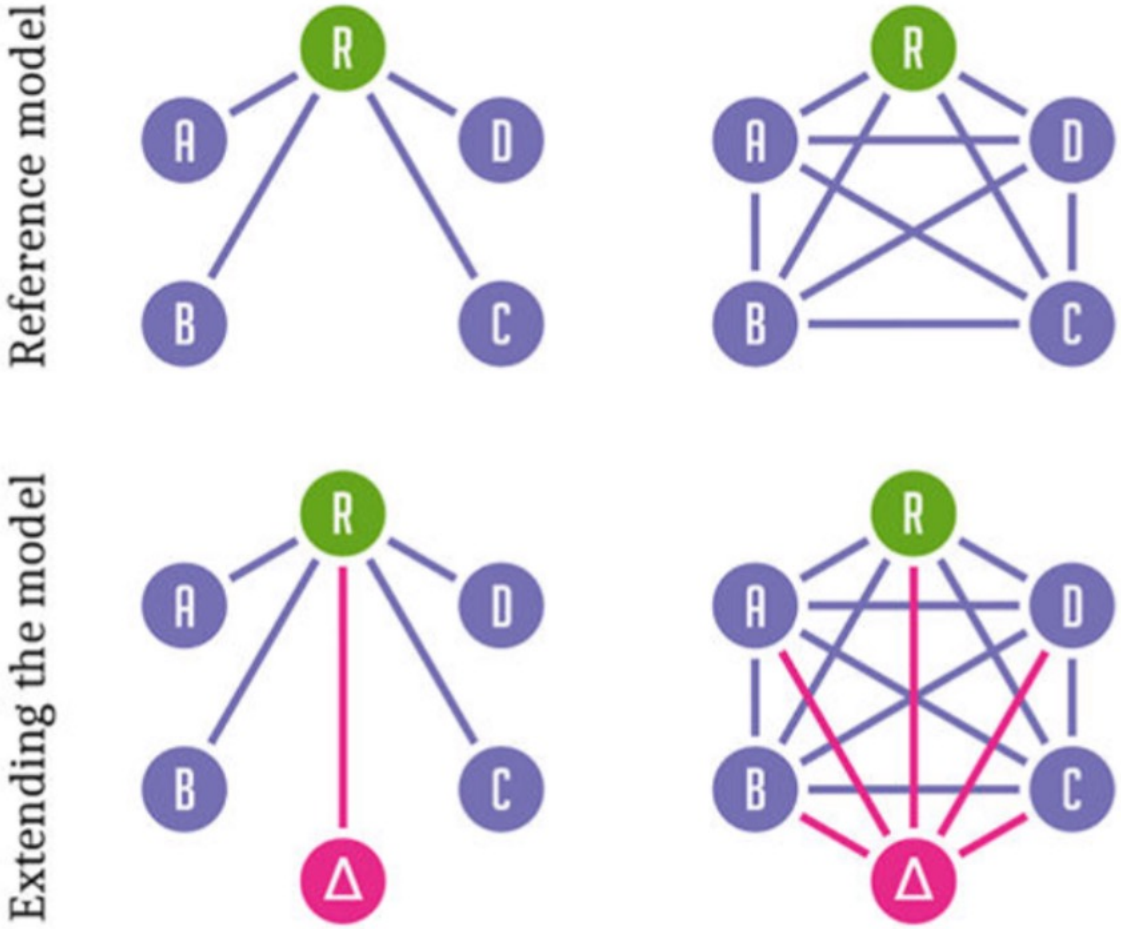






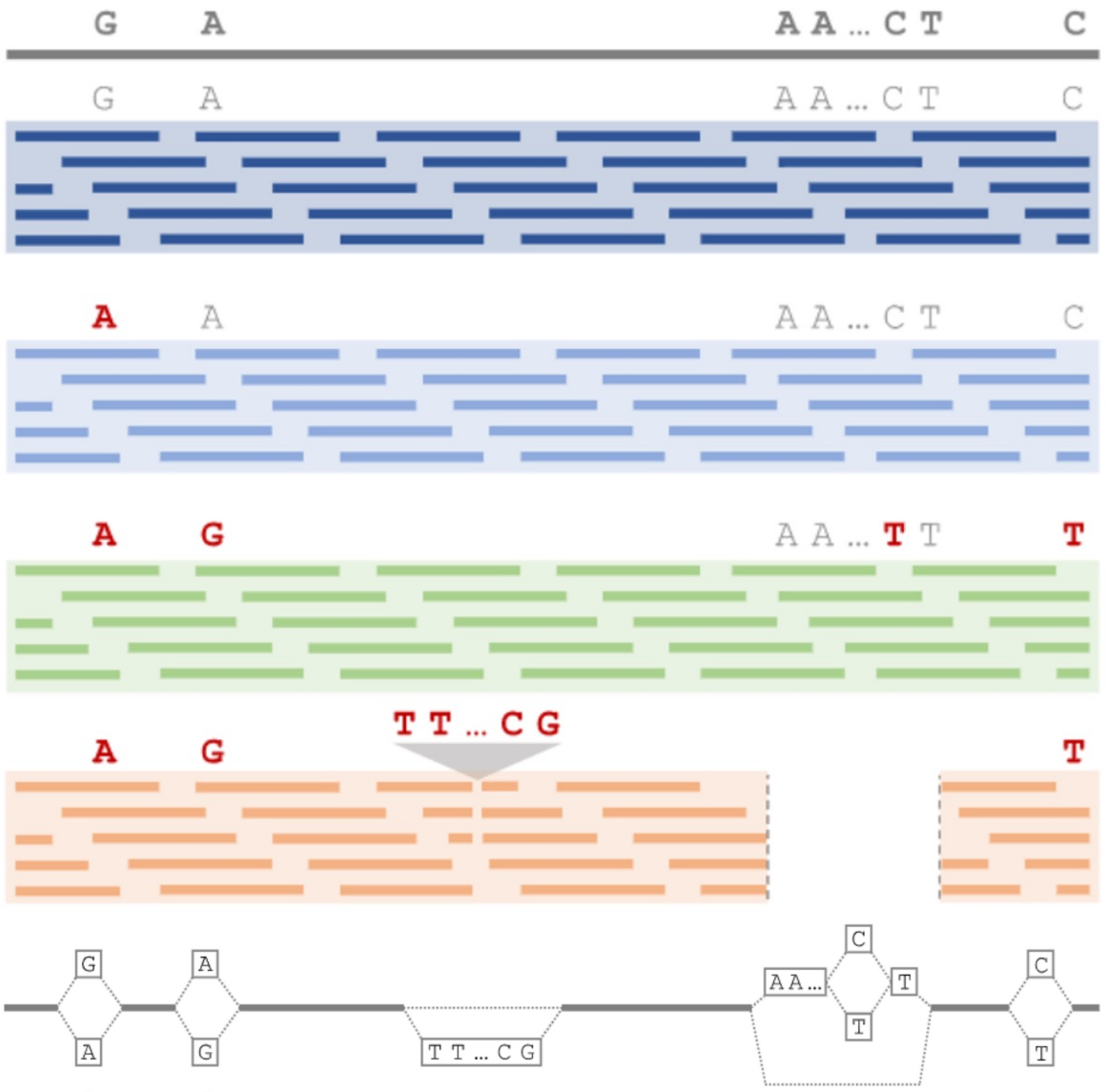
Coletta et al., Genome Biol, 2021

Pangenomics



Eizenga et al., 2020, Annual Review of Genomics and Human Genetics, 21:139-162

Construct a pangenome to facilitate genotyping

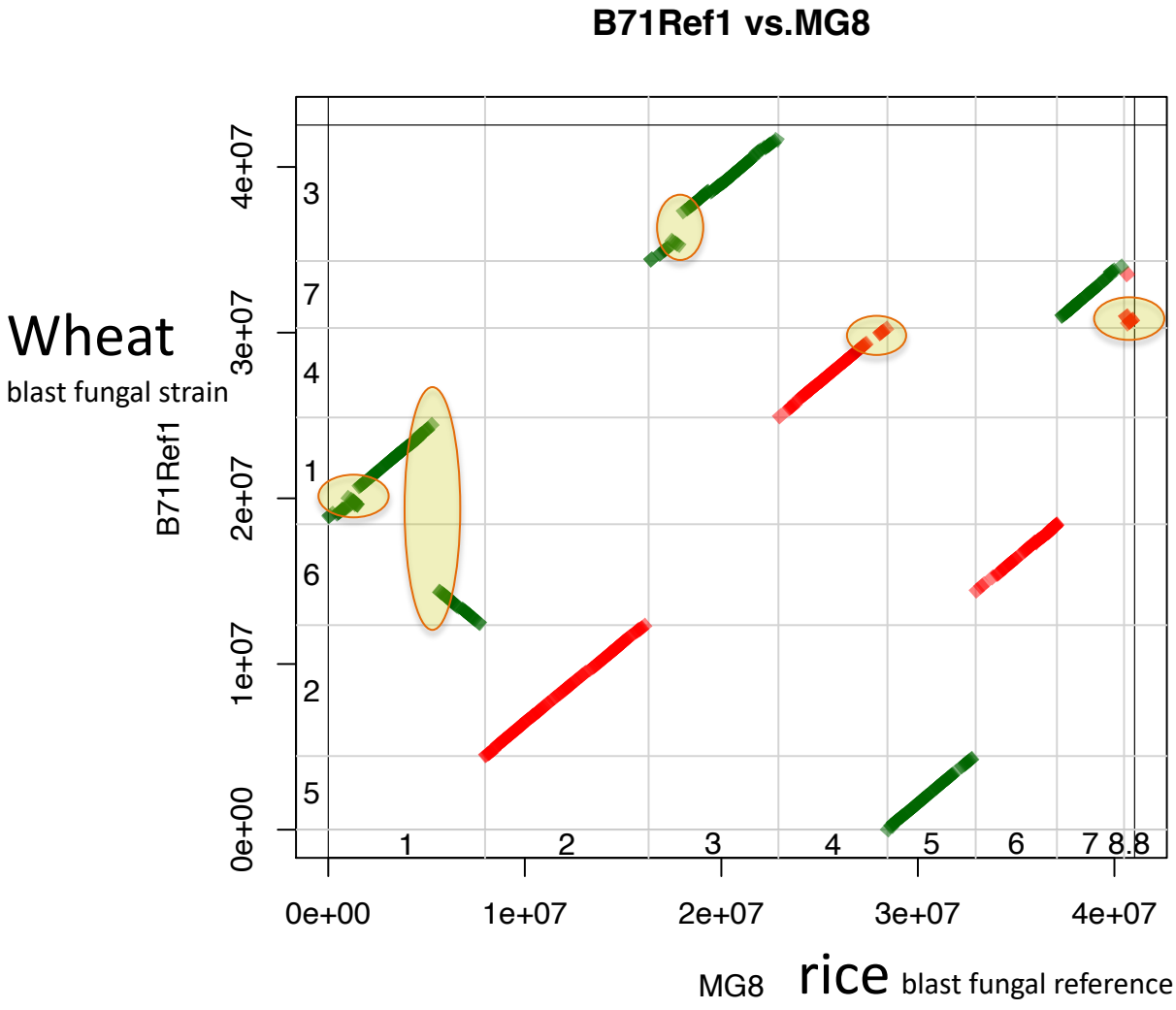


Outline

- Introduction of comparative genomics and structural variation
- Approaches
  1. Comparative genome hybridization
  2. Paired-end reads
  3. Read depth
  4. Whole genome assembly
- Pangenomics
- Case study: CGRD

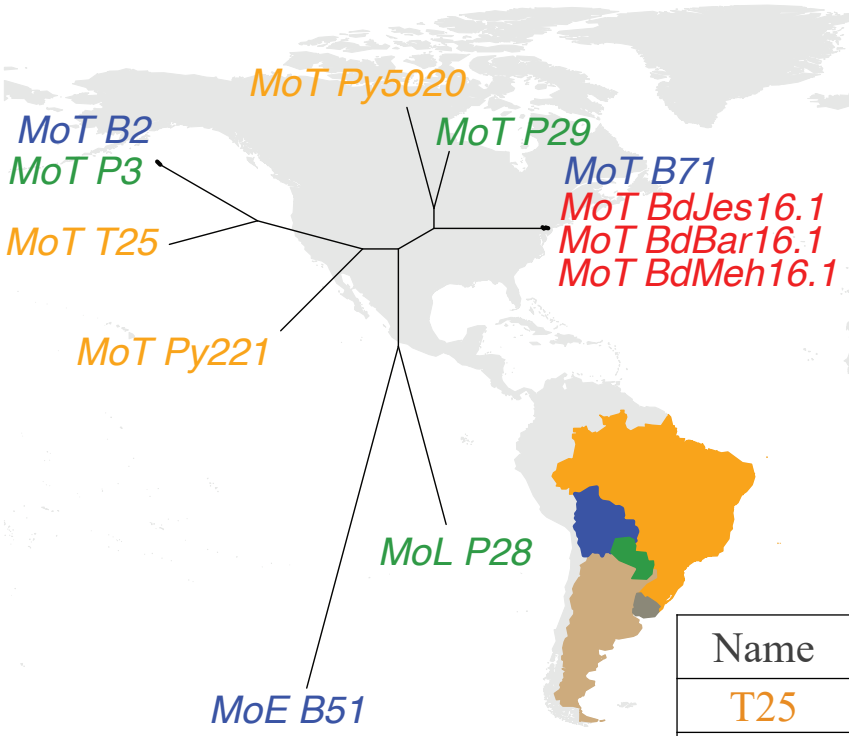
MoT B71 final assembly vs. MG8 (70-15, *rice* strain)

Chr	Length (bp)
chr1	6,442,091
chr2	7,902,655
chr3	8,206,304
chr4	5,402,116
chr5	4,442,877
chr6	6,090,985
chr7	4,042,640
scaf1	941,816
scaf2	739,928
scaf3	104,670
scaf4	74,396
scaf5	69,131



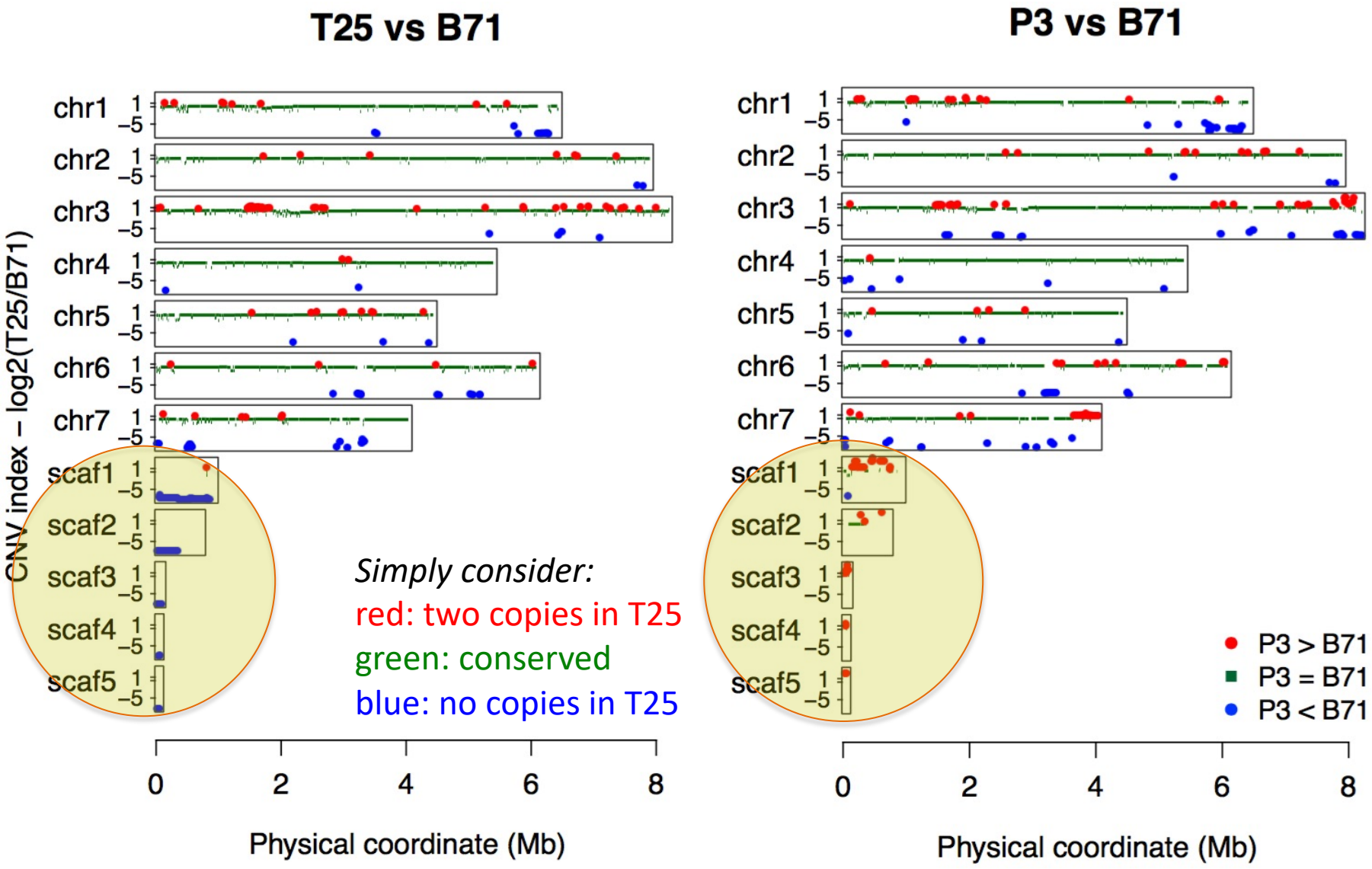
>10 kb overlap and >95% identity

Illumina sequencing of additional eight strains

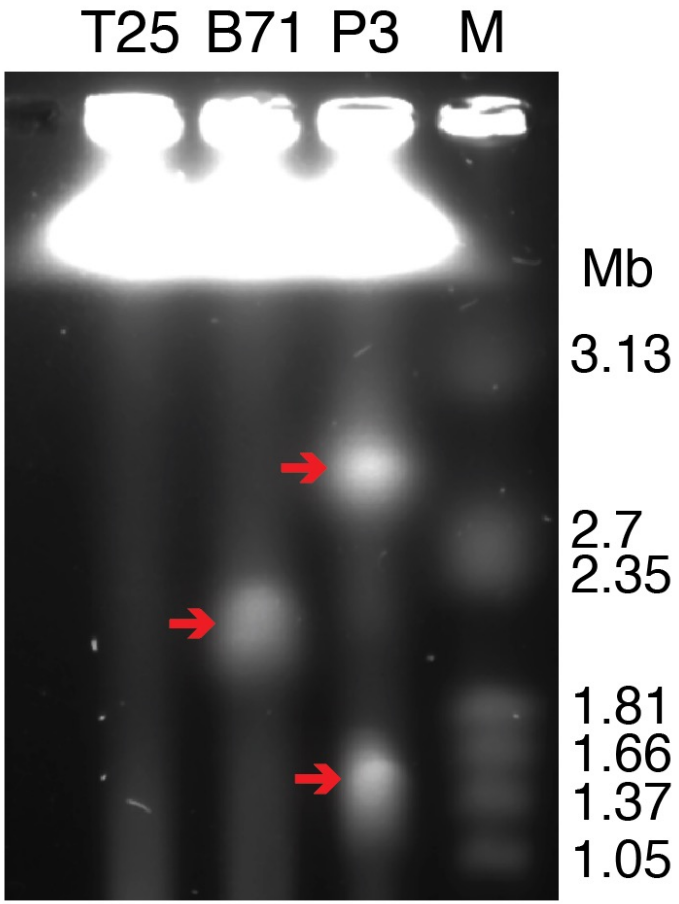


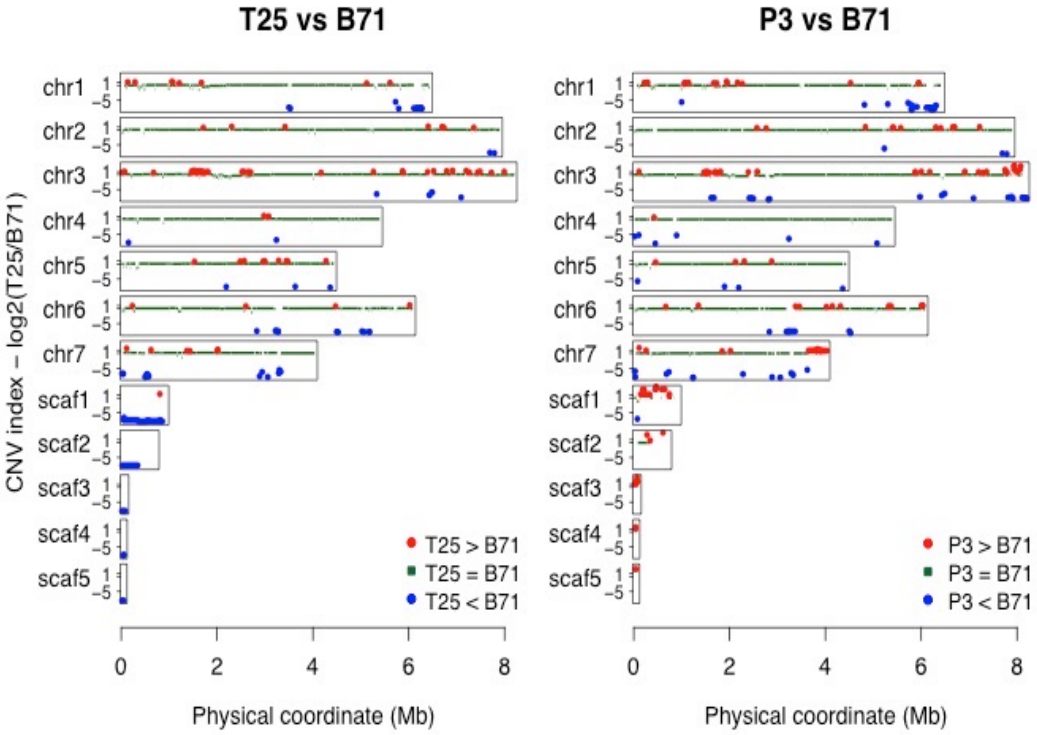
Name	Year isolated	Host species	Isolation location
T25	1988	*Triticum*	Paraná, Brazil
Py5020	2005	*Triticum*	Paraná, Brazil
Py22.1	2007	*Triticum*	Paraná, Brazil
B2	2011	*Triticum*	Quirusillas, Bolivia
B71	2012	*Triticum*	Okinawa, Bolivia
P3	2012	*Triticum*	Canindeyú,Paraguay
P28	2014	*Bromus*	Paraguay
P29	2014	*Bromus*	Paraguay
B51	2012	*Eleusine*	Quirusillas, Bolivia

Read depths to infer *copy number variation*



CHEF gel to separate chromosome-sized DNA





T25 has no mini-chromosomes

B71 has one and P3 has two

Additional References

Pinkel, D. & Albertson, D. G. Comparative genomic hybridization. Annu. Rev. Genomics Hum. Genet. 6, 331–354 (2005).

Miller, W., Makova, K. D., Nekrutenko, A. & Hardison, R. C. Comparative genomics. Annu. Rev. Genomics Hum. Genet. 5, 15–56 (2004).

Korbel, J. O. et al. Paired-end mapping reveals extensive structural variation in the human genome. Science 318, 420–426 (2007).

Armstrong, J., Fiddes, I. T., Diekhans, M. & Paten, B. Whole-Genome Alignment and Comparative Annotation. Annu Rev Anim Biosci (2018). doi:10.1146/annurev-animal-020518-115005

Kronenberg, Z. N. et al. High-resolution comparative analysis of great ape genomes. Science 360, (2018).