

to Graph Pangenomes

- introduction
- genome evolution with linear assemblies
- graph pangenome approaches
- perspectives

why genome variation?

genotypes

phenotypes

individuals

GTCGTGTACGTCAGTCGTGCTAGTCAG
 TCGTCATTGACTGCAGTCAGTCAGTCA
 GTCGTGTACGTCAGTCGTGCTAGT
 CGGCTAGTCATCGCCAAATCATGCGTT
 GACTGCAGTTCTGACGTGTACGTCAGT
 CAGTCGAGTCCGTTGACGTCATGCAGT
 CGTGTACGTCAGTCGTGCTAGTCAGTC
 GTCATTGACTGCACCAGTCAGTCAGTC
 AGTCGTGTACGTCAGTCGTGCTAG
 TCGGCTAGTCATGACGTACCAGTGTCA
 TGCGTGTACGTCATGCGATGTCAGTCA
 GTACTGTCAGTCAGTACGTCATCGCCA

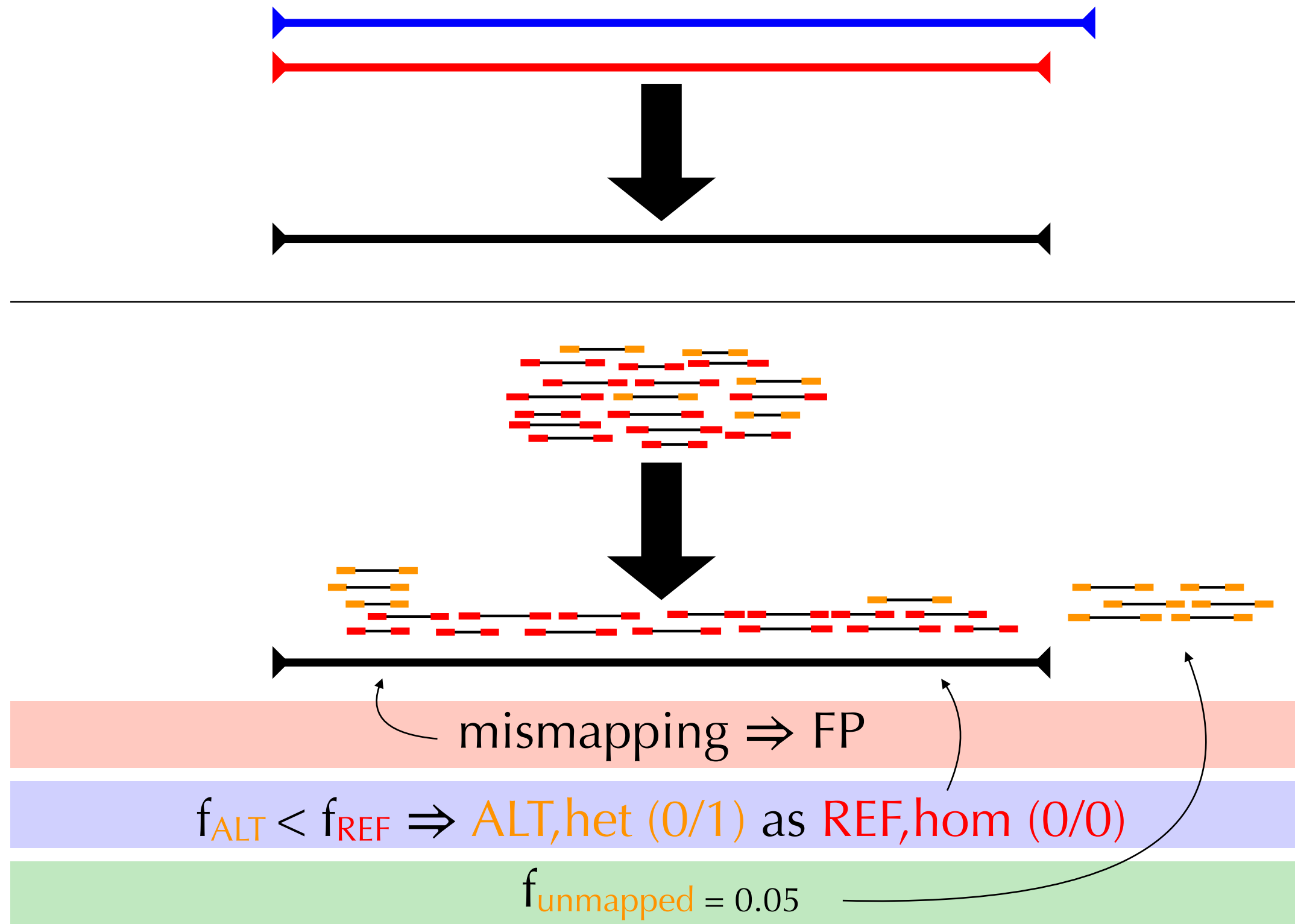
GGCGTGTACGTCTGTCGTGCTAGTCAG

reference



the reference bias

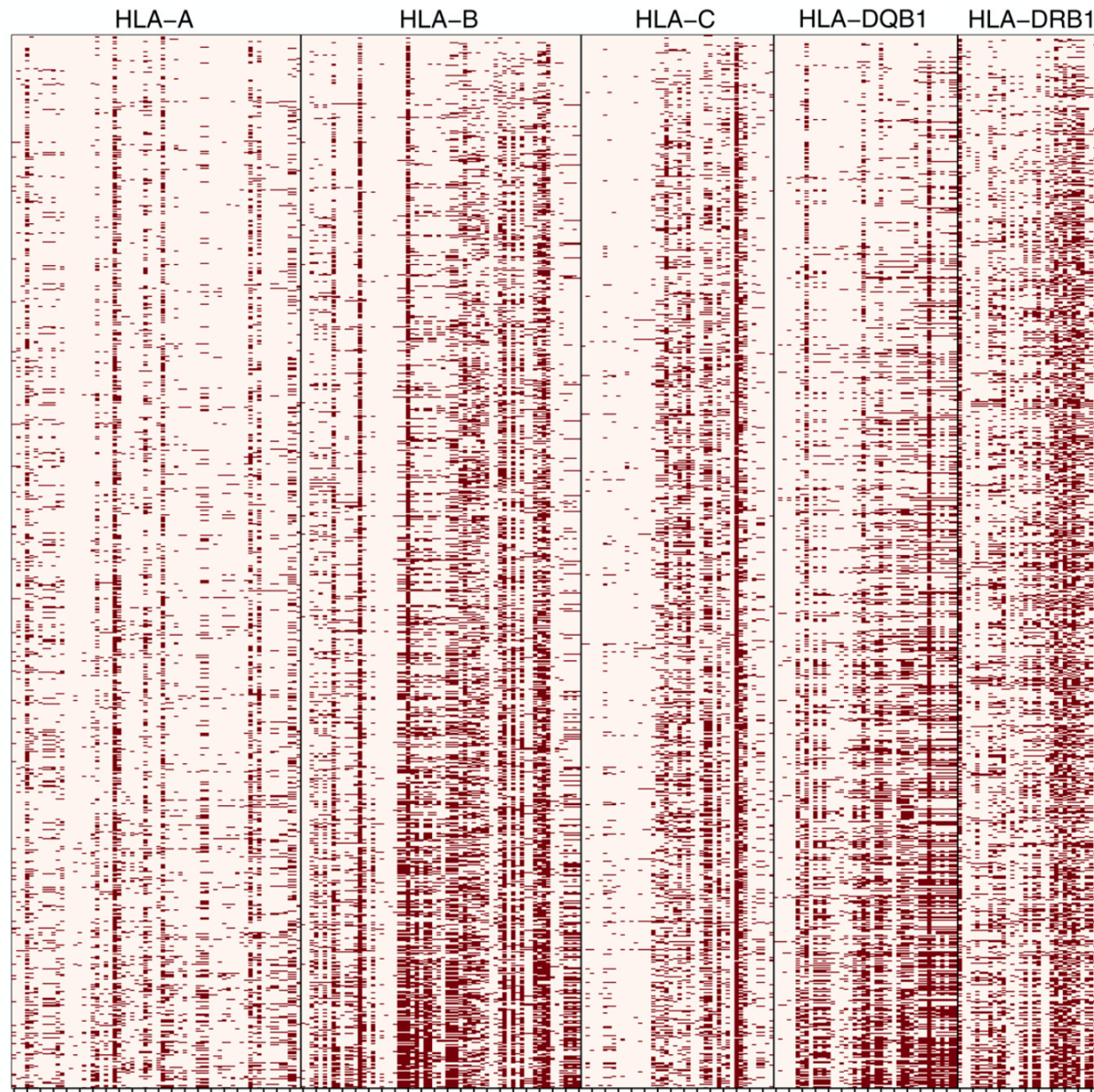
origin and consequences



the clinical reference bias

genotyping: SGS vs Sanger sequencing

individual

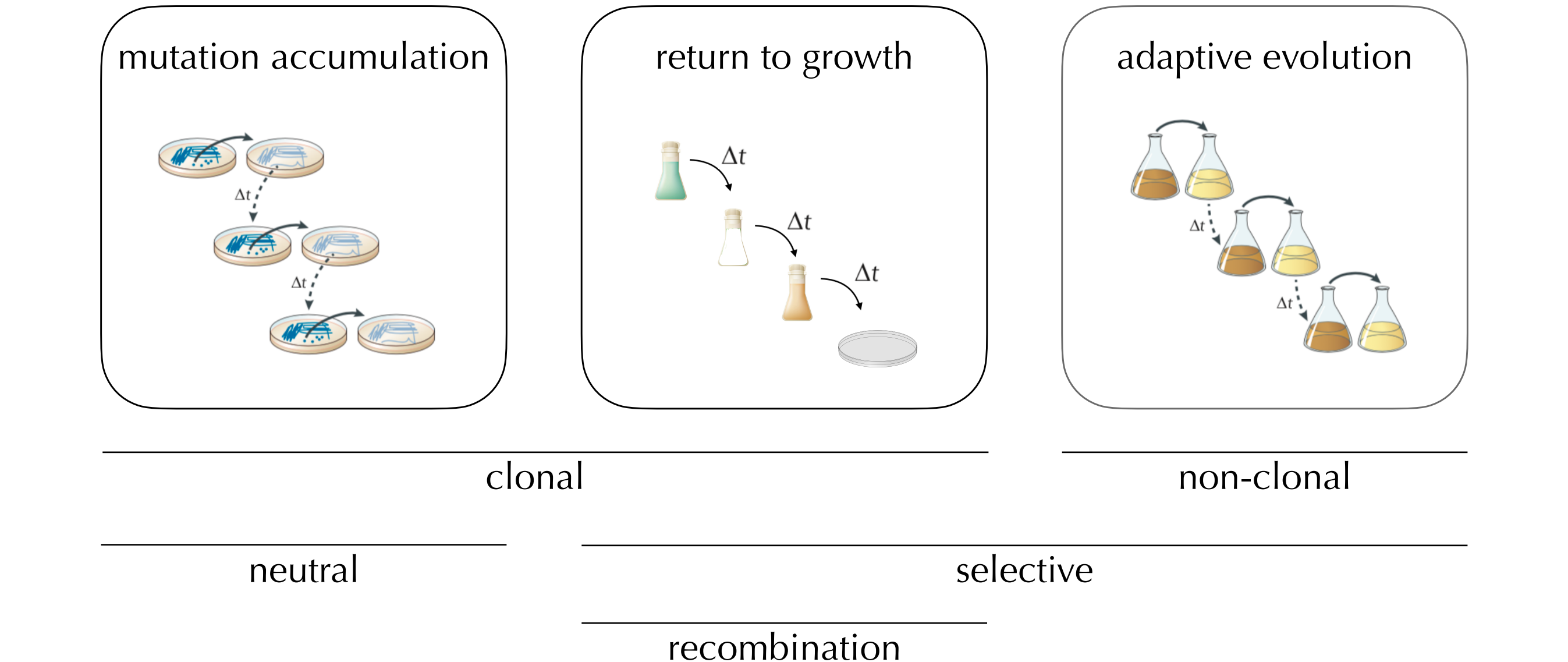
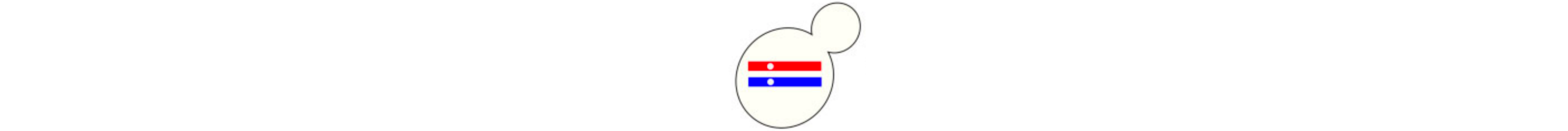


legend:
match
mismatch

locus

pushing the heterozygosity

genome variation in yeast hybrids using telomere-to-telomere assemblies



which assembly?

Tattini L, Maret T, ..., Liti G & Falque M. In preparation

Jain A, Tattini L, ..., Liti G & Coudreuse D. In preparation

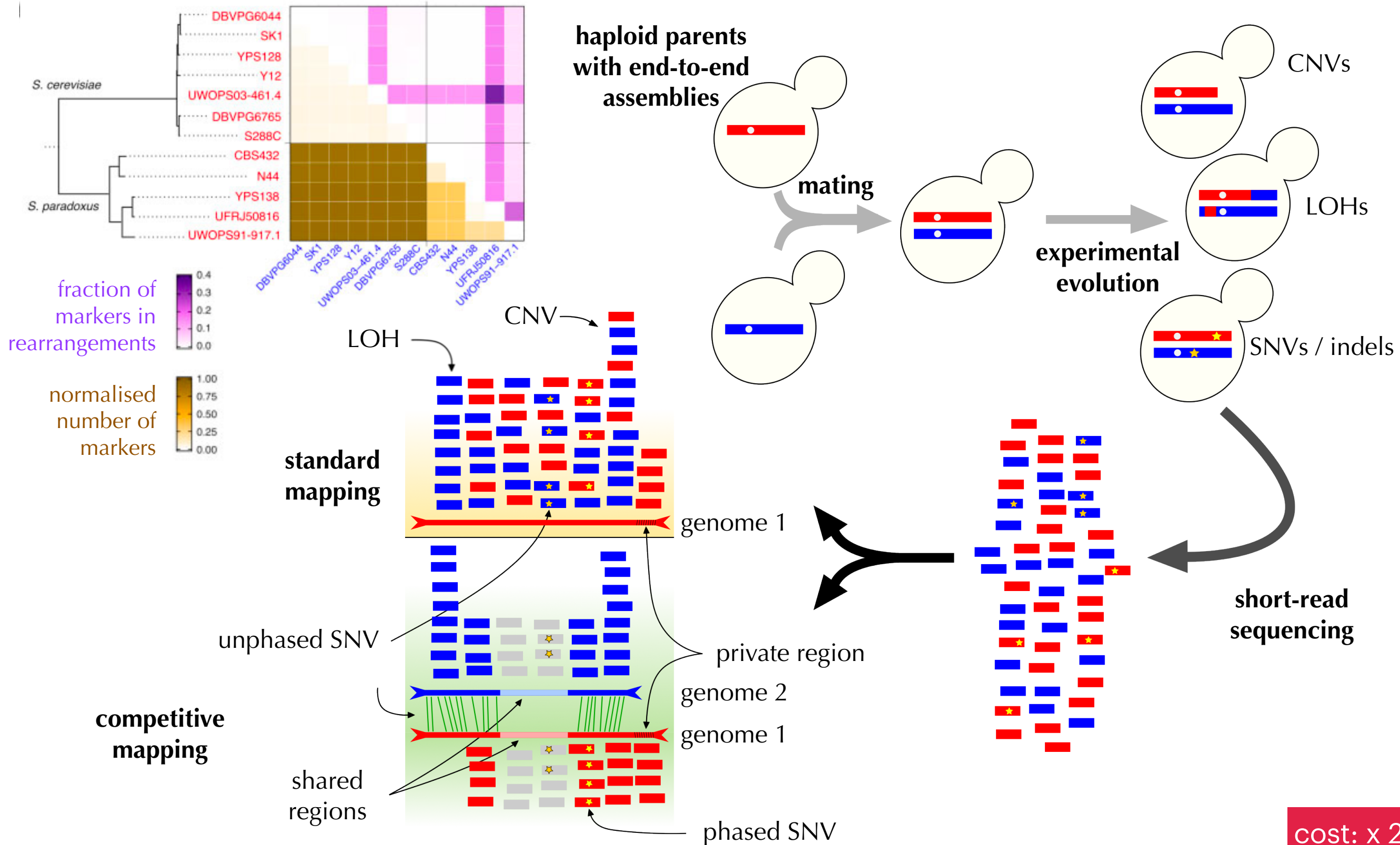
Tattini L, Mozzachiodi S, Tellini N, ..., & Liti G. In preparation

Mozzachiodi S*, Tattini L*, ..., & Liti G. Aborting meiosis allows recombination in sterile diploid yeast hybrids. NAT COMMUN. 2021.

Tattini L@, ..., & Liti G@. Accurate tracking of the mutational landscape of diploid hybrid genomes. MOL BIOL EVOL. 2019.

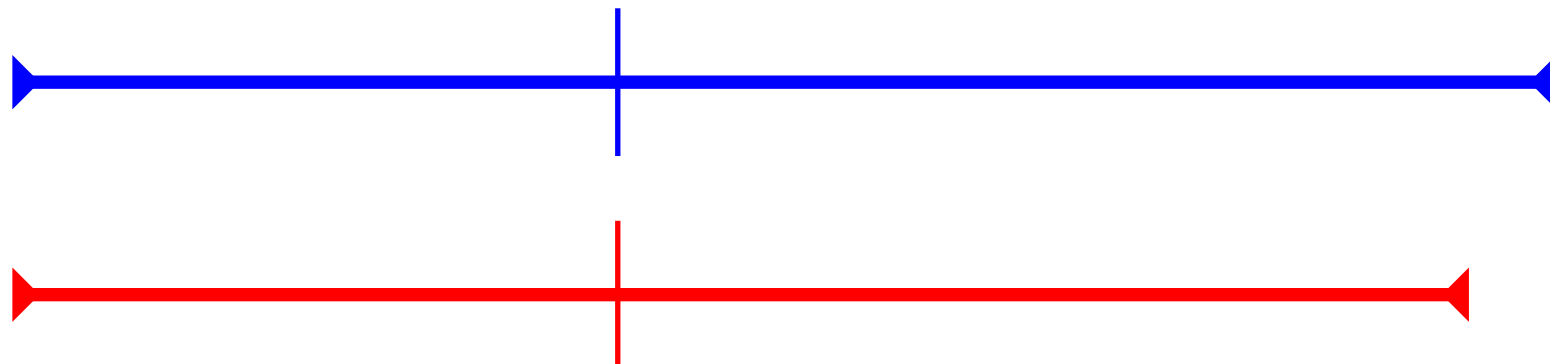
competitive mapping

a solution for mutation accumulation lines in hybrids



what's a variation graph?

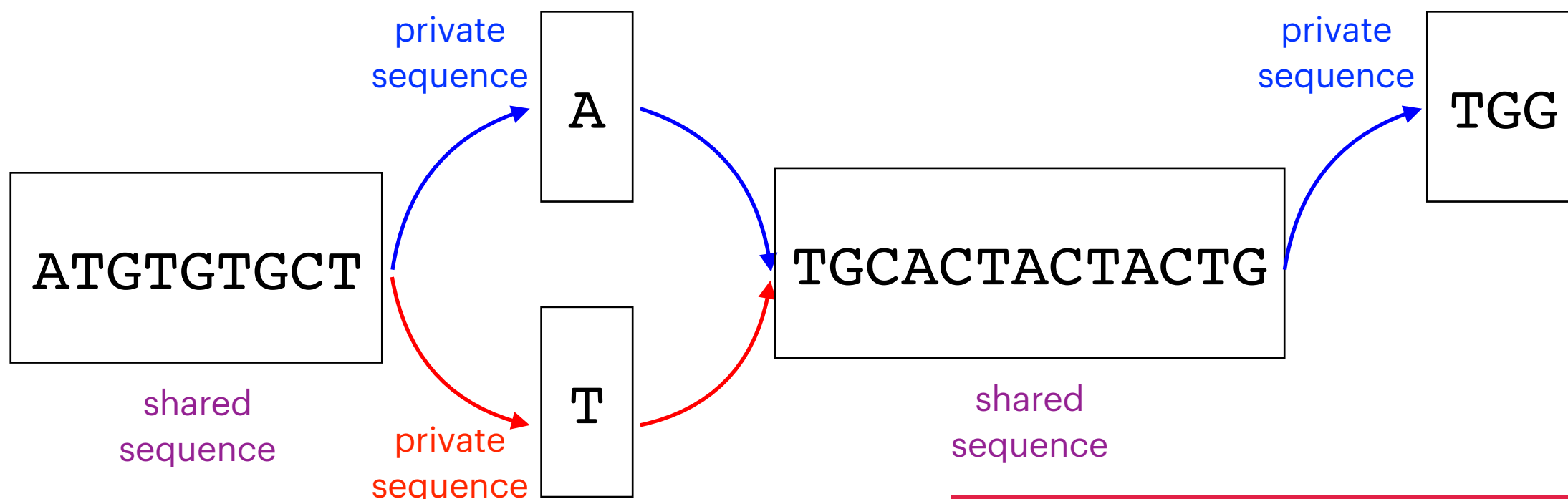
assemblies



alignment

ATGTGTGCT	A	TGCACTACTACTG	TGG
ATGTGTGCT	T	TGCACTACTACTG	

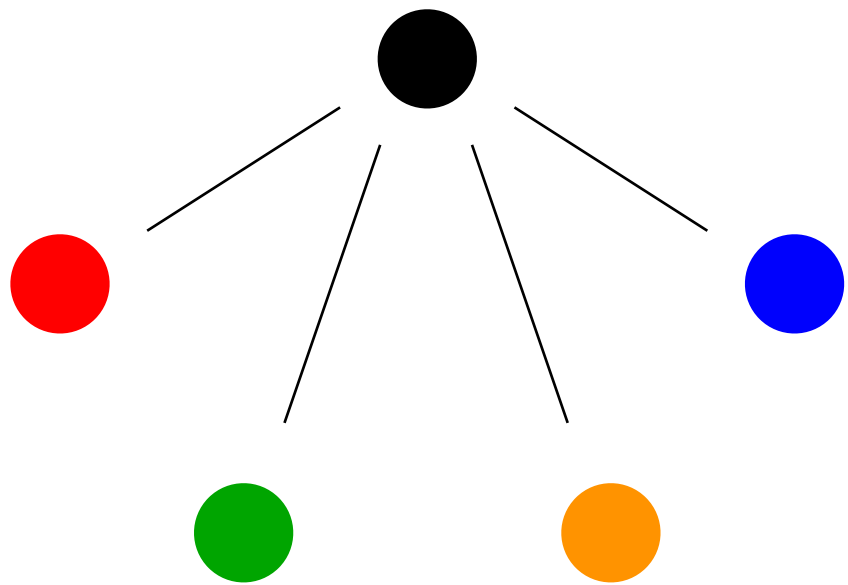
graph



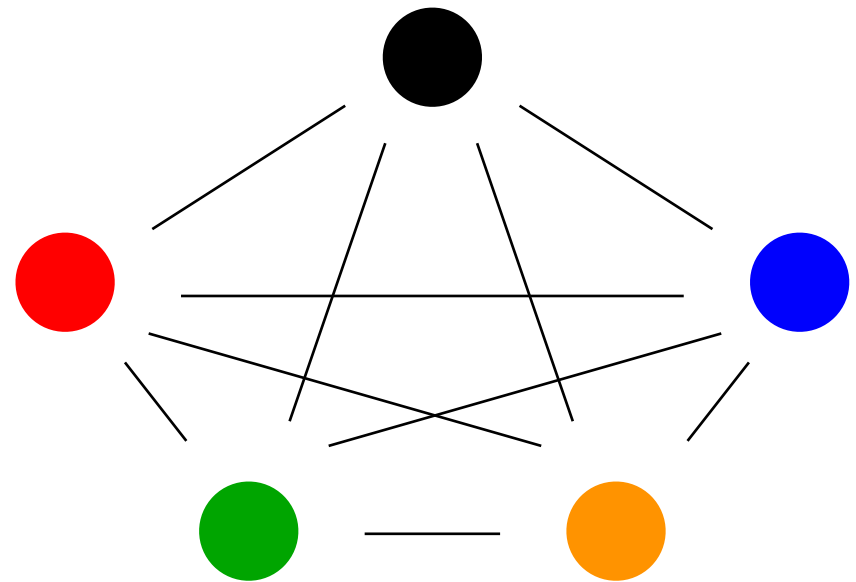
shared sequences appears only once

conceptual aspects

linear reference



all-vs-all



building a variation graph

reference + VCF

long reads



materials & methods

- pgggb
- vg
- odgi
- impg
- distBWT2

- sgp-eval
- panda



reference + VCF

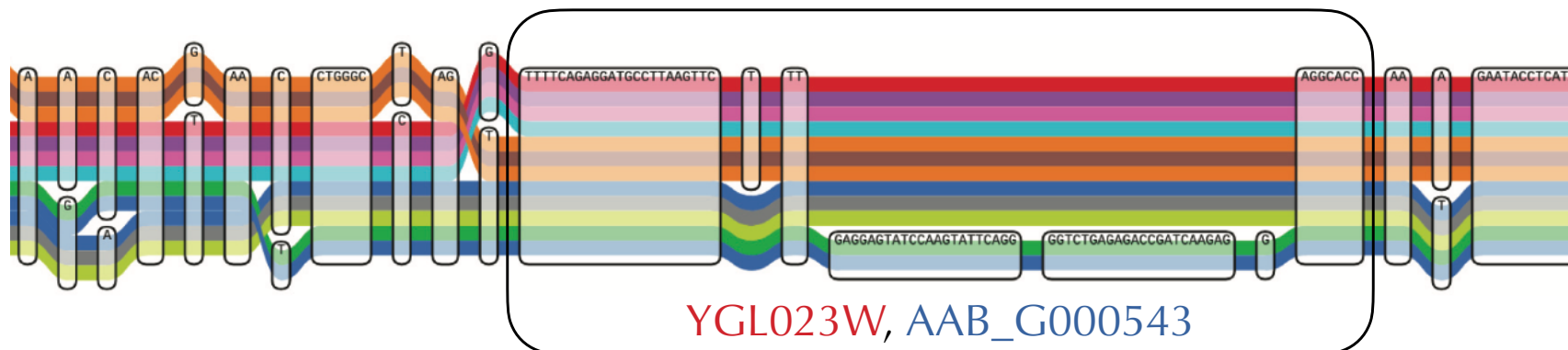
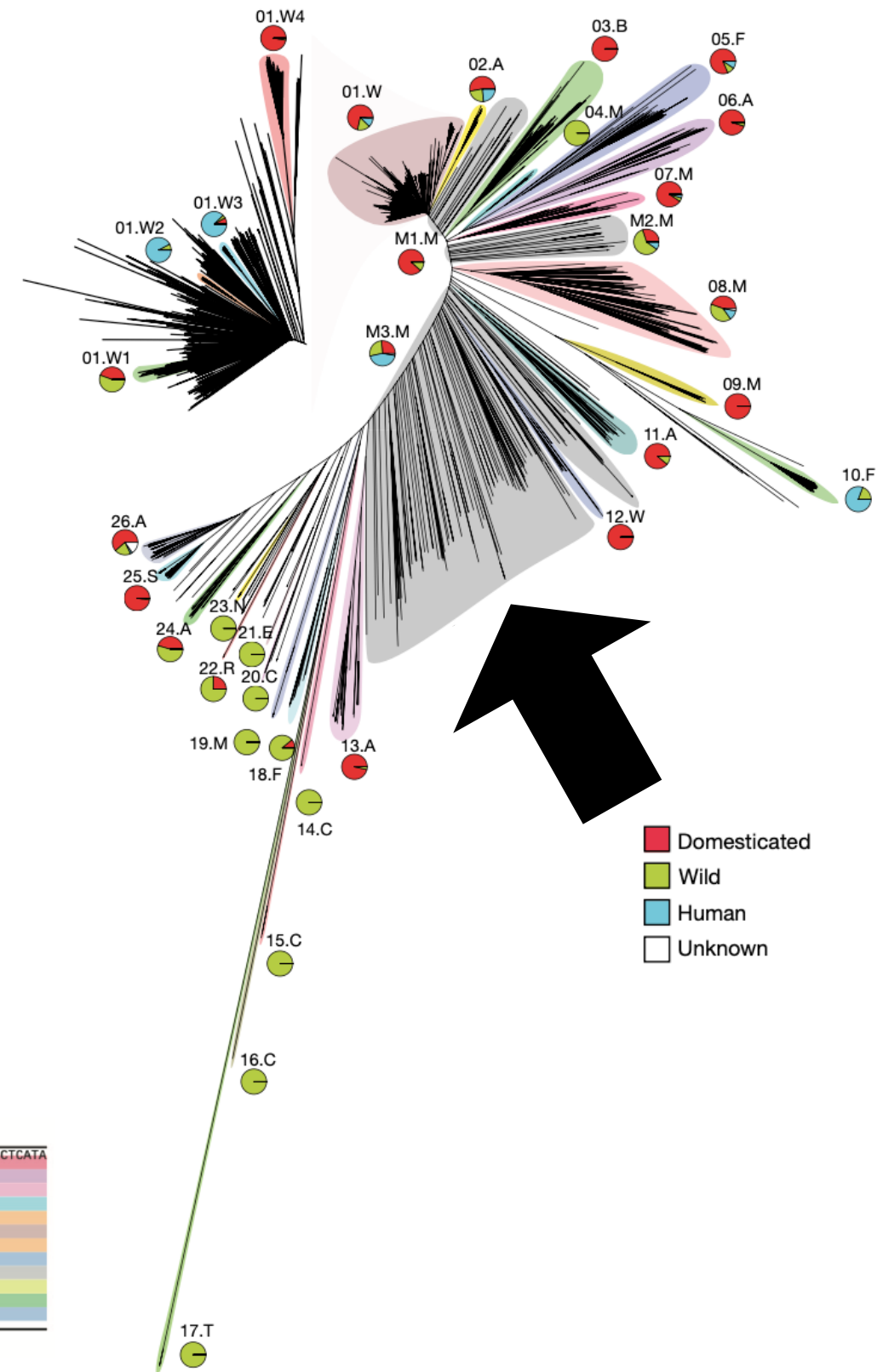
long reads

investigations on the scrap graph

- pangenome (annotations)
- chromosome-ends
- origin
- admixture
- pangenome (variants)
- phylogeny



SGP
VA



the *Saccharomyces* graph phylogeny evaluation tool

A diagram of a star graph. It features a central black circular node at the top. Four straight black lines radiate from this central node to four peripheral circular nodes. The peripheral nodes are colored red (left), blue (right), green (bottom-left), and orange (bottom-right).

variants

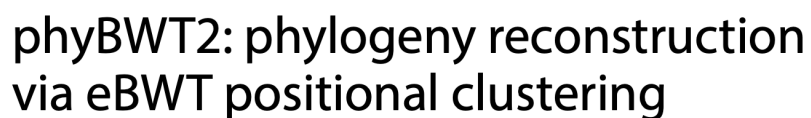
A diagram of a complete graph K_5 with 5 nodes and 10 edges. The nodes are arranged in a regular pentagon pattern: a black node at the top, a red node at the middle-left, a blue node at the middle-right, a green node at the bottom-left, and an orange node at the bottom-right. Every node is connected to every other node by a straight line edge, resulting in a total of 10 edges.

phyBWT2/distBWT2

cosine distance + neighbour-joining algorithm

Algorithms for Molecular Biology

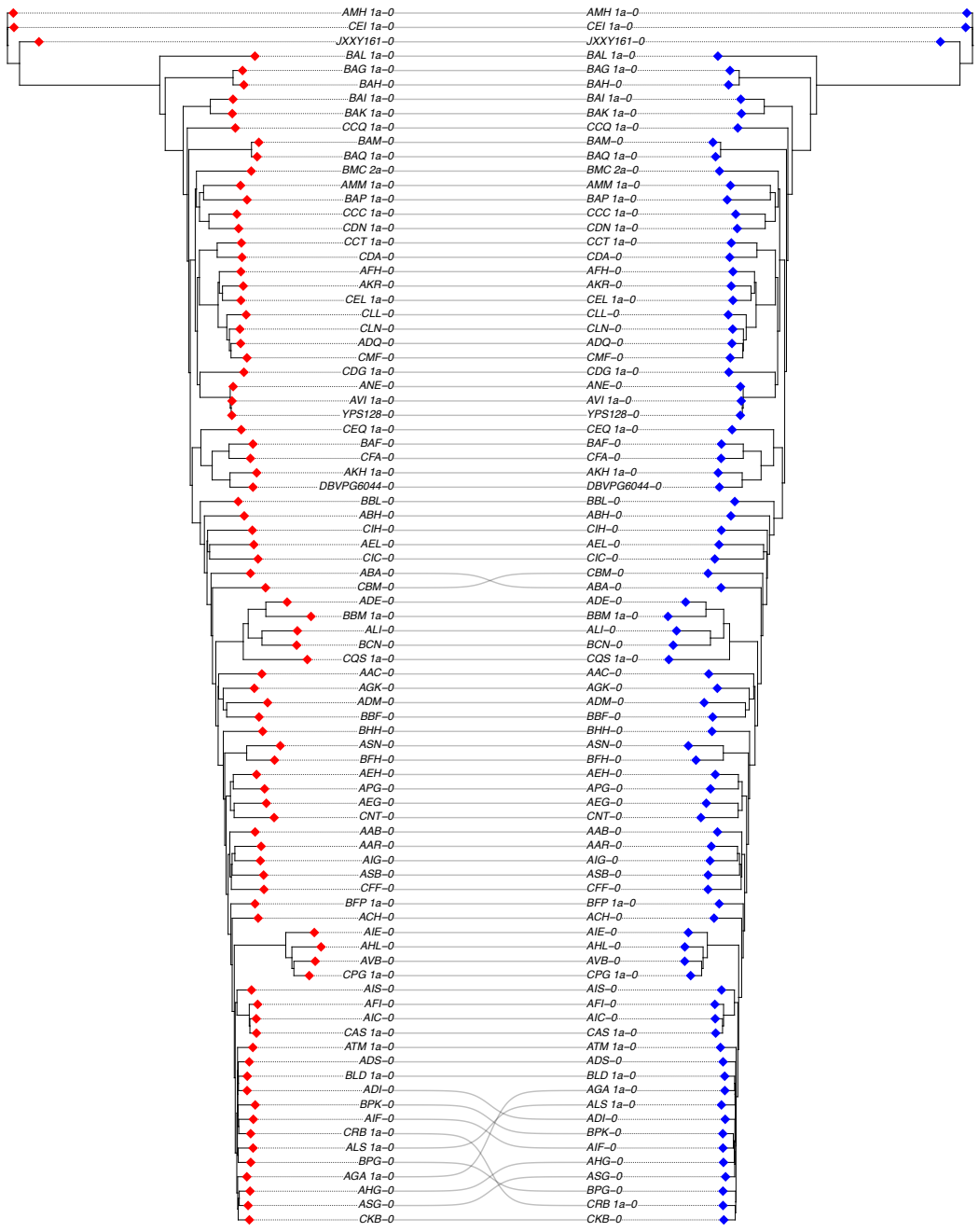
Open Access



cosine distance + neighbour-joining algorithm

comparing phylogenies

pggb + vg (vs SGD)



$$d_{\text{GRF}} = 0.08$$

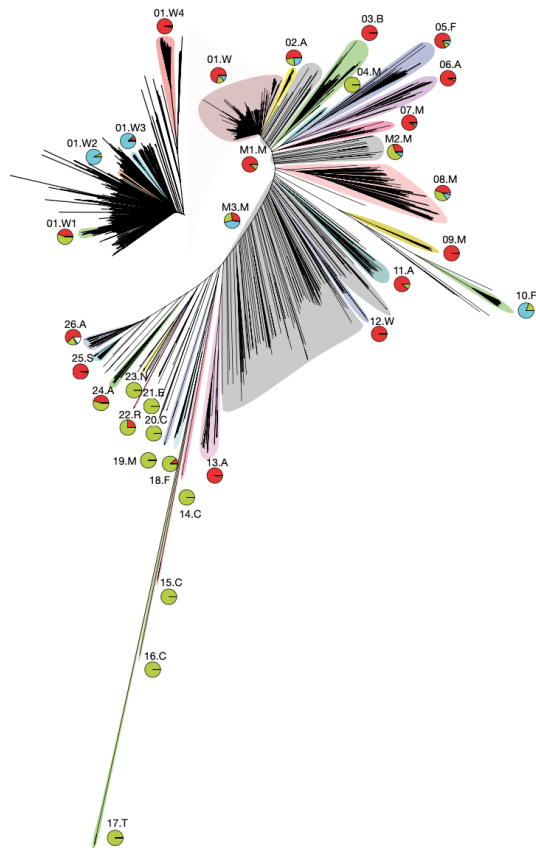
pggb (SNVs) benchmark

pangenome	time (hours)	memory (GB)	compression	N(SNV) / N(genomes)	F ₁ score (mean)
ecoli500	41.39	210.87	23.75	60,243	0.967
scerevisiae142.hc	20.46	119.68	57.12	61,439	0.972
athaliana82	204.06	130.59	36.28	572,924	0.920
tomato23	22.34	42.06	19.73	861,654	0.968
mouse36.chr19	3.80	28.66	11.68	166,734	0.940
hsapiens90.chr6	18.46	135.52	137.72	86.88	0.976

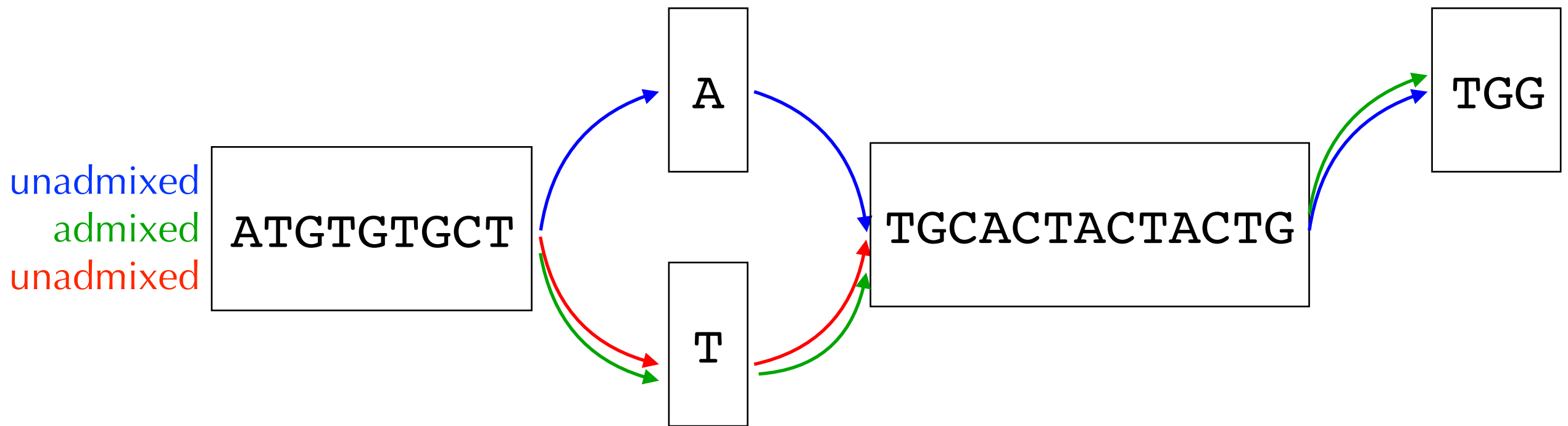
graph phylogenies

pggb + vg (vs SGD)

pggb + odgi


$$d_{\text{GRF}} = 0.44$$


origin & admixture on a graph



only 80% of chrIX can be traced to unadmixed strains

the missing origin



branch lengths suffer reference bias!

what is the unaligned material?

the missing origin



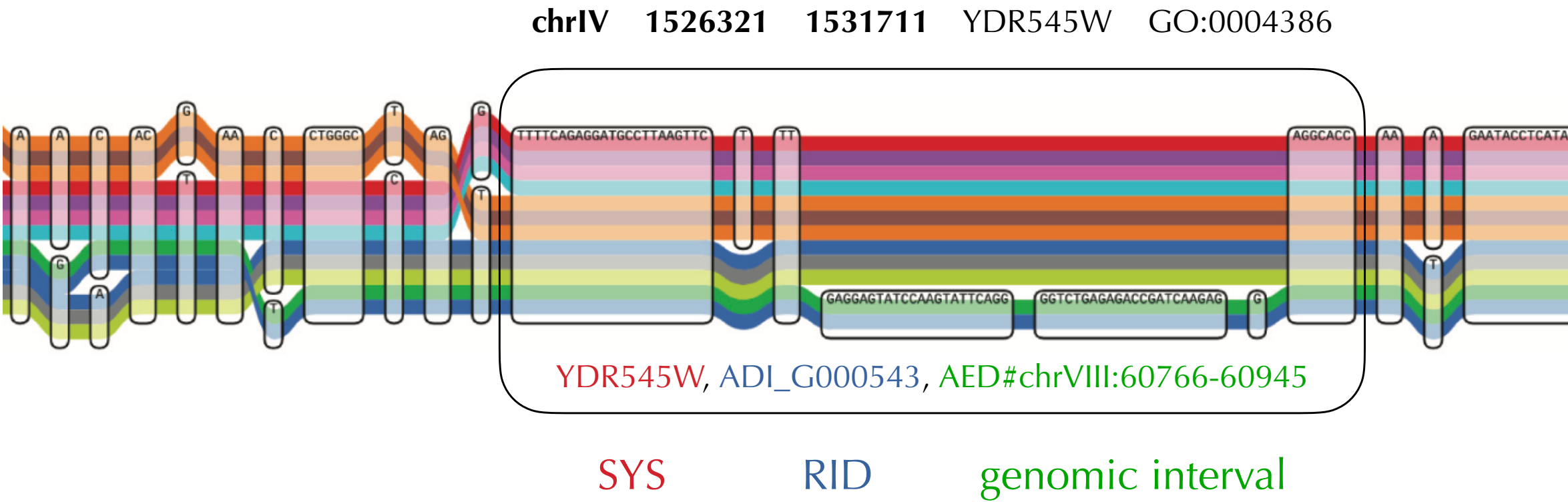
10011101CTCTTCGACCTGCATGGCTTCGGTTTCTA11001010CTTAAGTCCTTAATCGCTCTCGGGATCCTCCCGCCTCTTCTTTTATCAGACTATTTACATCTCGCATATGTCCCCTTTACGCAAACACCACAGCTCCTATTACCTTCTCAATCCCAGGTCTTCACTCTTGTAACAATTCTTTTCTTTTACCCCTTCTTACGGCCCTCATCTCTCTTGTGTCTTTCC
TTCTCGTACTCTTTCCGCTCTATCATTCGCTCATAGGTCCATCTCCCGCTGGGTACCTTTTACCCCTTTCCTCTCTCGCTATT10100011CTCTCTTATCCCCGTACGCCCAATCCCCTTGACCAGAGCTGCTATCCGTTTCAACTCGGCACCCCAACAGAGTATAATTGATCTTCTCCCACTGCTTTTACTTGCTTCCCTTGTCGCCCTCGTCGCCCTCATTTCTTTTGT0011100
1TTTCACAGTGTCTAGGCTTTTACTACACTCCACCTCAGAGGAATCGCTTCTATATCGATTTTCGTTTCCCAGGCAGTGACCGCAATCTTGGCCTCGTGCTTCCCTCTAATGTGCGTCTGTCTCCAGGGTCTTCACTAGCCTTAGTGGGCGTGTCTGGCCCTTGTGGCACTTCAACTTATCCCTAAGCAGTCCGCTTTTCCCTTTGTAGTGGTCAATCCAGAGTCTTCAGC
CGGAACCACCGCTTCTACCATCCTGGCGCACTTCGCGGCCGTGGGACAGTTTCTATGTTTCTCGGCCTCCACTTCTCCGAACACAAAGGCAGTGG00111101GCTTAGTTGTCCATTACTGTTCCTACAGCCGCTGTGCTTGGTTCAATAACAGGTGCTCCCCCACTTATGCGCTTCTCACTTCTTCCCGCGGTCTCTTCTCAAACTTCTACTCCCTCCAGTGTCAGTATC
TTTACCCAGCTTCTCCCTTCCCGCCTCCCGTGTCTCTCCCATCTAACTAGCAGACTCTTCGACCTGCATGGCTTCCGTTTCTAACCGCACCC01010111GTCTTTTACGCTTCTTCTTTTATCAGACTATTTACATCTCGCATATGTCCCTTATTACGCAAACACCACAGCTCCTATTACCTTCTCAATCCCCAGGTCTTTCAGTCTTGTACCAAT
TCTTTTTCTTTTACCCCTTCTTACGGCCTCATCTCTCTTGTGTCTTTCCTTCTCGTACTCTTTCGCTCATATTCCGTCTAGGTCCATCTCCGCTGGGTACCTTCTTATCTCTGCTGCTATTATTCTTGTG11000010TCAACTTATCCCTAAGCAGTCCGCTTTTCCCTTTGTAGTGGTCAATCCAGAGTCTTCAGCCGGAACCCCGCTTCTACCATCC
TGGCGCACTTCGCGGCCGTGGGACAGTTTCTATGTTTCTCCGCCTCCACTTCTCCGAACACAAAGGCAGTGGTTTCGTGCTTAGTTGTCCATTACTGTCCGCACTACAGCTTCTTACAGGTCGTCGCCCACTTATGCGCT01101010CTTCTTCCCGCGGTCTTCTCTCAAATCTTACTCCCTCCAGTGTCAGTATCTTTTACCCAGCTTCTCCCTTCCCG
CCTCCCTGCTCTCTCCCATCTAACTAGCAGACTCTTCGACCTGCATGGCTTCCGTTTCTAACCGCACCCCTTAAGTCTTAAATCGAACTCGGGATCCTCCCGCCTCTTCTTCTAAGTCTTCTCGCATATGTCCCTTATTACGCAAACACCACAGCTCCTATTACCTTCTCAATCC01010001CTTCACTCTTGTACCAATCTTTTCTTTTACCCCTTCTTAC
GGCTCATCTCTCTTGTGTCTTTCCTTCTCGTACTCTTTCCGTCTATCATTCGCTCATAGGTCCATCTCCCGCTGGGTACCTTTTACCCCTTTCCTTCTCGCTTTATTCTTGTGGCACTTCAACTTATCCCTAAGCAGTCCGCTTTTCCCTTTGTAGTGGTCAATCCAGAGTCTTCAGCCGGAACCCCGCTTCTACCATCTTGGCGCACTTCGCGGCCGTGGGACAGTT
CCTCTATGTTTCTCCGCTCCACTTCTATGTTTCTCTATGTTTCTCATCTCGATCGACTAGCATCAGGCGAGGCTAGCATATCTACGACTGCATCTACGAGCAGCTACGAGAGGCATACTAGCGACGAGCGATCATCAGCACGAGCAGCGATCAGCAGCATCACGAGCATCAGCAGCTACGACATCATCGACGATCAGCGACGATTATATTACGAATTATACGCG10110101

panda

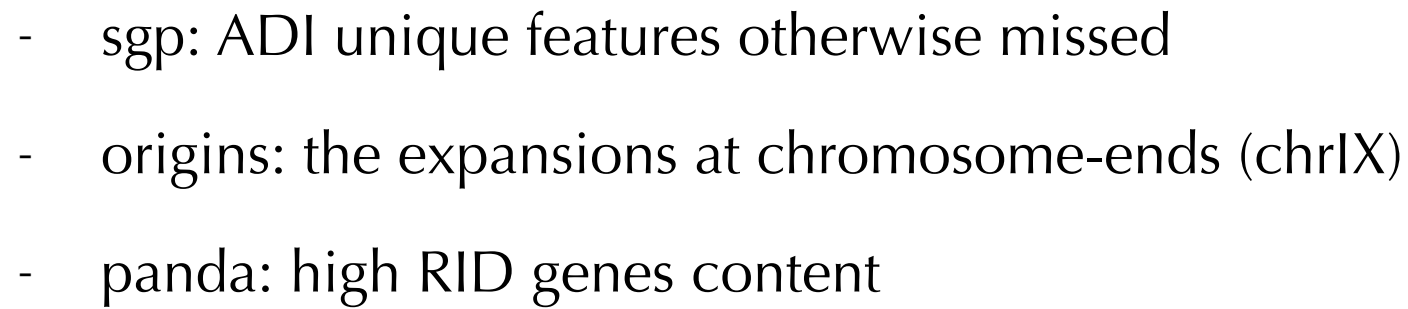
pangenome annotations data analysis

- which strain bears the largest number of RID genes?
- how many RID genes are on chrIX?
 - 65
- how many are at chromosome-ends?
 - all

strain	N(RID)
ADI	263
CQS_1a	65
ATM_1a	62
AIF_HP2	60
BBM_1a	49
DBVPG6044	37
CNT_HP1	36
CEQ_1a	35
...	...



summary



- pggb + odgi



Lundblad V & Blackburn EH. An alternative pathway for yeast telomere maintenance rescues est1- senescence. CELL. 1993.

perspectives

acknowledgements



Wilfried Hearty

Roberto Grossi

Gianni Liti
Raja Appuswamy

Damien Coudreuse

“Progress in science depends on new techniques, new discoveries, and new ideas, probably in that order.”

- Sydney Brenner, 2002 Nobel Prize Winner -