Comparative genomics

Bioinformatics Applications (PLPTH813)

Sanzhen Liu

4/11/2019

Outline

- Introduction of comparative genomics
- Structural variation
  1. Copy number variation
  2. Translocation and inversion
- Homology
- Approaches
  1. Comparative genome hybridization
  2. Paired-end reads
  3. Read depth
  3. Whole genome assembly

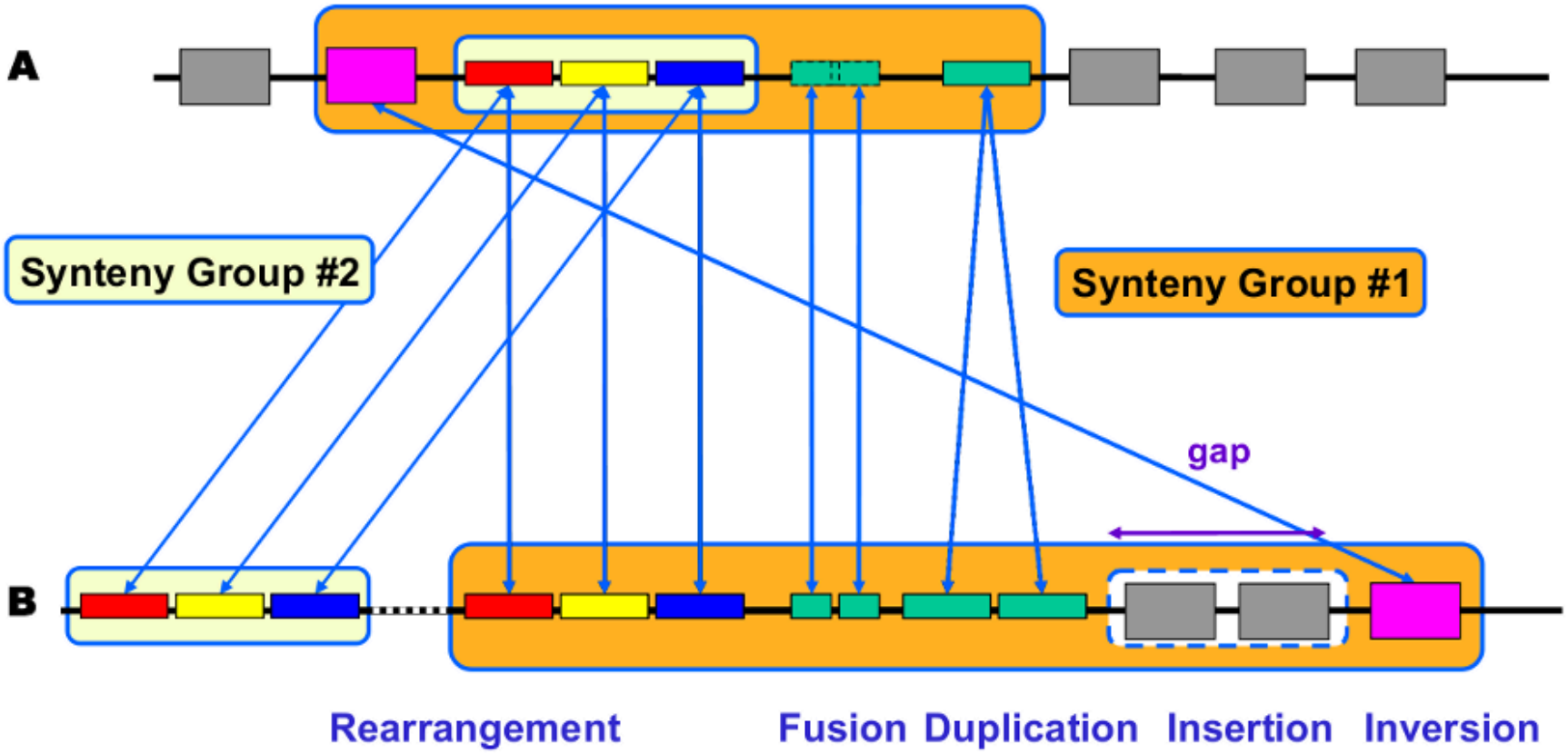
Comparative genomics among genomes

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

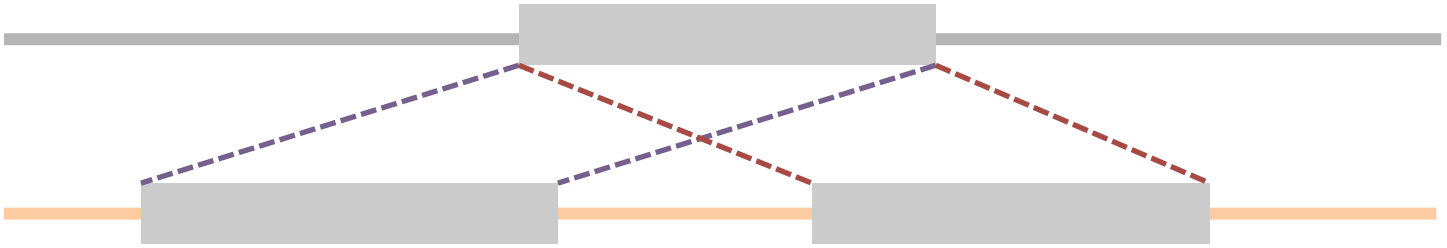
- Inversion

- Translocation

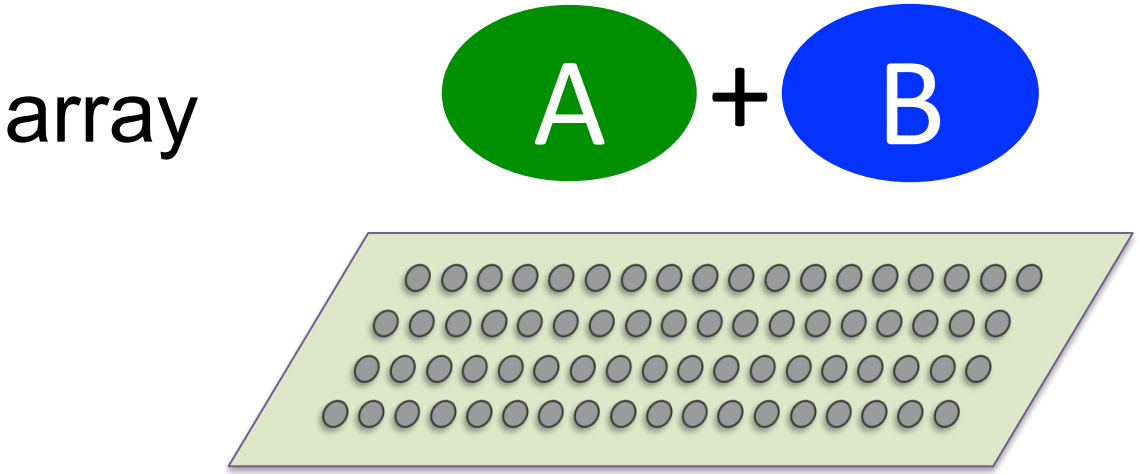
Unbalanced variation:

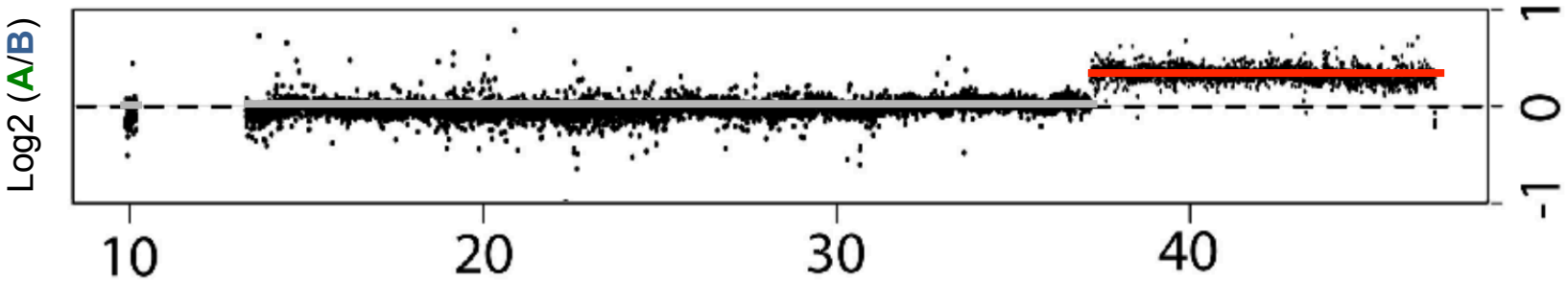
- Copy number variation (CNV)

(Presence/Absence variation, PAV)



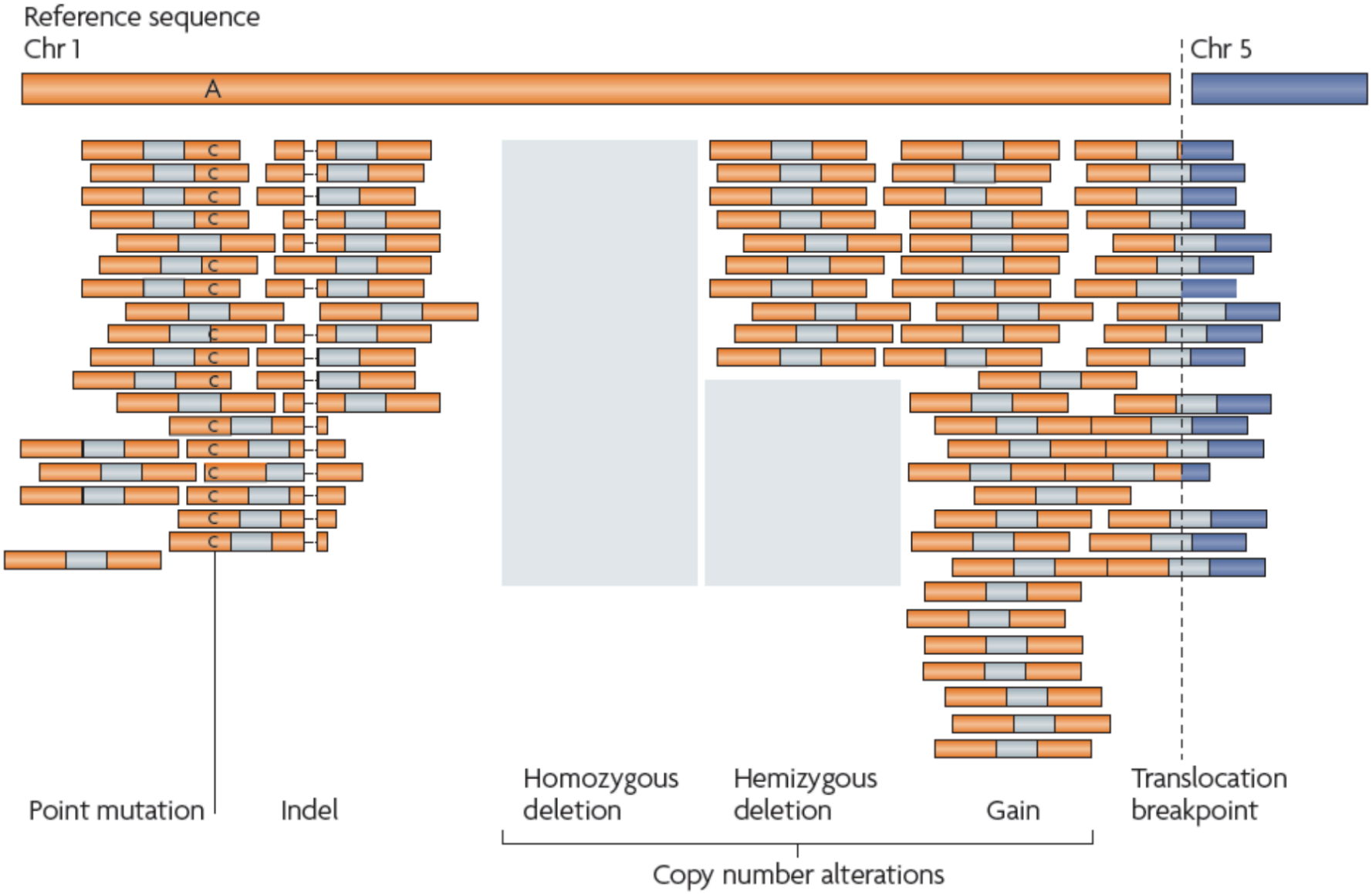
array Comparative Genomic Hybridization (aCGH)



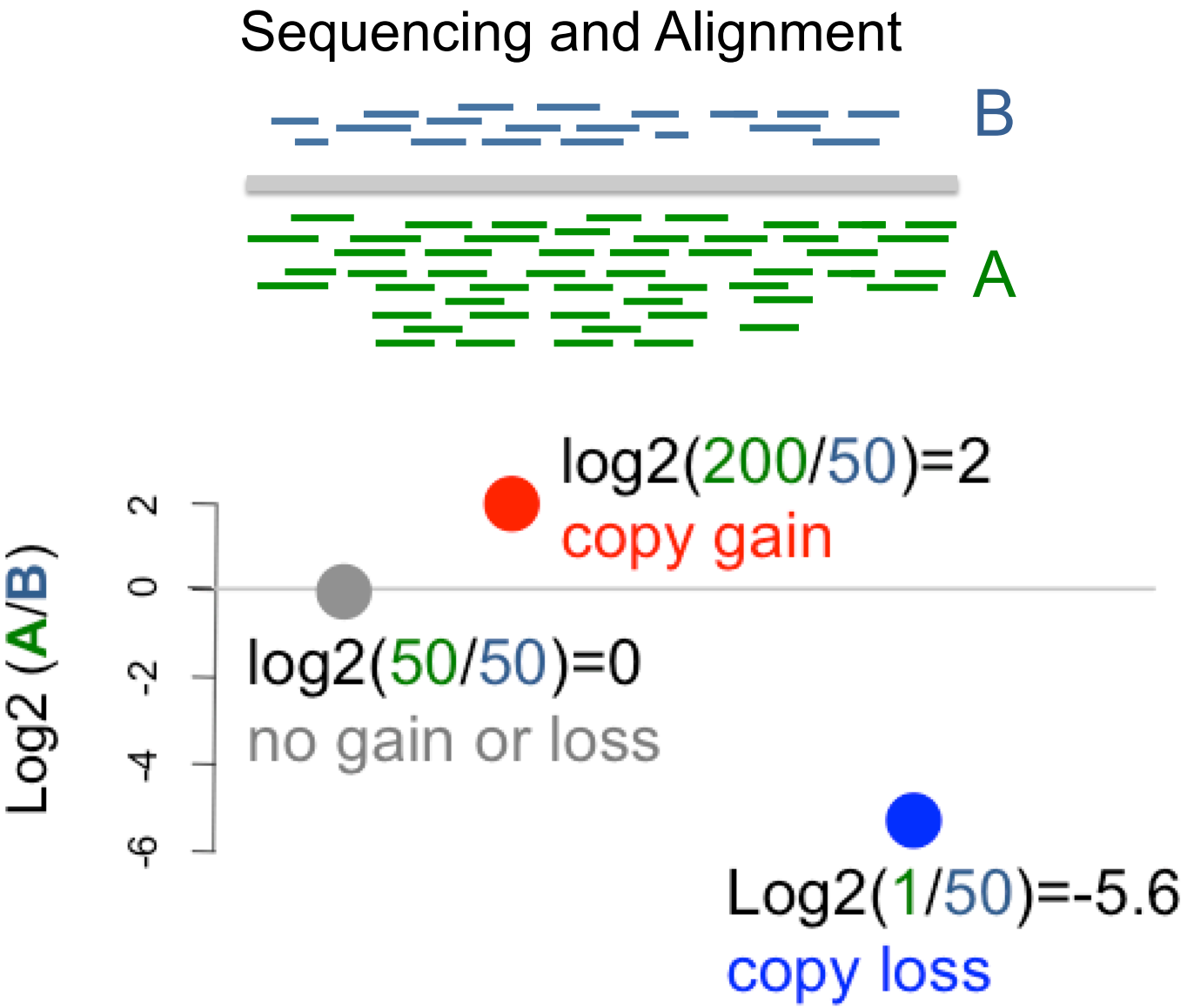


NGS provides information for the discovery of variants, including structural variation

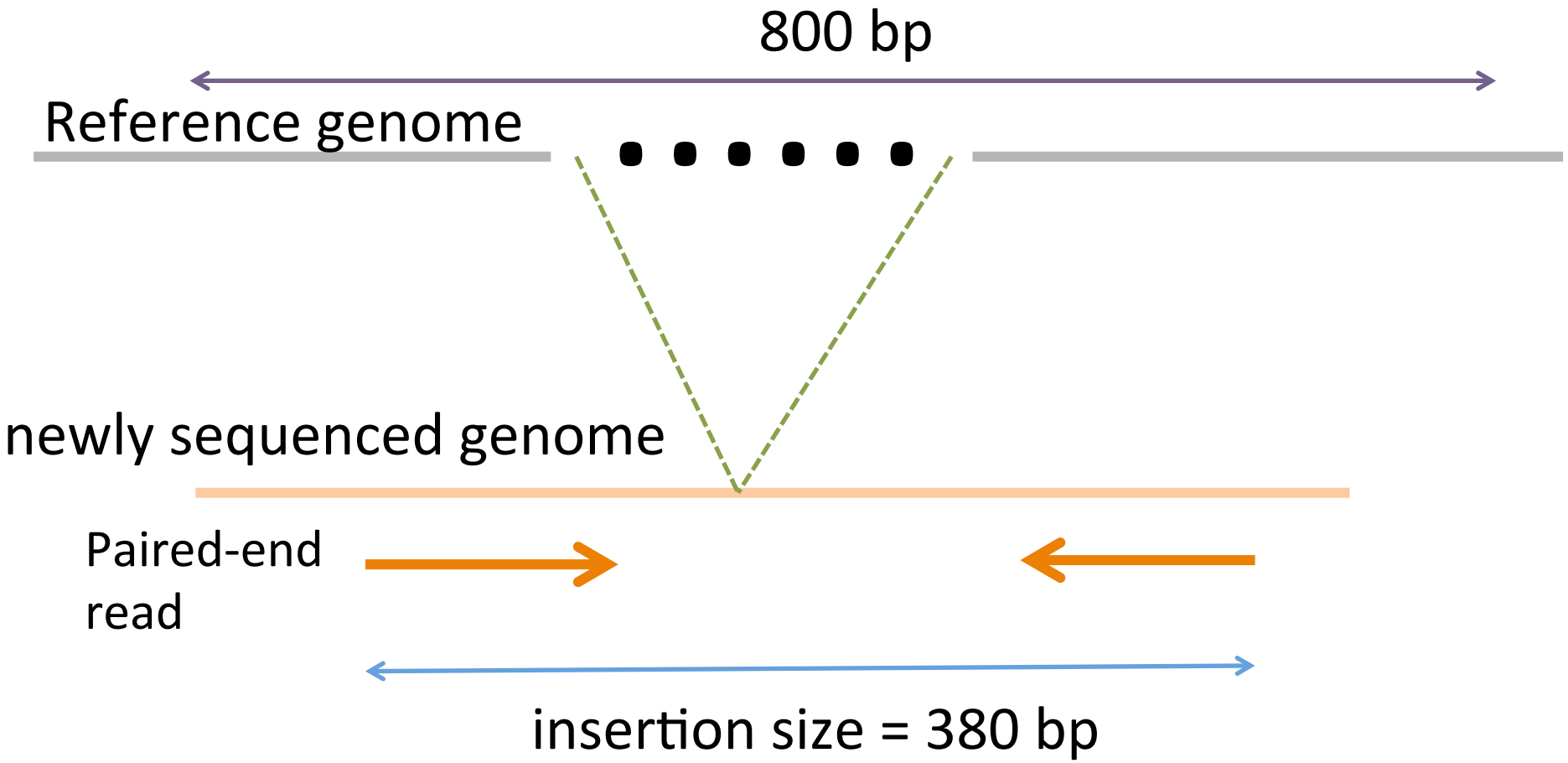
Variants in sequencing reads



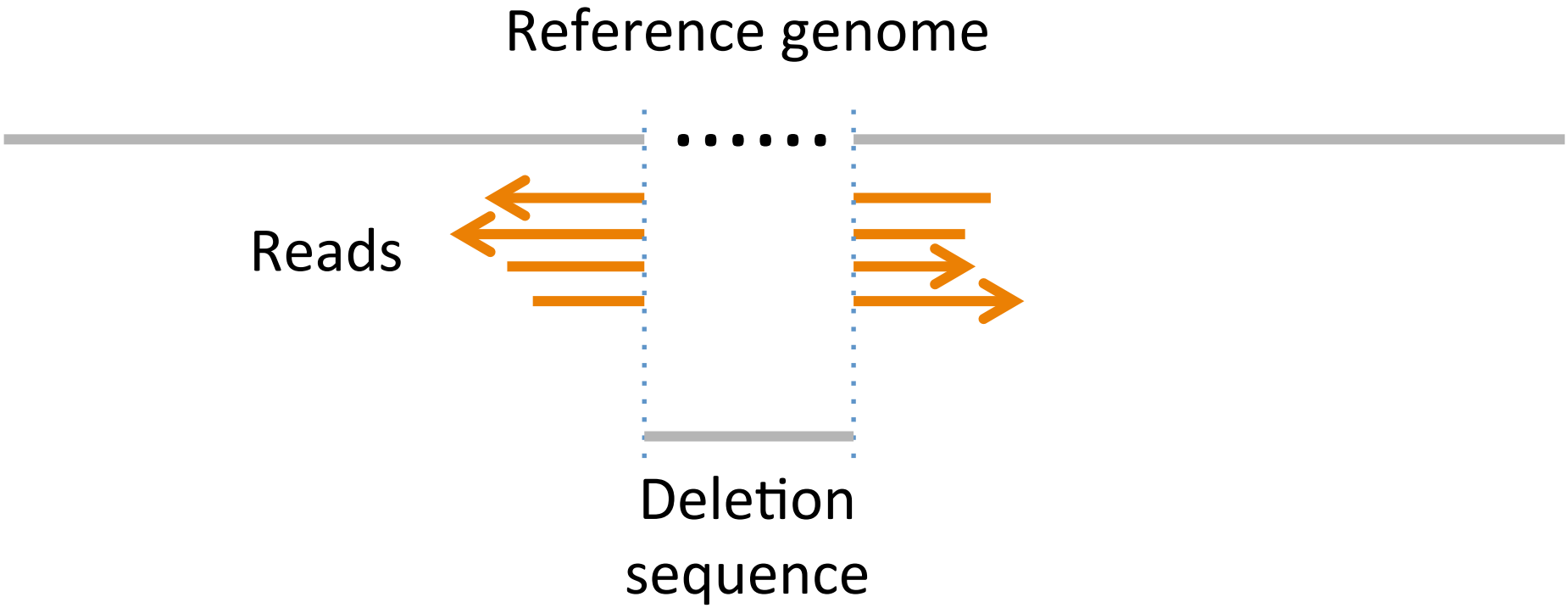
Read depths


Sequencing and Alignment
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

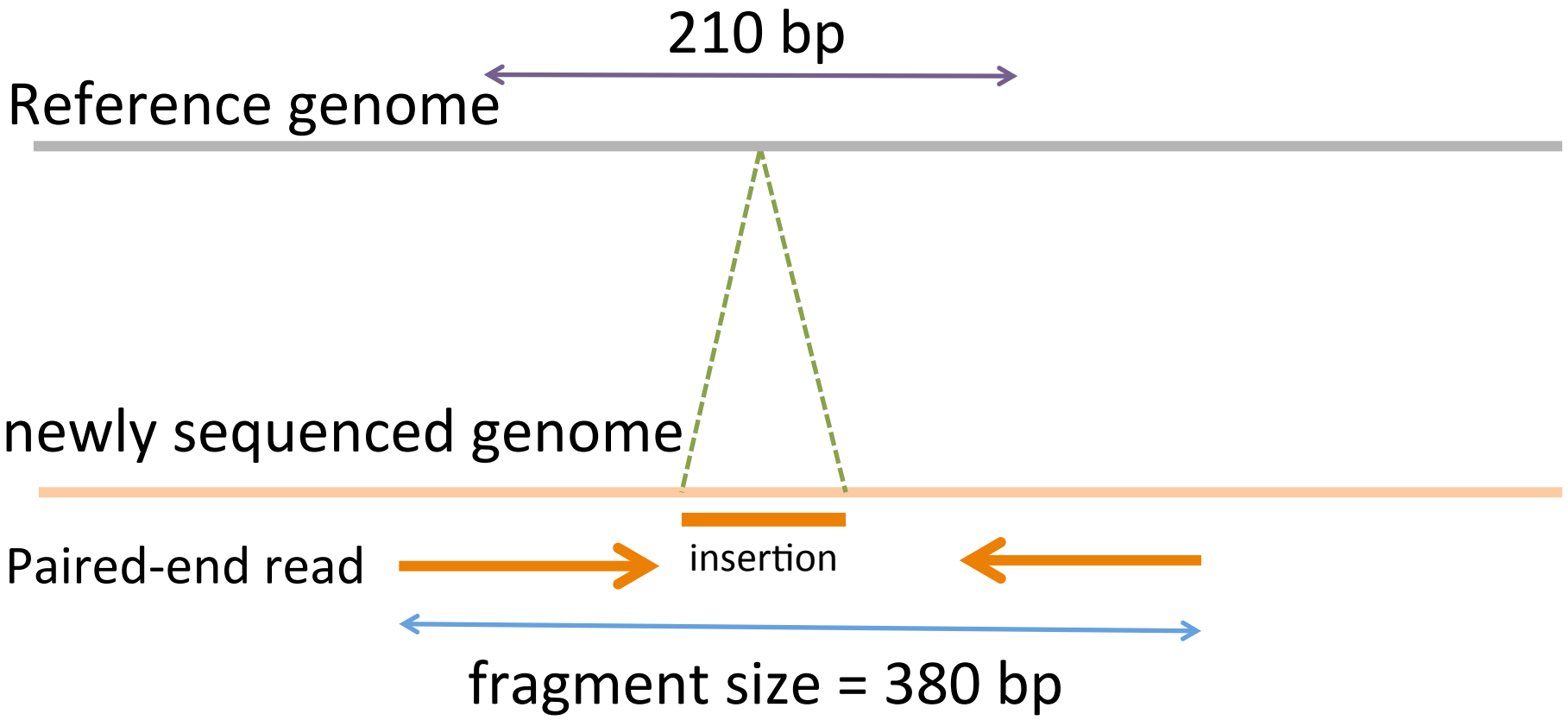
Paired-end reads to find "deletion" relative to the reference



Split reads to find "exact deletion sequence"

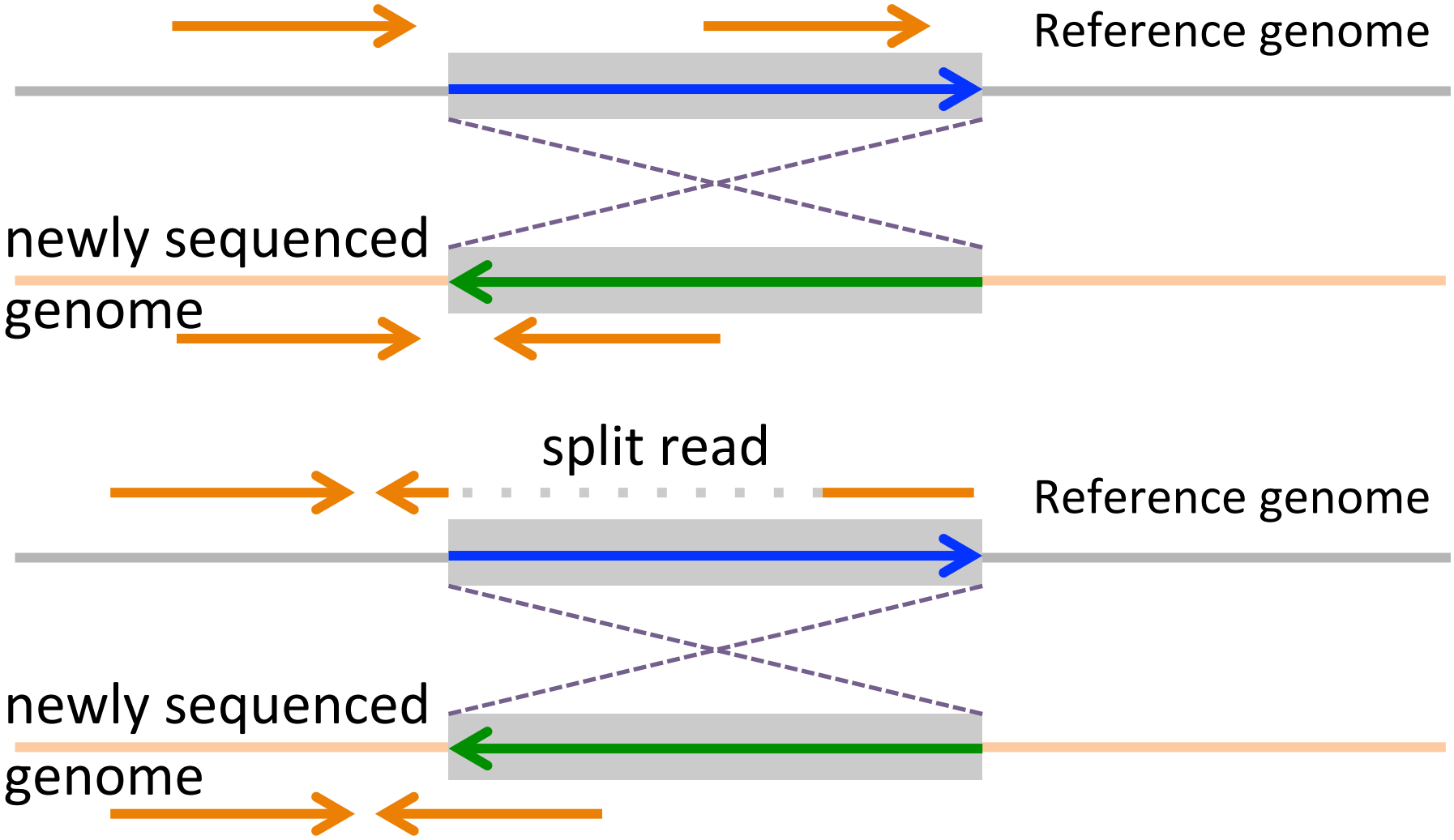


Paired-end reads to find "insertion" relative to the reference

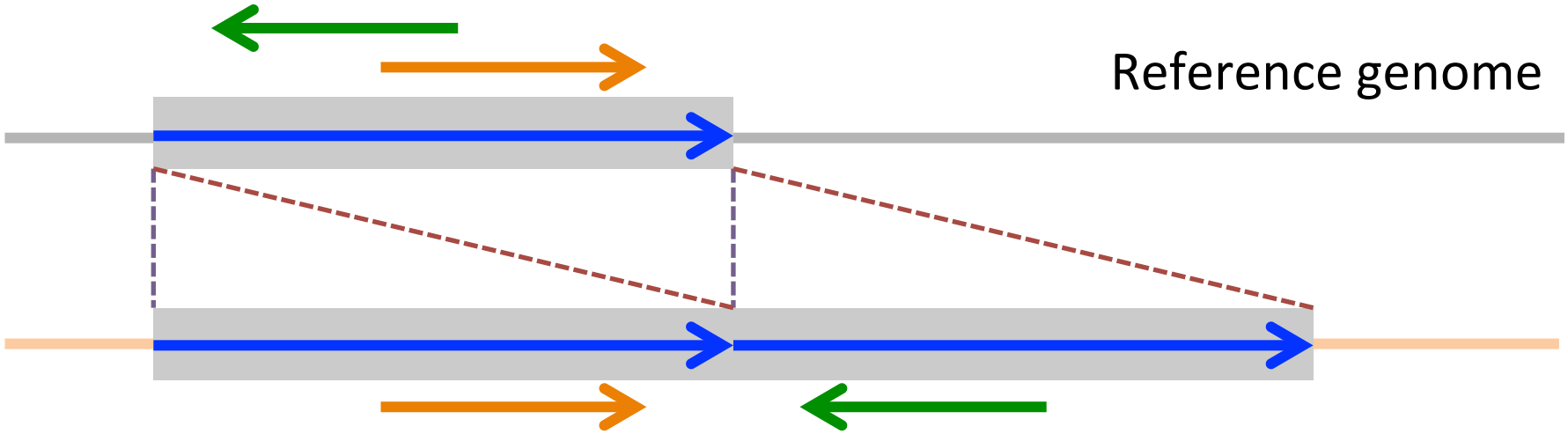


The size of insertions that can be identified by PE reads is determined by fragment sizes and read lengths.

inversion

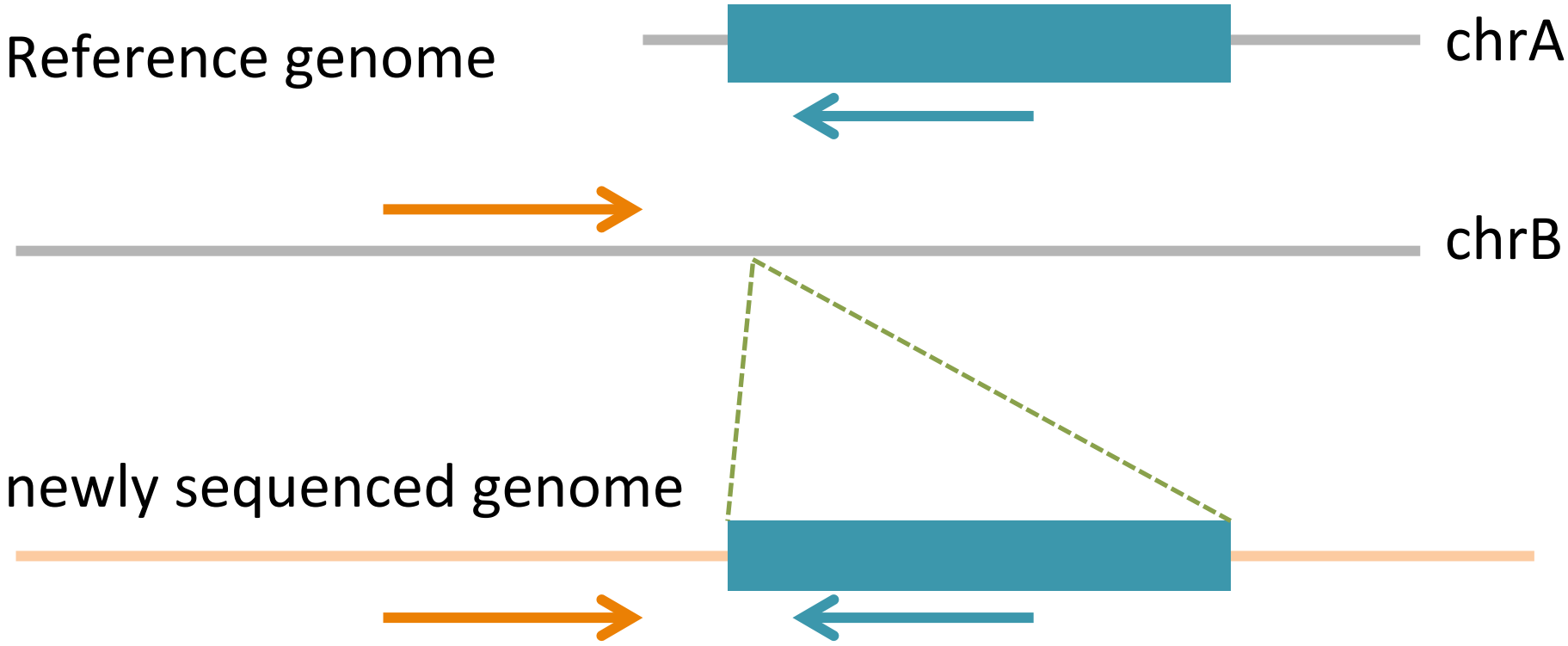

Reference genome
newly sequenced genome
split read
Reference genome
newly sequenced genome

Tandem duplication

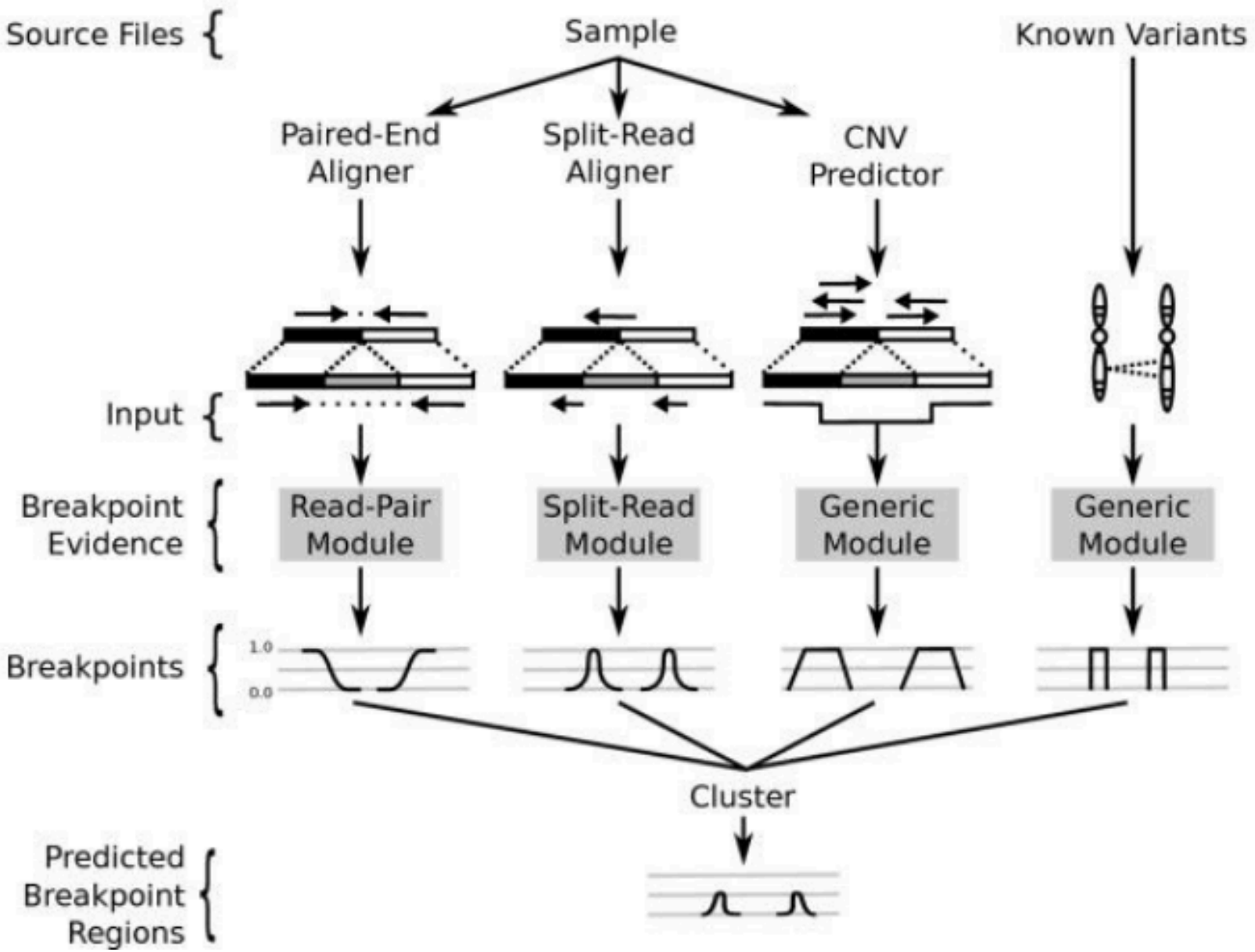


newly sequenced genome

Translocation



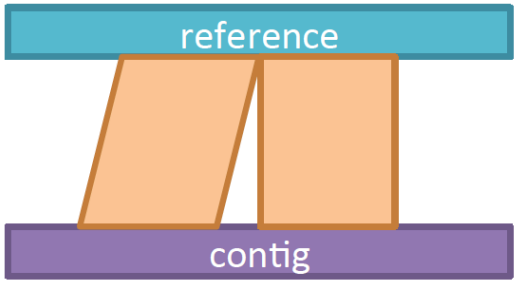
LUMPY: a integrative framework for SV discovery



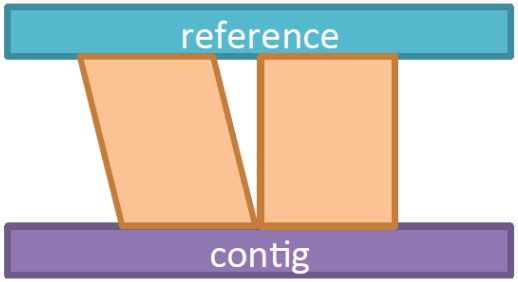
Genome assembly and genome-wide SV discovery

Defined point

Insertion

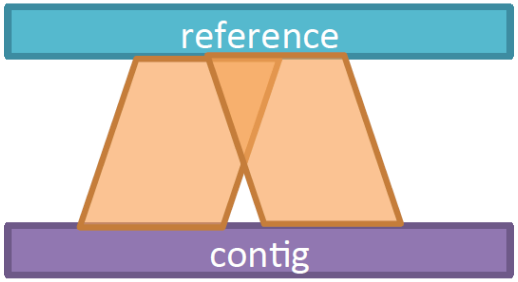


Deletion

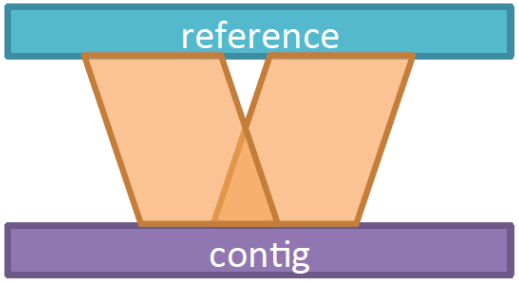


Overlapping alignments suggest tandem repeat

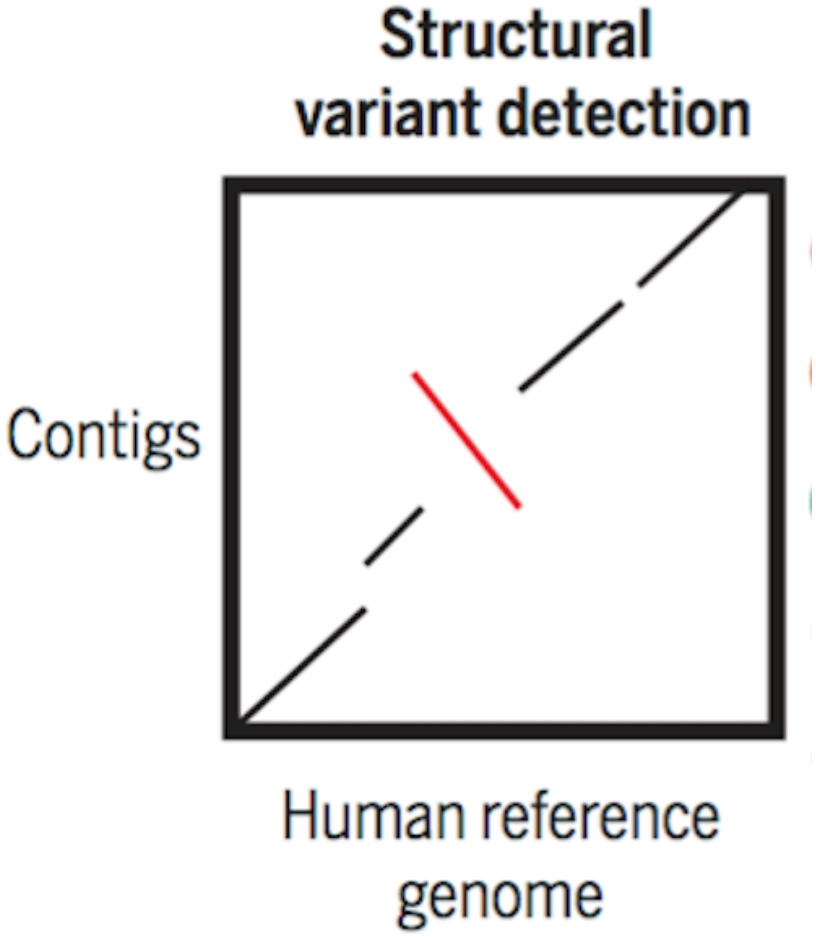
Tandem Expansions



Tandem Contractions



Genome assembly and macro-scale comparison

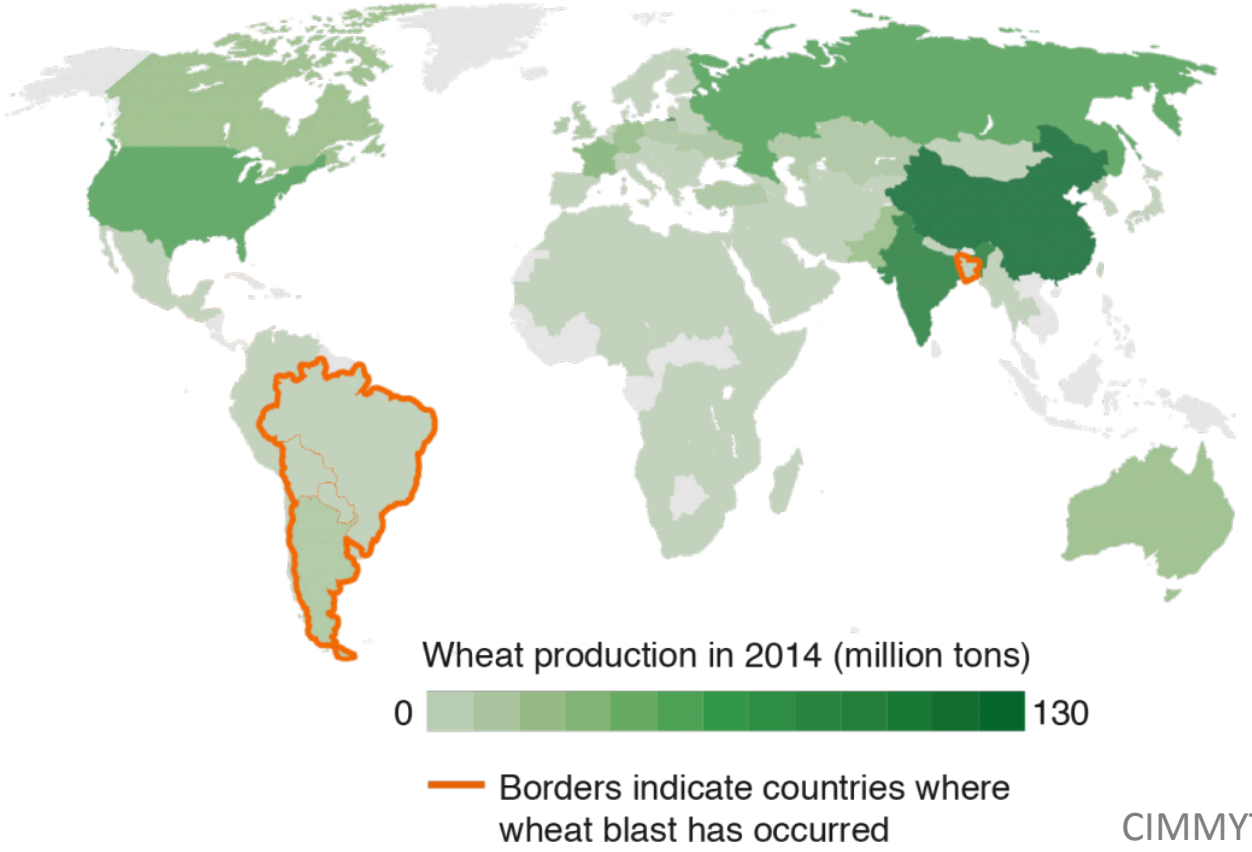


Wheat blast

Wheat blast *identified* in Brazil in 1985 ; *spread* to other countries in South America soon after that; *jumped* to Bangladesh in 2016.

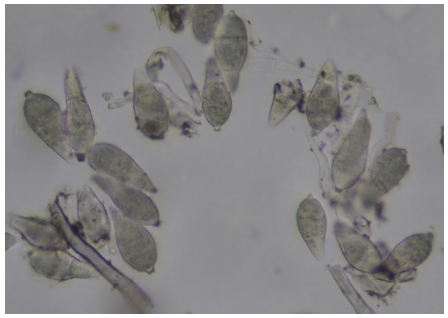
courtesy of Dr. Barbara Valent

Photo by Paritosh Malaker
Bangladesh Agricultural Research Institute

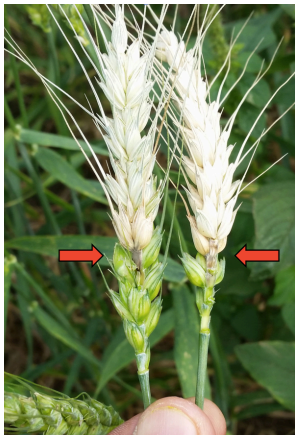


Wheat blast affected ~15% of total wheat area in Bangladesh

Wheat blast caused by the fungus *Magnaporthe oryzae*



M. oryzae



courtesy of Dr. Barbara Valent

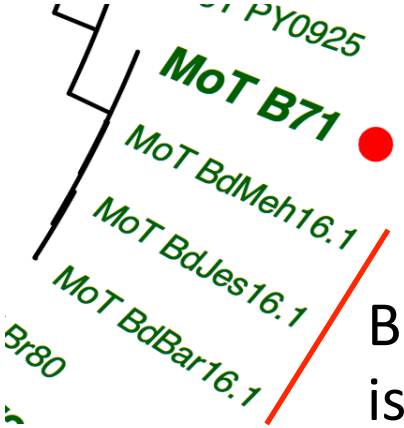
Warm rainy weather at heading results in 100% empty blasted heads.

Select B71 to generate the reference genome

Strain	Year isolated	Isolation location
B71	2012	Okinawa, Bolivia

B71 is an aggressive field isolate.

B71 is almost identical to isolates identified in Bangladesh 2016.

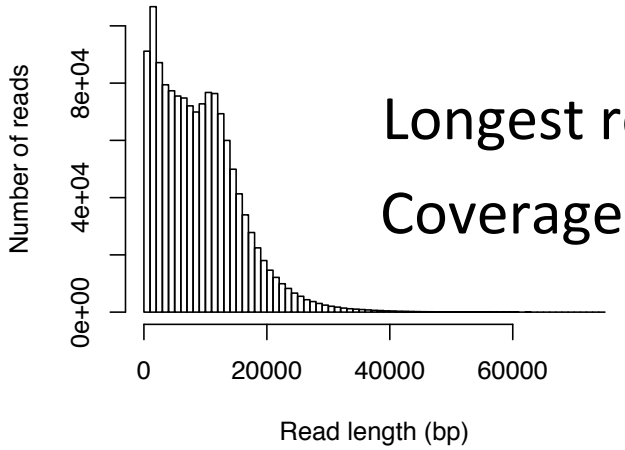


Bangladesh isolates

Genome sequencing for B71

- 10 SMRTCell PacBio data (P6-C4)

PacBio RS II



Longest reads: 74,474 bp

Coverage*: 276x

- PCR-free Illumina TruSeq data

HiSeq 2500

>2x10^7 pairs of 2x250 paired-end reads

Coverage*: 222x

* assuming the genome size is 45 Mb

Summary of the assembly

Statistics of the assembly*

Item	Statistics
total contig number	**31**
total contig length	44,522,920
longest contig	**7,902,655**
shortest contig	12,906
N50	**5,402,116**
overall GC	0.50
min contig GC	0.28

* *Canu* for PacBio assembly
Quiver polishing
Further error correction using Illumina data

mitochondial sequence

List of contigs (N=31)

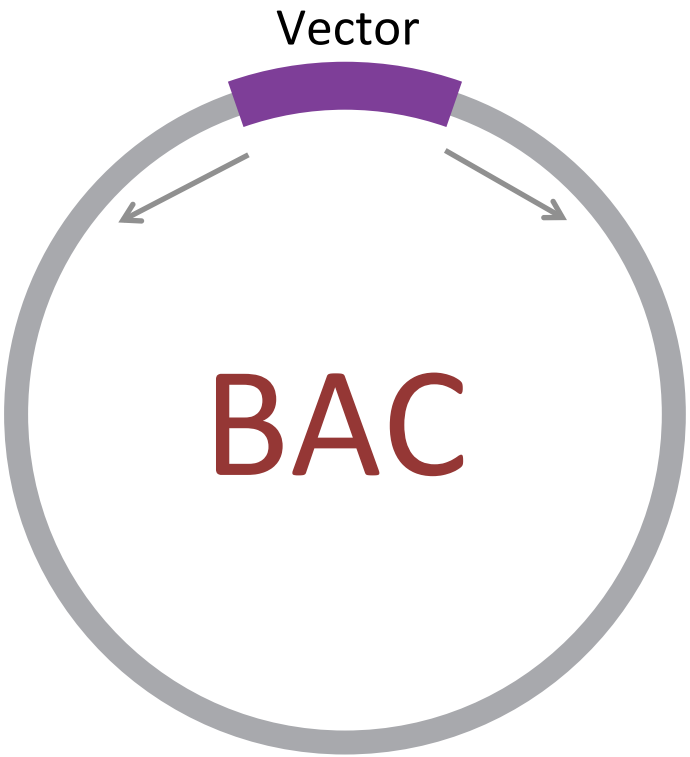
Contig	Length (bp)	GC%
tig00000000	7,902,655	51.2
tig00000024	7,531,155	49.1
tig00000030	6,090,985	50.2
tig00000003	5,402,116	51.2
tig00000011	4,657,194	49.2
tig00000004	4,442,877	51.2
tig00000005	4,042,640	49.2
tig00000001	1,778,515	46.4
tig00000014	461,599	49.4
tig00000025	398,015	44.3
tig00000039	253,502	45.1
tig00000026	240,676	43.8
tig00000018	237,459	47.4
tig00000016	184,678	48
tig00000007	138,010	49.9
tig00000022	110,918	46
tig00000020	78,737	49.6
tig00000006	69,533	51.1
tig00000033	69,131	49.2
tig00000043	57,199	28.4
tig00000015	51,862	48.5
tig00000008	47,134	47.6
tig00000031	42,261	48.6
tig00000041	36,641	52.5
tig00000023	36,455	44.2
tig00000037	35,037	48.9
tig00000032	32,608	49.7
tig00000017	28,099	51.7
tig00000038	27,162	49.5
tig00000019	25,161	51.2
tig00000027	12,906	47.3

BAC-end sequencing for assembly improvement?

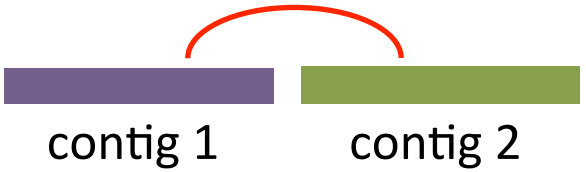
Vector

+ insert

long genomic DNA fragment (e.g., 200 kb)



scaffolding



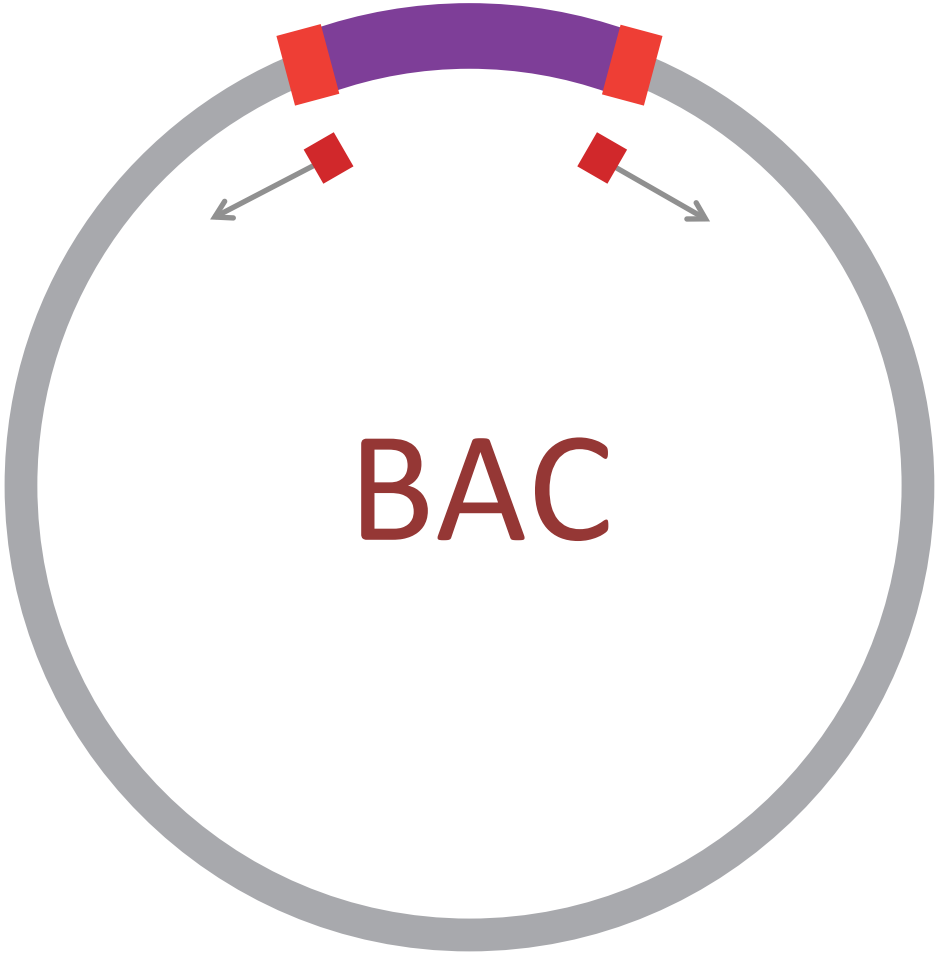
Sequence BAC ends of clones ***one by one***, which is very low-throughput and cost-inefficient.

Idea: build random barcodes in vectors to enable a high throughput (I)

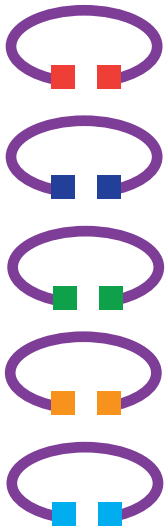
code (Red)

Vector with barcode (Red)

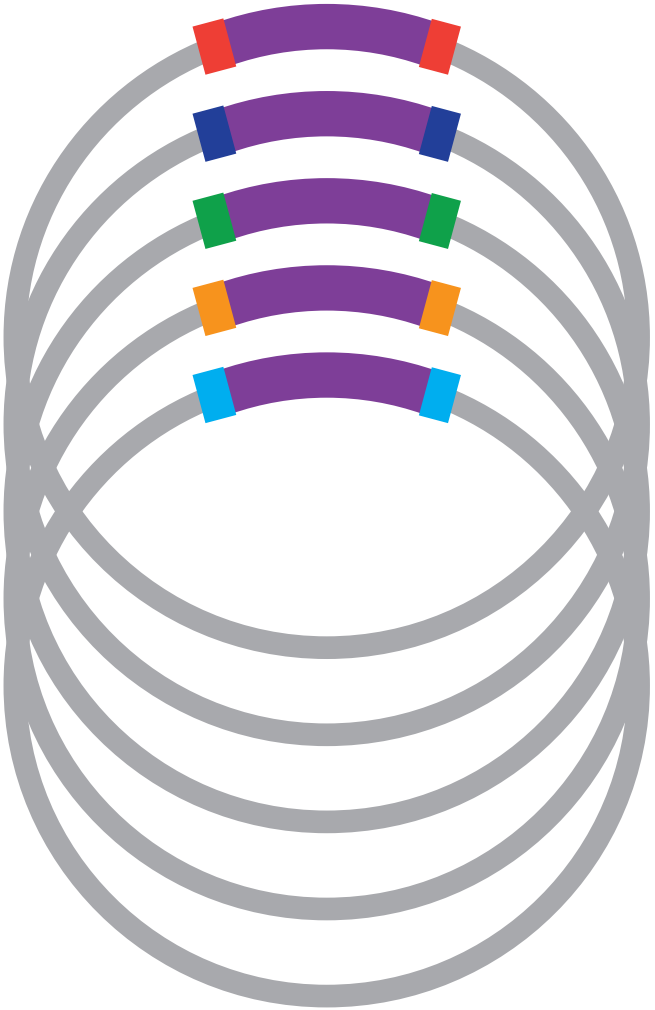
Sequences with barcode



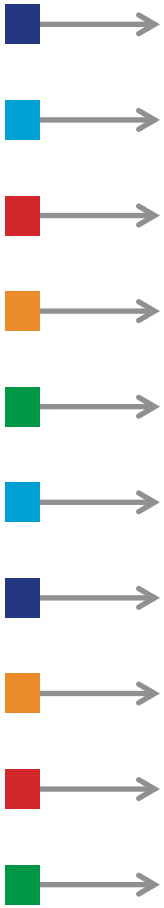
Idea: build random barcodes in vectors to enable a high throughput (II)



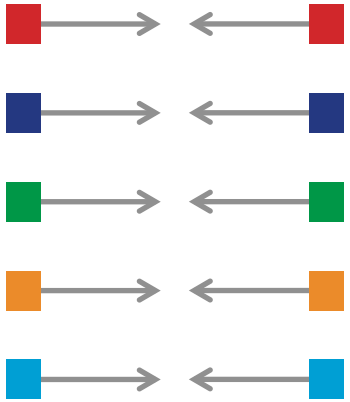
millions of barcoded vectors in a pool



tens of thousands of clones



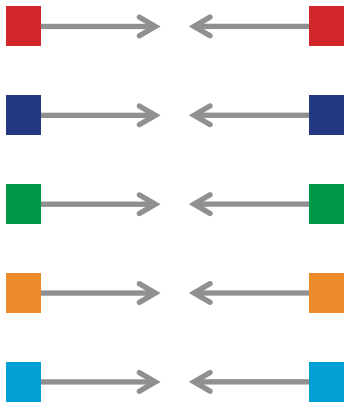
High-throughput end sequencing



Paired ends based on barcodes

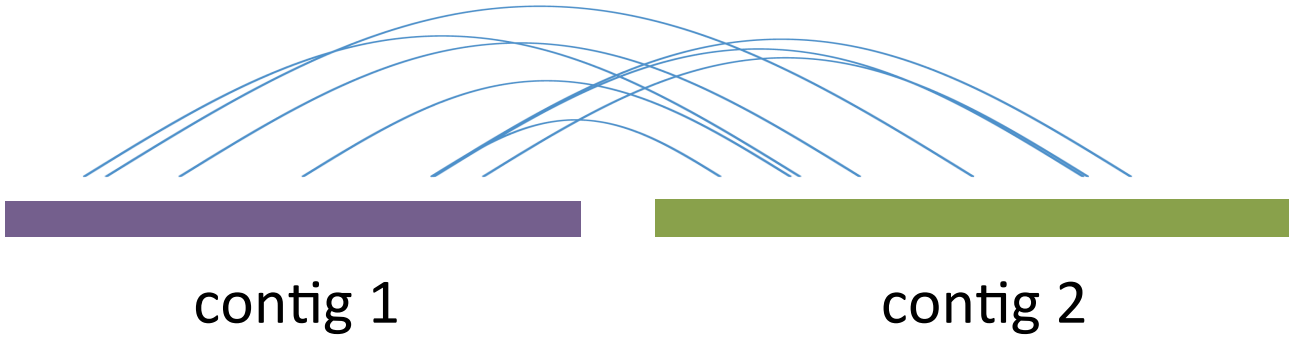
LIEP Scaffolding to improve the B71 assembly

Long Distance End Pairs (LIEP)



~95% are long-distance read pairs

LIEP scaffolding

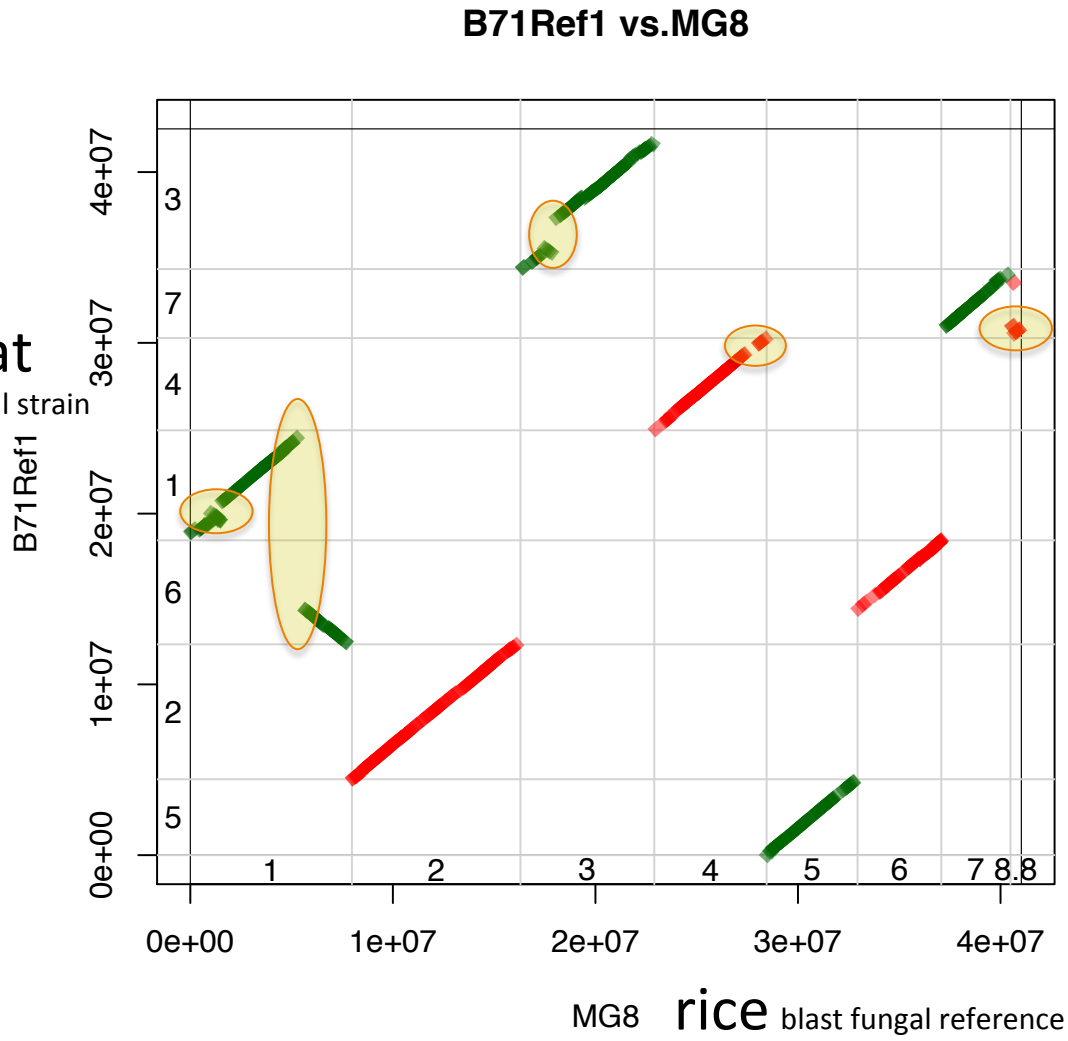


scaffolding two contigs to a sequence

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

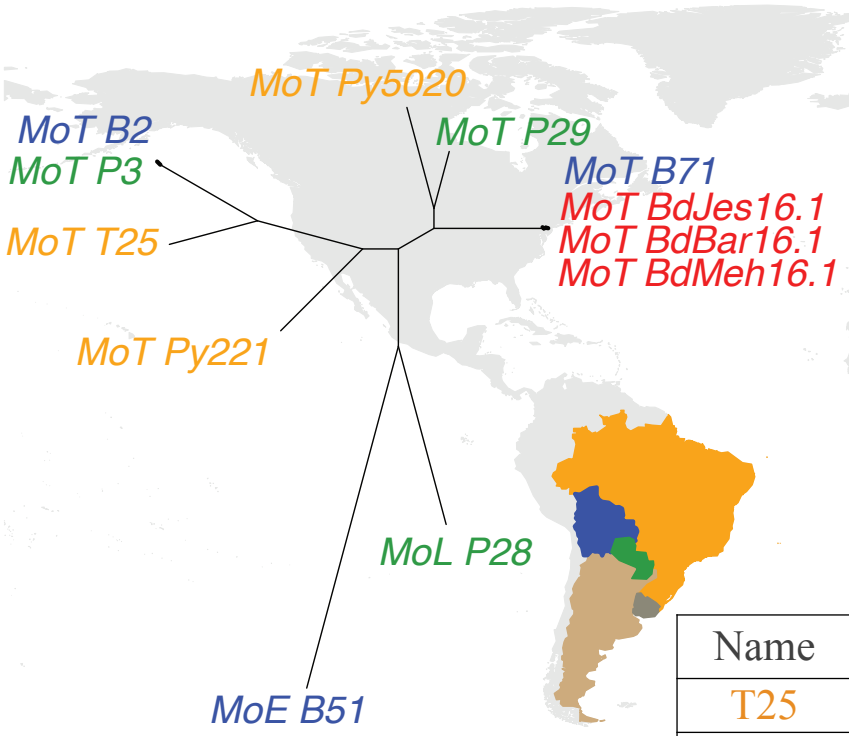
Chr	Length (bp)
chr1	6,442,091
chr2	7,902,655
chr3	8,206,304
chr4	5,402,116
chr5	4,442,877
chr6	6,090,985
chr7	4,042,640
scaf1	941,816
scaf2	739,928
scaf3	104,670
scaf4	74,396
scaf5	69,131

Wheat
blast fungal strain



>10 kb overlap and >95% identity

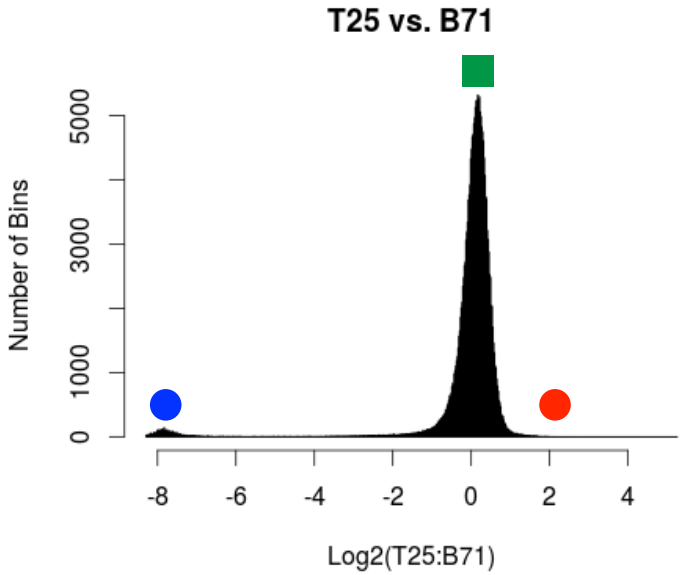
Illumina sequencing of additional eight strains



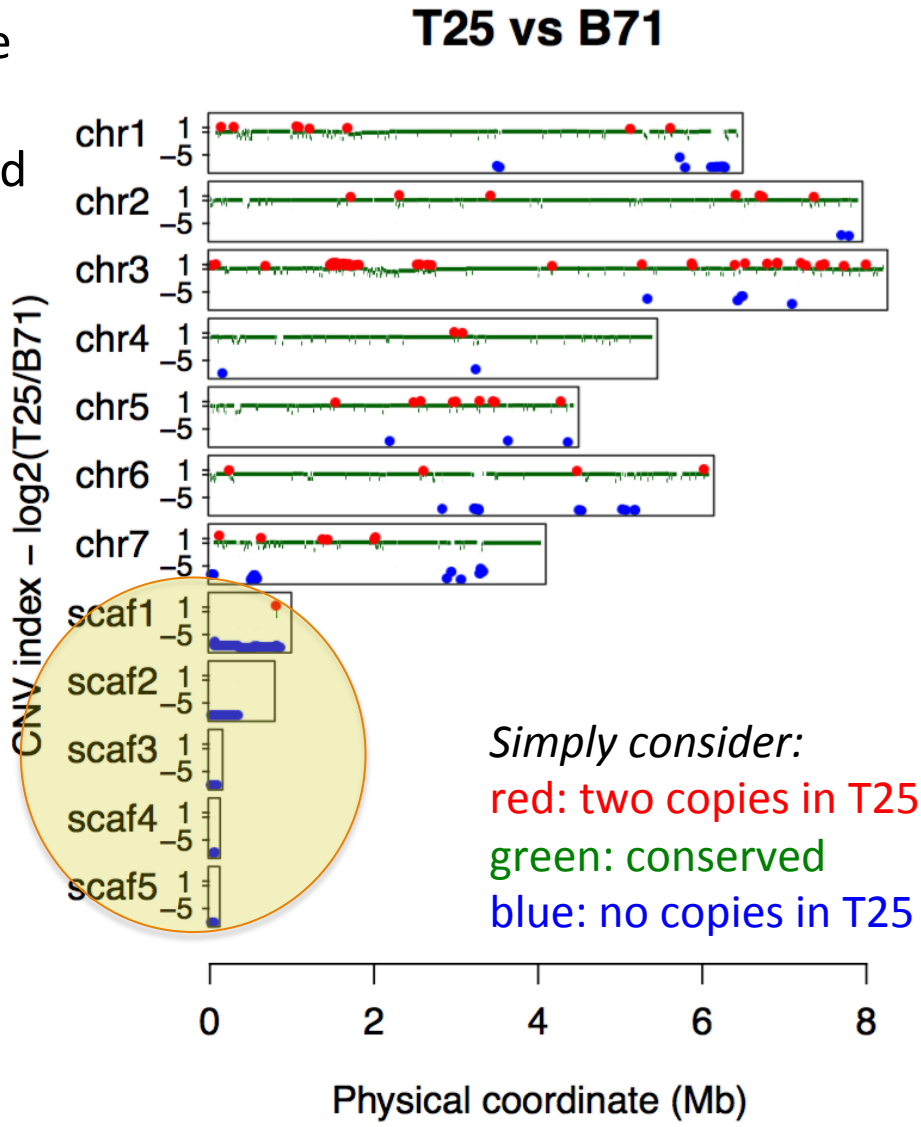
Name	Year isolated	Host species	Isolation location
T25	1988	*Triticum*	Paraná, Brazil
Py5020	2005	*Triticum*	Paraná, Brazil
Py22.1	2007	*Triticum*	Paraná, Brazil
B2	2011	*Triticum*	Quirusillas, Bolivia
B71	2012	*Triticum*	Okinawa, Bolivia
P3	2012	*Triticum*	Canindeyú,Paraguay
P28	2014	*Bromus*	Paraguay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

Read depths to infer *copy number variation*

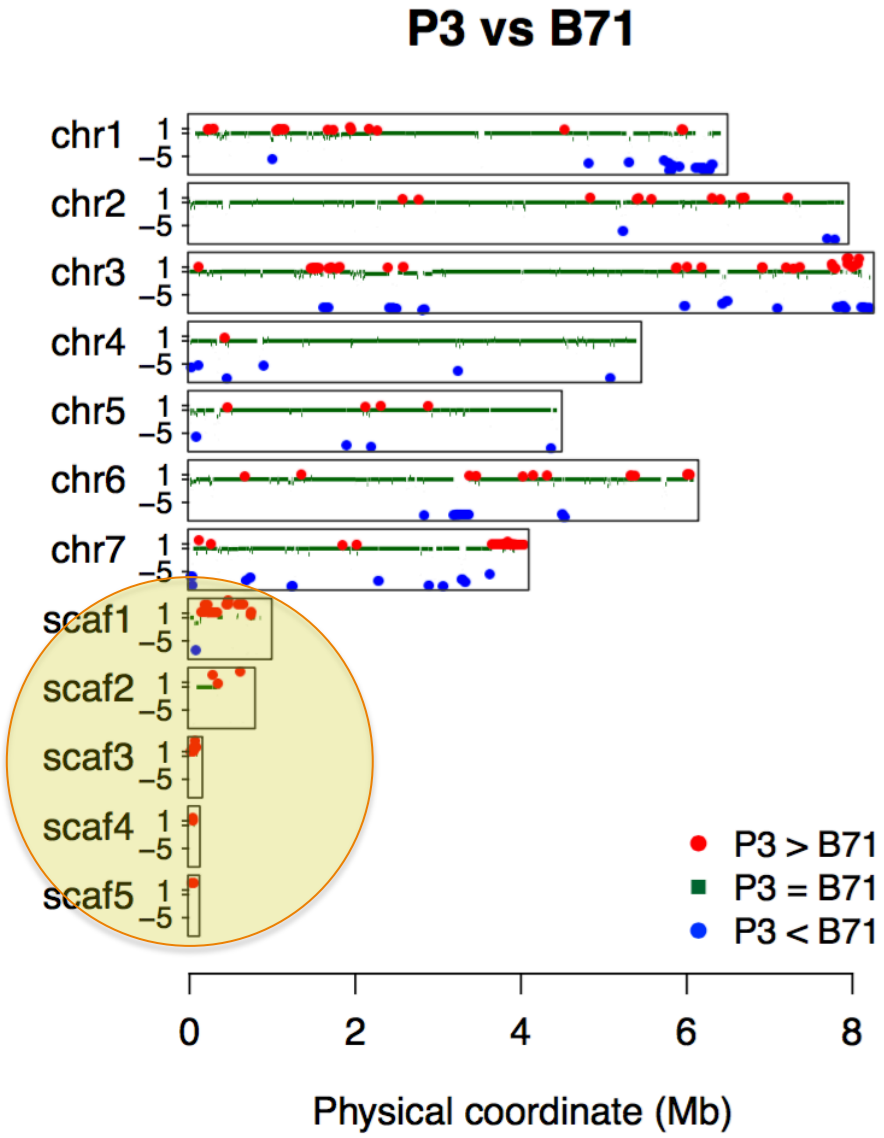
1. Aligned reads to the reference genome
2. Divided genome divided to bins
3. Determined counts of uniquely mapped reads per bin
4. **Compared read depths in T25 *vs* B71**



5. Merged neighboring bins with similar signals into **segments**



P3 vs. B71



Scaf1-5

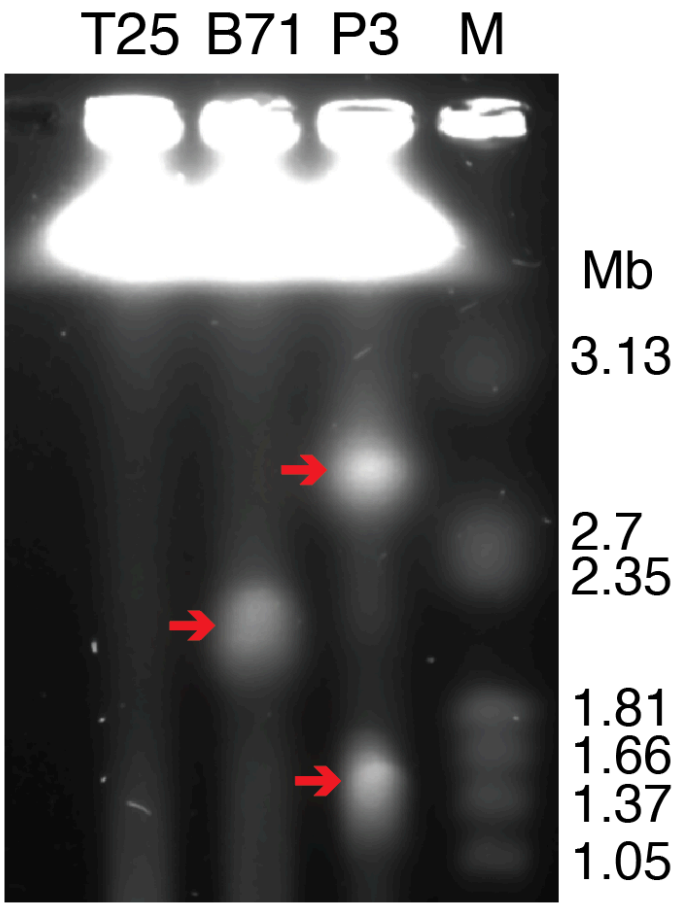
- Highly dynamic
- Highly repetitive
- Not anchored to chr1-7

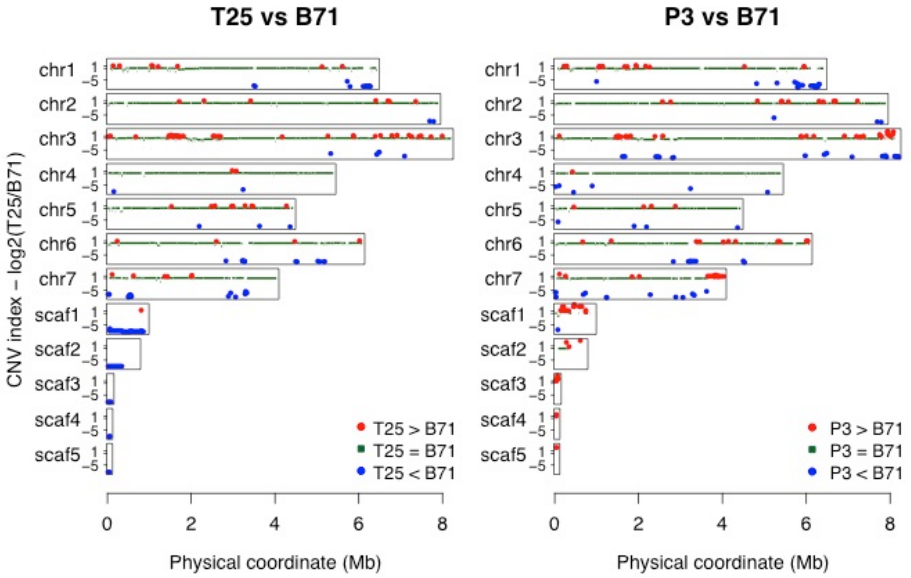
of a rice strain (*dispensable*)

Dispensable mini-chromosome?

extra chromosome; supernumerary chromosome

CHEF gel to separate chromosome-sized DNA



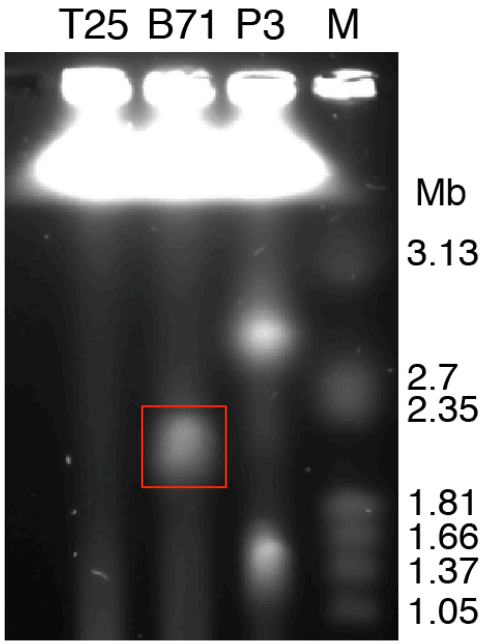


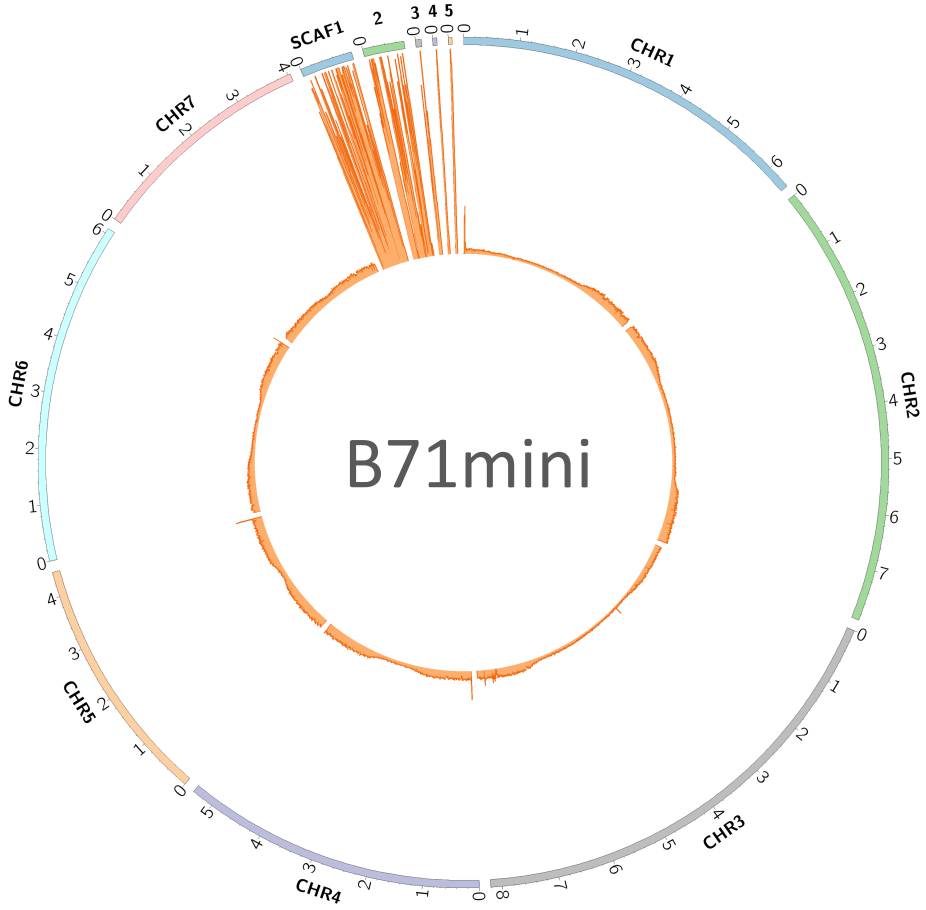
Are some of the scaf1-5 or all from the mini-chromosome?

T25 has no mini-chromosomes

B71 has one and P3 has two

Illumina sequencing of mini-chromosomes (mini) recovered from CHEF gels





1. DNA recovered from gels
2. Whole mini sequencing
3. Alignment to the B71 genome
4. Genomic coverage and depth of uniquely mapped reads

All five scaffolds (scaf1-5) are from the mini.

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).