Evolutionary Pangenomics in the Era of Complete Genomes

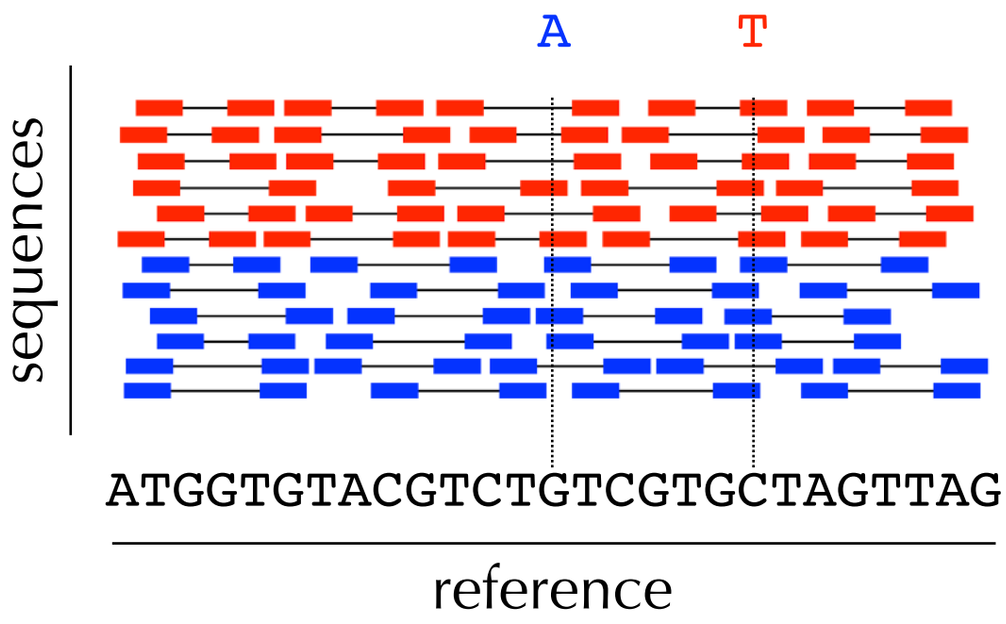
Lorenzo Tattini

EURECOM
Sophia Antipolis

why genome evolution?

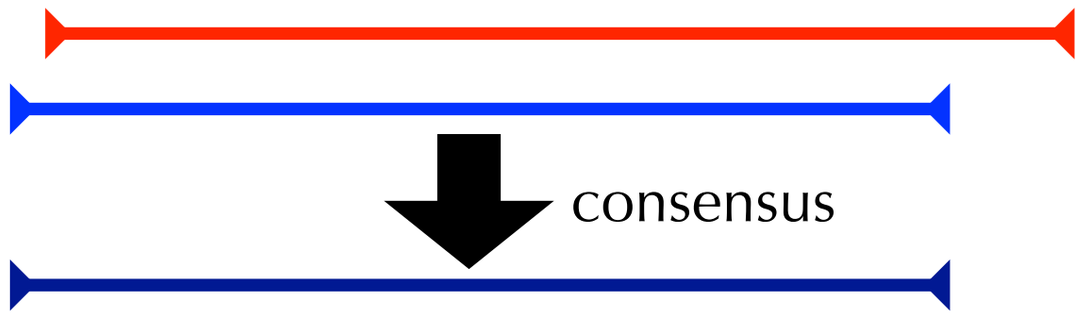
genotypes

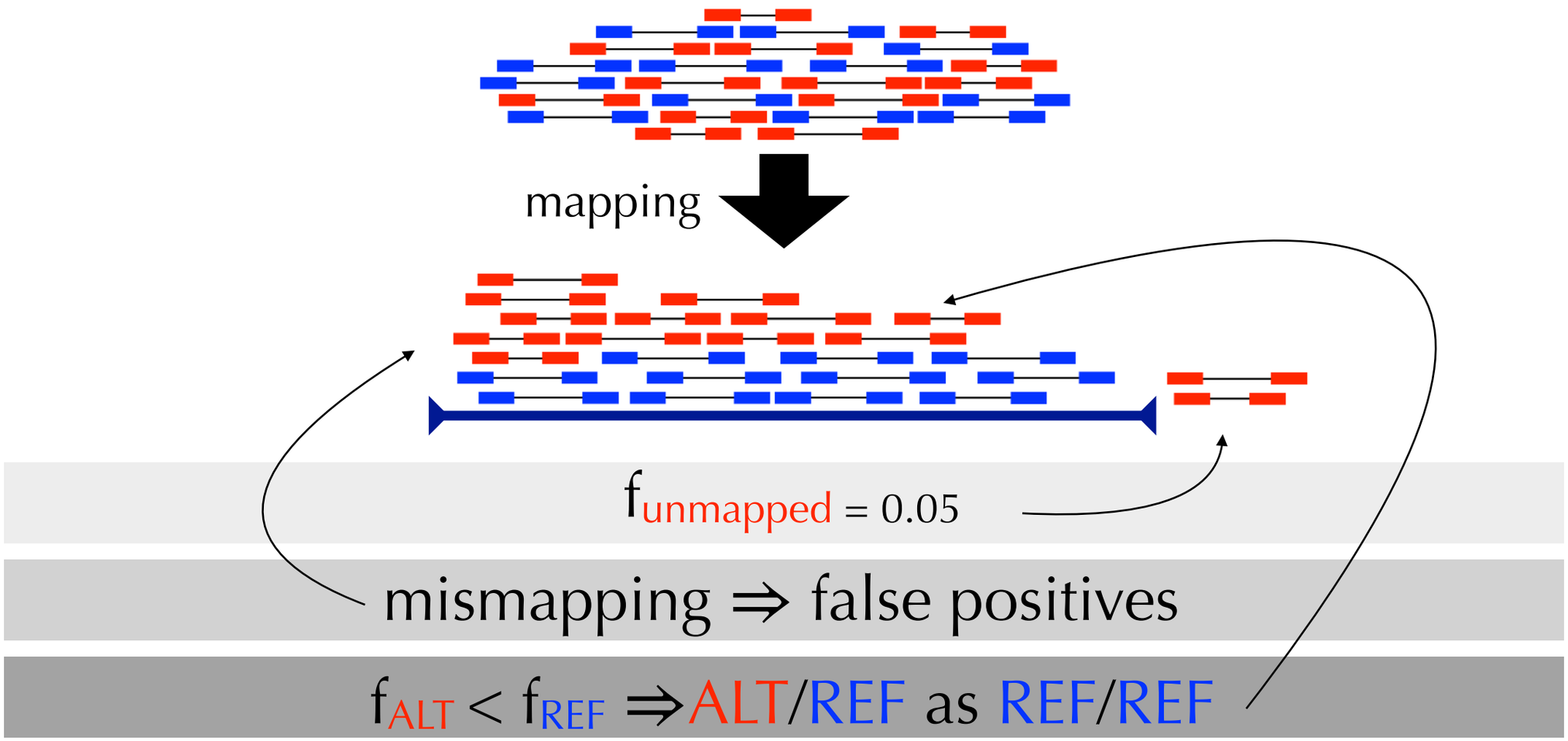
phenotypes



the reference bias

origin and consequences

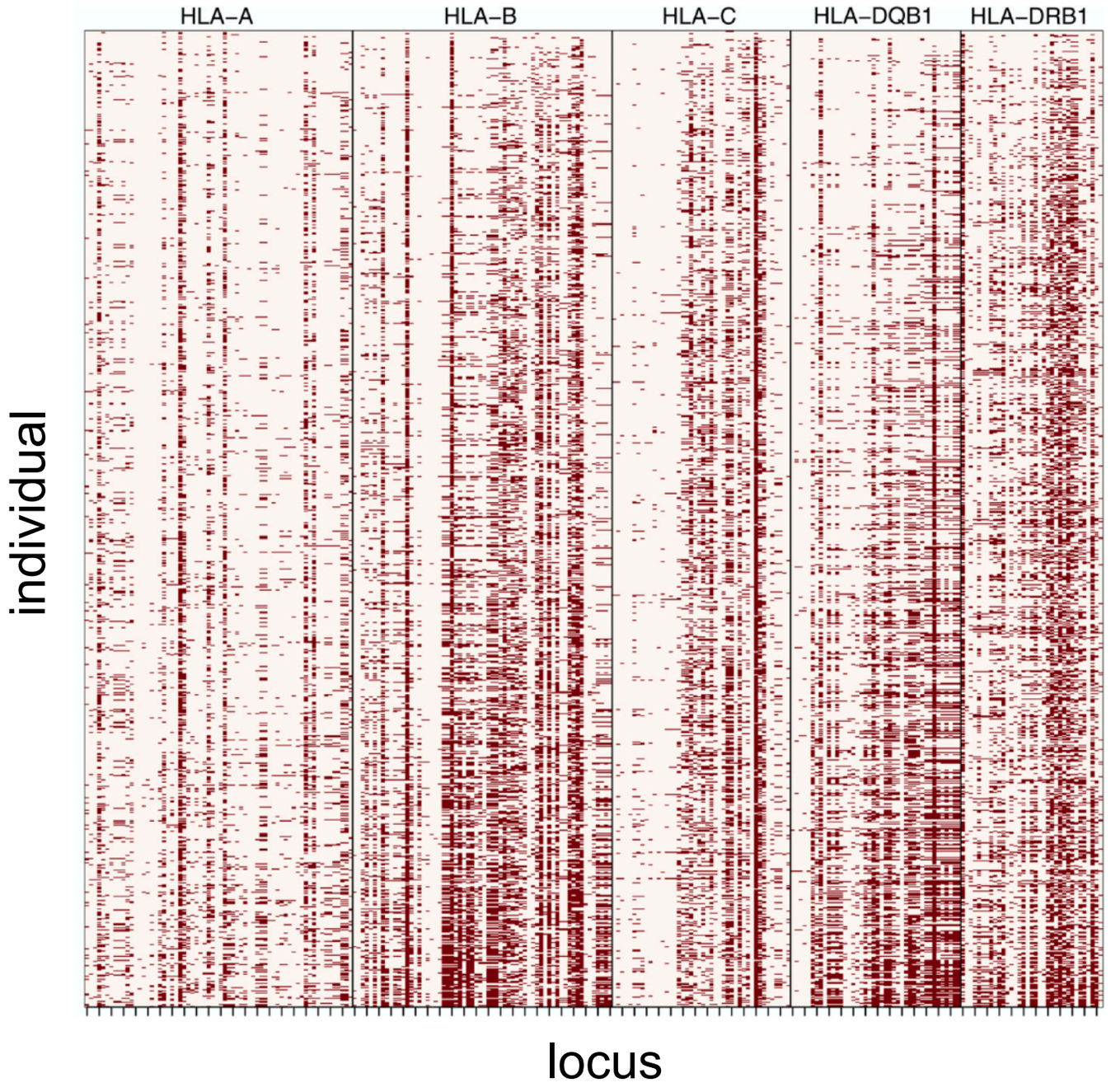




Eizenga JM, … & Garrison E. ANNU REV GENOMICS HUM GENET. 2020.

the clinical reference bias

genotyping: short reads vs Sanger sequencing



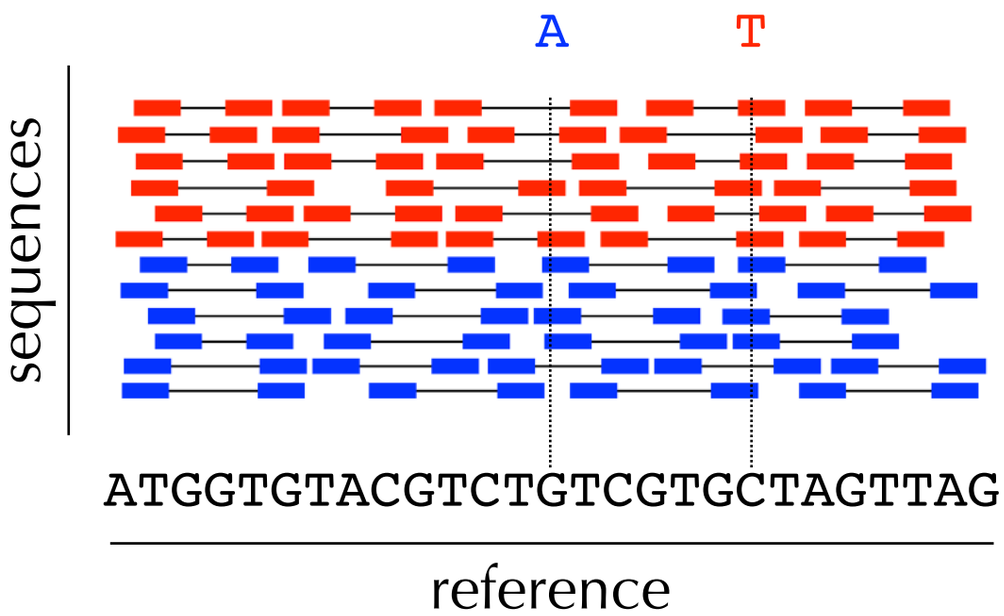
legend:
match (82%)
mismatch (18%)

Brandt DYC, …, & Meyer D. G3 (BETHESDA). 2015.

the missing heritability

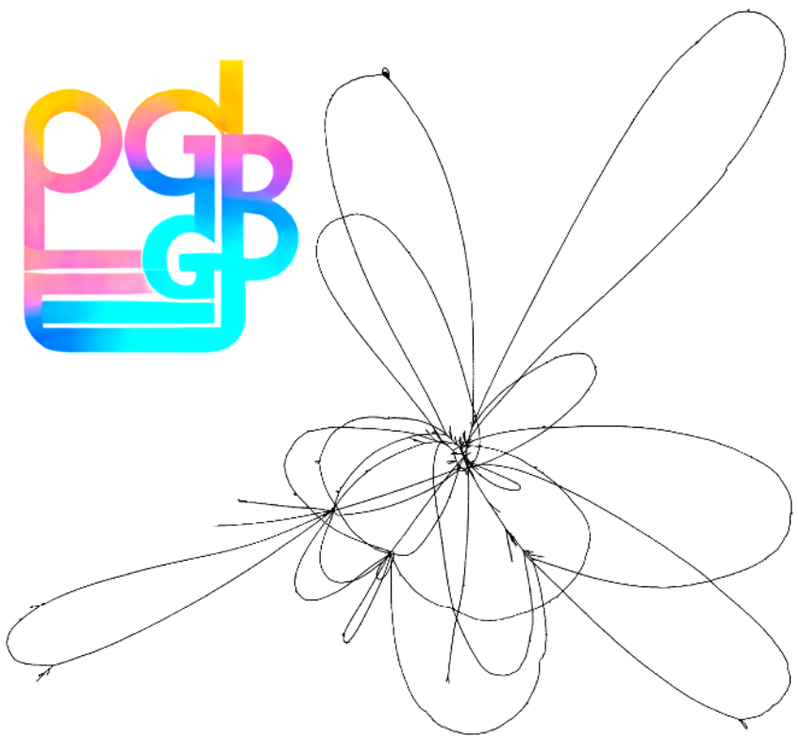
genotypes

phenotypes



variation graphs: a new paradigm

genotypes



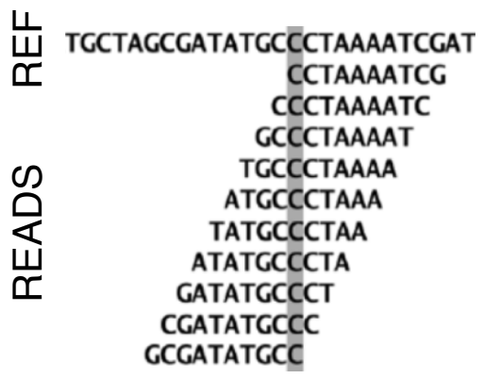
phenotypes

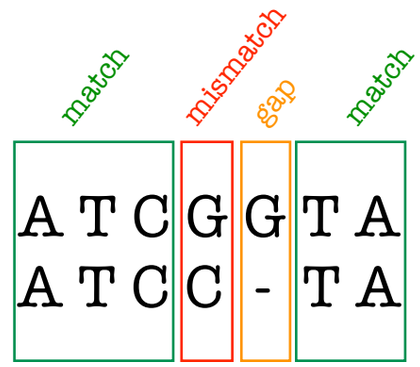
Garrison E, …, & Prins P. NATURE METHODS. 2024.

building a variation graph

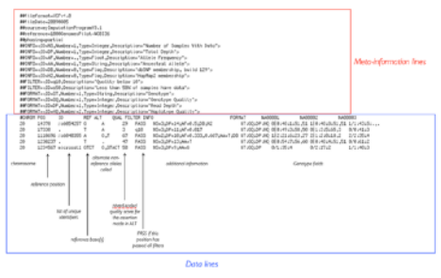
reference + variants

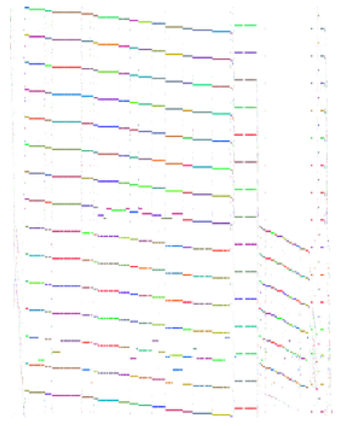
pairwise all-against-all
T2T assemblies





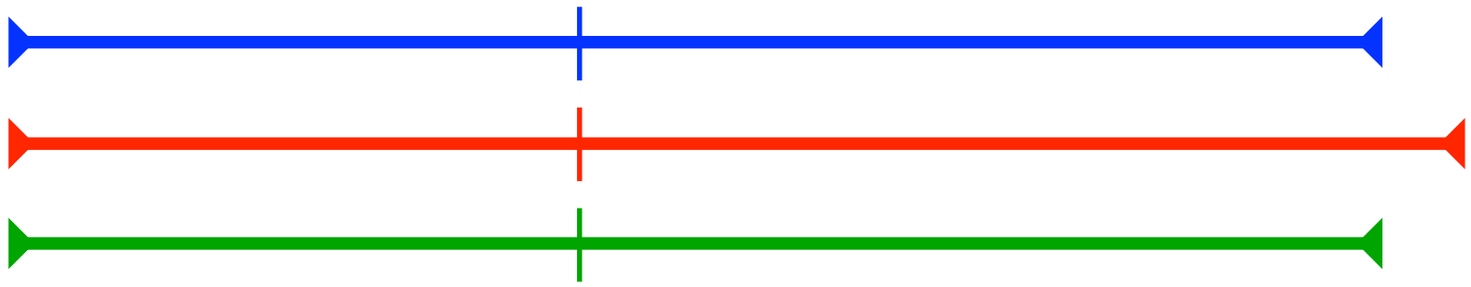
VCF





what's a variation graph?

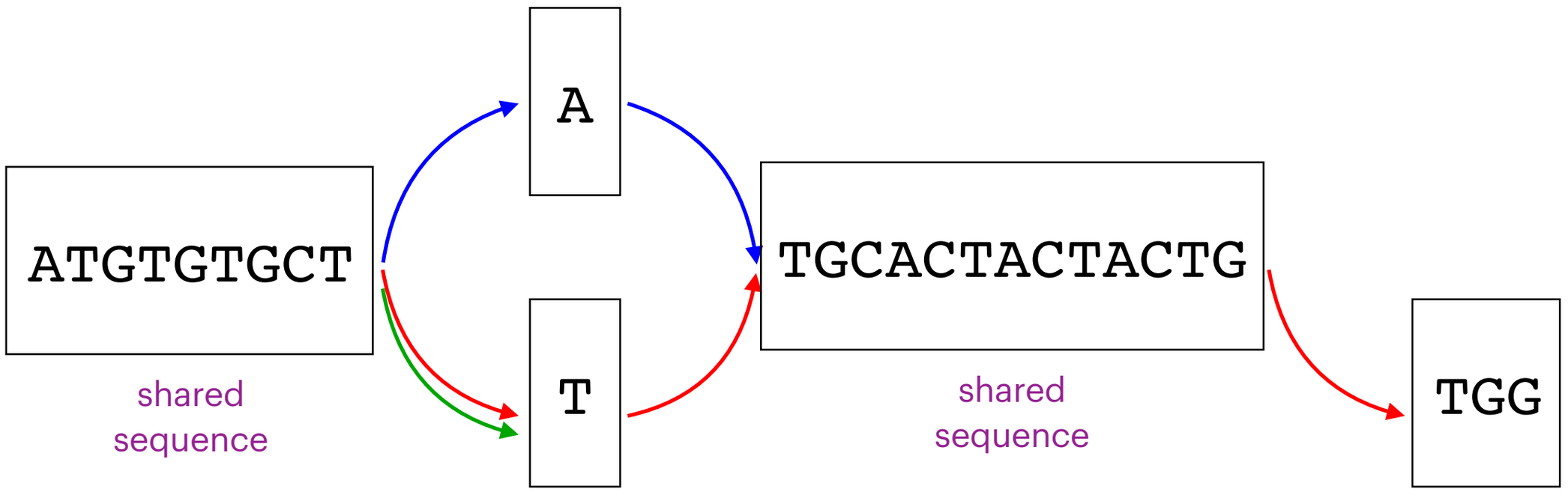
assemblies



alignment

```
ATGTGTGCT   A   TGCACTACTACTG
ATGTGTGCT   T   TGCACTACTACTGTGG
ATGTGTGCT   T   TGCACTACTACTG
```

graph

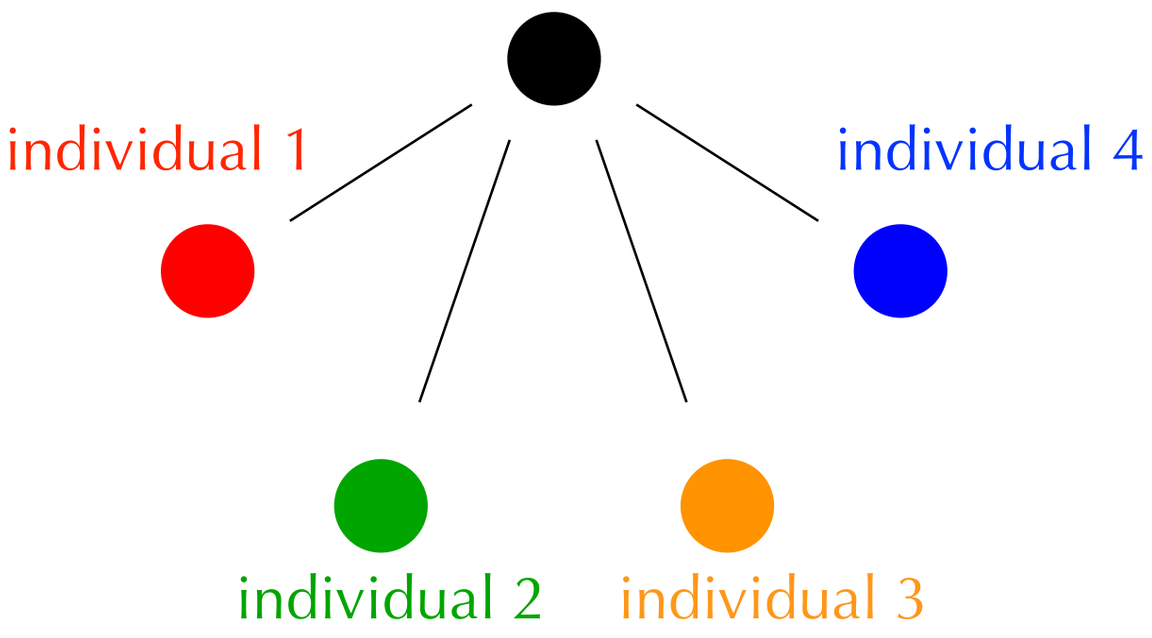


why a variation graph?

practical but also conceptual aspects

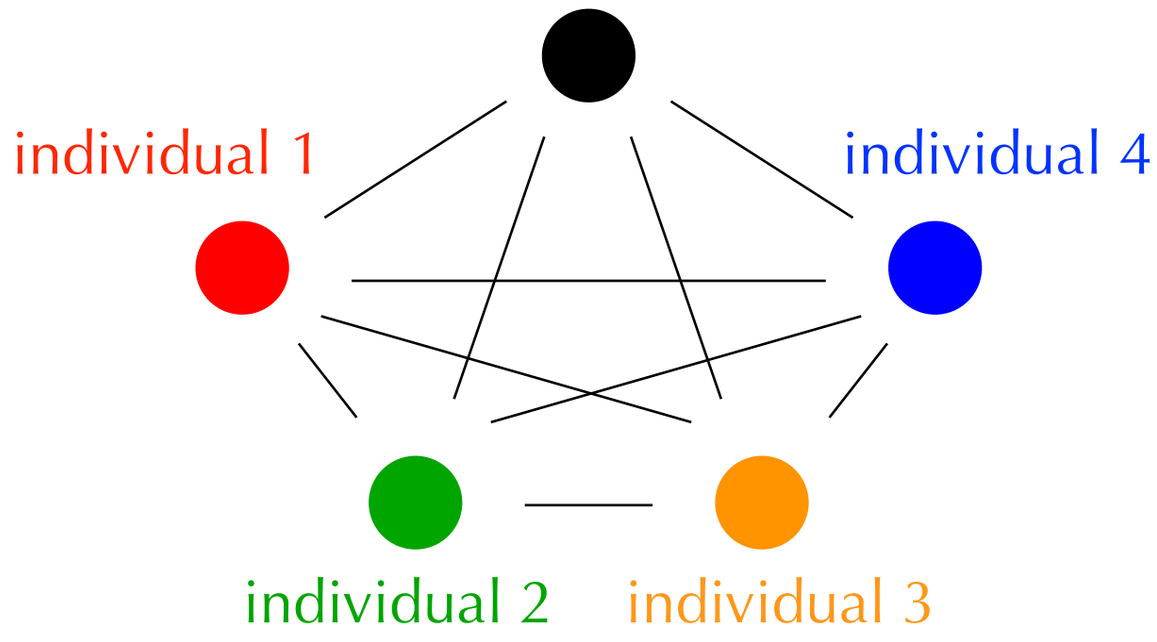
reference-based

linear reference

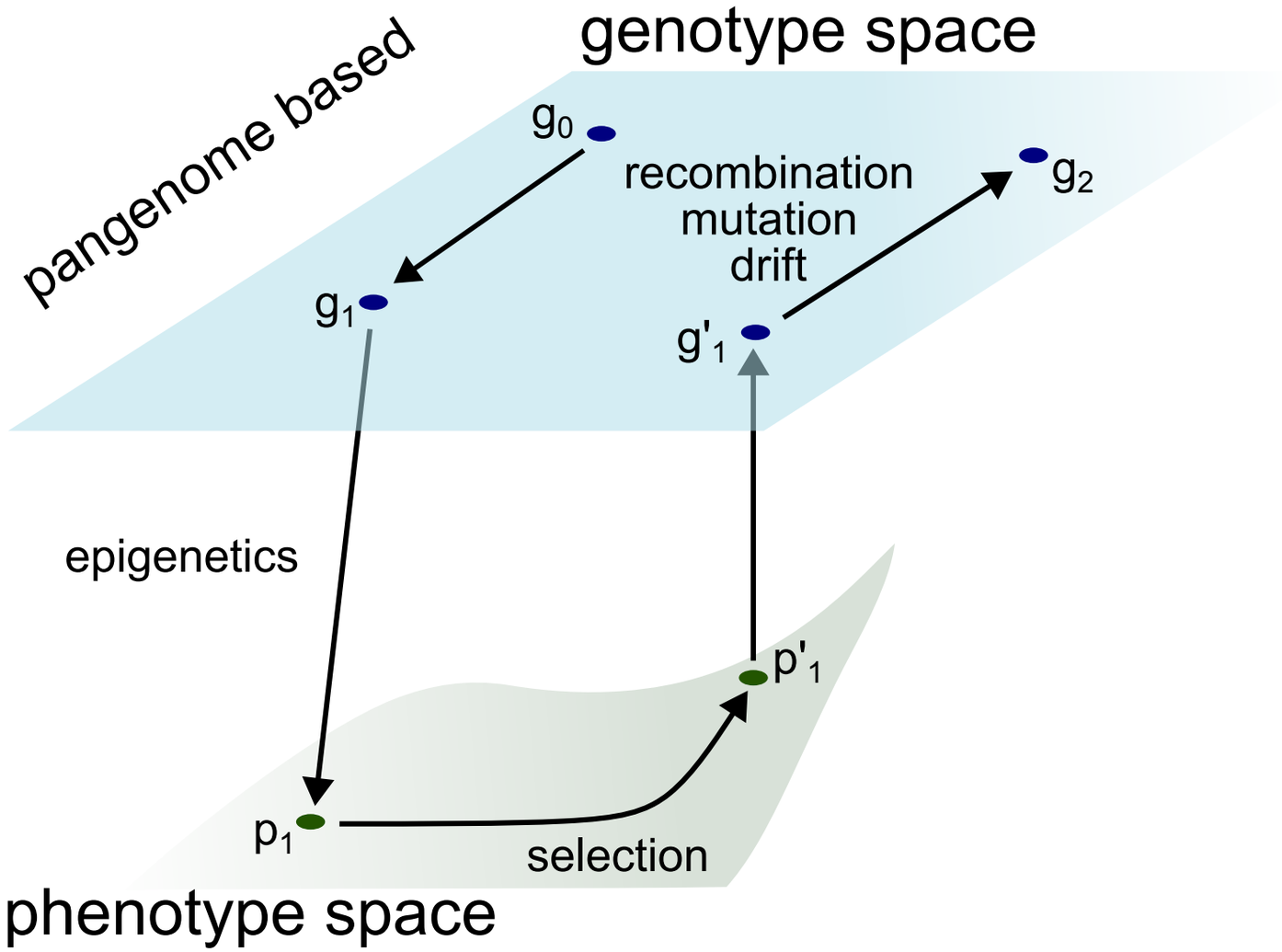


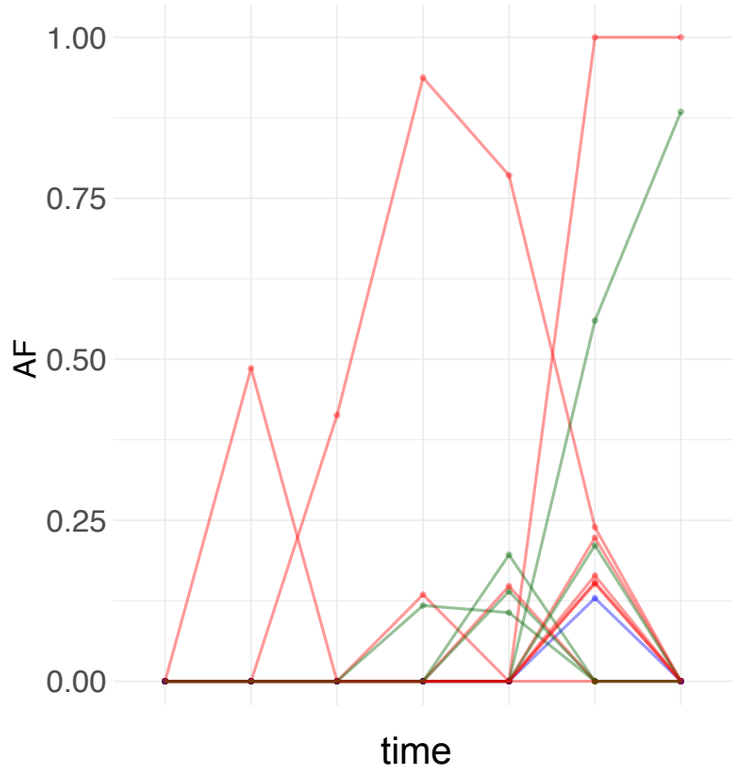
graph-based

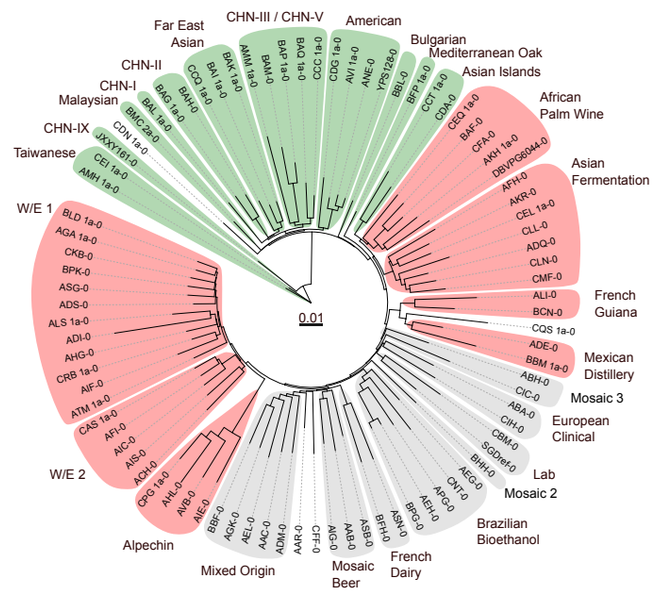
linear reference



reference vs pangenome spaces

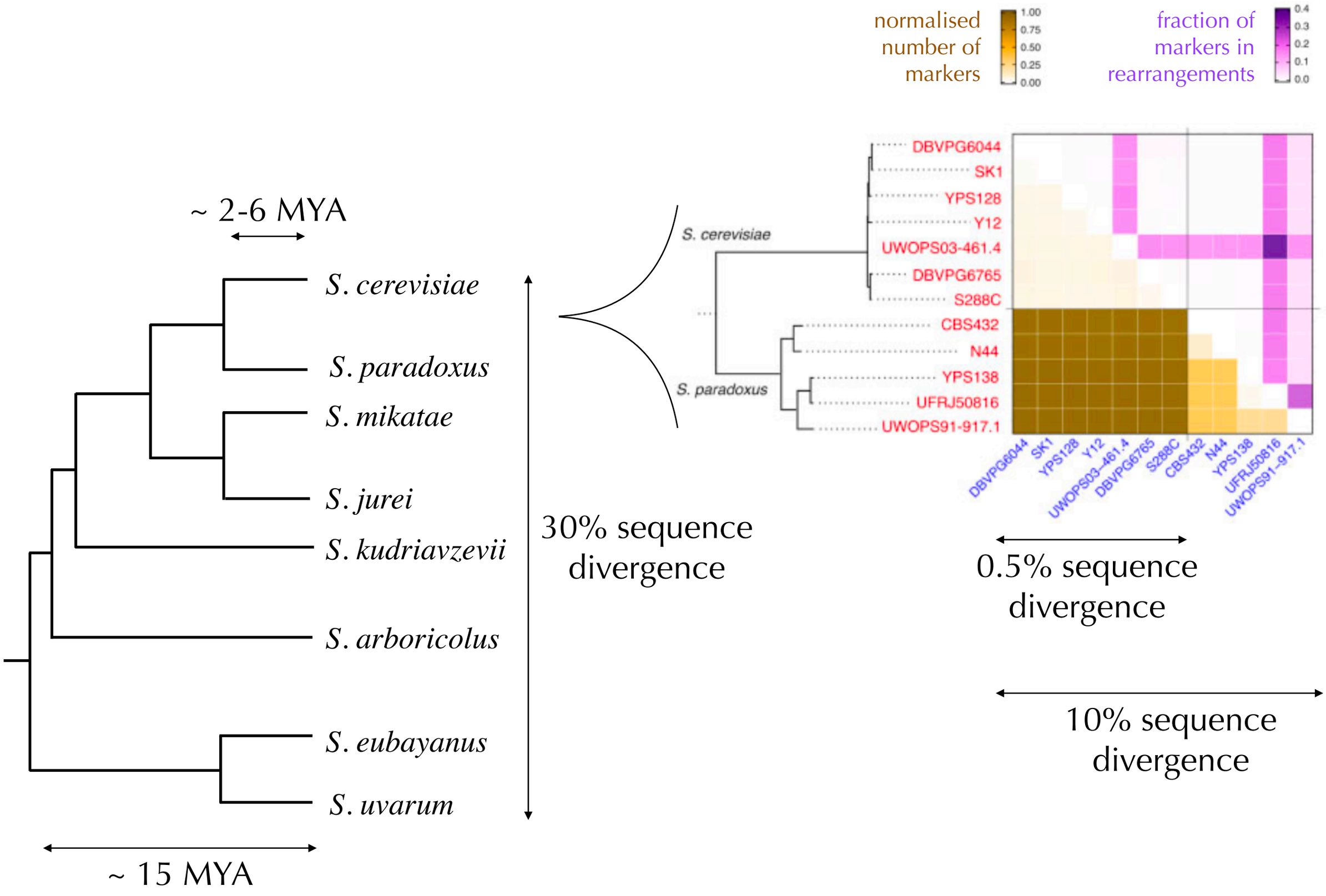






Fang B & Edwards SV. TRENDS ECOL EVOL. 2025.

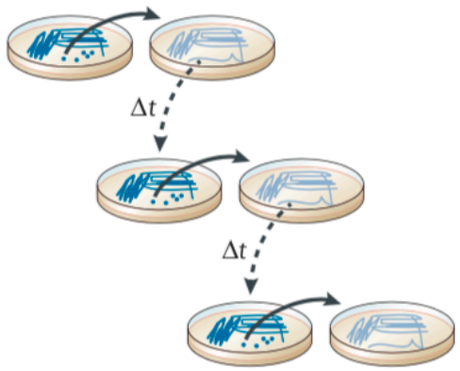
the *Saccharomyces* sensu stricto



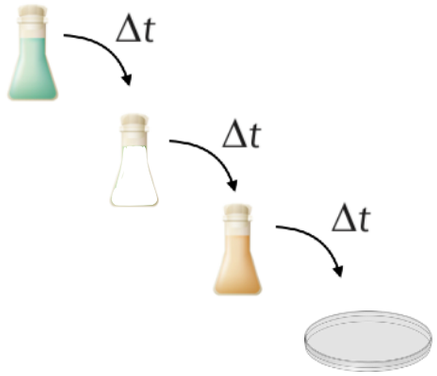
experimental evolution in hybrids

evolutionary dead ends vs sources of hybrid vigour

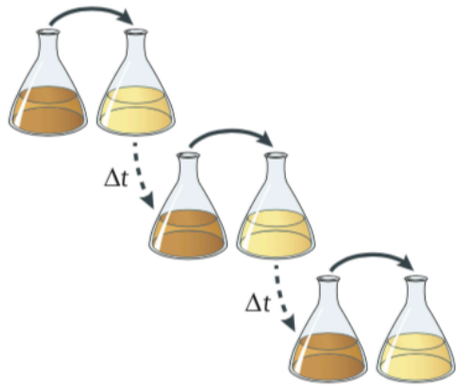
mutation accumulation



return to growth



adaptive evolution



clonal

non-clonal

neutral

selective

recombination

Mozzachiodi S*, Tattini L*, …, & Liti G. NAT COMMUN. 2021.

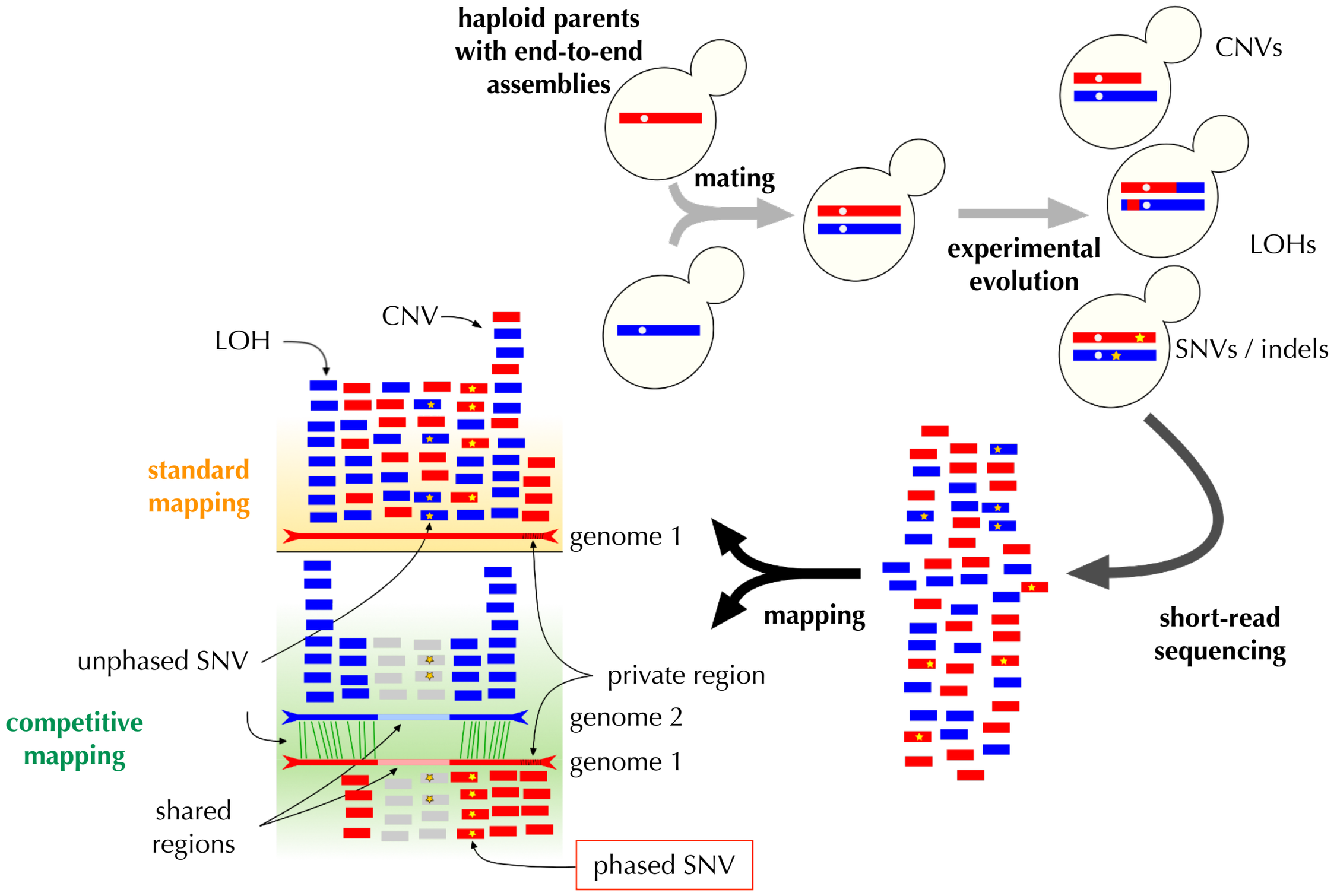
Tattini L@, …, & Liti G@. MOL BIOL EVOL. 2019.

Tattini L*, Mozzachiodi S*, …, & Liti G. In preparation

Tattini L, Maret T, …, & Falque M. In preparation

Jain A, Tattini L, …, & Coudreuse D. In preparation

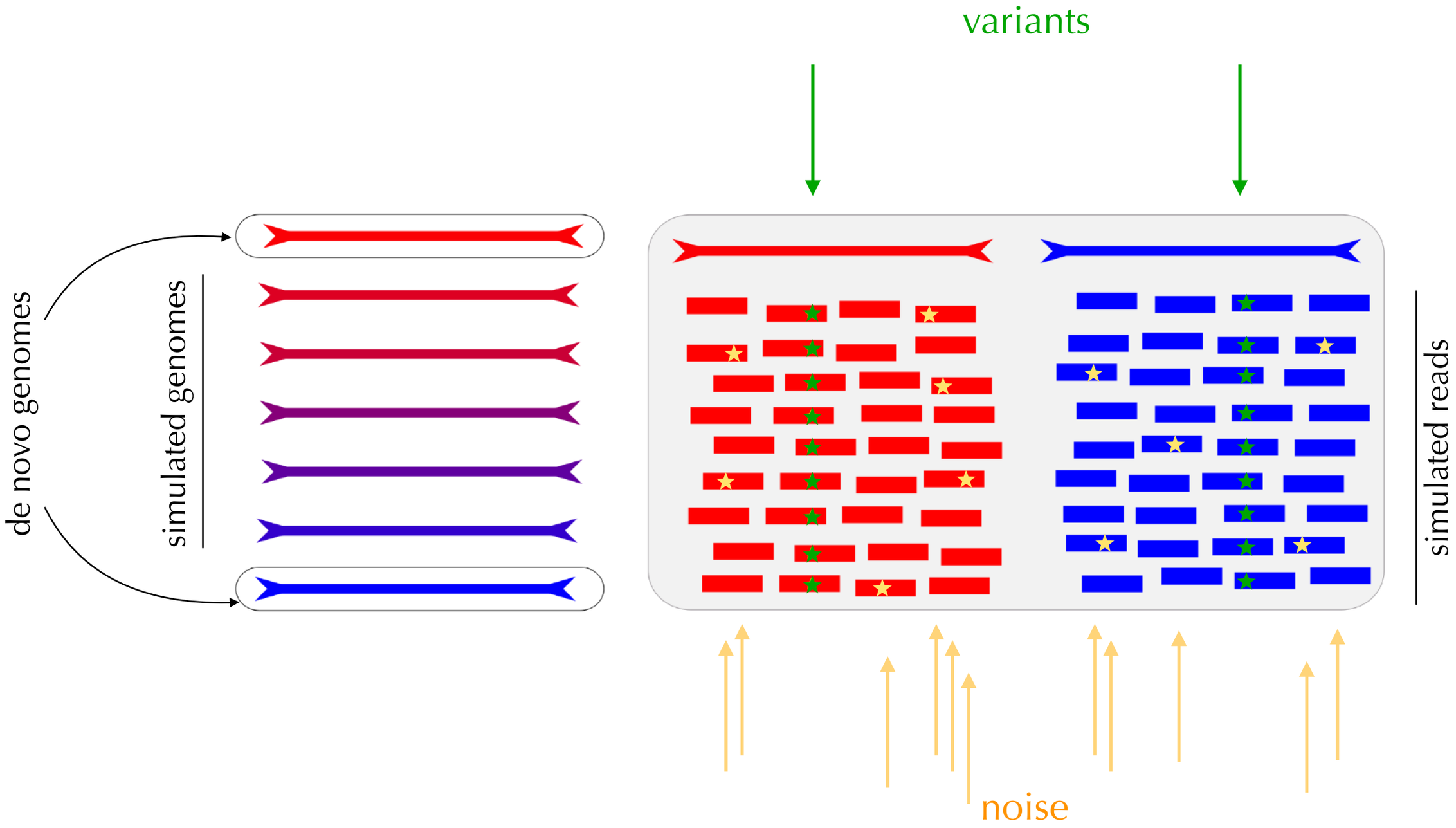
mutation accumulation lines



cost: x 2

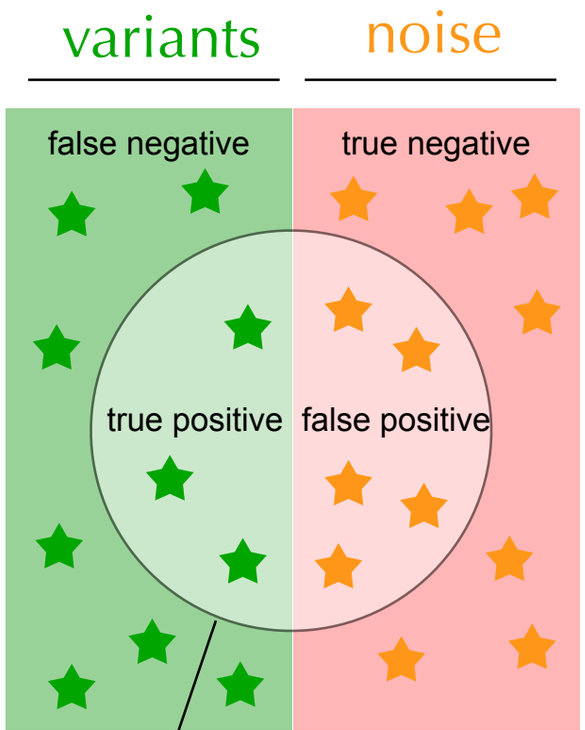
Tattini L@, …, & Liti G@. MOL BIOL EVOL. 2019.

synthetic validation data set

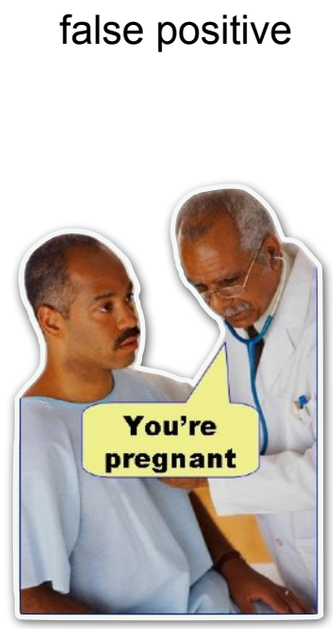


Tattini L@, …, & Liti G@. MOL BIOL EVOL. 2019.

validation metrics



selected elements



$$F_1 = 2 \cdot \frac{R \cdot P}{R + P}$$

how many selected items are relevant?

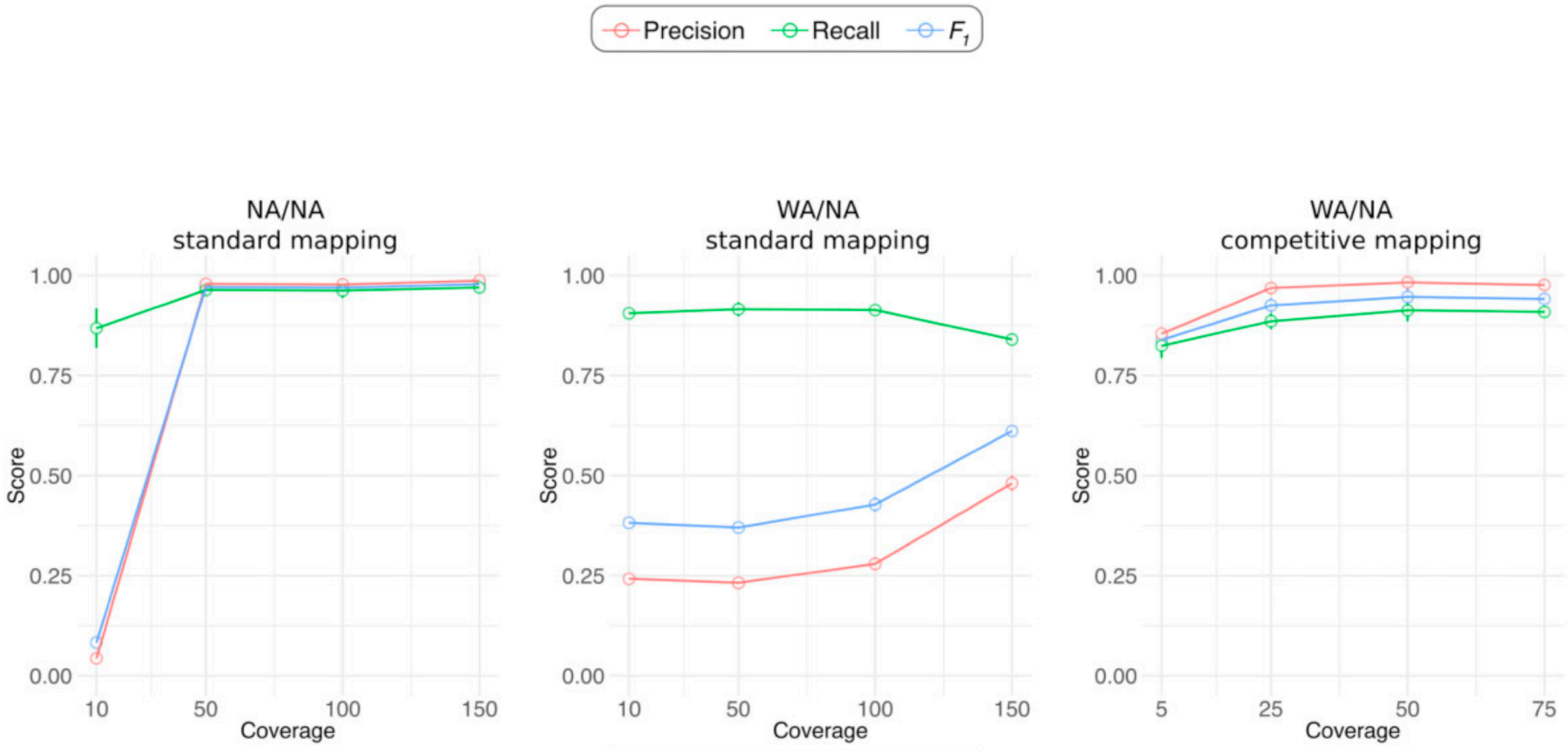
$$P = \frac{}{}$$

how many relevant items were selected?

$$R = \frac{}{}$$

Tattini L@, …, & Liti G@. MOL BIOL EVOL. 2019.

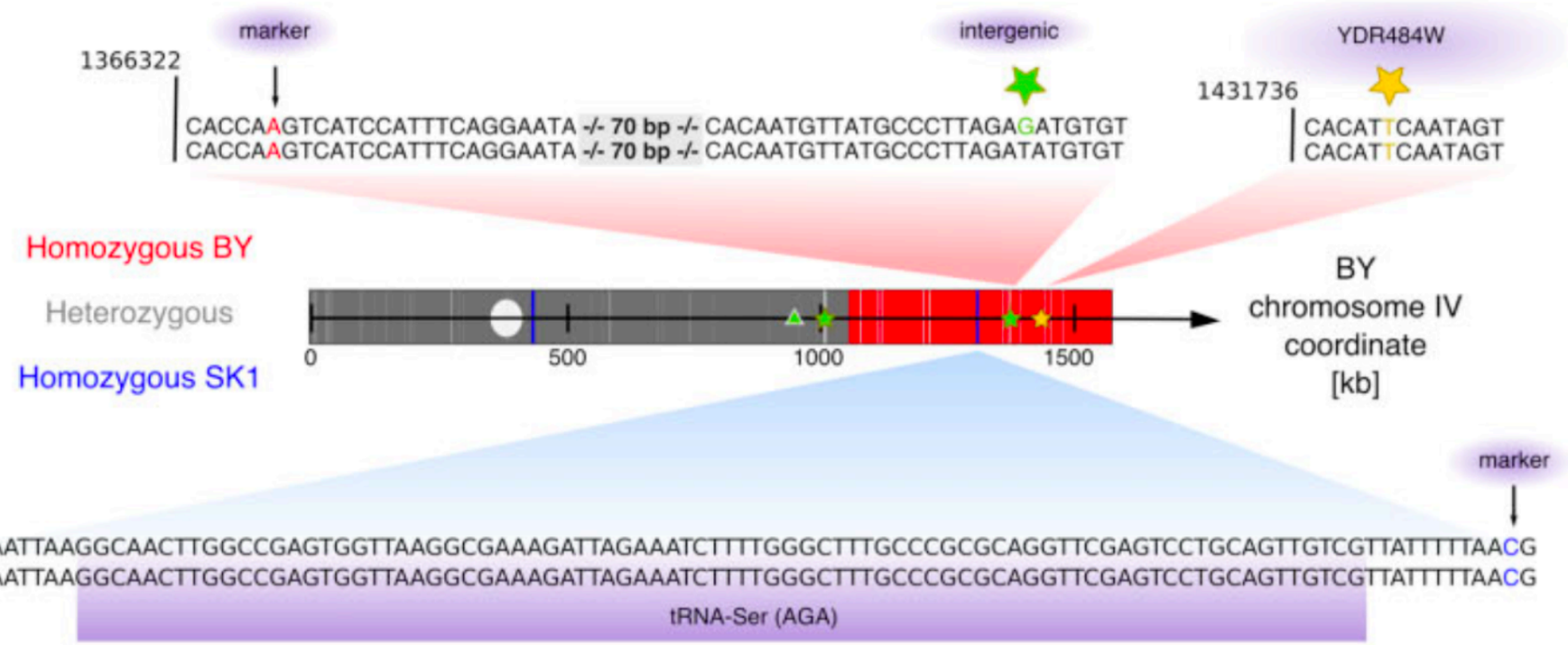
validation of competitive mapping

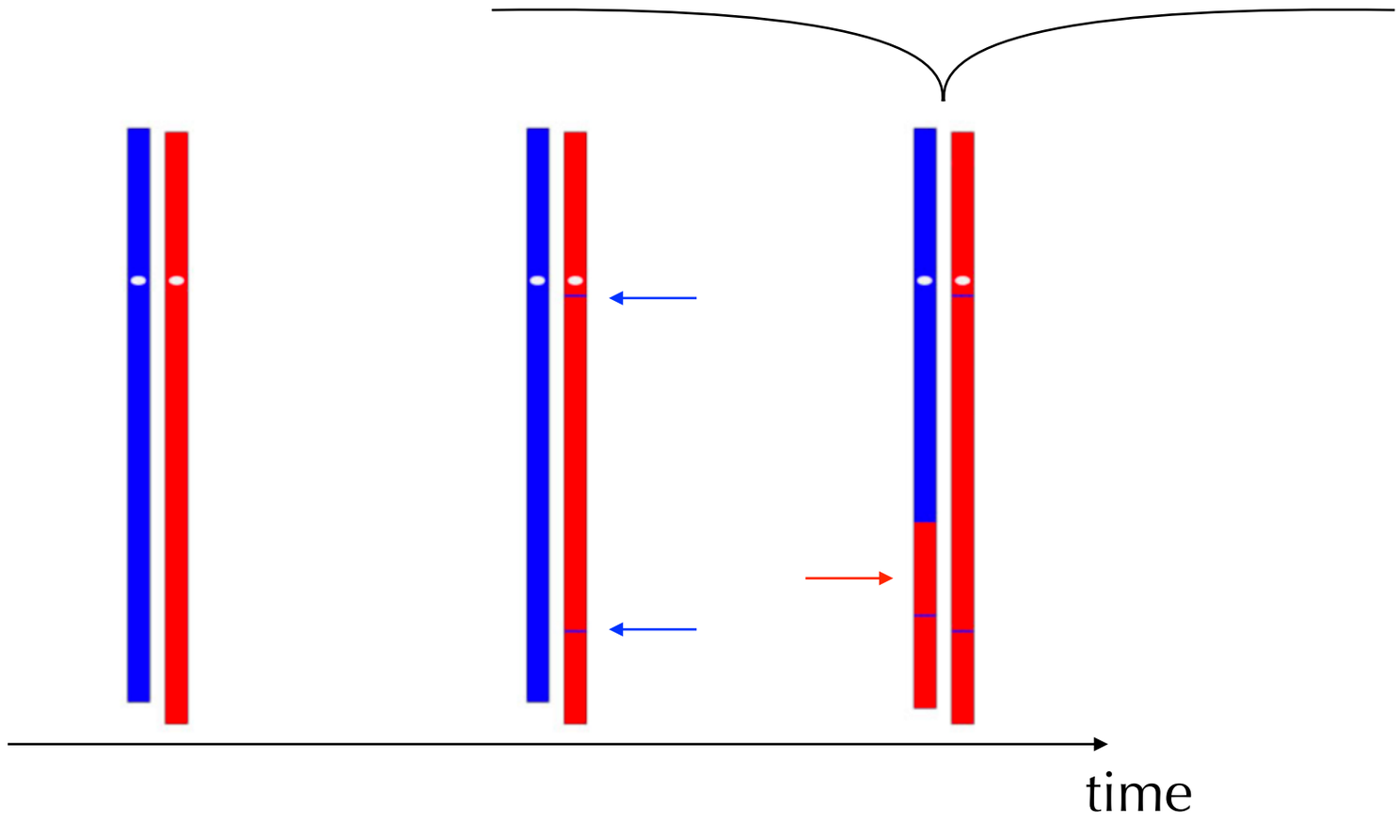


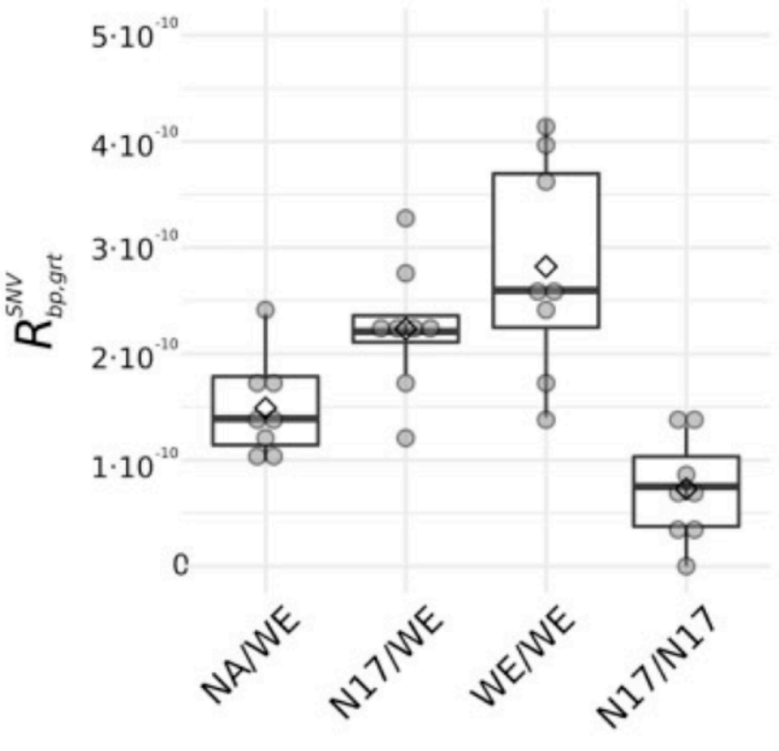
standard + competitive

Tattini L@, …, & Liti G@. MOL BIOL EVOL. 2019.

mutational landscape



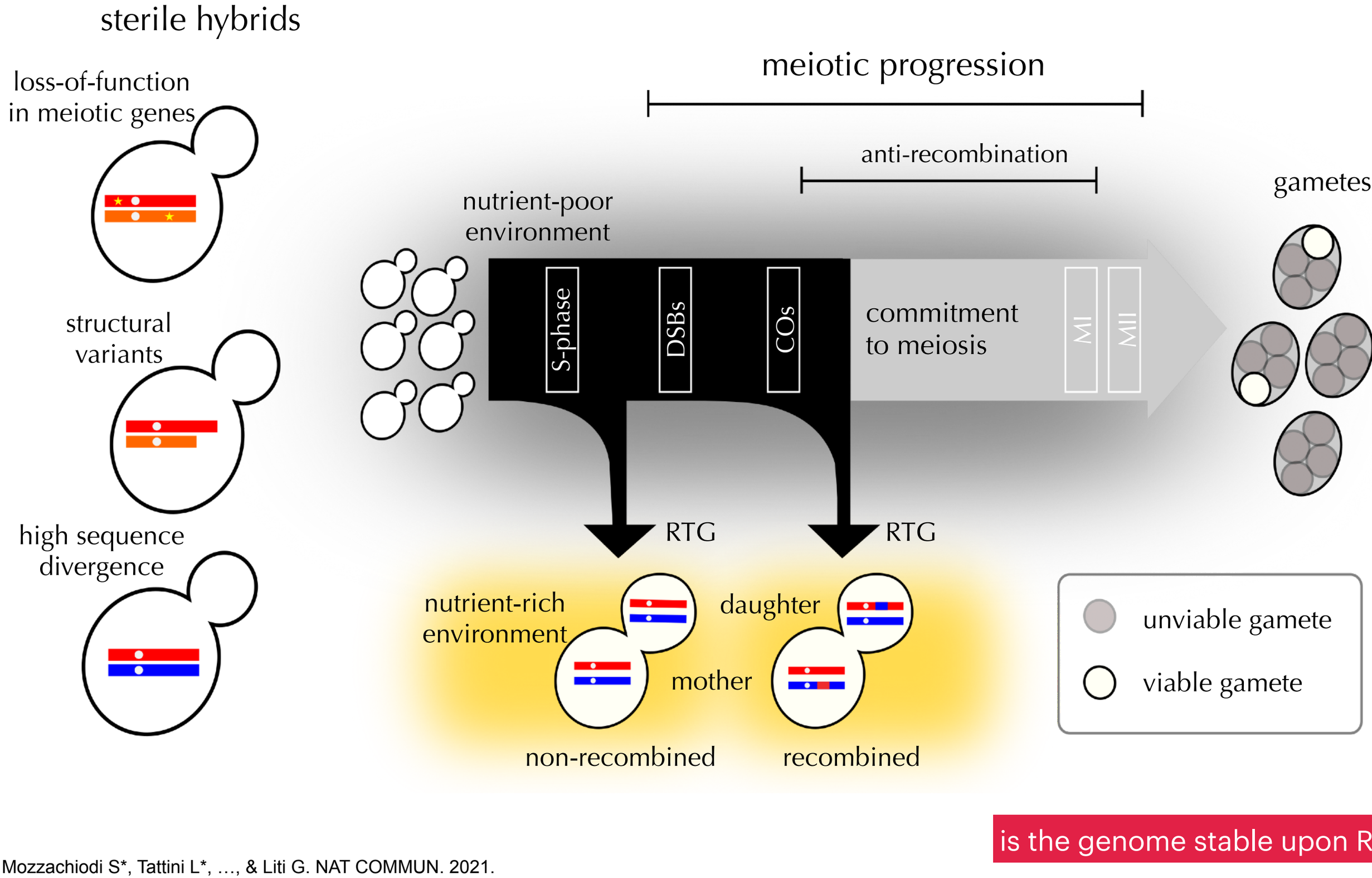




stable

dead-ends?

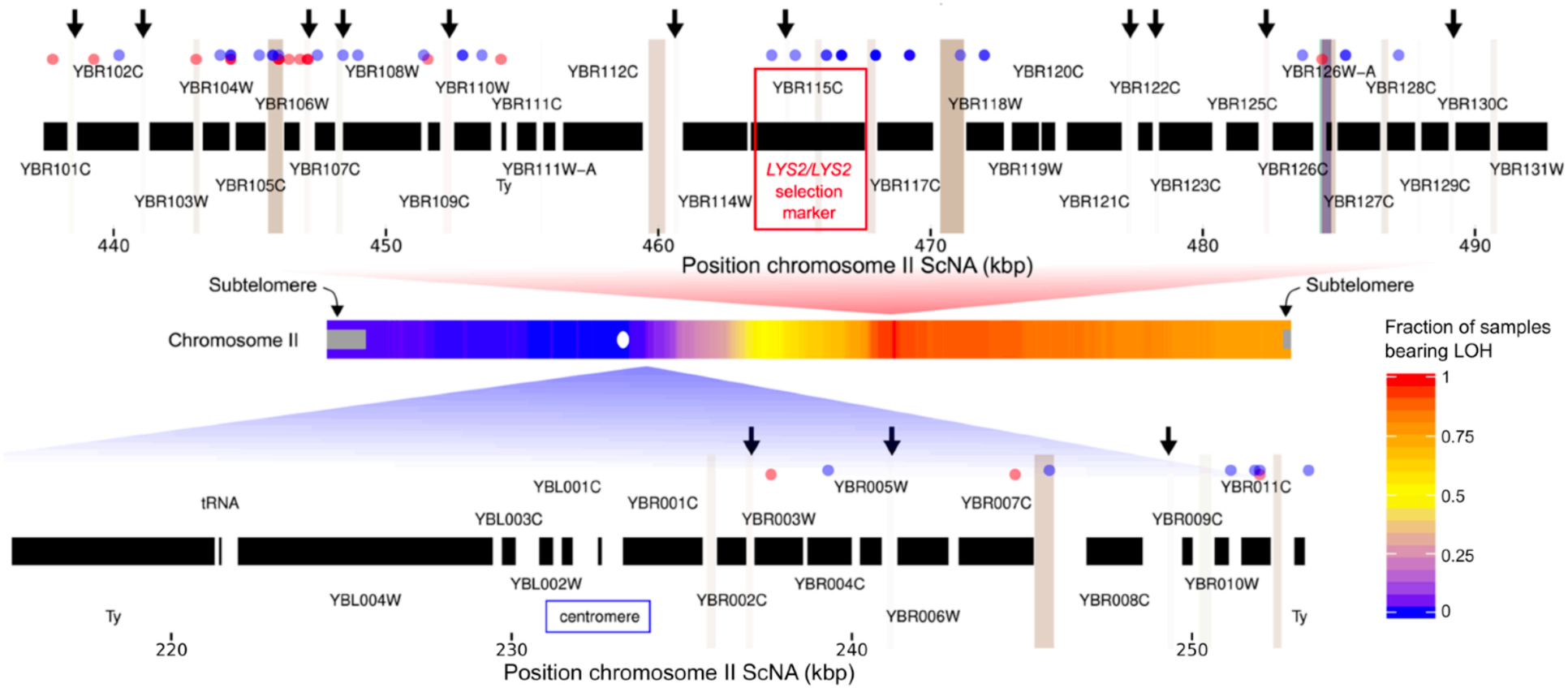
return-to-growth



is the genome stable upon RTG?

Mozzachiodi S*, Tattini L*, …, & Liti G. NAT COMMUN. 2021.

LOH landscape

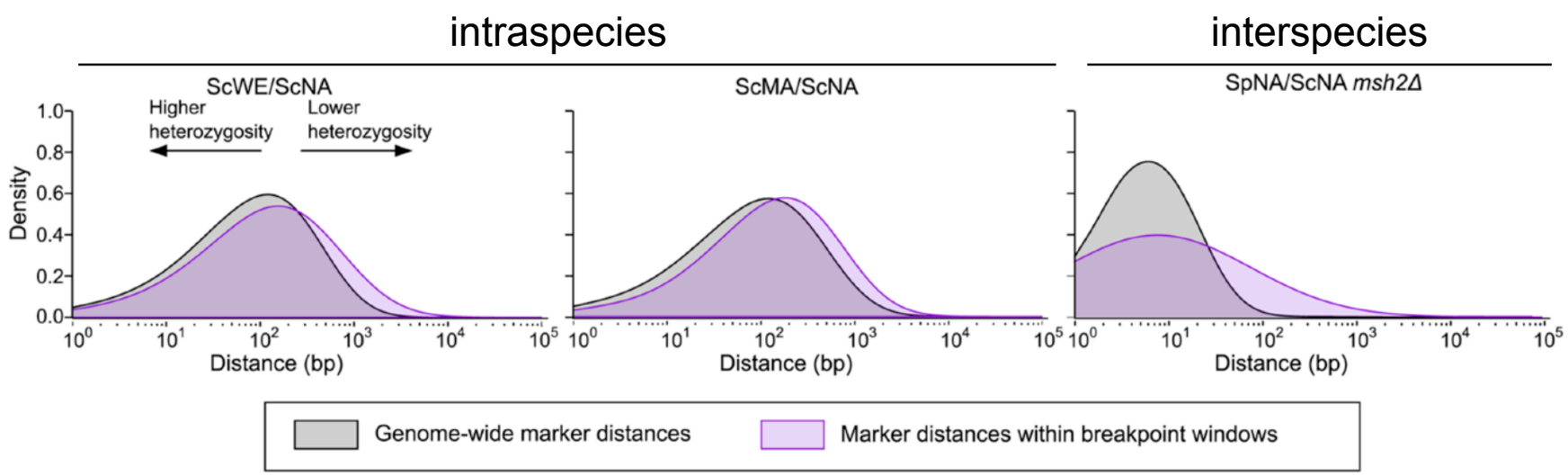


massive LOH (30%)

breakpoints ~ hot-spots

local hetorozygosity

genome stability



D'Angiolo, …, & Liti G. NATURE. 2020.

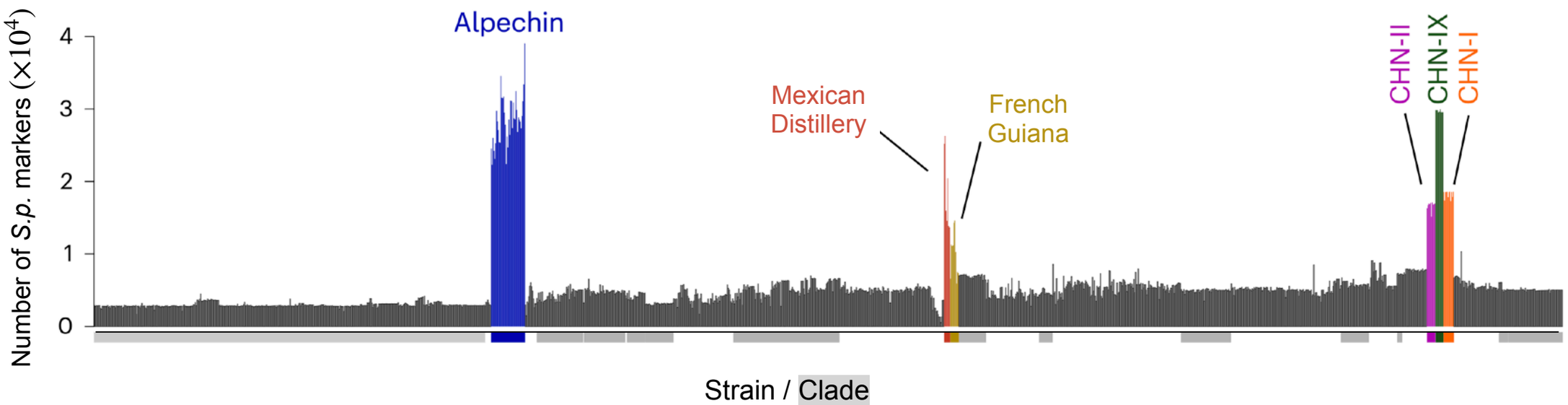
Mozzachiodi S*, Tattini L*, …, & Liti G. NAT COMMUN. 2021.

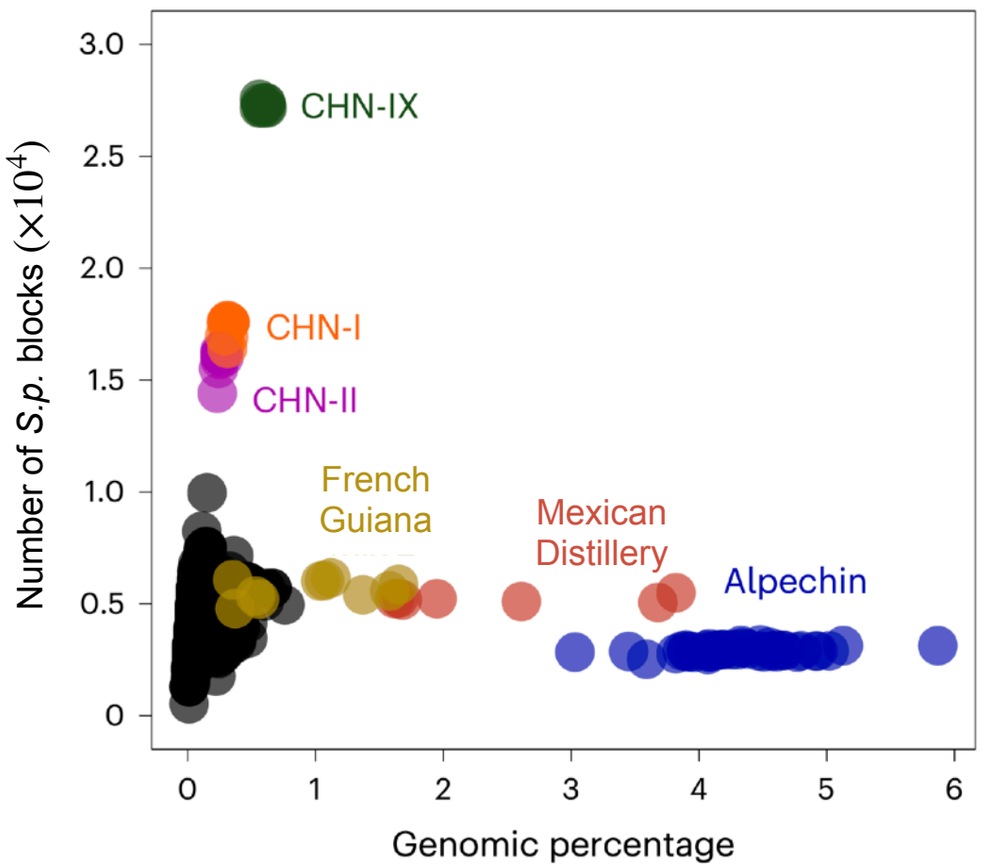
Pan J, …, & Keeney S. CELL. 2011.

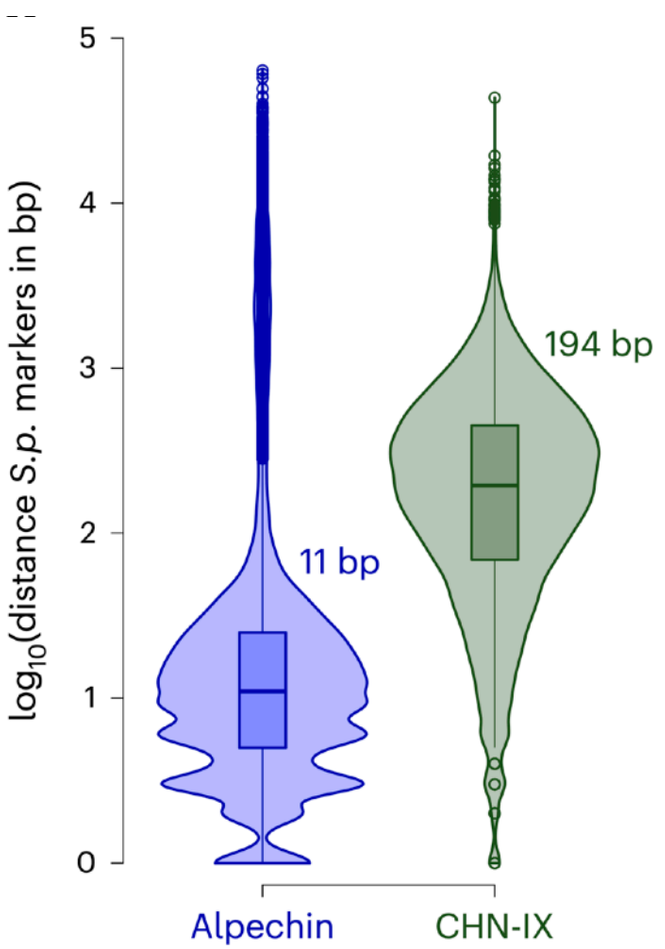
Mancera E, …, & Steinmetz LM. NATURE. 2008.

ancient and recent origins of polymorphism

in natural populations

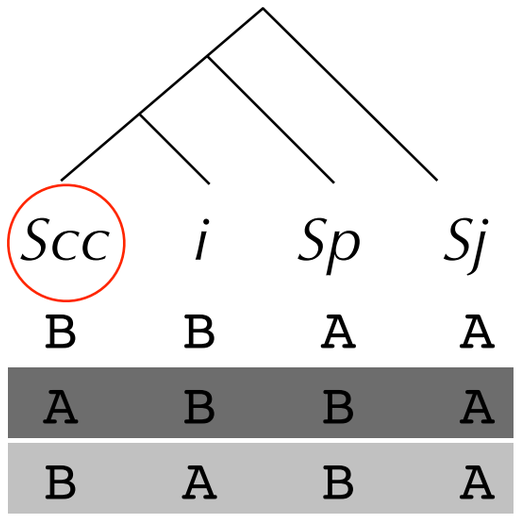


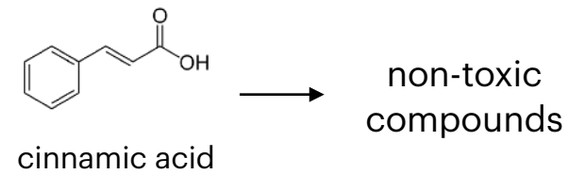


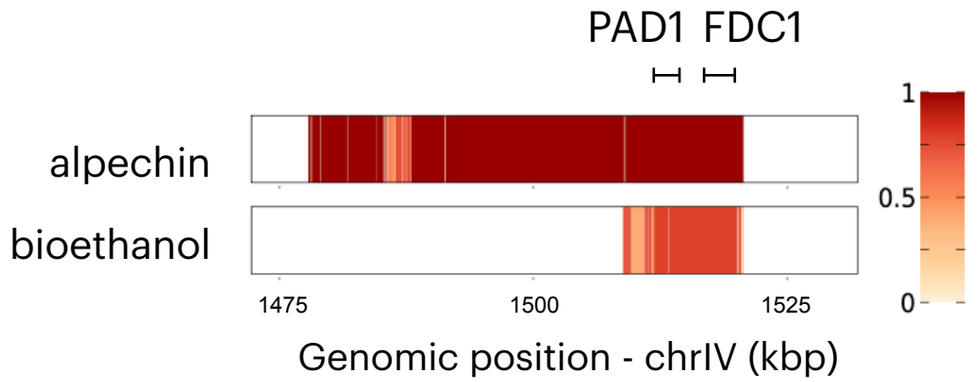


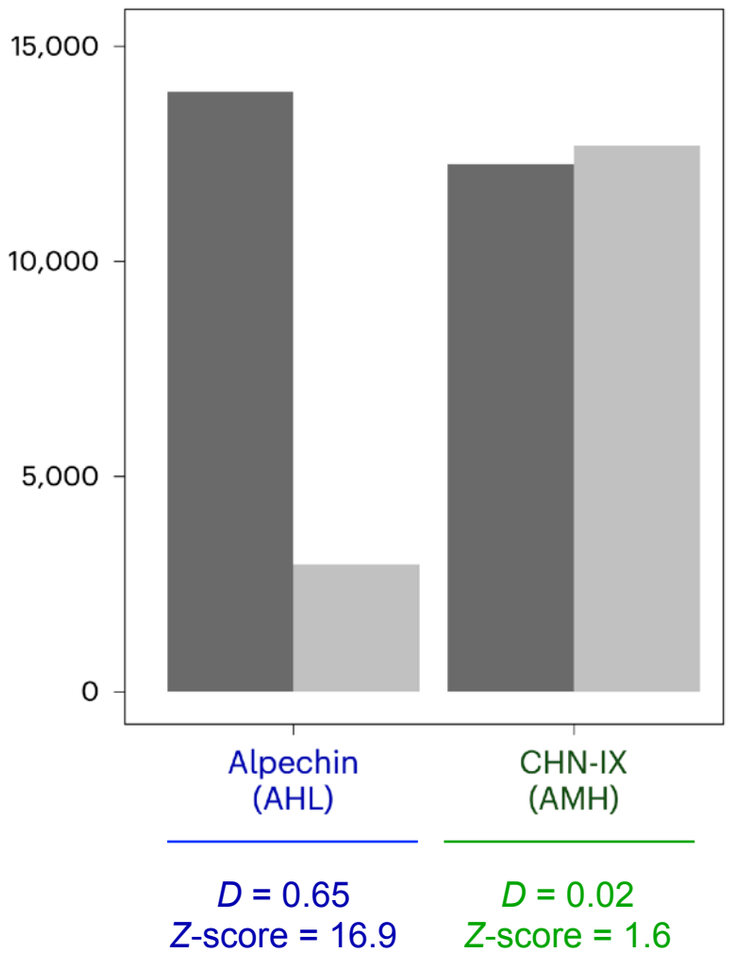
Tellini N, …, & Liti G. NAT ECOL EVOL. 2024.

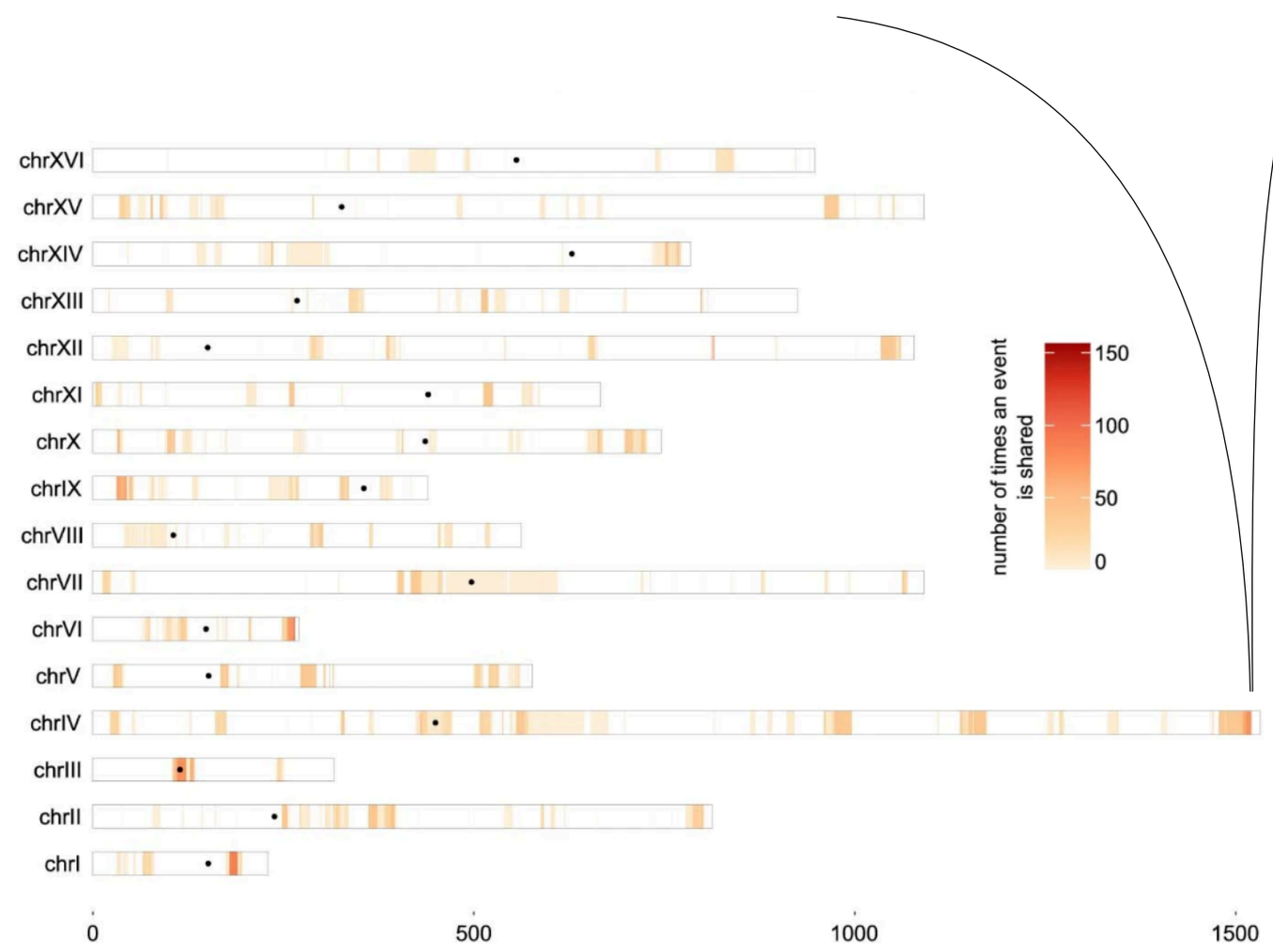
convergent adaptation





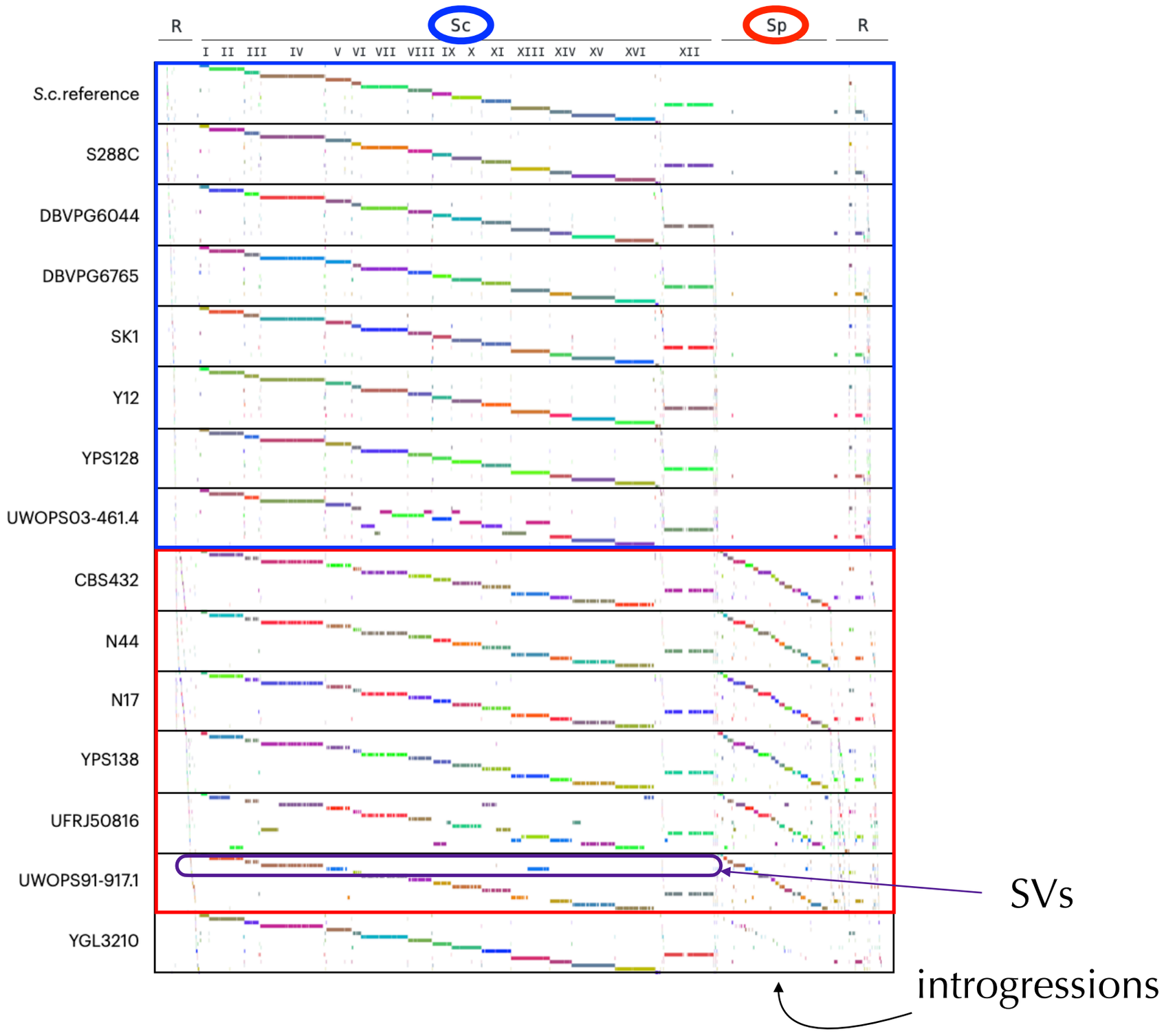






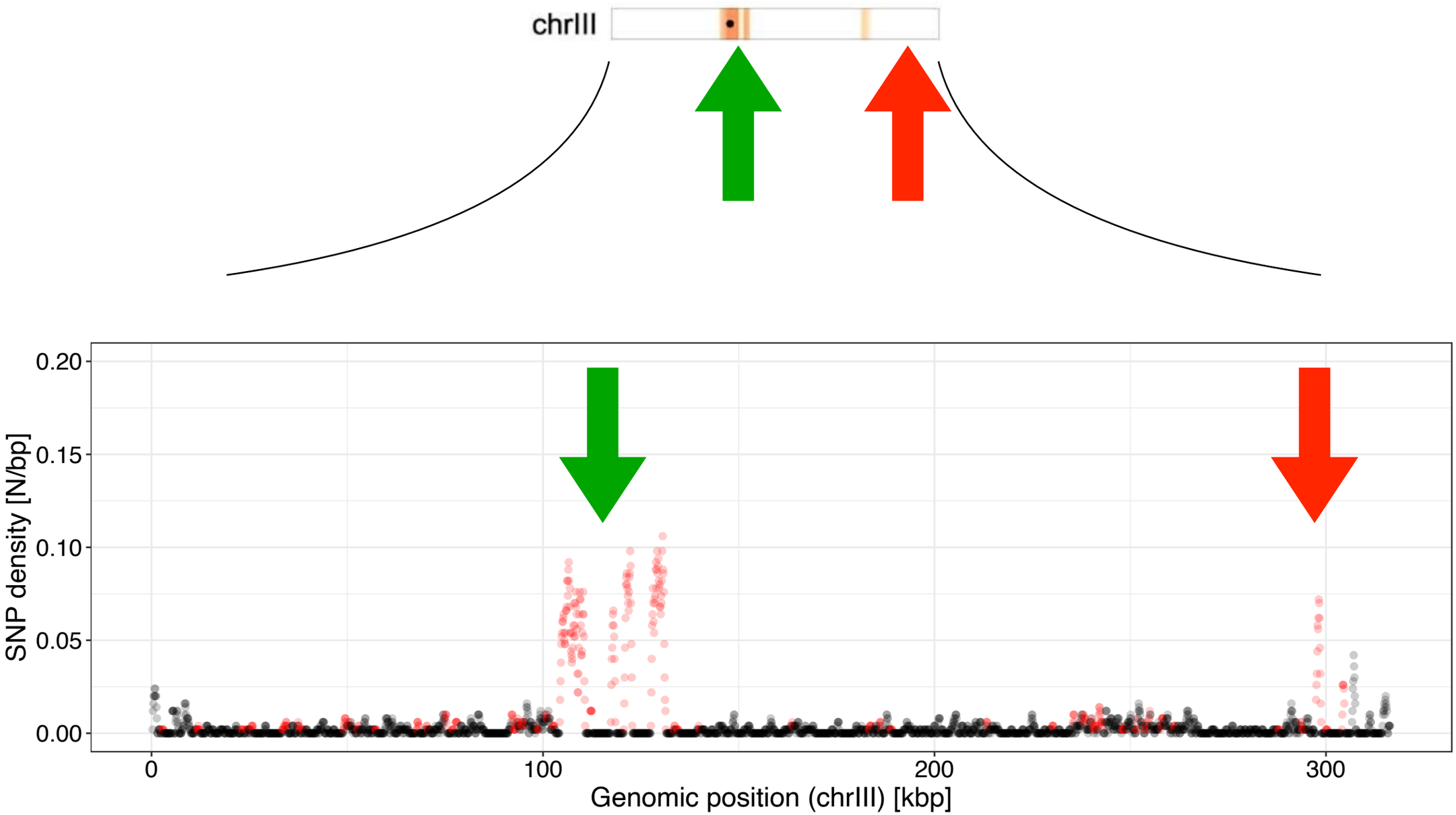
Tellini N, …, & Liti G. NAT ECOL EVOL. 2024.

pangenomes with variation graphs



Tattini L, … & Liti G. EMBO WORKSHOP. 2022.

introgressions on a graph



the scrap/pggb variation graph

nature genetics

Article

https://doi.org/10.1038/s41588-023-01459-y

Telomere-to-telomere assemblies of 142 strains characterize the genome structural landscape in *Saccharomyces cerevisiae*

Received: 1 November 2022

Accepted: 26 June 2023

Published online: 31 July 2023

Check for updates

Samuel O'Donnell[1,7], Jia-Xing Yue[2,3,7], Omar Abou Saada[4], Nicolas Agier[1], Claudia Caradec[4], Thomas Cokelaer[5,6], Matteo De Chiara[3], Stéphane Delmas[1], Fabien Dutreux[4], Téo Fournier[4], Anne Friedrich[4], Etienne Kornobis[5,6], Jing Li[2,3], Zepu Miao[2], Lorenzo Tattini[3], Joseph Schacherer[4,8] ✉, Gianni Liti[3,8] ✉ & Gilles Fischer[1,8] ✉

Saccharomyces cerevisiae reference assembly panel

O'Donnell S*, Yue JX*, …, Tattini L, Schacherer J, Liti G & Fischer G. NAT GENET. 2023.

nature methods

Brief Communication

https://doi.org/10.1038/s41592-024-02430-3

Building pangenome graphs

Received: 30 July 2023

Accepted: 26 August 2024

Published online: 21 October 2024

Check for updates

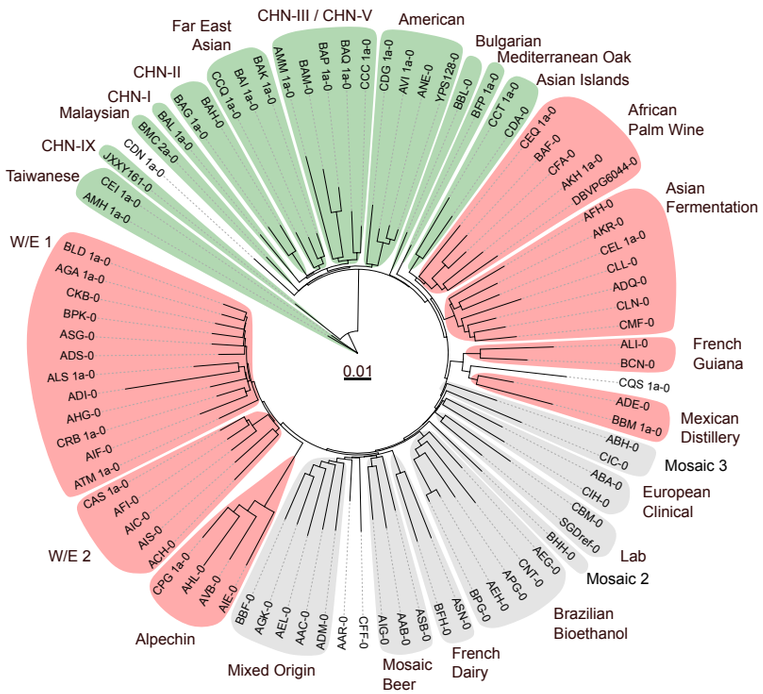
Erik Garrison[1,30] ✉, Andrea Guarracino[1,2,30], Simon Heumos[3,4,5], Flavia Villani[1], Zhigui Bao[6,7], Lorenzo Tattini[8,9], Jörg Hagmann[10], Sebastian Vorbrugg[6], Santiago Marco-Sola[11,12], Christian Kubica[6], David G. Ashbrook[1], Kaisa Thorell[13], Rachel L. Rusholme-Pilcher[14], Gianni Liti[8], Emilio Rudbeck[15], Agnieszka A. Golicz[16], Sven Nahnsen[3,4,5,17], Zuyu Yang[18], Moses Njagi Mwaniki[19], Franklin L. Nobrega[20], Yi Wu[20], Hao Chen[21], Joep de Ligt[22], Peter H. Sudmant[23], Sanwen Huang[7], Detlef Weigel[6,24], Nicole Soranzo[2,25,26,27,28], Vincenza Colonna[1,29], Robert W. Williams[1] & Pjotr Prins[1]

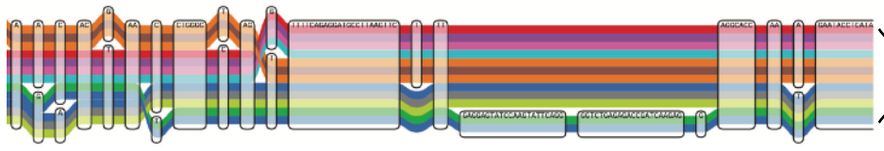
pan-genome graph builder

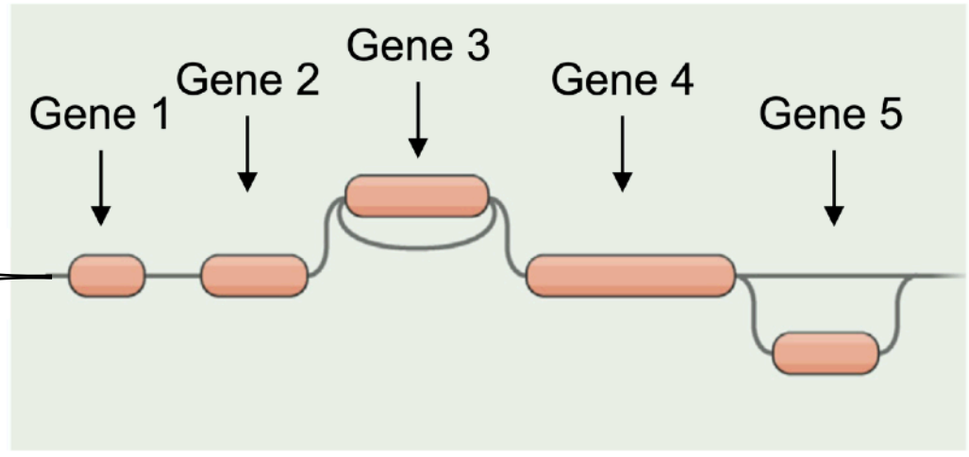
Garrison E*@, Guarracino A*, …, Prins P. NATURE METHODS. 2024.

phylogeny/admixture, & pangenome content

SGP
EVA



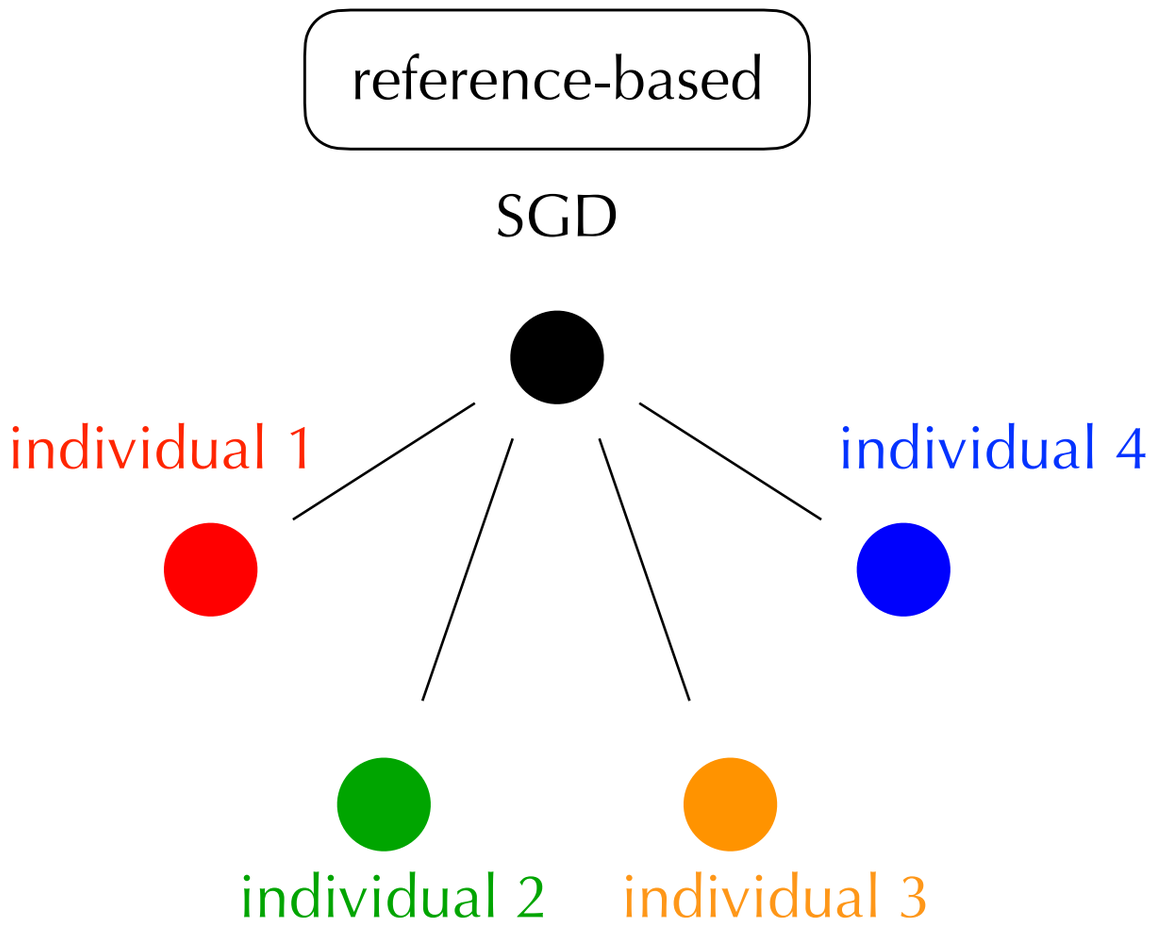


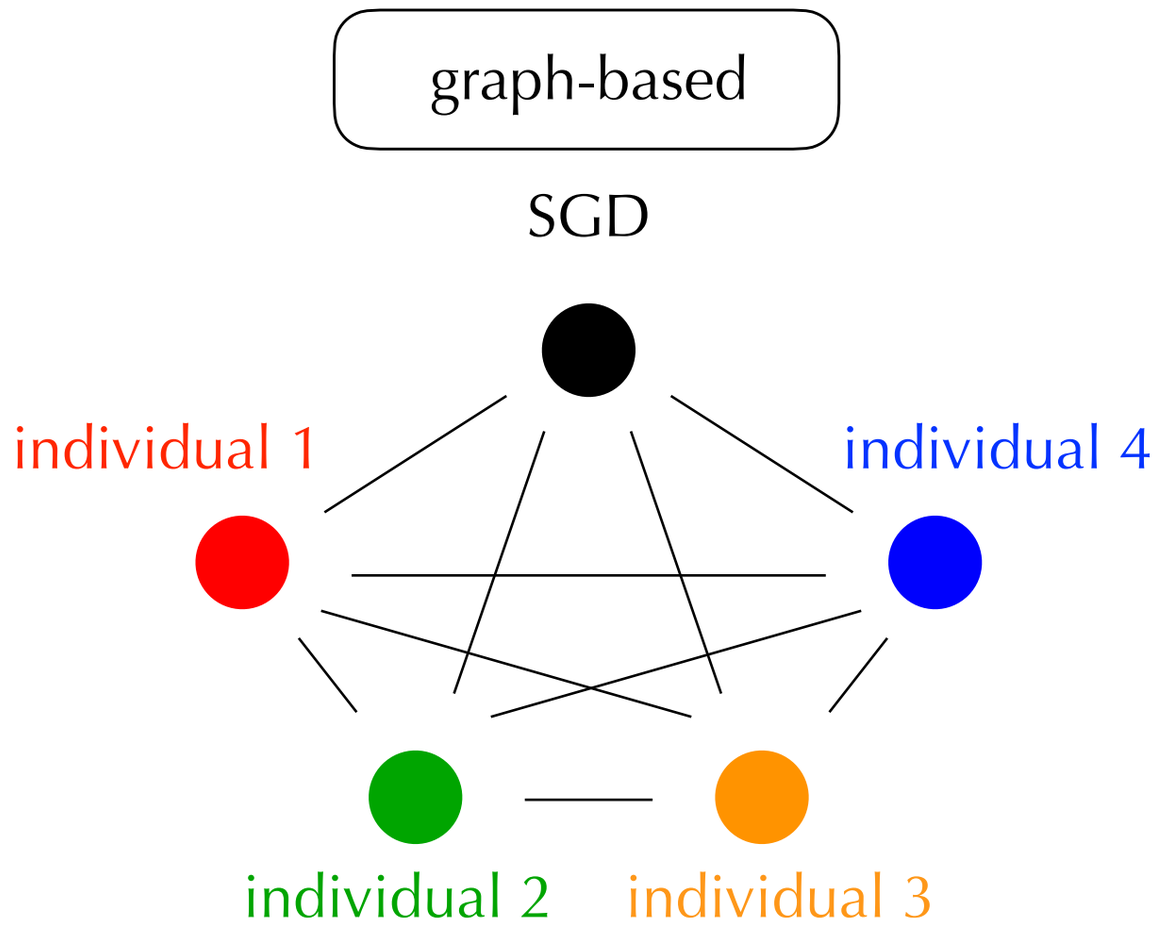


https://github.com/lt11/panda

https://github.com/lt11/sgp-eva

the *Saccharomyces* graph phylogeny evaluation tool





nucmer	variants	pggb + vg (vs SGD)
	nodes	pggb + odgi
	clusters	phyBWT2/distBWT2

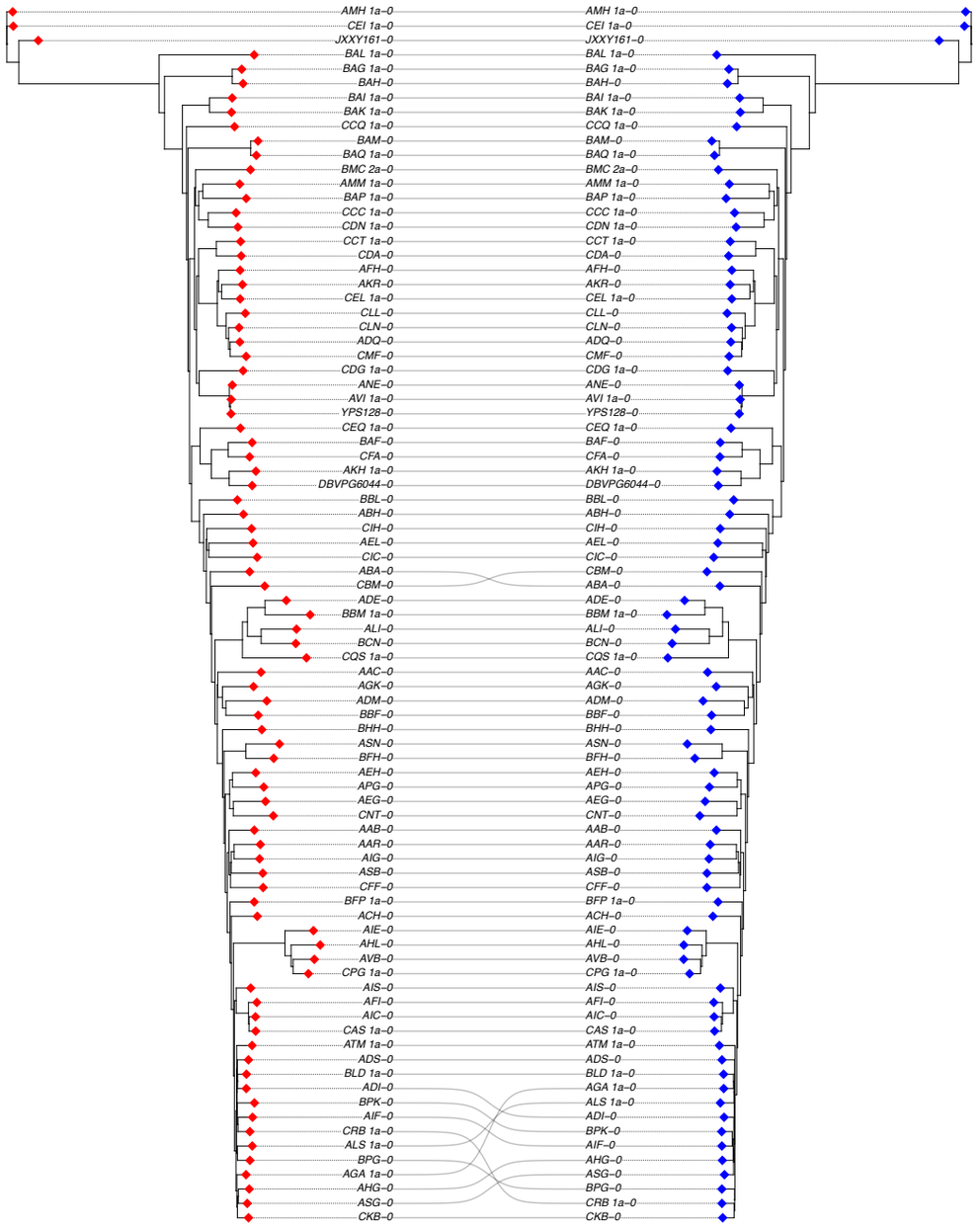
cosine distance

neighbour-joining algorithm

comparing phylogenies

using reference-based variants

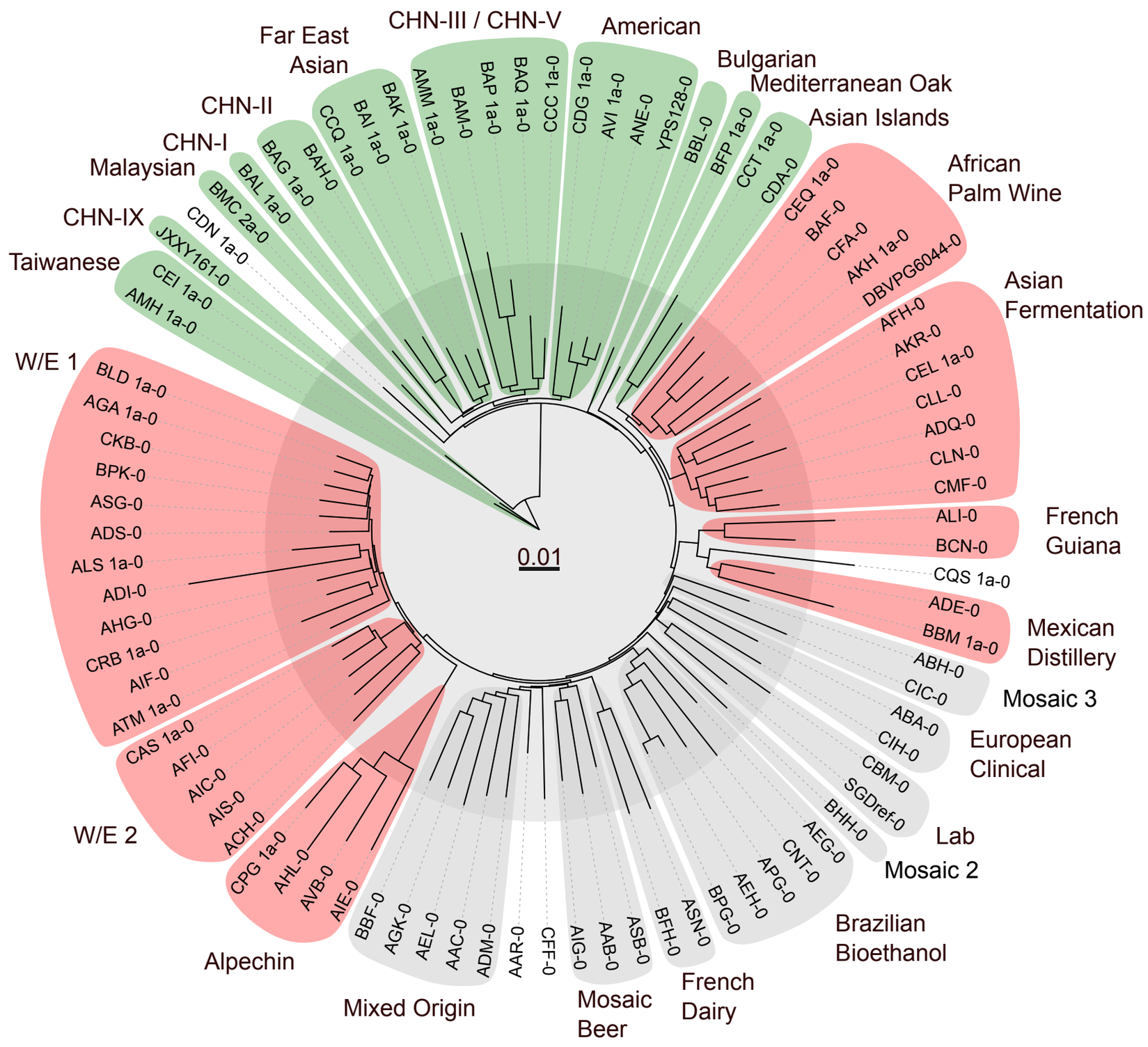
nucmer



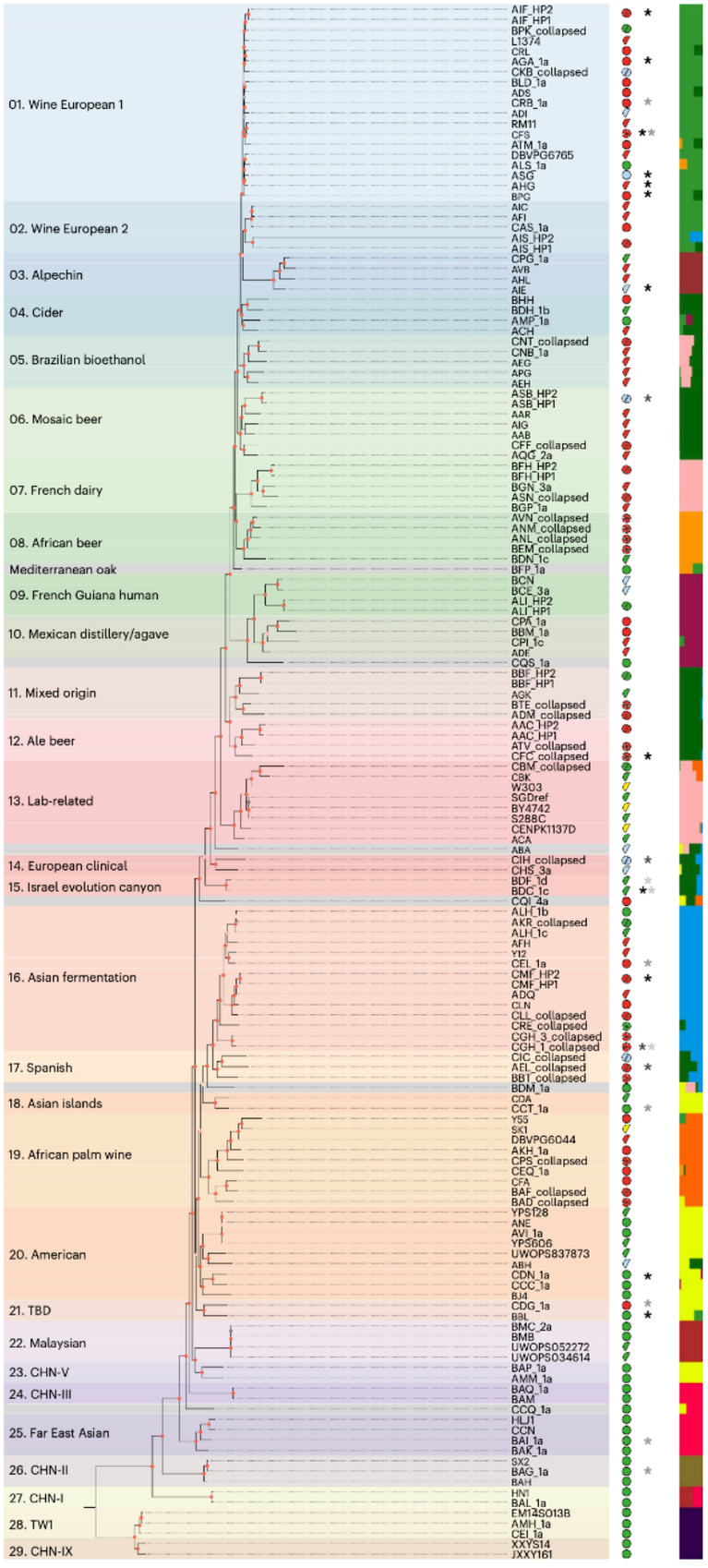
pggb + vg (vs SGD)

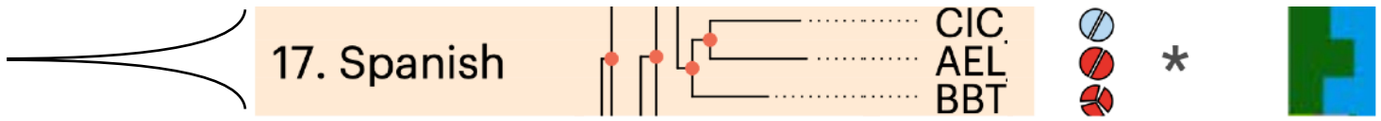
$d_{GRF} = 0.08$

phylogeny overview



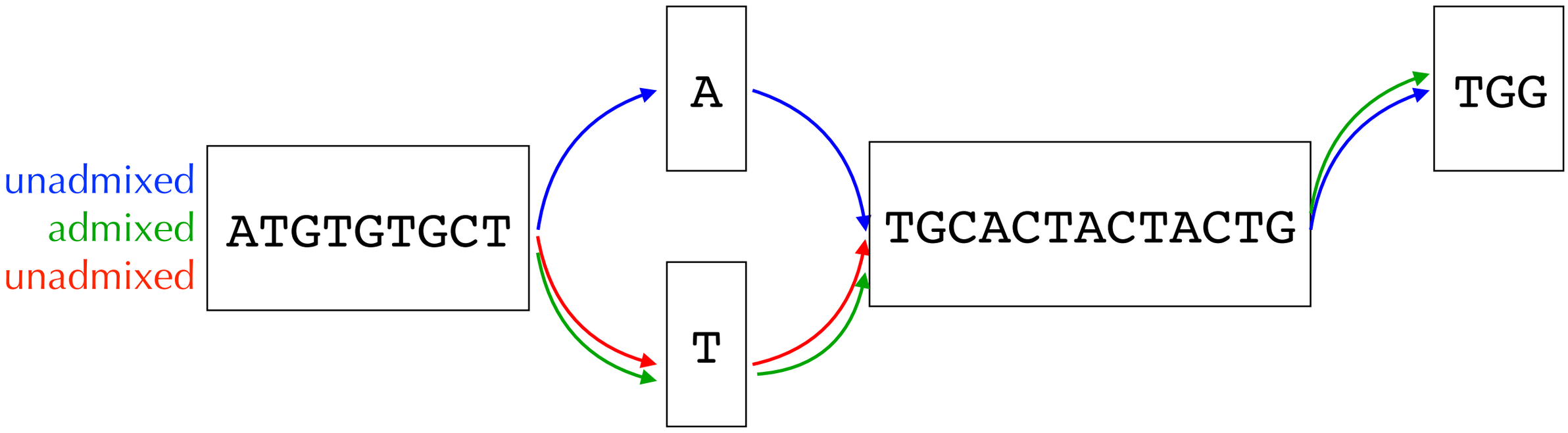
unsolved clades and admixture levels





O'Donnell S*, Yue JX*, …, Tattini L, Schacherer J, Liti G & Fischer G. NAT GENET. 2023.

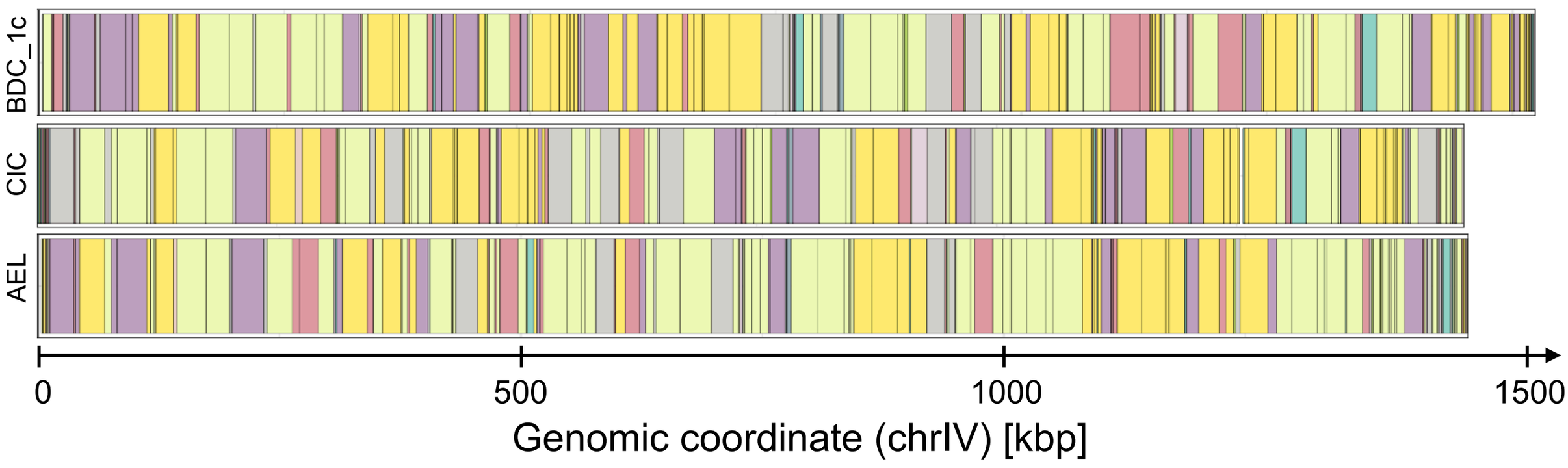
origin & admixture on a graph



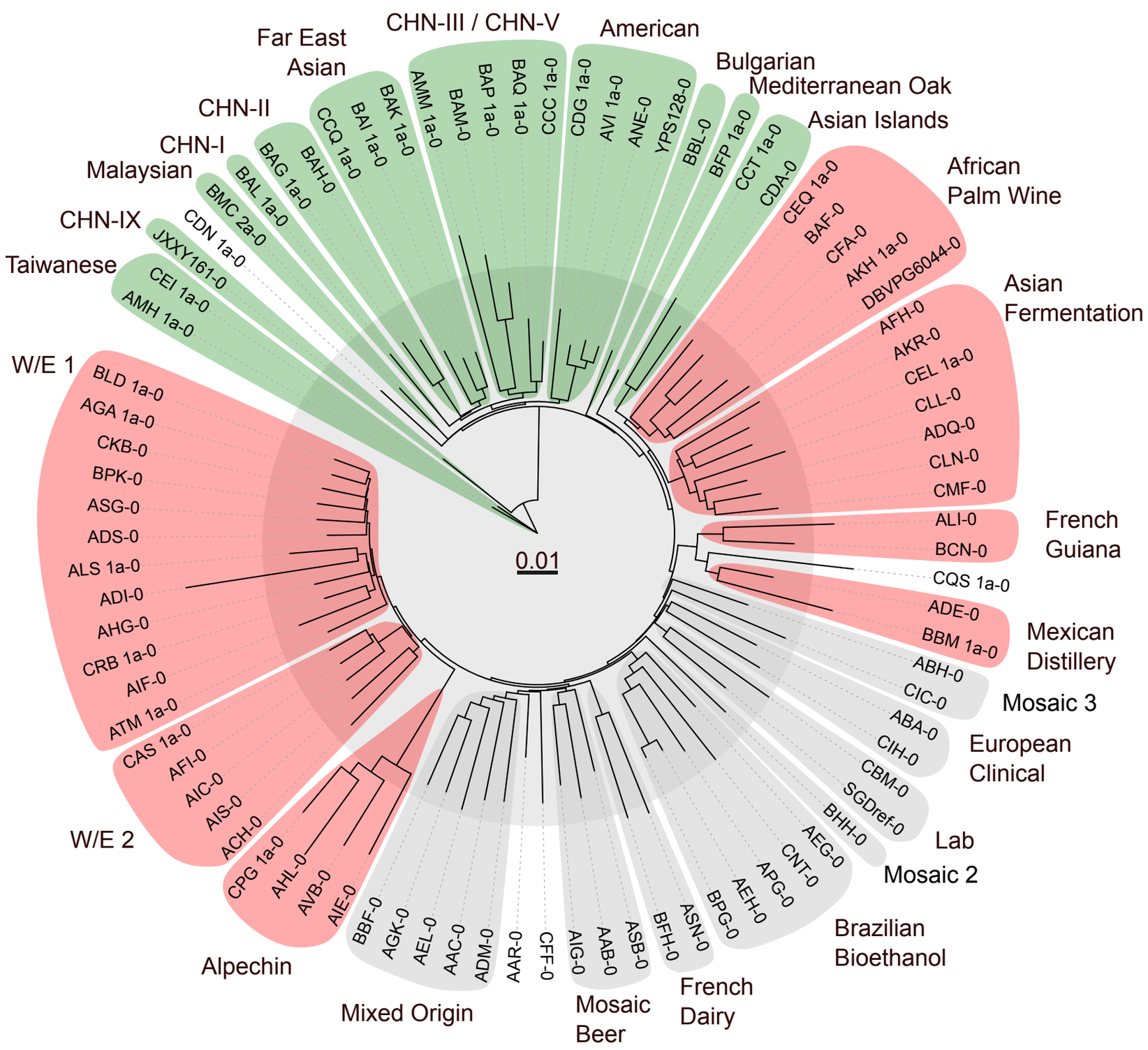
strain	blue	red
green	75.0	25.0

admixture on a graph

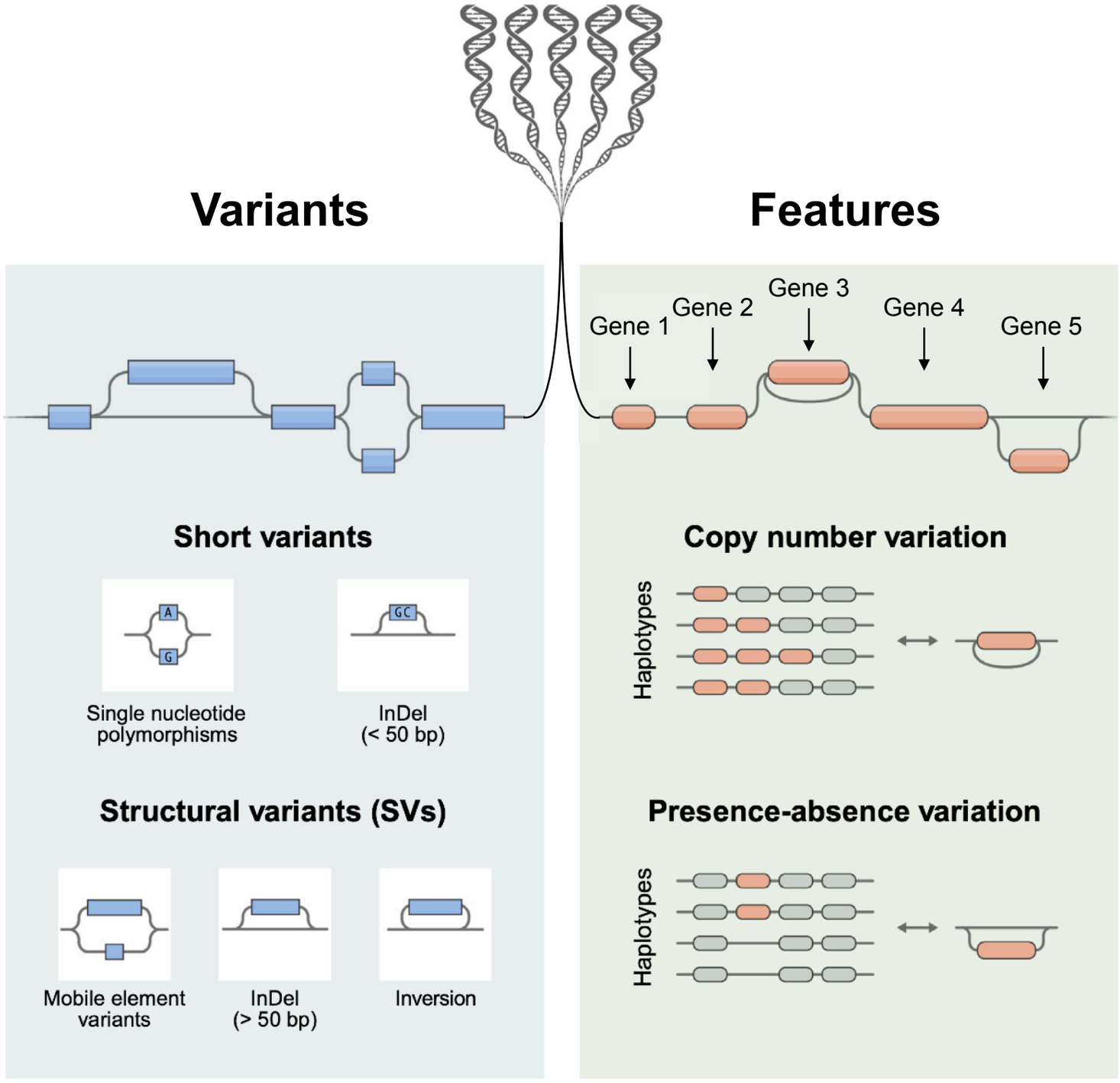
strain	tai	c9	mal	c2	fea	c3	ame	bul	medo	afe	fgu	mda	afb	we1
ADI	0.6	0.1	0.3	0.3	0.2	0.3	0.4	1.5	6.7	0.6	4.3	0.3	0.9	83.7
AGA_1a	0.1	0.0	0.1	0.4	0.2	0.2	0.0	0.9	9.5	0.2	0.8	0.2	0.9	86.4
ABH	0.2	0.1	2.0	1.0	1.3	2.0	36.0	5.1	5.3	13.0	1.0	0.8	0.4	31.7
BDC_1c	0.3	0.1	1.9	0.9	2.5	1.7	10.9	4.9	10.8	34.7	1.3	0.7	2.2	27.0
CIC	0.3	0.2	1.1	0.9	1.4	1.2	15.1	4.6	9.9	39.3	1.0	0.5	1.0	23.4
AEL	0.4	0.4	0.7	0.5	1.1	0.6	5.7	6.3	17.1	41.1	0.9	0.3	1.5	23.5
ADM	0.3	0.3	1.0	0.4	0.9	0.5	5.5	3.7	17.4	25.5	0.8	0.4	1.0	42.4
BBF	0.5	0.1	1.2	0.5	1.0	0.9	3.5	3.2	18.0	20.9	1.5	0.5	1.0	47.2



the ADI case


0.01
Far East Asian
CHN-III / CHN-V
American
Bulgarian
Mediterranean Oak
Asian Islands
CHN-II
CHN-I
Malaysian
CHN-IX
Taiwanese
African Palm Wine
Asian Fermentation
W/E 1
French Guiana
Mexican Distillery
Mosaic 3
European Clinical
Lab
Mosaic 2
Brazilian Bioethanol
French Dairy
Mosaic Beer
Mixed Origin
Alpechin
W/E 2
BAK 1a-0
BAI 1a-0
CCQ 1a-0
BAH-0
BAG 1a-0
BAL 1a-0
BMC 2a-0
CDN 1a-0
JXXY161-0
CEI 1a-0
AMH 1a-0
AMM 1a-0
BAM-0
BAP 1a-0
BAQ 1a-0
CCC 1a-0
CDG 1a-0
AVI 1a-0
ANE-0
YPS128-0
BBL-0
BFP 1a-0
CCT 1a-0
CDA-0
CEQ 1a-0
BAF-0
CFA-0
AKH 1a-0
DBVPG6044-0
AFH-0
AKR-0
CEL 1a-0
CLL-0
ADQ-0
CLN-0
CMF-0
ALI-0
BCN-0
CQS 1a-0
ADE-0
BBM 1a-0
ABH-0
CIC-0
ABA-0
CIH-0
CBM-0
SGDref-0
BHH-0
AEG-0
CNT-0
APG-0
AEH-0
BPG-0
ASN-0
BFH-0
ASB-0
AAB-0
AIG-0
CFF-0
AAR-0
ADM-0
AAC-0
AEL-0
AGK-0
BBF-0
AIE-0
AVB-0
AHL-0
CPG 1a-0
ACH-0
AIS-0
AIC-0
AFI-0
CAS 1a-0
ATM 1a-0
AIF-0
CRB 1a-0
AHG-0
ADI-0
ALS 1a-0
ADS-0
ASG-0
BPK-0
CKB-0
AGA 1a-0
BLD 1a-0

variants vs features



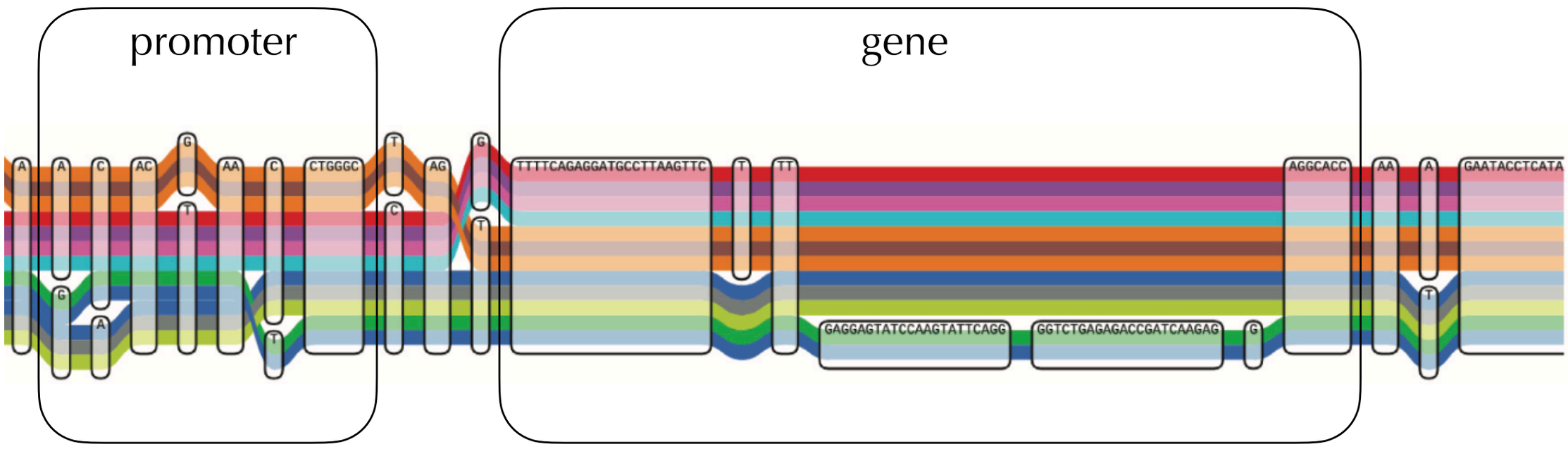
Tonkin-Hill G, …, & Parkhill J. GENOME BIOL. 2020.

Li H, Marin M & Farhat MR. BIOINFORMATICS. 2024.

Bergström A, …, & Liti G. MOL BIOL EVOL. 2014.

Gautreau G, …, & Vallenet D. PLOS COMPUT BIOL. 2020.

pangenomics on a graphs: feature presence/absence



systematic-id-**YGL023W**-reference
systematic-id-**YGL023C**-S288C
random-id-**AAB0543**-AAB
sequence-**chrI:123-456**-SK1

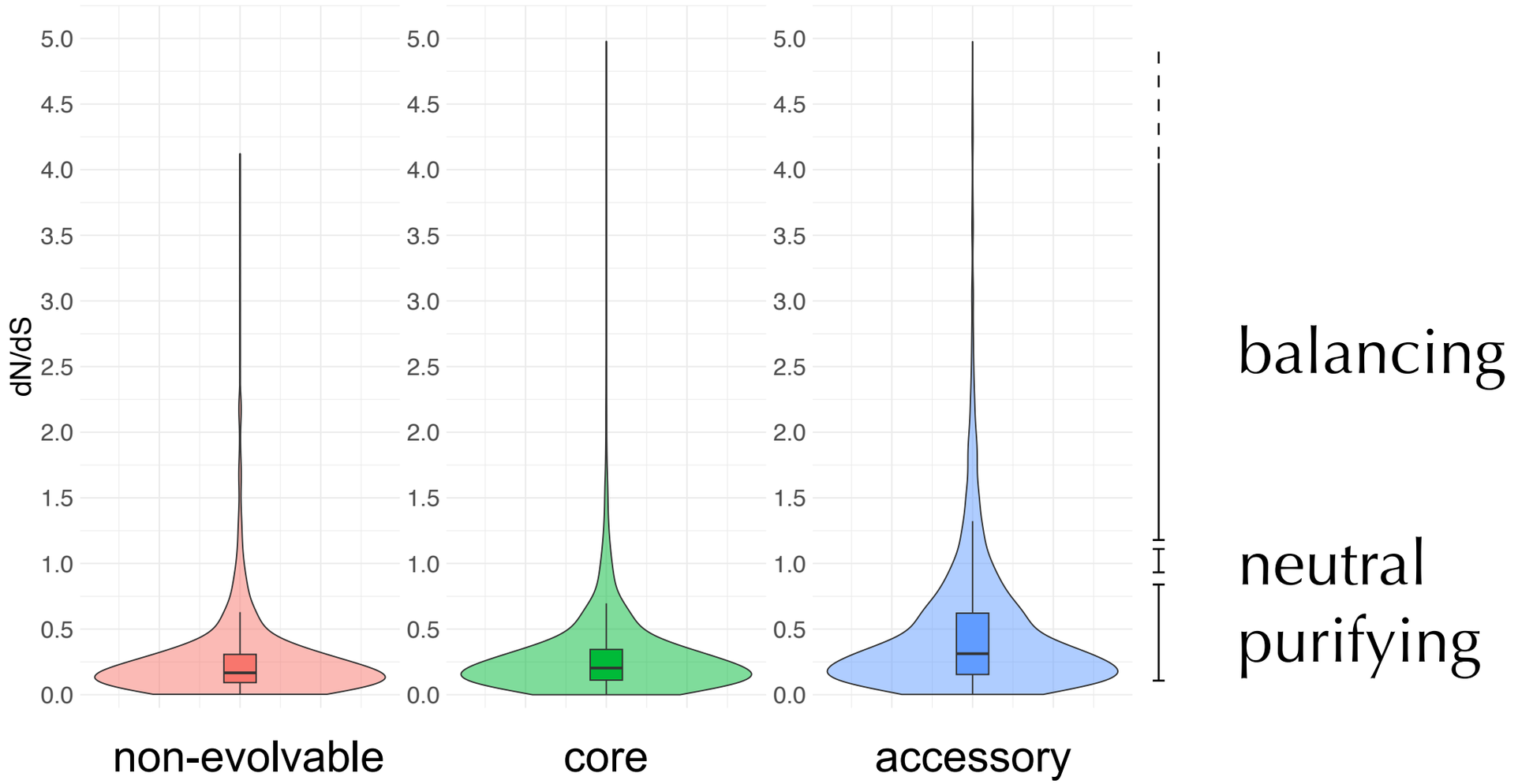
validation
- 4 genes (no sequence)
+ 976 genes

type	sequences	gene families
non-evolvable	91,499	854
core	471,450	4,647
accessory	37,706	4,291

https://github.com/lt11/panda

Liu G, …, & Rancati G. CELL. 2015.

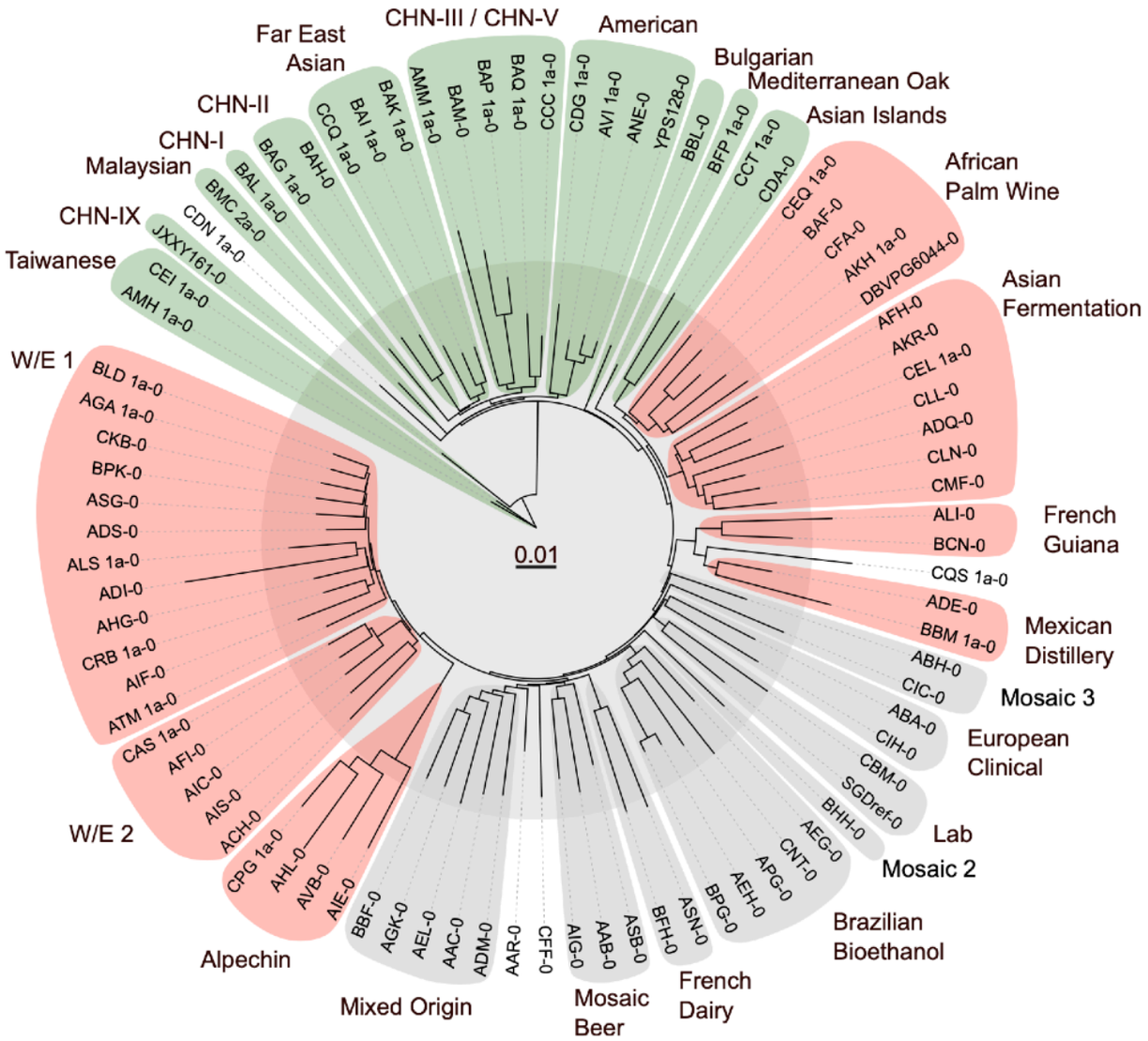
quantifying selection



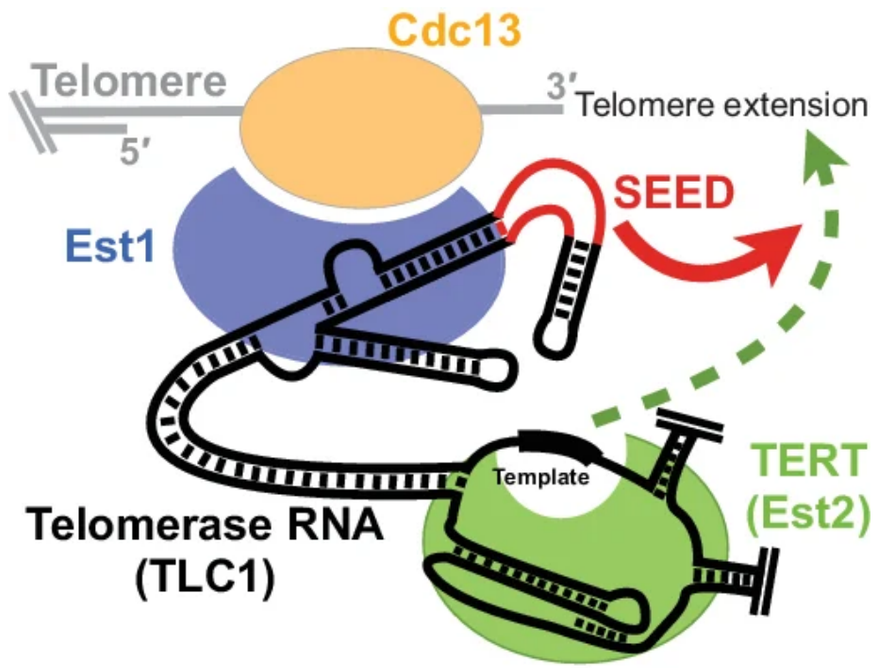
Peter J*, De Chiara M*, …, Liti G@ & Schacherer J@. NATURE. 2018.

Liu G, …, & Rancati G. CELL. 2015.

pangenome gene content

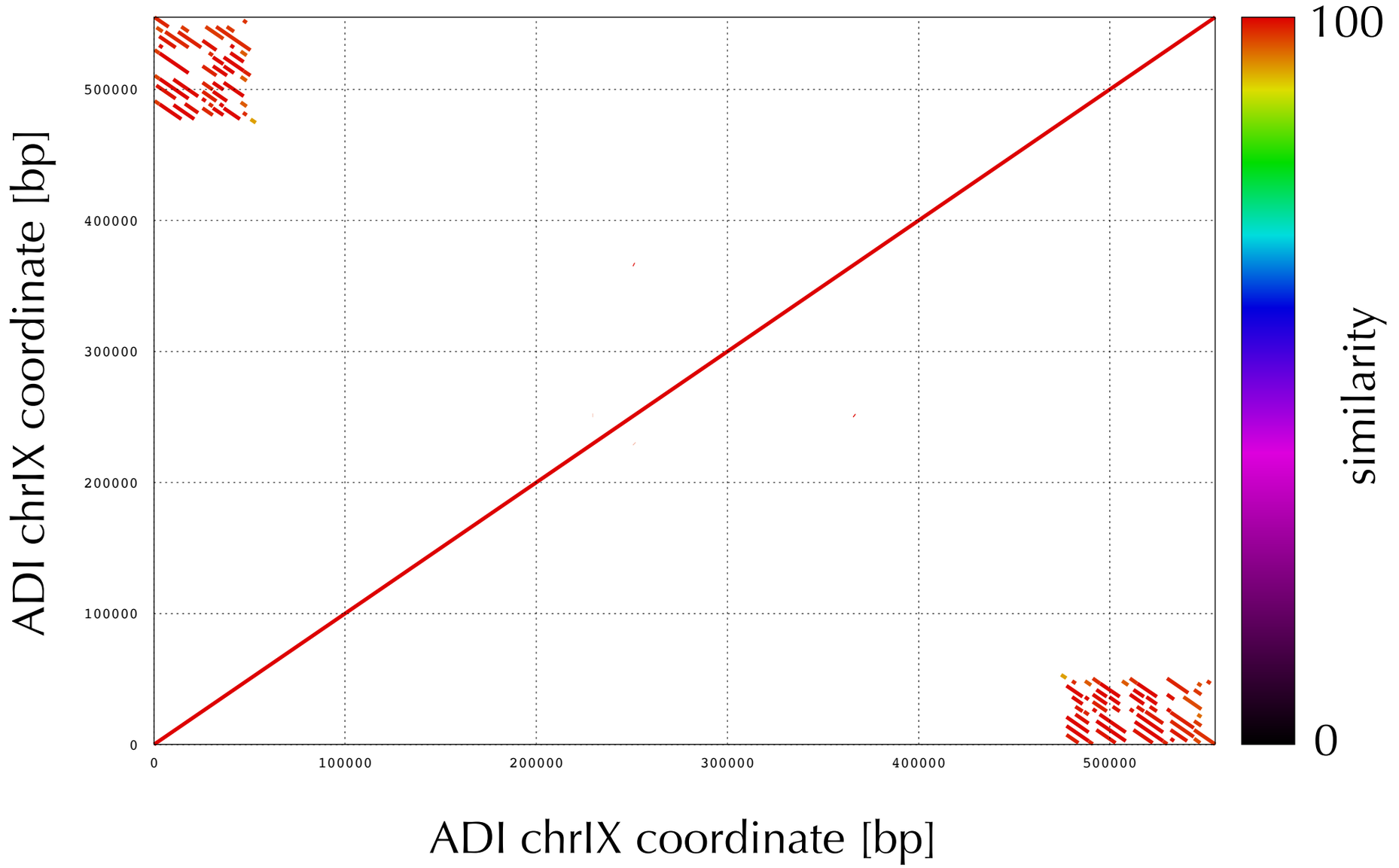


strain	RID genes
ADI	263
CQS_1a	65
ATM_1a	62
AIF_HP2	60
BBM_1a	49
DBVPG6044	37
CNT_HP1	36
CEQ_1a	35
...	...

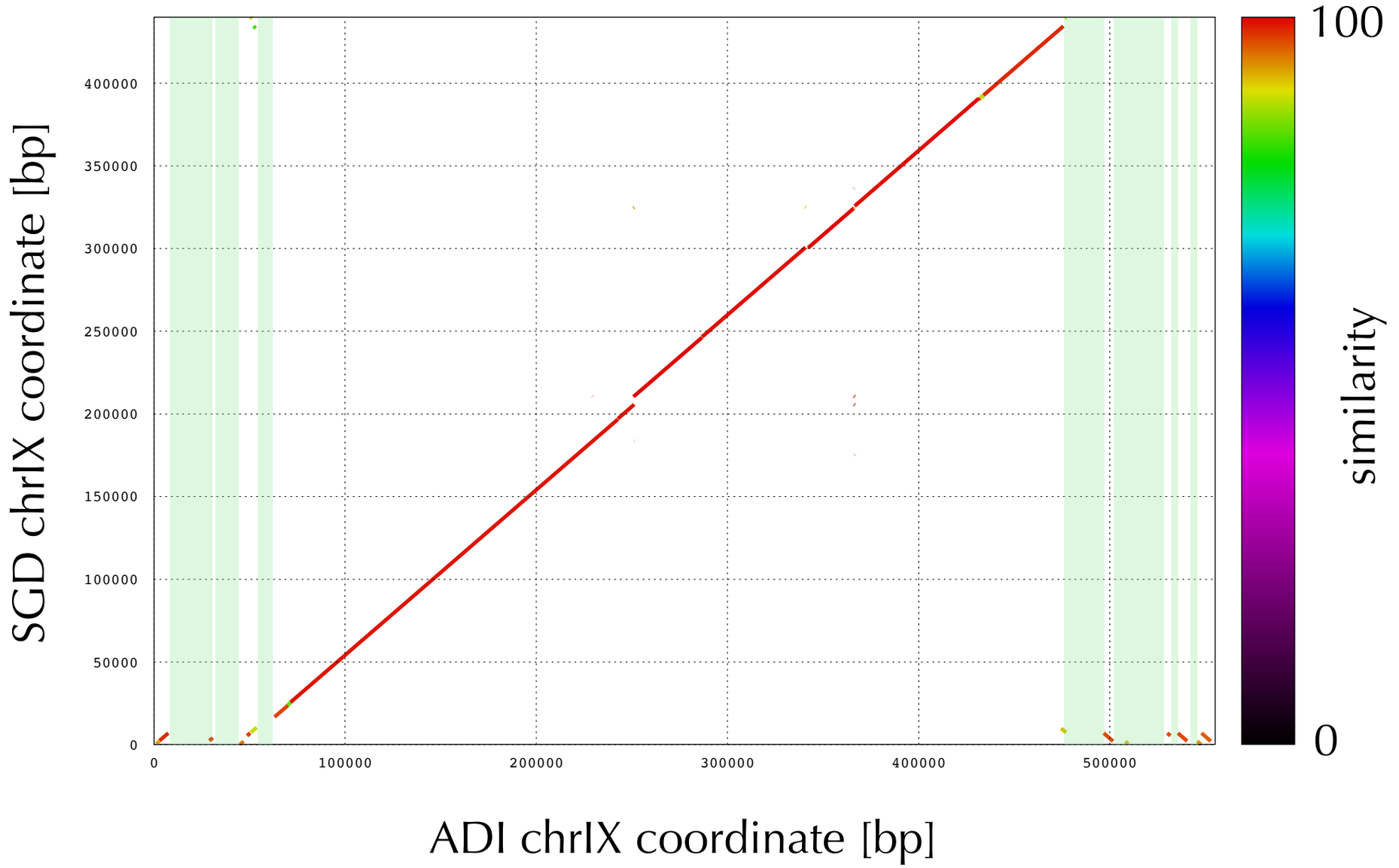


Lundblad V & Blackburn EH. CELL. 1993.

genomic loci of RID genes

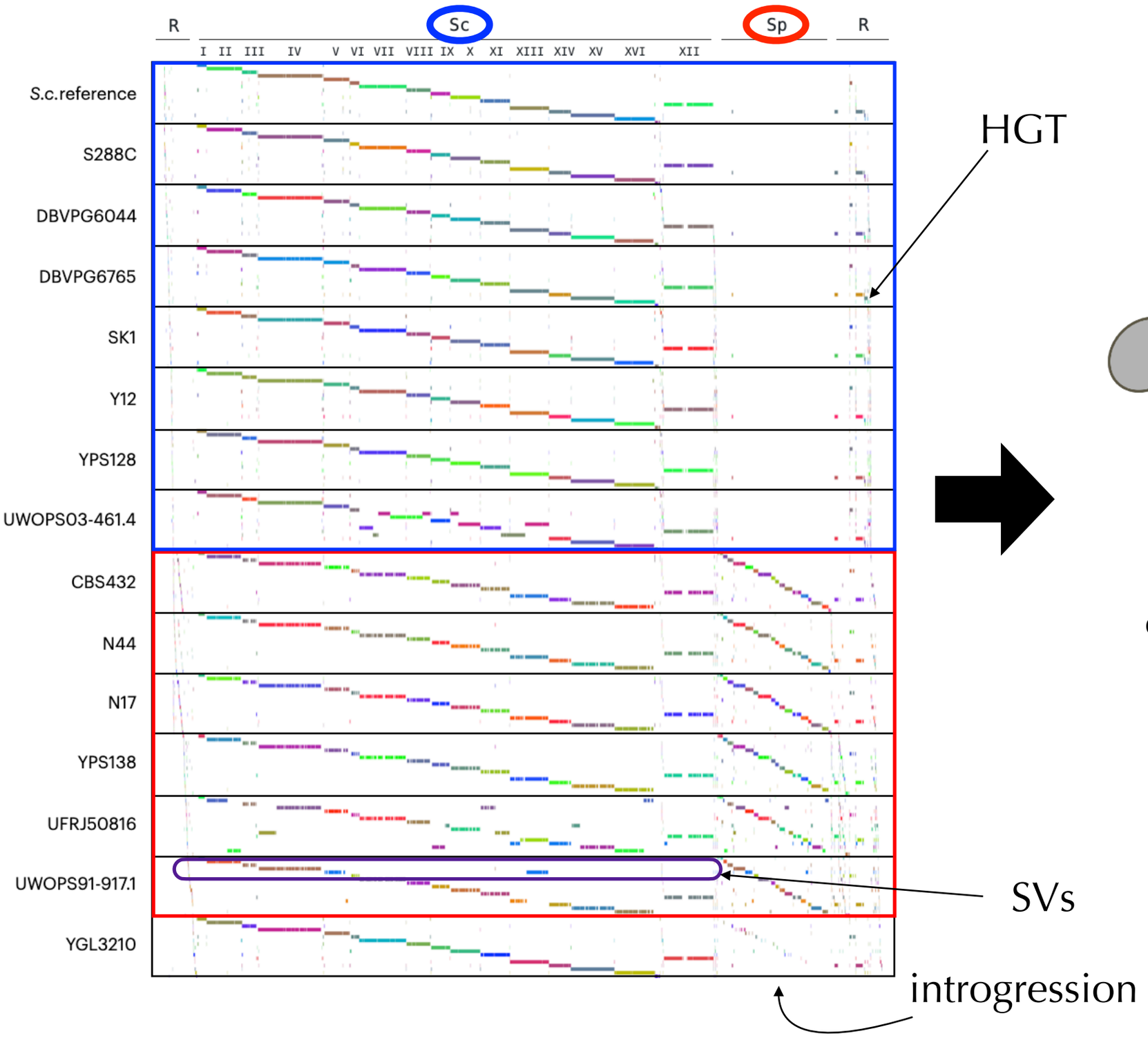


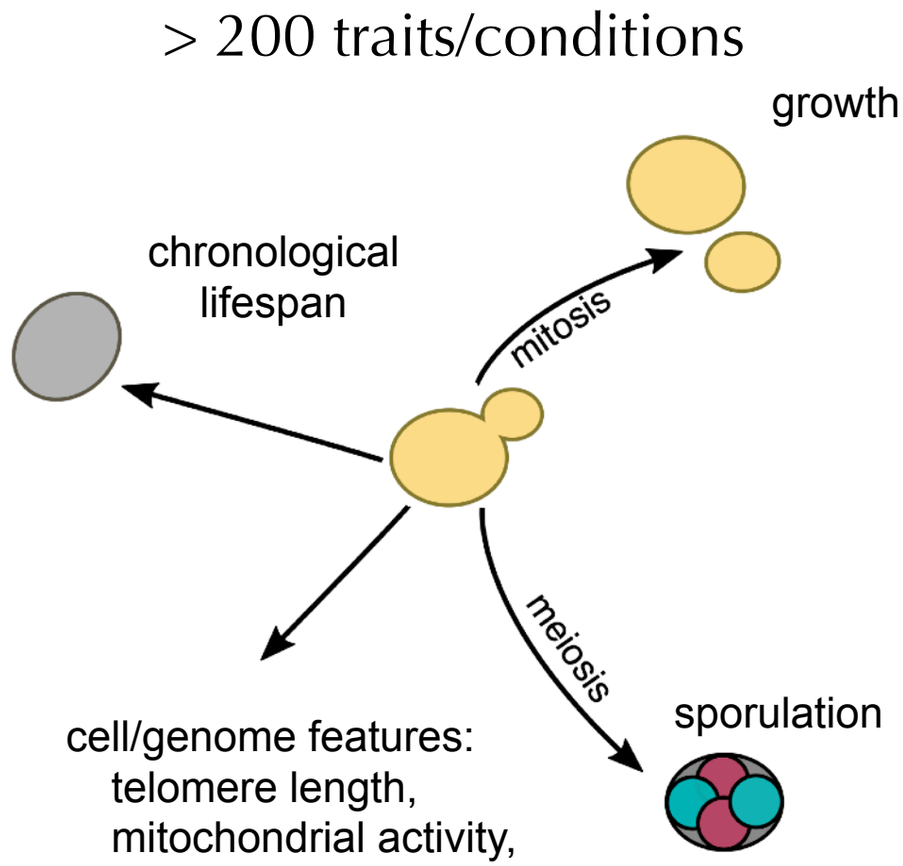
extra material



summary & perspetives

competitive mapping vs variation graphs





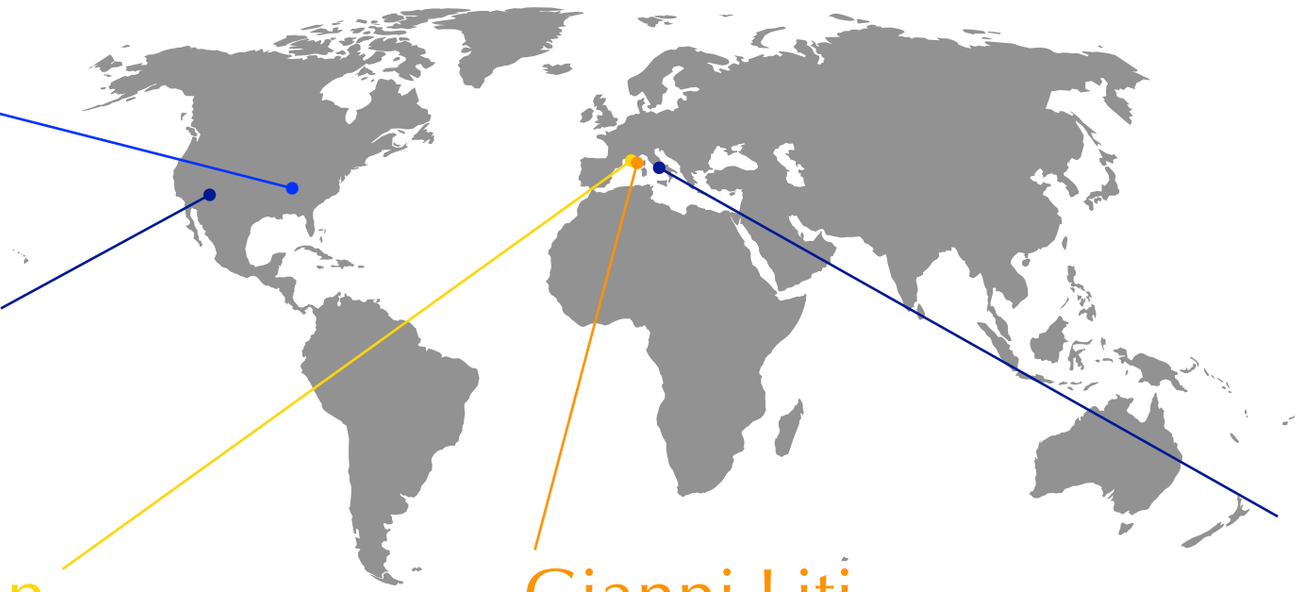
Tattini L, … & Liti G. EMBO WORKSHOP. 2022.

Khaiwal S, … & Liti G. MOL SYST BIOL. 2025.

collaborations

Erik Garrison
UTHSC, USA

Andrea Guarracino
TGen Institute, USA



Etienne Danchin
ISA

Gianni Liti
IRCAN

Roberto Grossi
UniPi, Italy