



THE UNIVERSITY
of EDINBURGH



Ancestral recombination Graphs (ARGs)

Gregor Gorjanc, Chris Gaynor, Jon Bancic, Daniel Tolhurst

UNE, Armidale

2024-02-09



Learning objectives

- Motivate tree-thinking as a way forward in genomics
- Revisiting the fundamentals of DNA events and how to efficiently encode DNA variation
- Understand ARGs & tree sequence format
- Showcase ongoing agricultural applications

Warning disclaimer

Active area of learning, exploration, & research in our lab!!!

- Ideas & work in progress (pre-publication stage!)
- Building experience with applications at this stage

What is ancestral recombination graph?

- Description of events that generated DNA of our samples
- Ultimate population genetics object :)
- Theoretical/Impractical concept for a long time :(

Intro literature

REVIEW





The era of the ARG: An introduction to ancestral recombination graphs and their significance in empirical evolutionary genomics

Alexander L. Lewanski ^{1,2,3,4*}, Michael C. Grundler ⁴, Gideon S. Bradburd ^{2,4}

Inferring whole-genome histories in large population datasets

Jerome Kelleher ^{*}, Yan Wong, Anthony W. Wohns , Chaimaa Fadil , Patrick K. Albers 
and Gil McVean 

The Promise of Inferring the Past Using the Ancestral Recombination Graph

Débora Y.C. Brandt ^{1,*}, Christian D. Huber ^{2,*}, Charleston W.K. Chiang ^{3,4,*},
and Diego Ortega-Del Vecchyo ^{5,*}

A general and efficient representation of ancestral recombination graphs

Yan Wong¹, Anastasia Ignatieva^{2,3*}, Jere Koskela^{4,5*}, Gregor Gorjanc⁶,
Anthony W. Wohns^{7,8}, and Jerome Kelleher^{1†}

What drives our interest in ARGs?

- Many branches of genetics
 - Molecular genetics
 - Population genetics
 - Phylogenetics
 - Conservation genetics
 - ...
 - Quantitative genetics
 - Animal and Plant Breeding
 - ...

It often feels as if different geneticists live on different planets (obviously focus&applications differ)

What drives our need/hope in ARGs?

- Can we effectively leverage haplotype information?
- Can we do quantitative genetics within and between families, populations, ...?
 - Plant breeding families (linkage vs linkage-disequilibrium information)
 - Pig & poultry breeding (many lines)
 - Cattle crossbreds, including Bos Taurus & Bos Indicus crosses
 - Admixed cattle in Africa (lots of breeds and populations)
- Partially missing pedigree and genetic groups
- Ancestral alleles, mutations, rare, & common alleles
- Limited cross-talk between population and quantitative genetics
- Genomic data is getting extremely big!

Intro literature

Population genetics

<https://doi.org/10.1038/s41588-023-01389-9>

Using enormous genealogies to map causal variants in space and time

Kelley Harris

 Check for updates

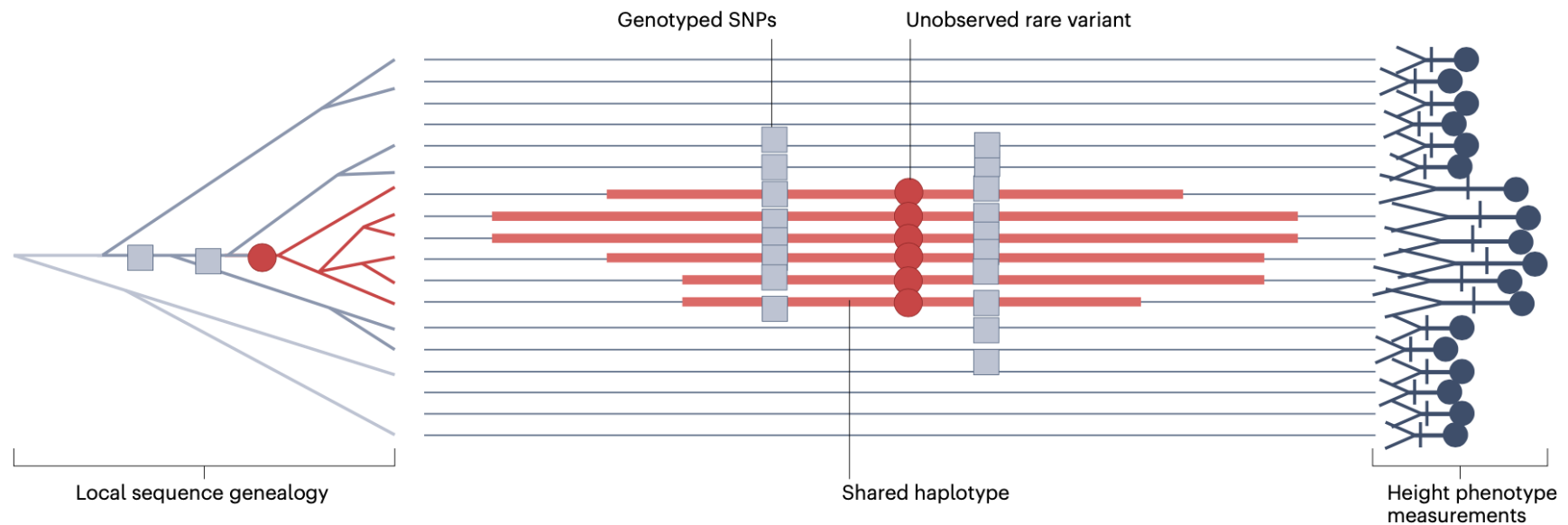
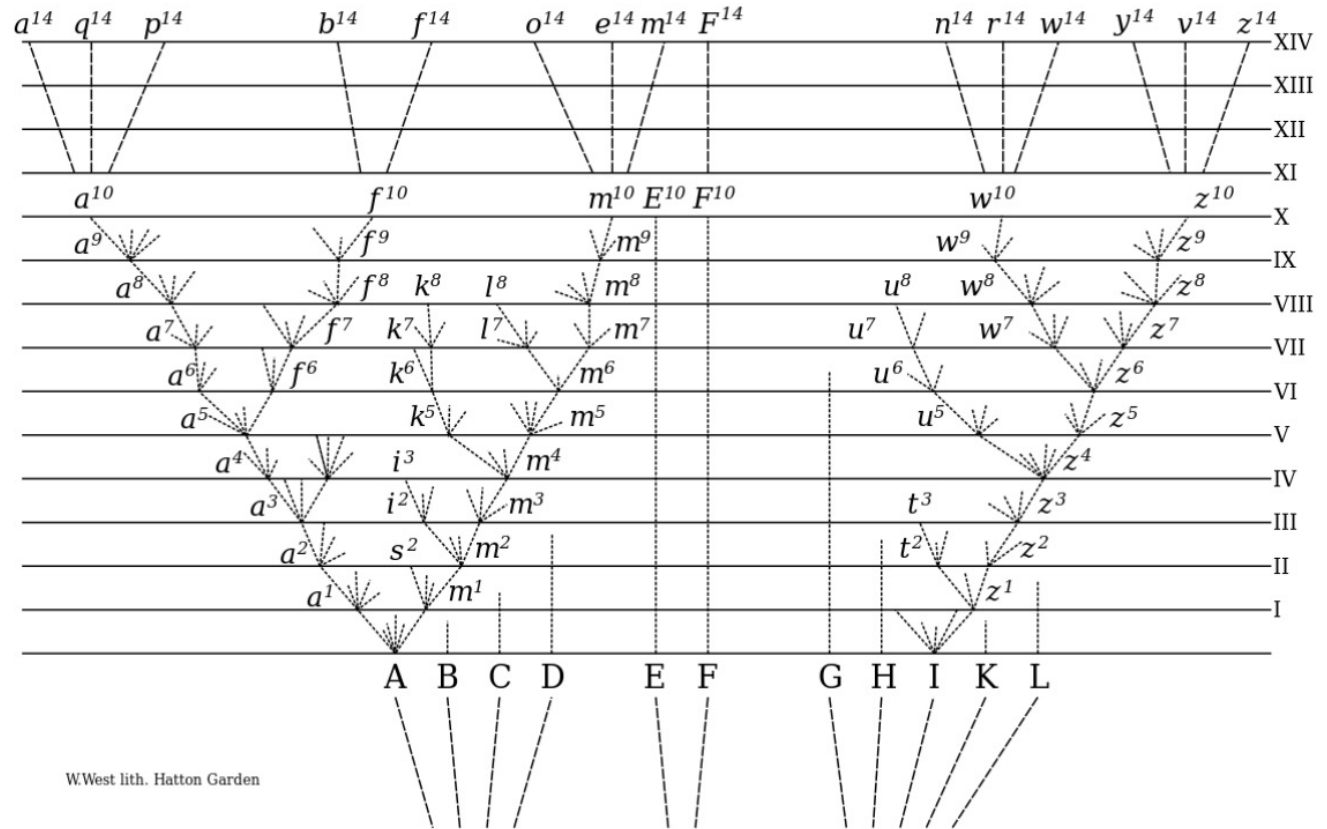
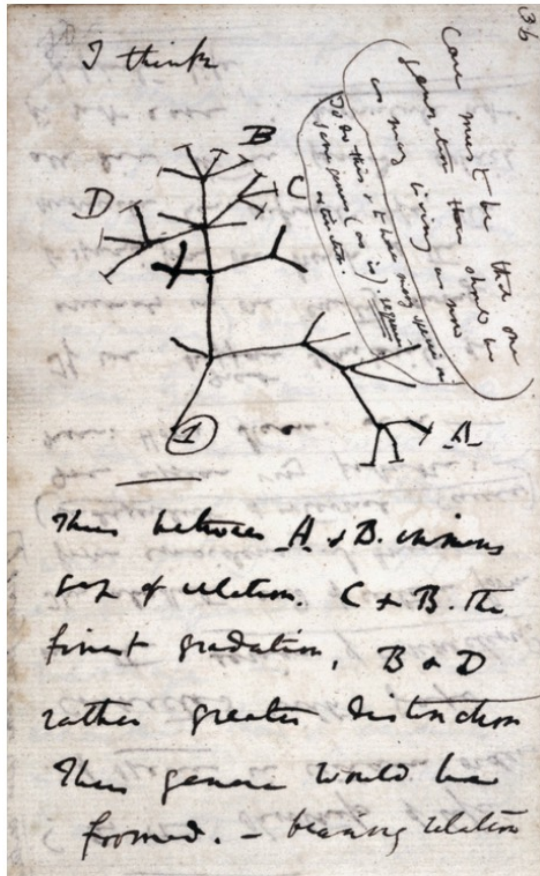


Fig. 1 | Mapping of a height-increasing allele using an ARG that was reconstructed from genotype array data. A hypothetical variant increases height in a large population, but is not directly genotyped. Nearly genotypes SNPs can be used to reconstruct the local sequence genealogy and discover that height is associated with the branch where the variant arose.

Darwin's pioneering tree-thinking



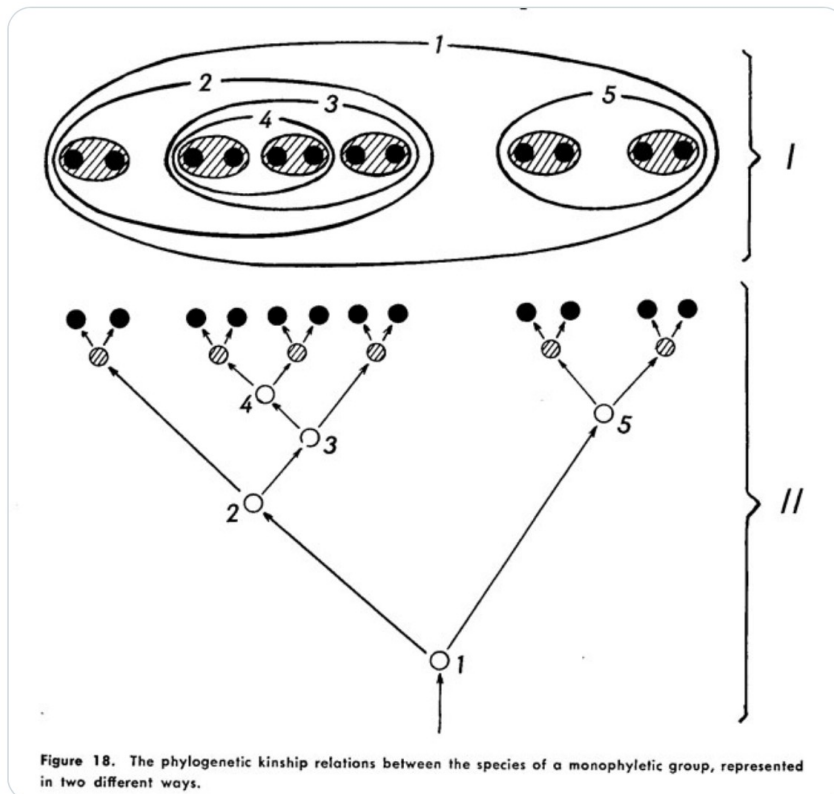
Tree-thinking central to much of biology



Leo 柿 Borges
@aquitemcaqui

...

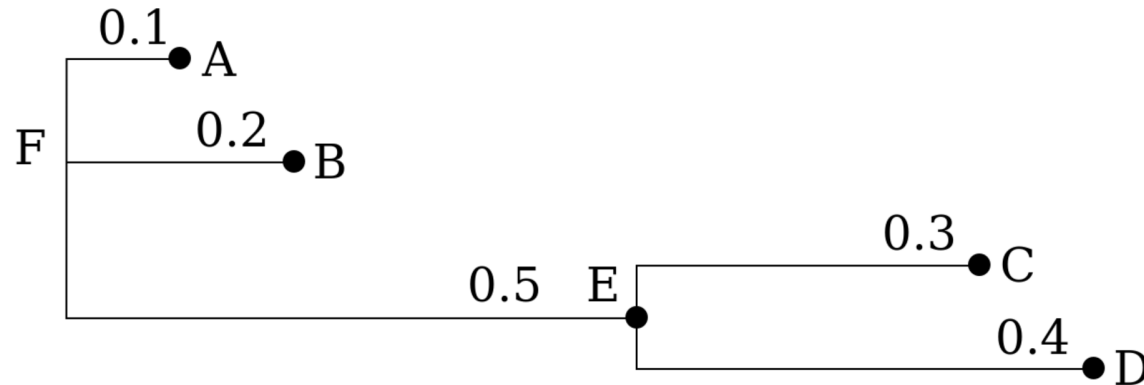
I always use this figure from Hennig's book when I am teaching undergrads how to read phylogenies



“Nothing in Biology Makes Sense
except in the Light of Evolution”
(Theodosius Dobzhansky, 1973)

Newick format for trees

- “Newick tree format is a way of representing graph-theoretical trees with edge lengths using parentheses and commas. It was adopted by ... after a meeting at **Newick's restaurant** in Dover, New Hampshire, US” (Wikipedia)

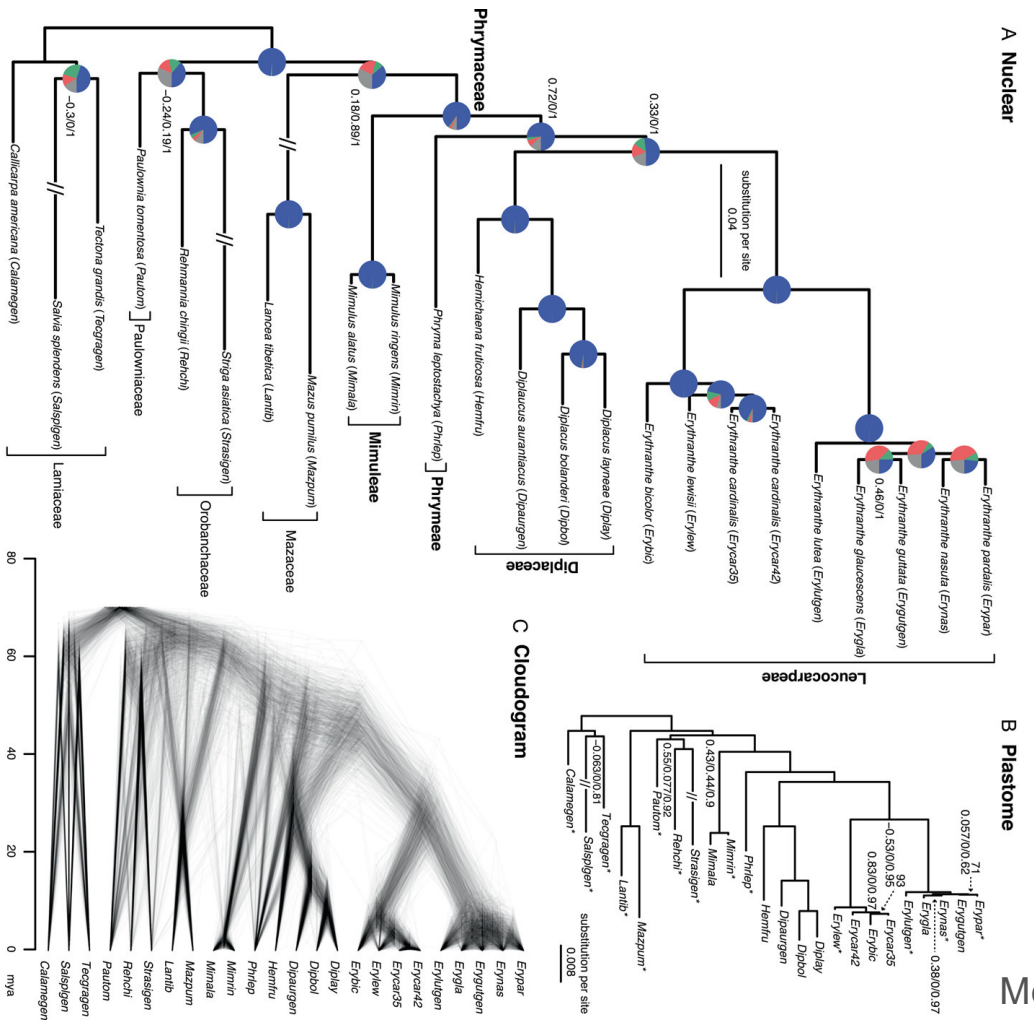


could be represented in Newick format in several ways

<code>(,,(,));</code>	<i>no nodes are named</i>
<code>(A,B,(C,D));</code>	<i>leaf nodes are named</i>
<code>(A,B,(C,D)E)F;</code>	<i>all nodes are named</i>
<code>(:0.1,:0.2,(0.3,0.4):0.5);</code>	<i>all but root node have a distance to parent</i>
<code>(:0.1,:0.2,(0.3,0.4):0.5):0.0;</code>	<i>all have a distance to parent</i>
<code>(A:0.1,B:0.2,(C:0.3,D:0.4):0.5);</code>	<i>distances and leaf names (popular)</i>
<code>(A:0.1,B:0.2,(C:0.3,D:0.4)E:0.5)F;</code>	<i>distances and all names</i>
<code>((B:0.2,(C:0.3,D:0.4)E:0.5)F:0.1)A;</code>	<i>a tree rooted on a leaf node (rare)</i>

Species/population trees can be uncertain

- Why:
- Inference uncertainty
 - Whole genome vs genomic regions
 - Other reasons



Morales-Briones et al. (2022)

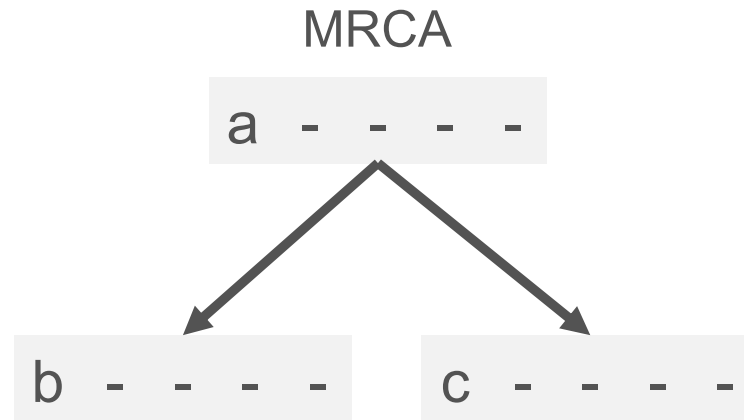
Trees for X

- Species → Species tree
((Human, (Chimp, Bonobos)), Gorilla);
- Populations → Population tree
(((Angus, Aryshire), Holstein), Gir);
- Gene/Locus → Gene tree
(Allele0,...);

Fundamentals of DNA

DNA

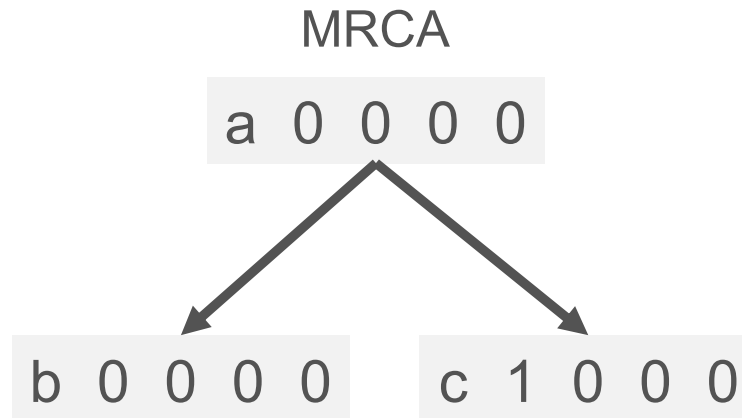
- copy
(forward-in-time)
 $a \rightarrow (b, c)$
- coalescence
(backward-in-time)
 $(b, c) \rightarrow a$



Fundamentals of DNA

DNA

- copy
(forward-in-time)
 $a \rightarrow (b, c)$
- coalescence
(backward-in-time)
 $(b, c) \rightarrow a$



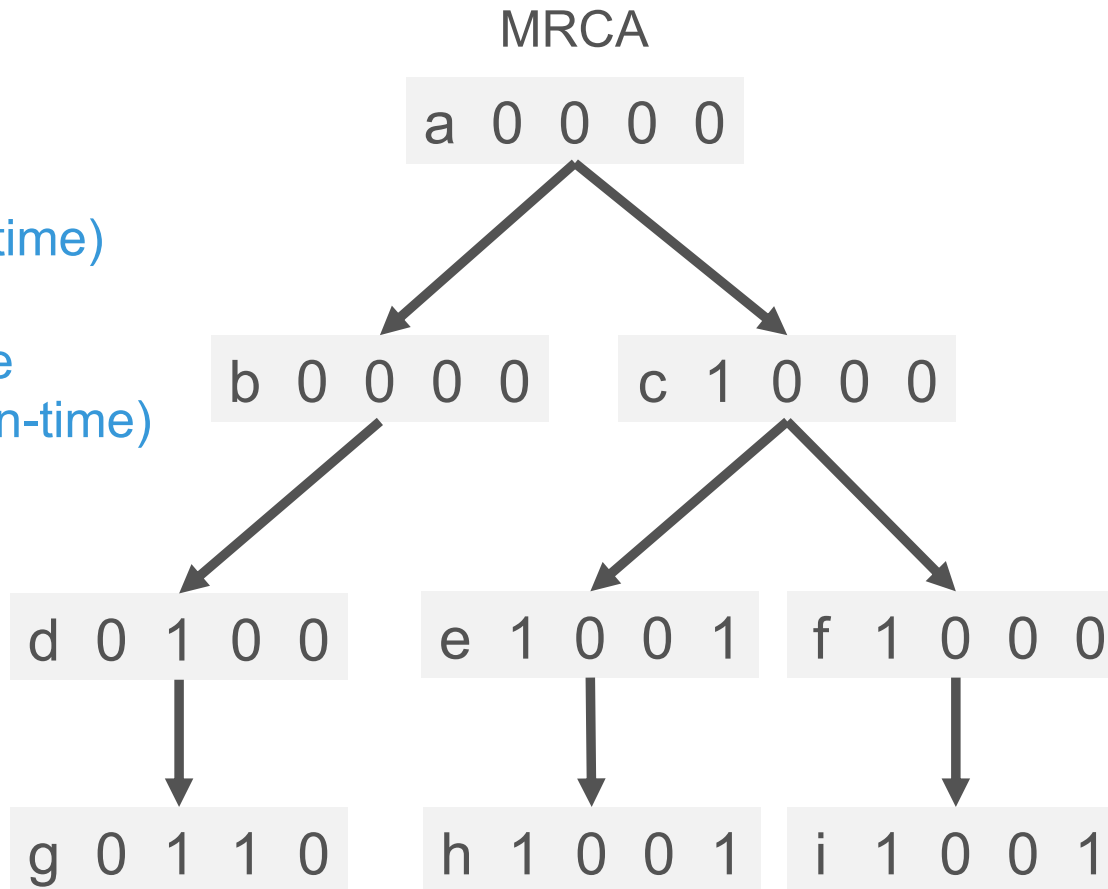
Ancestral state

- Ancestral allele 0
- Mutation 1

Fundamentals of DNA

DNA

- copy
(forward-in-time)
 $a \rightarrow (b, c)$
- coalescence
(backward-in-time)
 $(b, c) \rightarrow a$



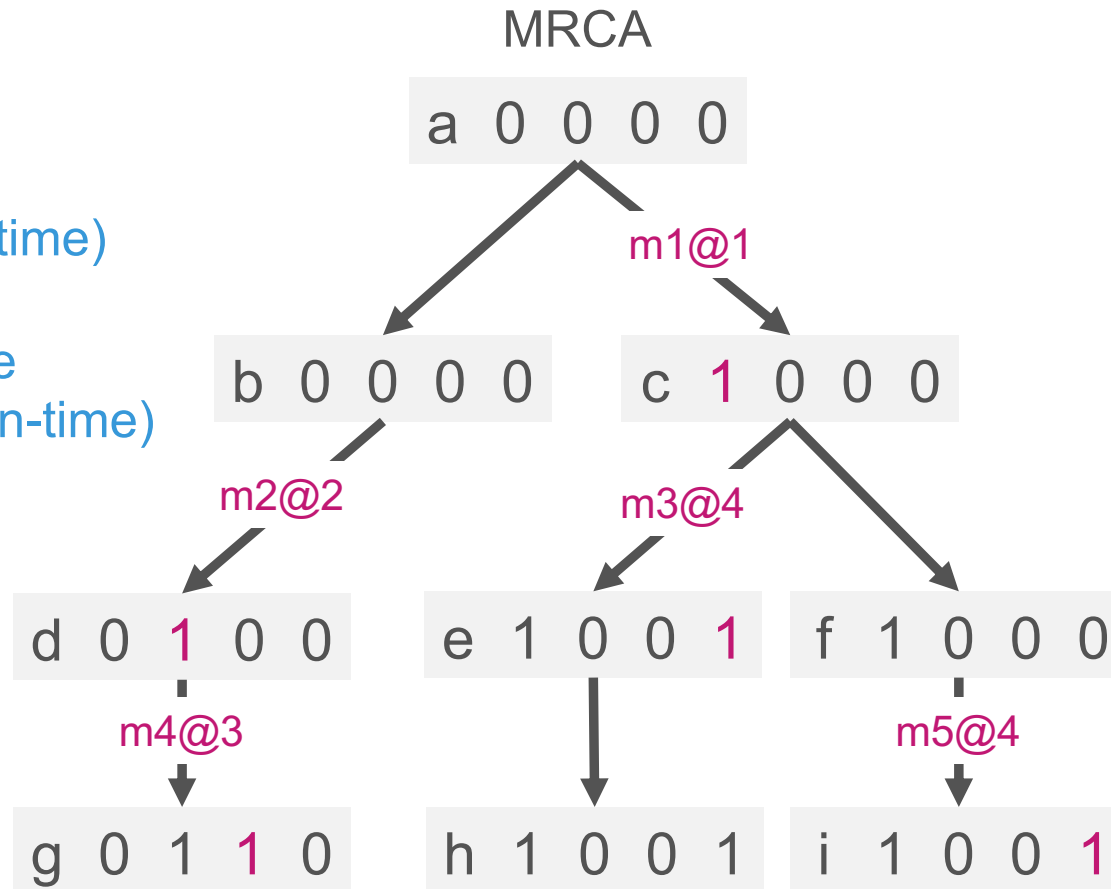
Ancestral state

- Ancestral allele 0
- Mutation 1

Fundamentals of DNA

DNA

- copy
(forward-in-time)
 $a \rightarrow (b, c)$
- coalescence
(backward-in-time)
 $(b, c) \rightarrow a$

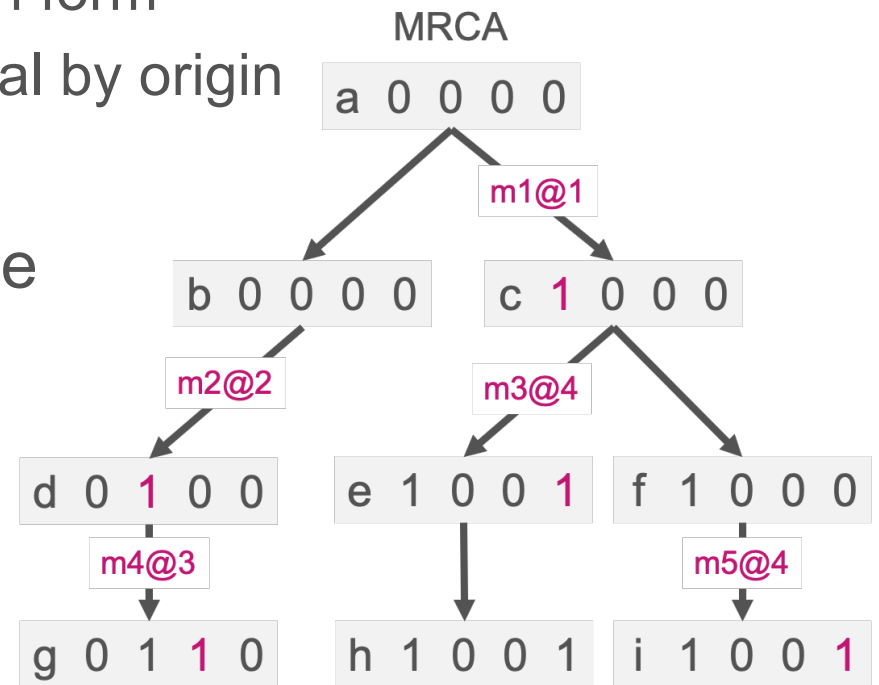


Ancestral state

- Ancestral allele 0
- Mutation 1
- De-novo mutation
(mutation event)

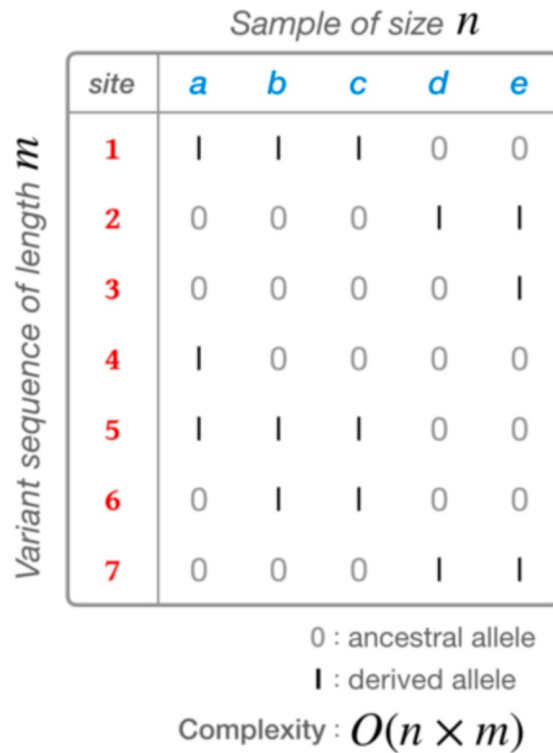
Identity by State and Descent

- Wright initiated work with pedigree relationships
- Malecot formalised identity of genomes
 - Identity by State (IBS) – identical in form
 - Identity by Descent (IBD) – identical by origin
- Define IBS & IBD for the example

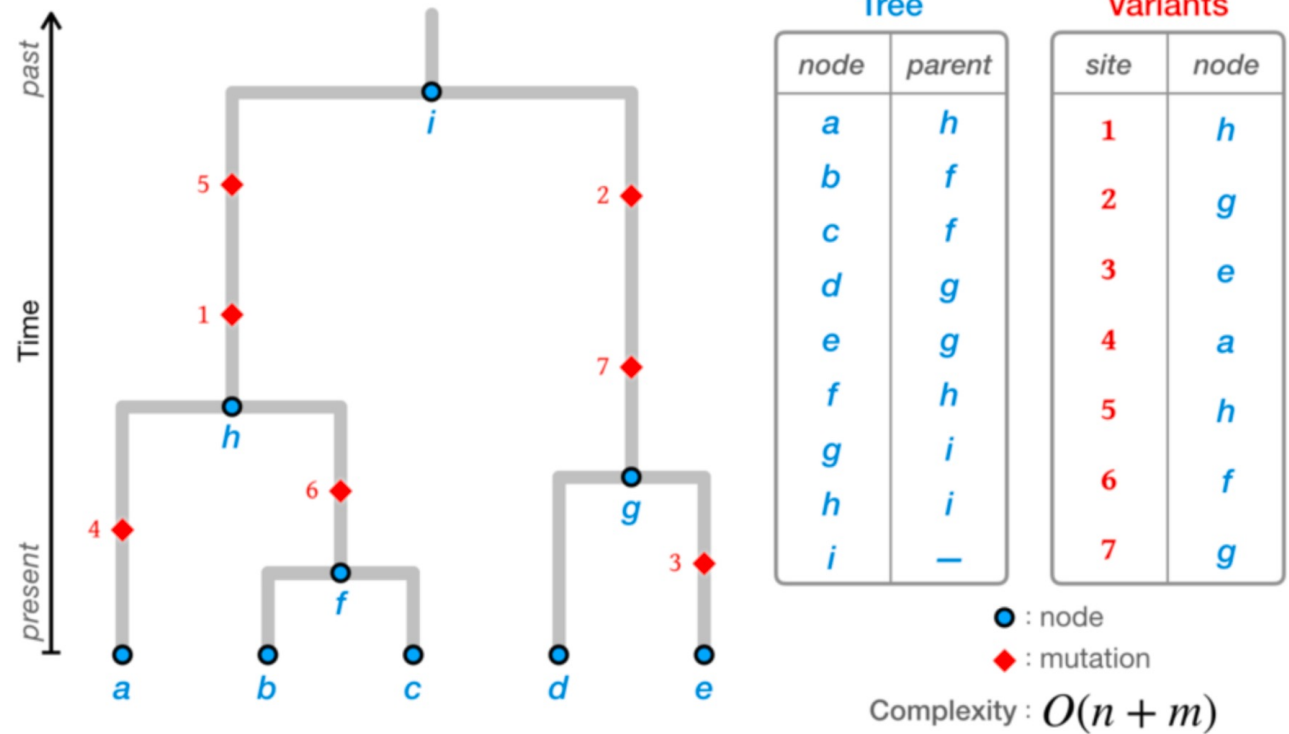


Efficient tree and mutation encoding

A Conventional data storage



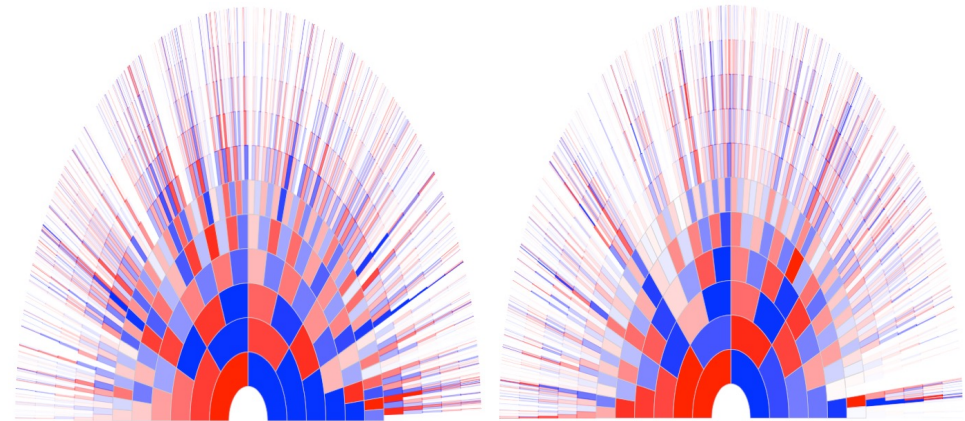
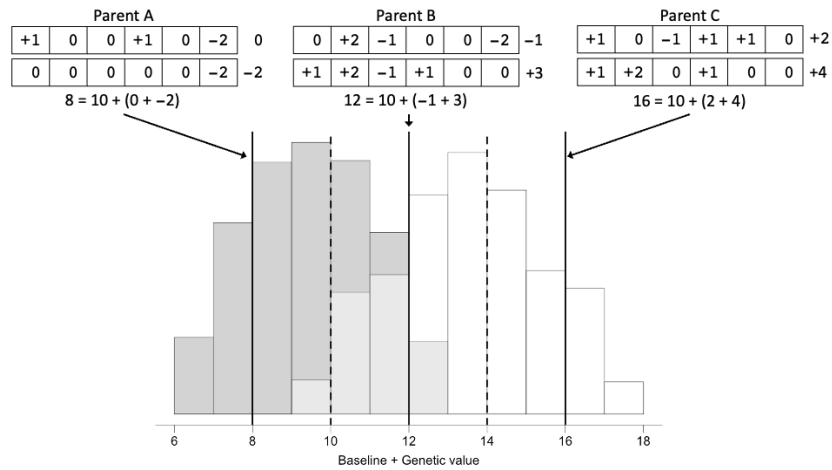
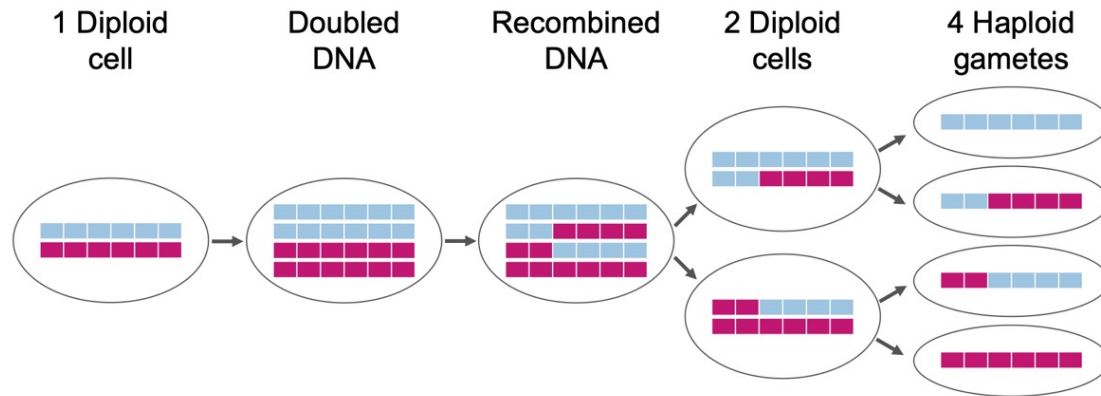
B Tree encoding



Kelleher et al. (2019)

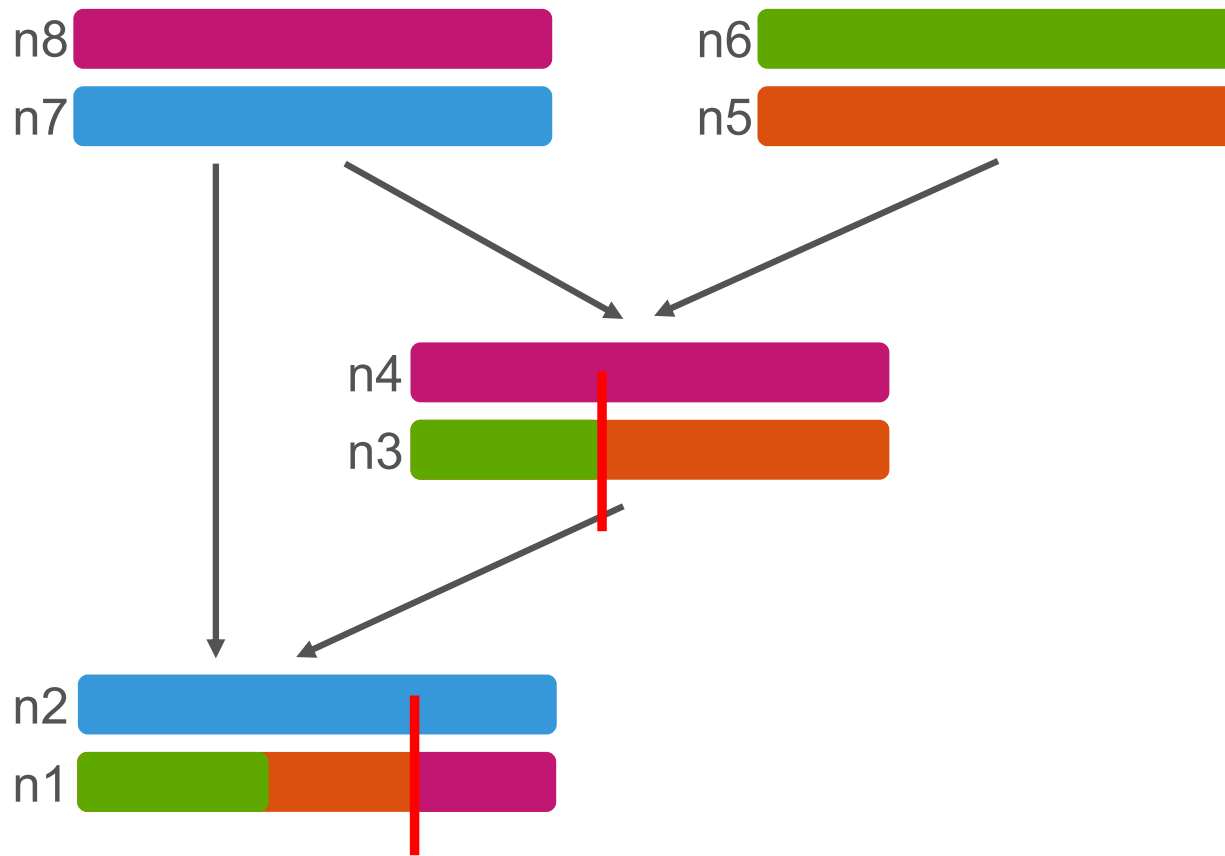
Questions?!

What about recombination & segregation?

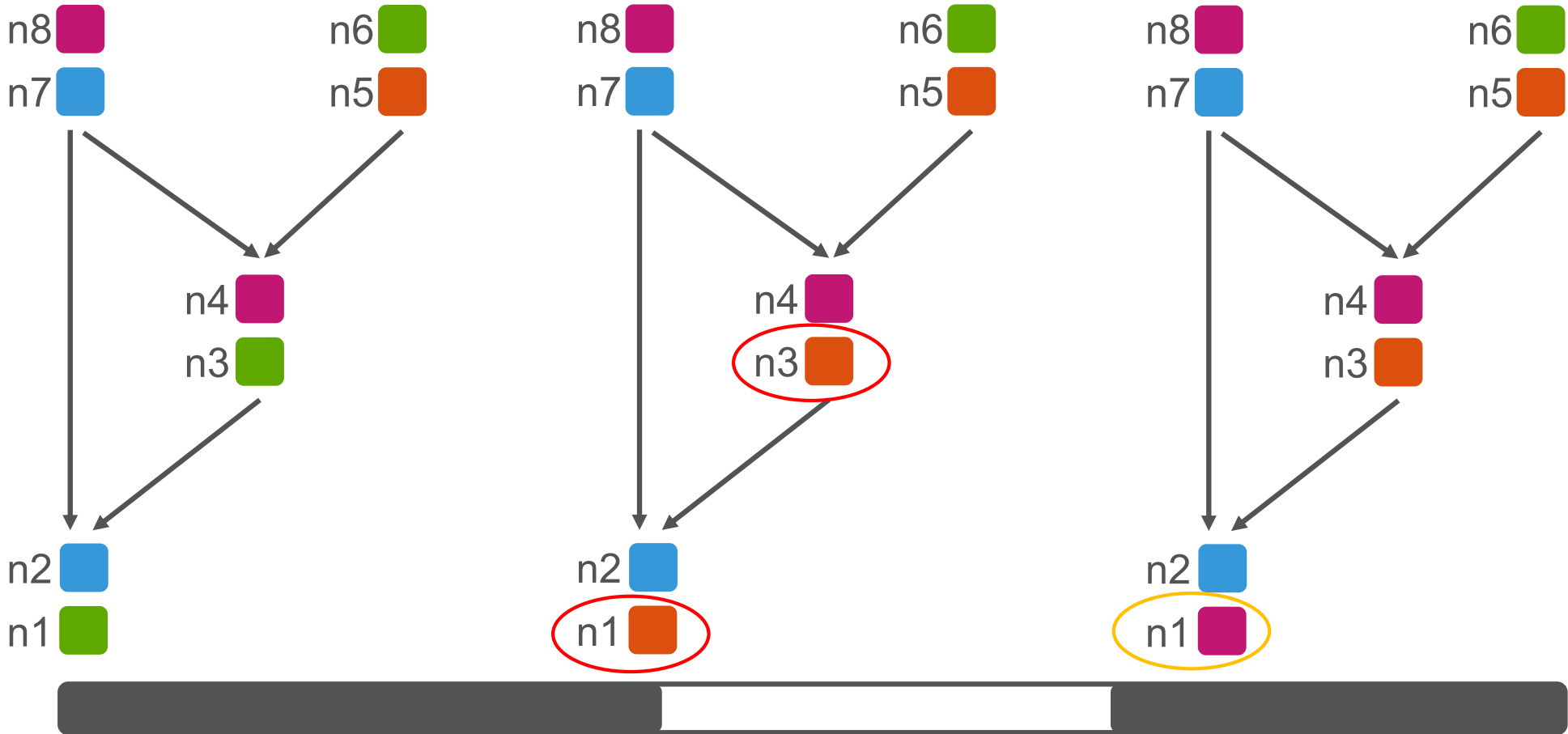


Coop (2013)

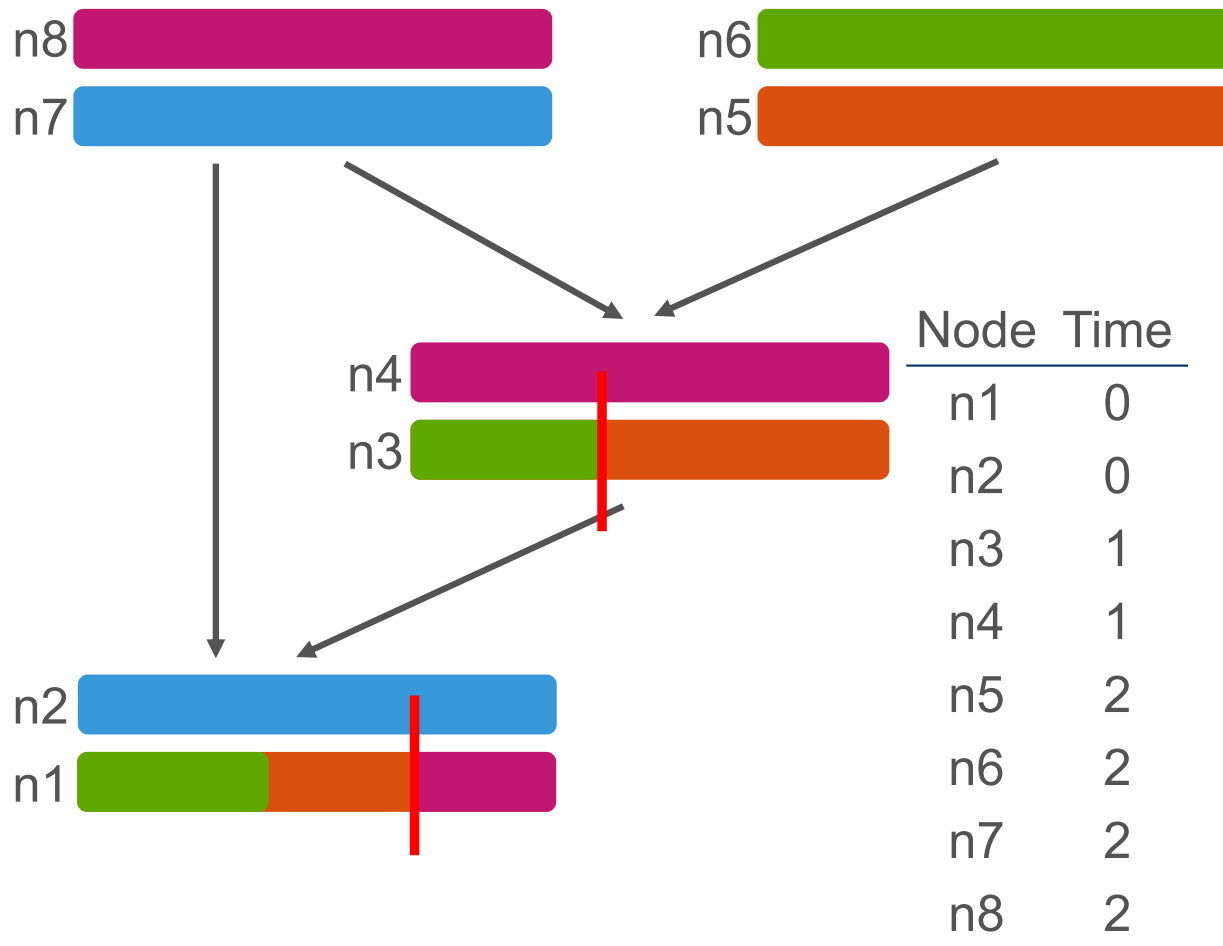
Tracking recombining DNA segments (pedigree)



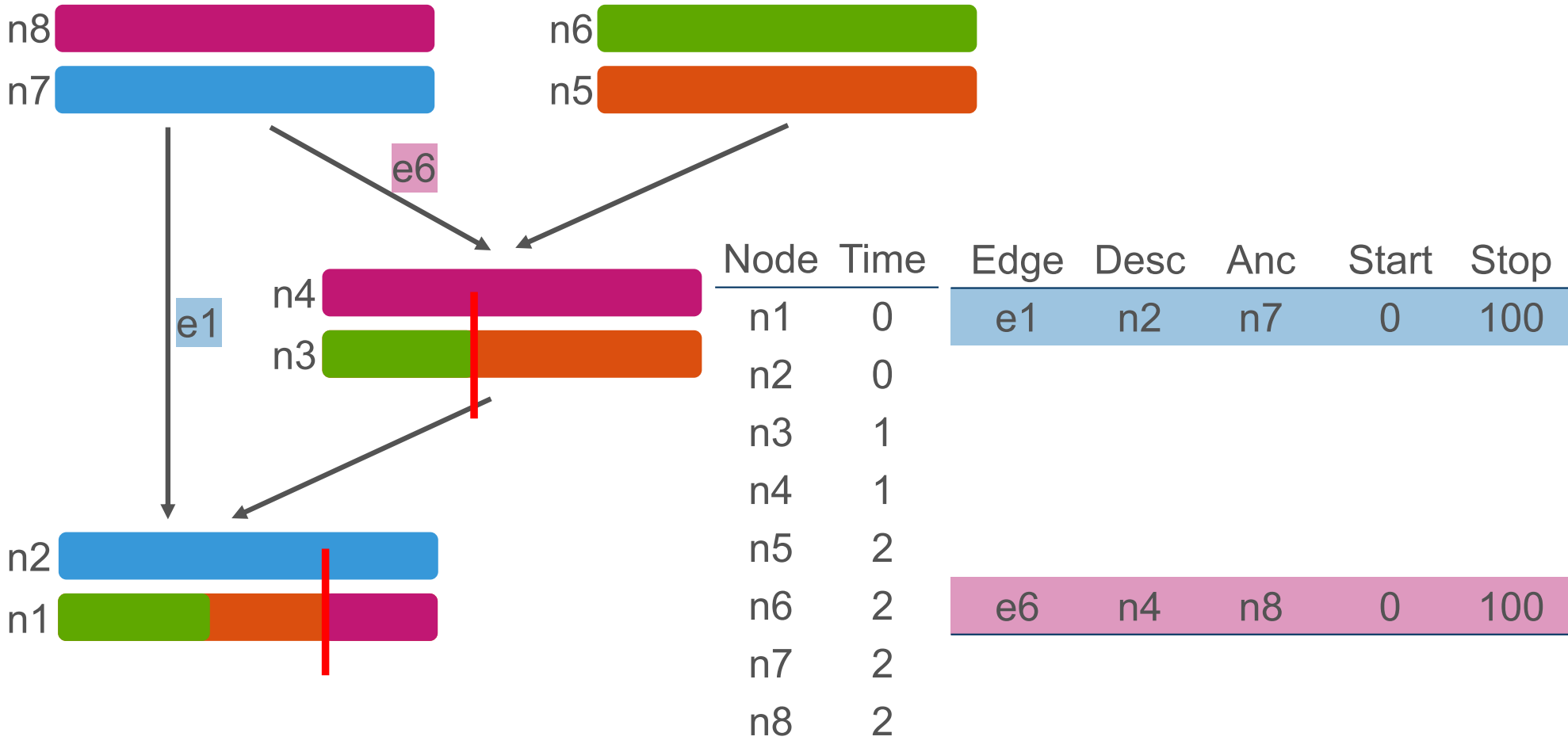
Local trees (pedigree)



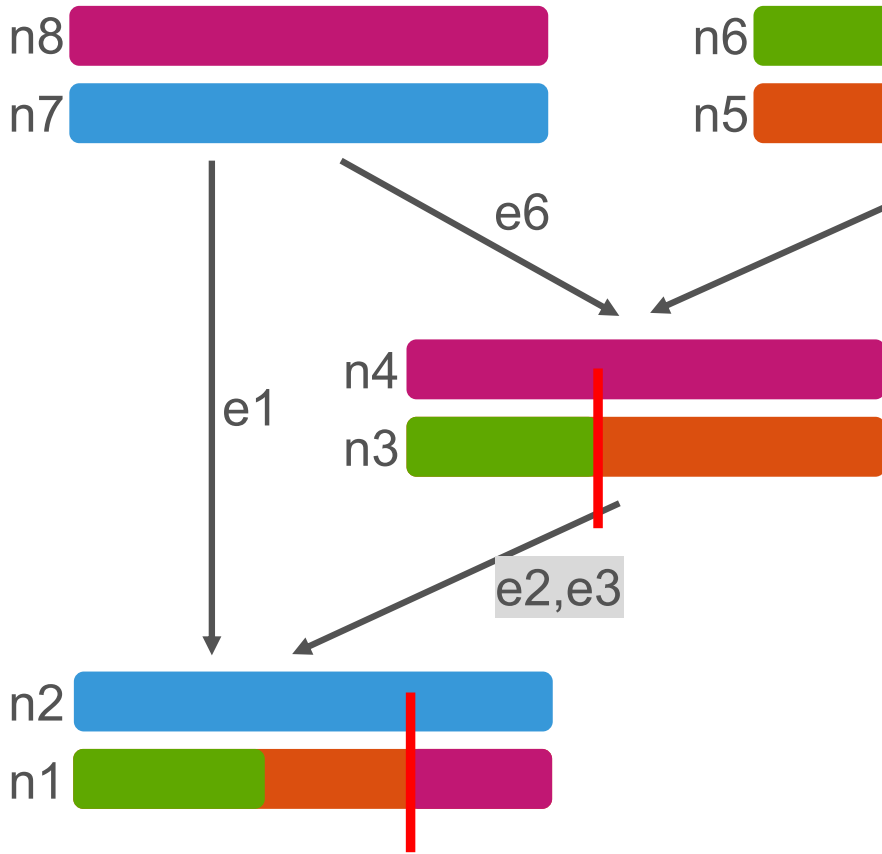
Tree sequence – Nodes (pedigree)



Tree sequence – Nodes & Edges (pedigree)

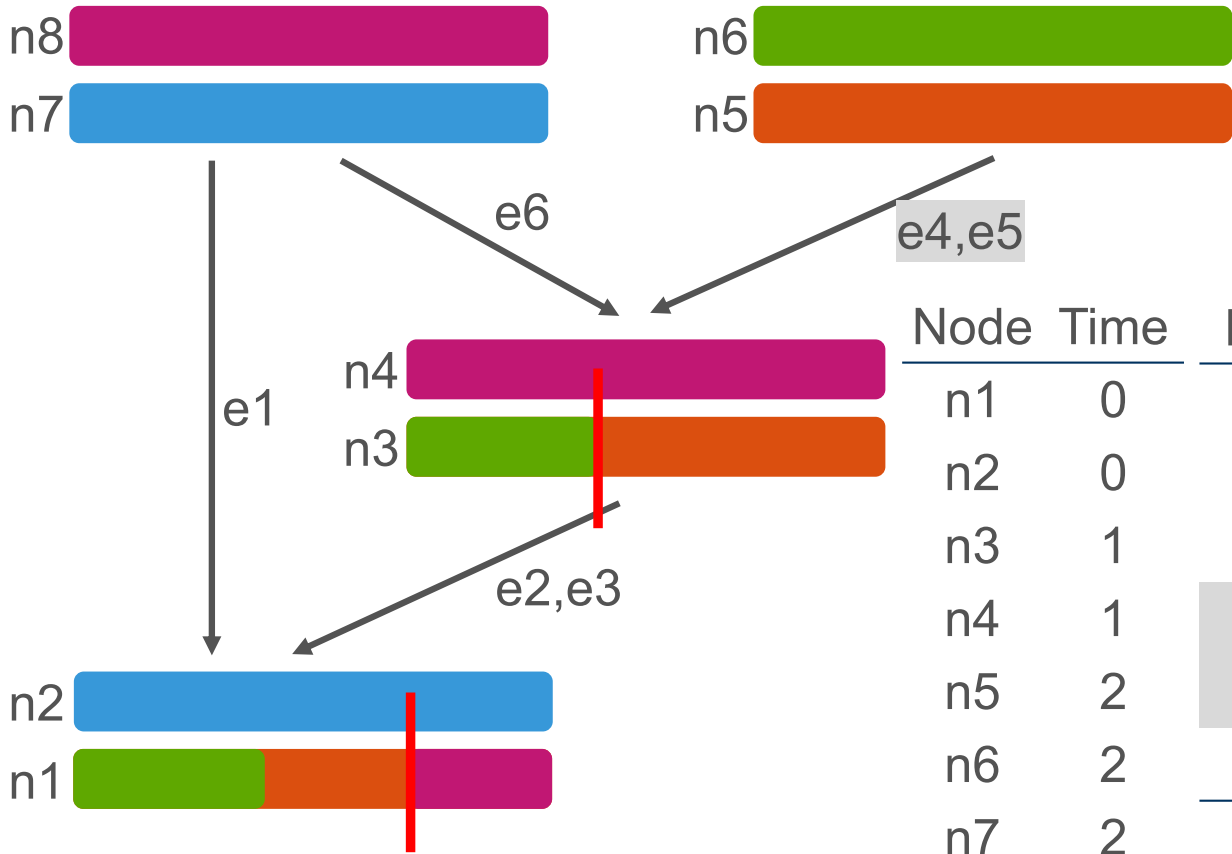


Tree sequence – Nodes & Edges (pedigree)



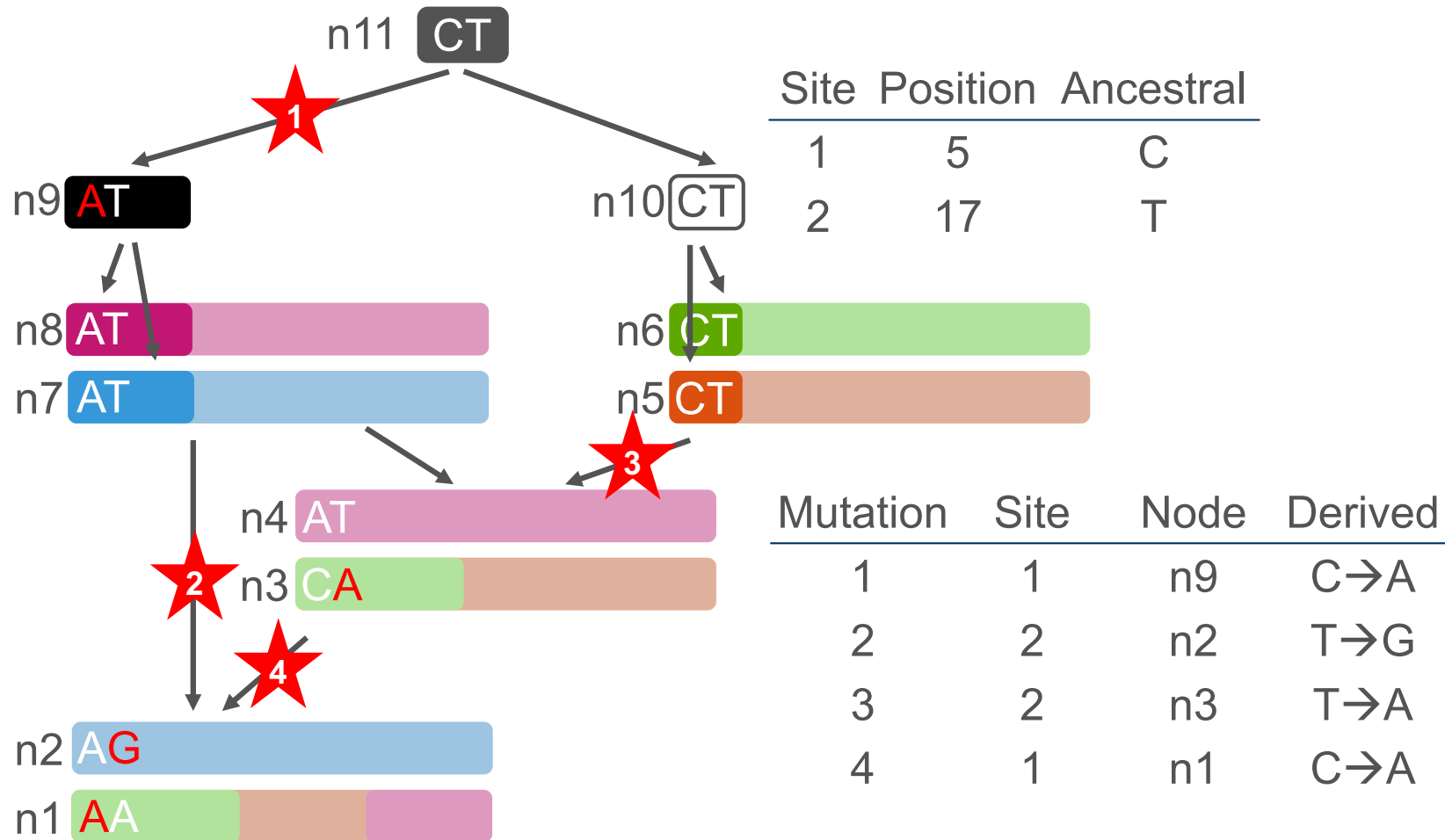
Node	Time	Edge	Desc	Anc	Start	Stop
n1	0	e1	n2	n7	0	100
n2	0	e2	n1	n3	0	70
n3	1	e3	n1	n4	71	100
n4	1					
n5	2					
n6	2	e6	n4	n8	0	100
n7	2					
n8	2					

Tree sequence – Nodes & Edges (pedigree)

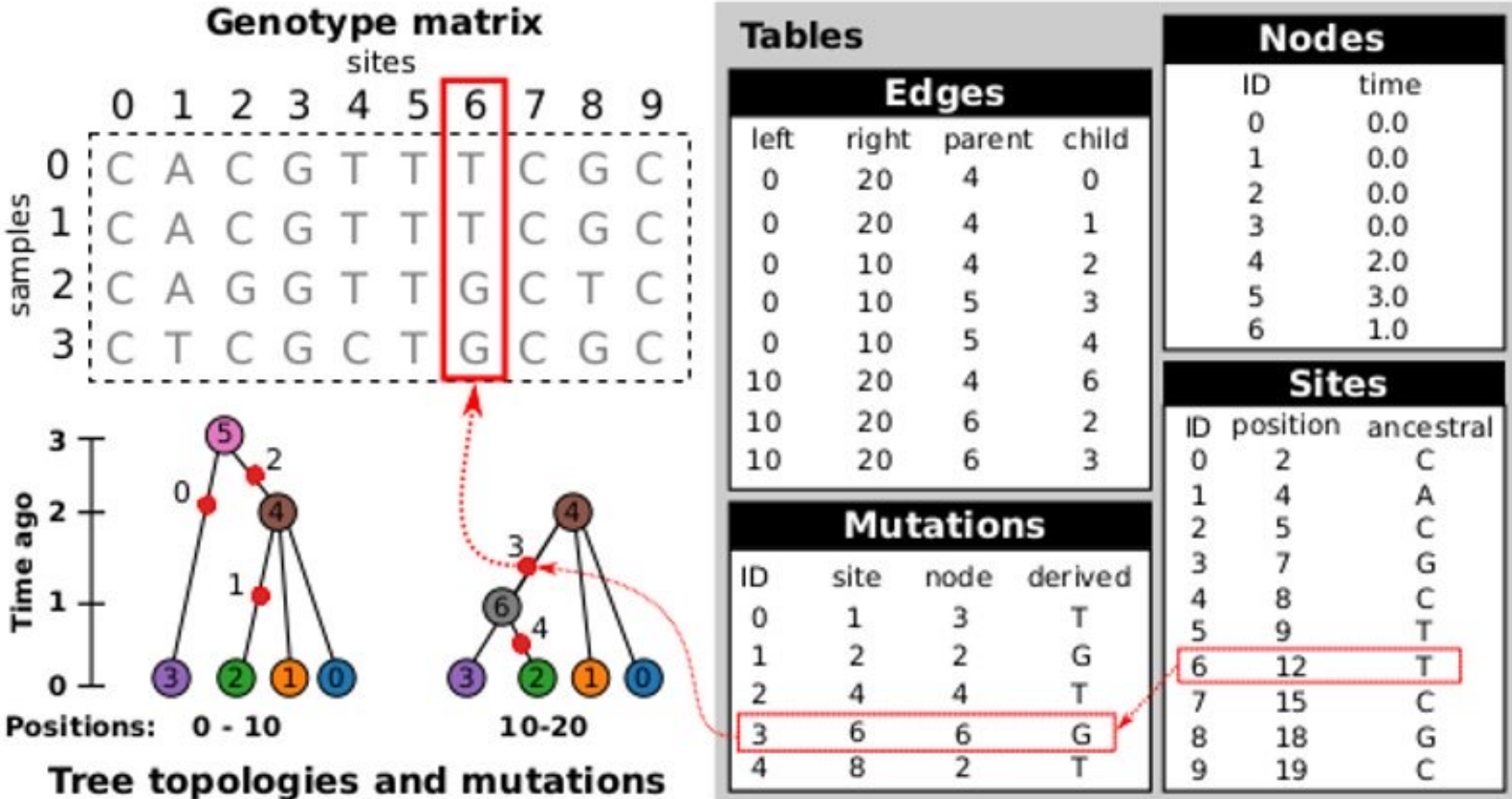


Node	Time	Edge	Desc	Anc	Start	Stop
n1	0	e1	n2	n7	0	100
n2	0	e2	n1	n3	0	70
n3	1	e3	n1	n4	71	100
n4	1	e4	n3	n6	0	40
n5	2	e5	n3	n5	41	100
n6	2	e6	n4	n8	0	100
n7	2					
n8	2					

Tree sequence – Nodes, Edges, Sites, & Mutations



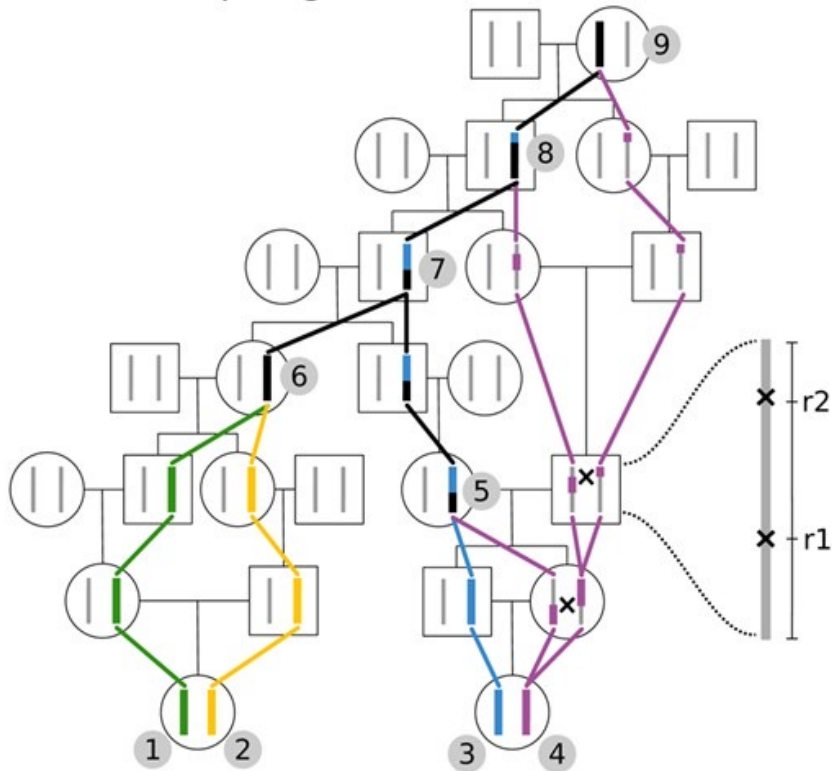
Tree sequence – another example



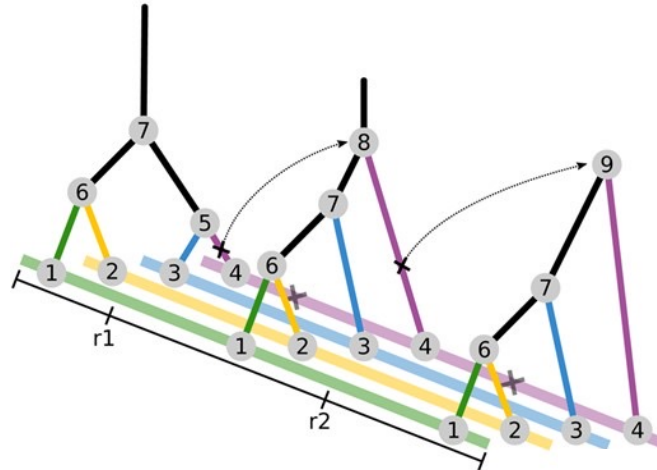
Baumdicker et al. (2013)

Another nice fig;)

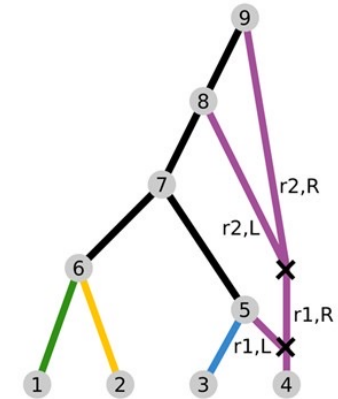
A - Ancestral Recombination Graph (ARG) in pedigree



B - Same ARG as in **A** and **C** represented as a set of local trees



C - Same ARG as in **A** and **B** represented as a graph

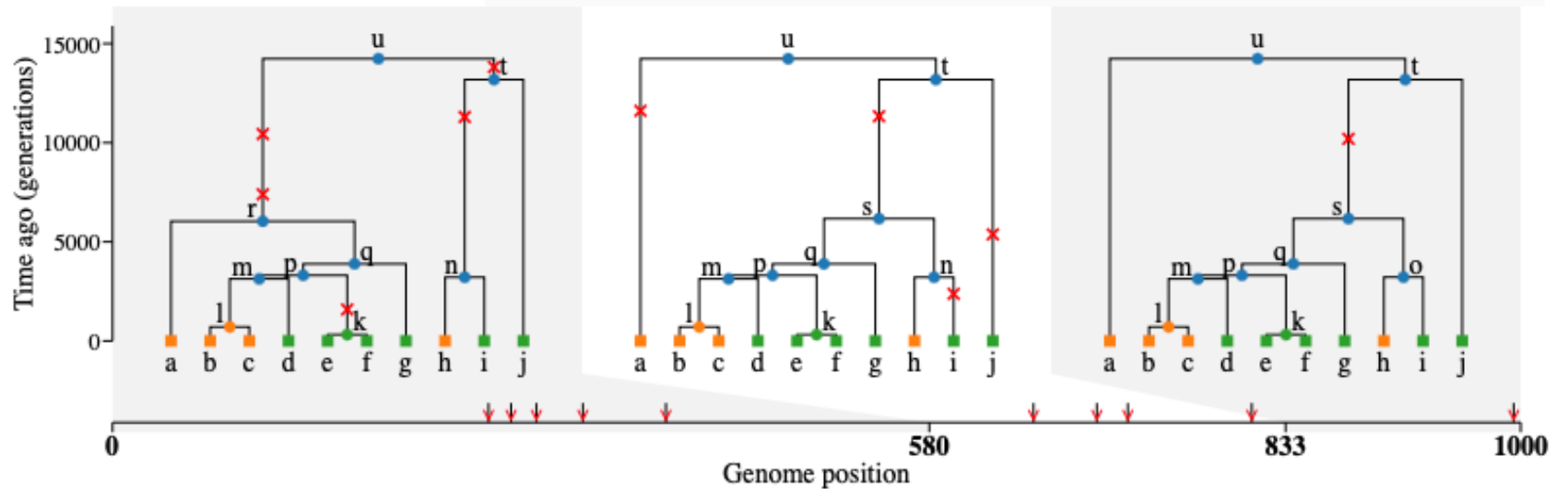


Brandt et al. (2013)

Another one



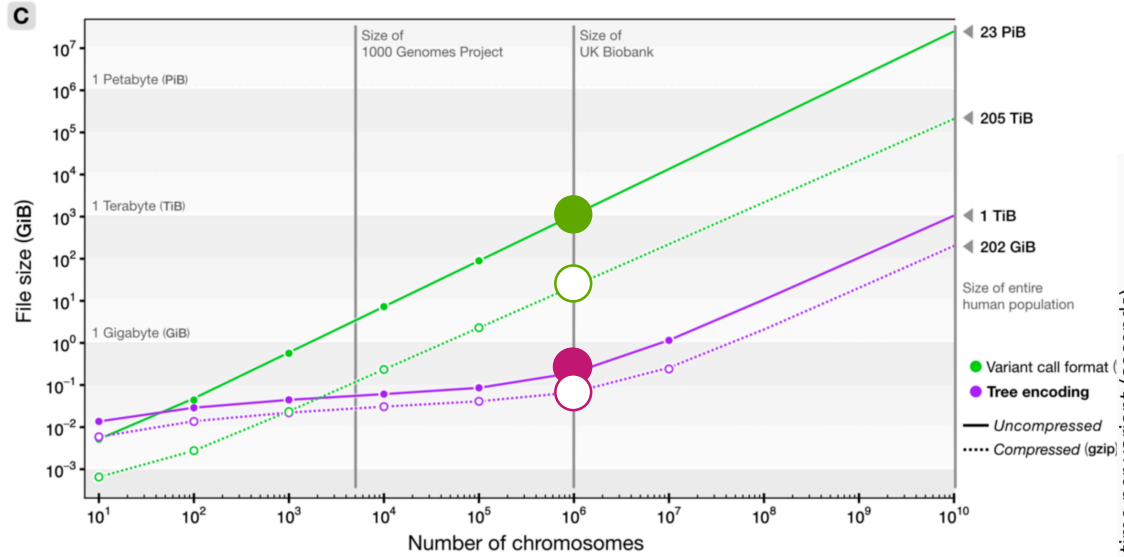
ANCESTRAL GENOMES		Position: 267	283	301	334	393	654	699	721	809	995
Genome u (time 14237.0 in the past):	G	C	G	T	C	G	G	C	A	G	
Genome t (time 13183.7 in the past):	G	G	G	T	C	G	G	C	A	G	
Genome s (time 6169.6 in the past):						G	G	C	T	A	
Genome r (time 6036.2 in the past):	A	C	G	T	T						
Genome q (time 3887.8 in the past):	A	C	G	T	T	G	G	C	T	A	
Genome p (time 3311.5 in the past):	A	C	G	T	T	G	G	C	T	A	
Genome o (time 3218.8 in the past):											A
Genome n (time 3213.6 in the past):	G	G	T	T	C	G	G	C	T		
Genome m (time 3131.7 in the past):	A	C	G	T	T	G	G	C	T	A	
Genome l (time 705.4 in the past):	A	C	G	T	T	G	G	C	T	A	
Genome k (time 318.8 in the past):	A	C	G	C	T	G	G	C	T	A	



https://tskit.dev/tutorials/what_is.html

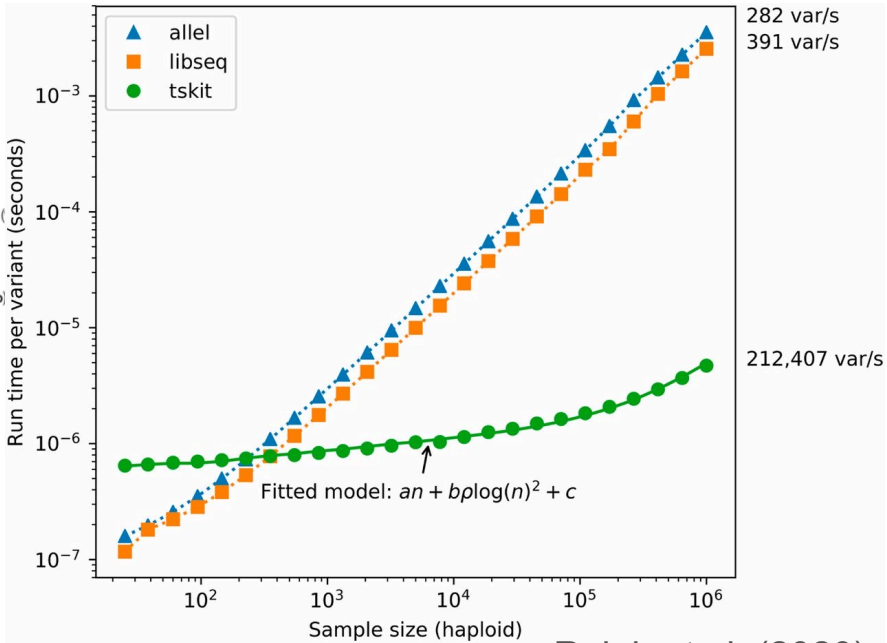
Computational power of tree sequences

STORAGE



Kelleher et al. (2019)

COMPUTE



Ralph et al. (2020)

Toolset

- msprime – backward-in-time simulation
- tsinfer – inferring (=estimating) tree sequences from data
- tskit – common toolset for tree sequences, including “stats”
- tsdate – dating the ancestors
- “ts.place” – placing the ancestors
- “ts.relatedness” – linear algebra on tree sequences (covariance, PCA, ...)
- SLiM – forward-in-time simulation
- Relate – inferring local trees (not tree sequence!)
- ARGinfer – inferring tree sequence ****distribution****
- ...

Population structure summary (Genealogical Nearest Neighbour)

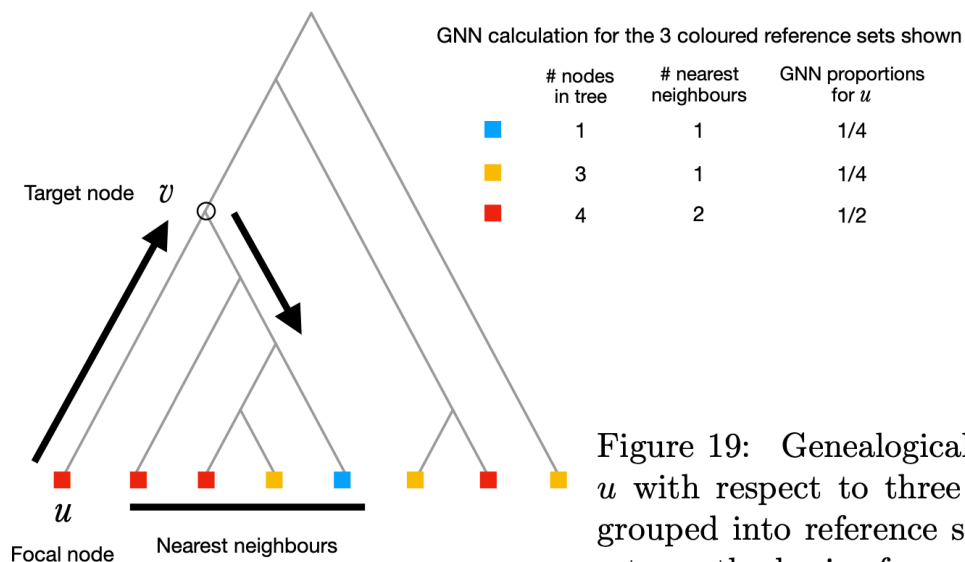
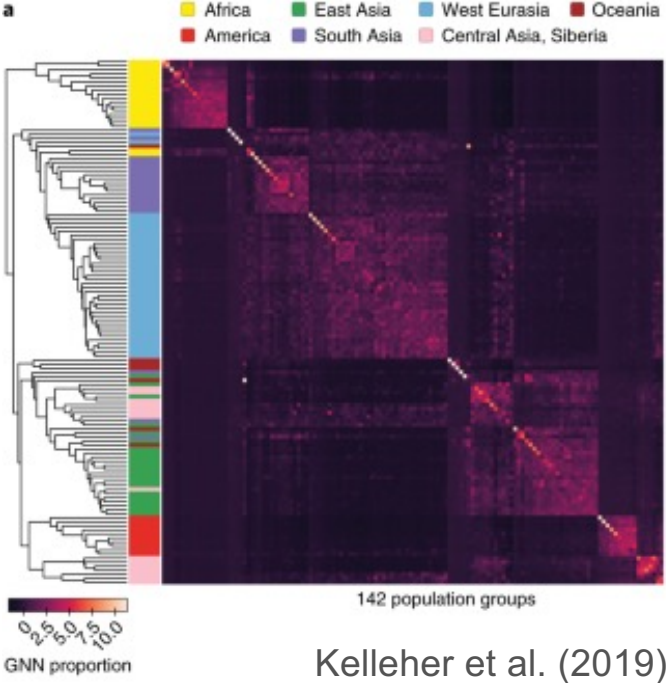


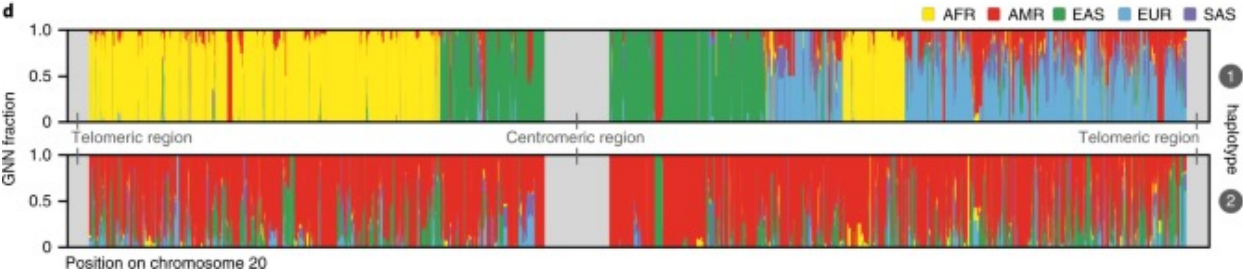
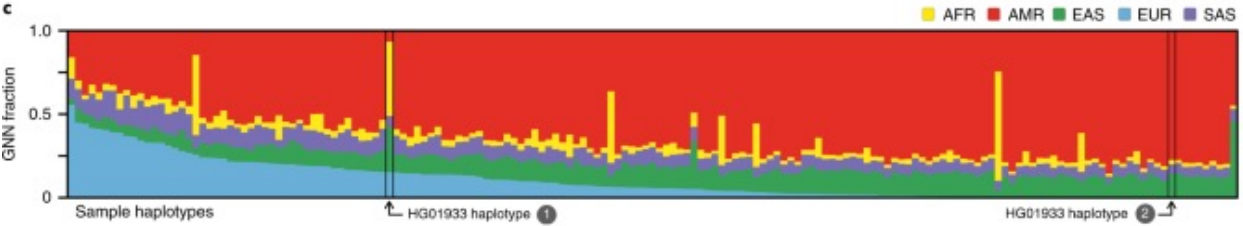
Figure 19: Genealogical Nearest Neighbours (GNN) example for a focal node u with respect to three reference sets (indicated by colours). Nodes can be grouped into reference sets arbitrarily, but we often assign nodes to reference sets on the basis of geography. For example, the colours blue, yellow, and red in this illustration could represent European, African, and Native American samples. To compute GNN values we first traverse upwards from focal node u to find the target node v . This node is the first node on the path from u to root which counts among its descendants at least one member of a reference set (not including u). The GNN values are then computed as the proportion of descendants of v (not including u) from each of the reference sets.

Kelleher et al. (2019)

Global human population structure

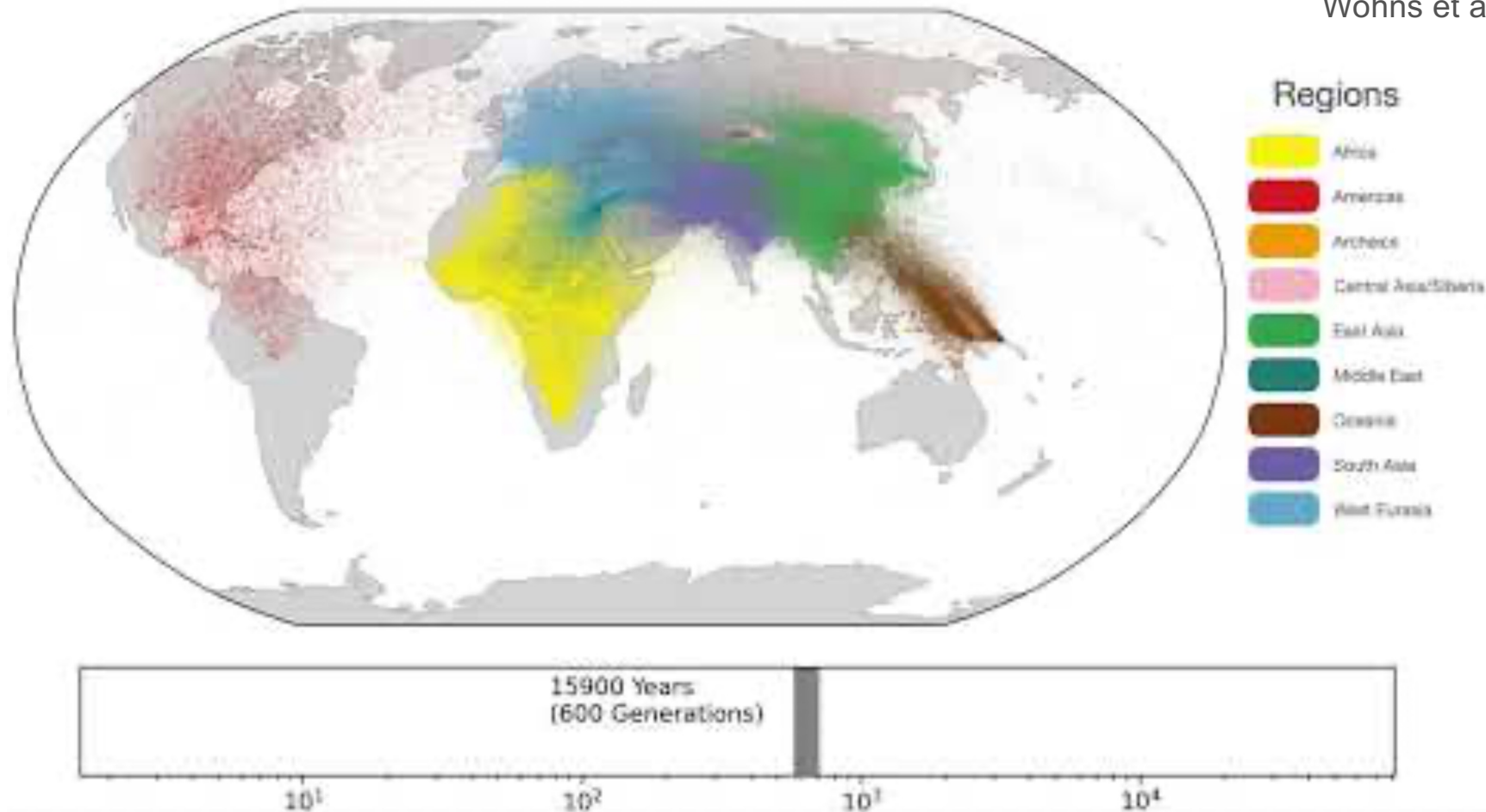


Kelleher et al. (2019)



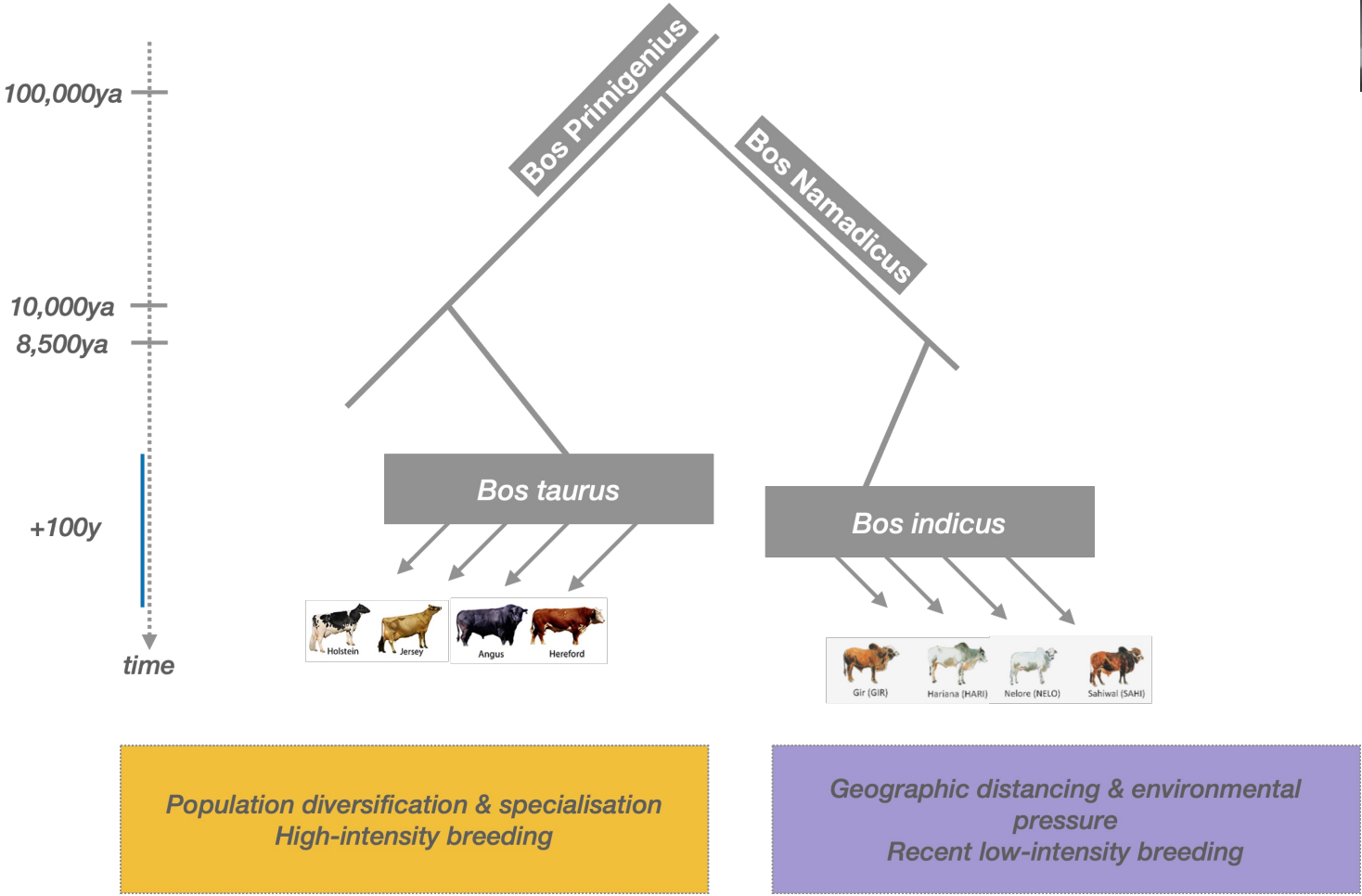
Dating and placing the unified genealogy of modern and ancient human genomes

Wohns et al. (2022)



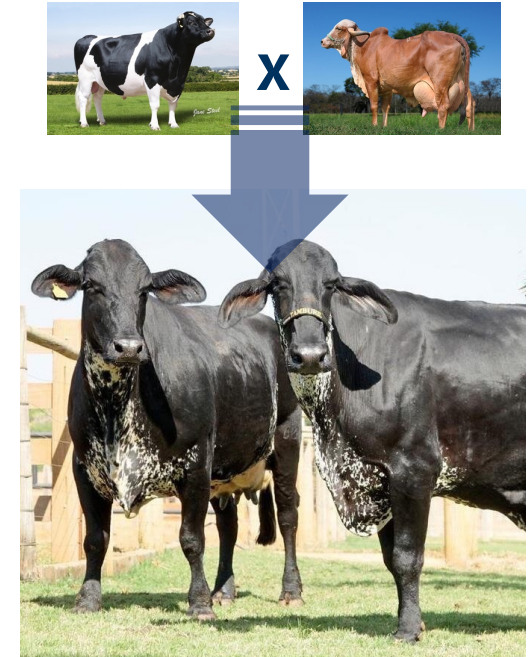
Questions?!

Genetics & breeding of Taurine-Indicine crossbred cattle

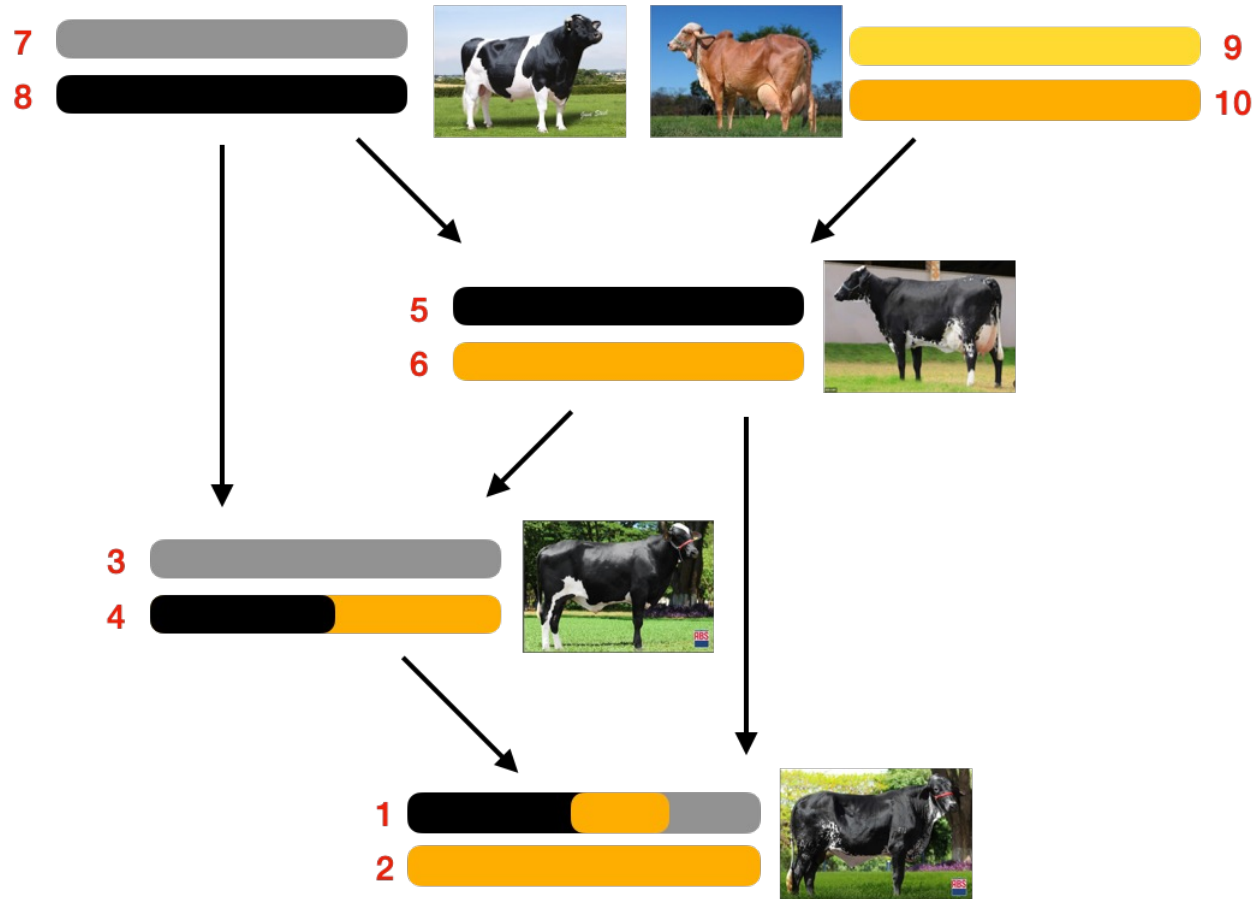


Girolando

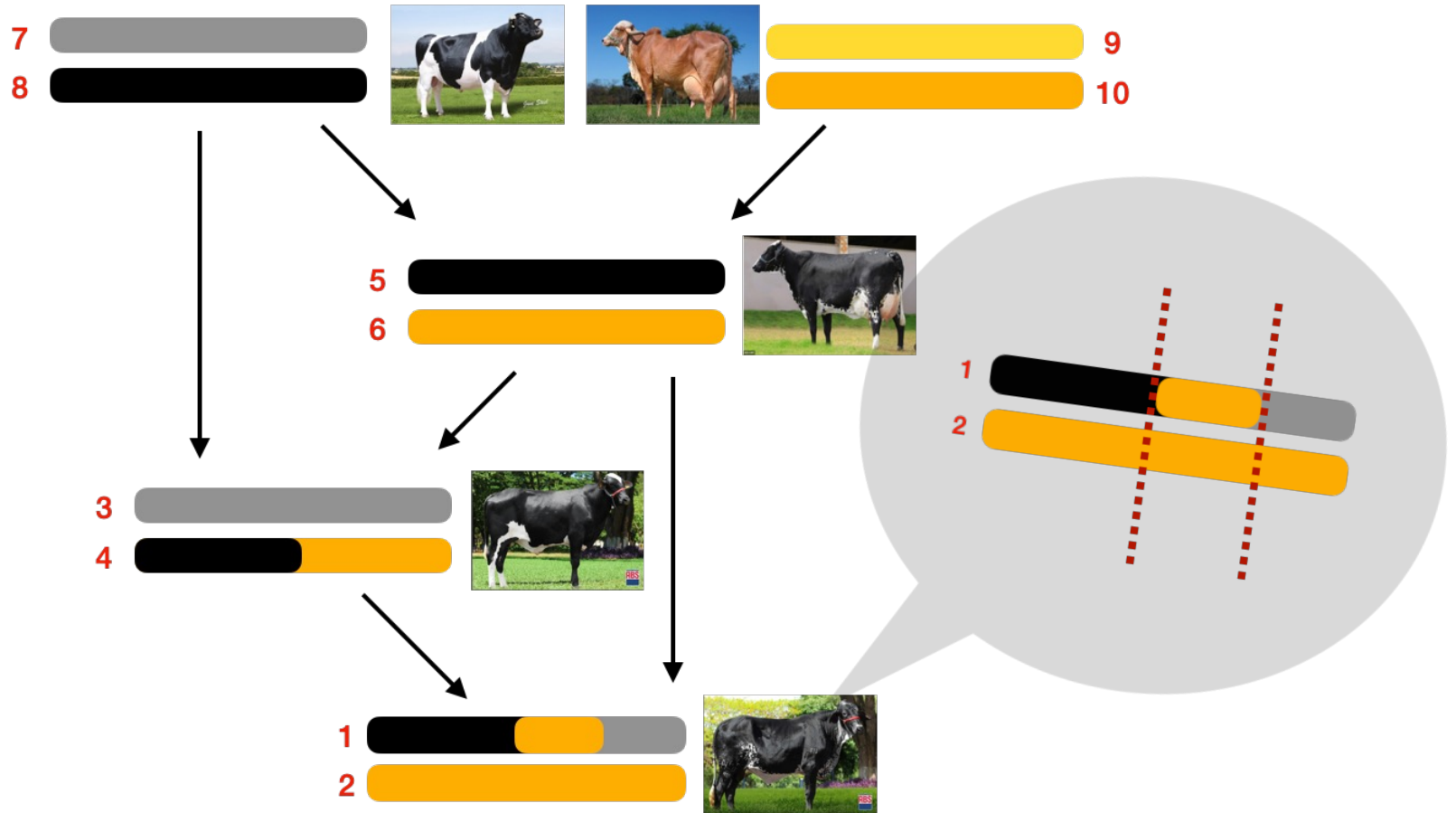
- Tropical dairy depends on crossbreds
- Brazil is 5th milk producer, yet 2nd largest dairy herd, 80% crossbreds
- How to improve their crossbred evaluation?
- Considering the environment variation, is it worth to consolidate Girolando as a breed? If so, how?



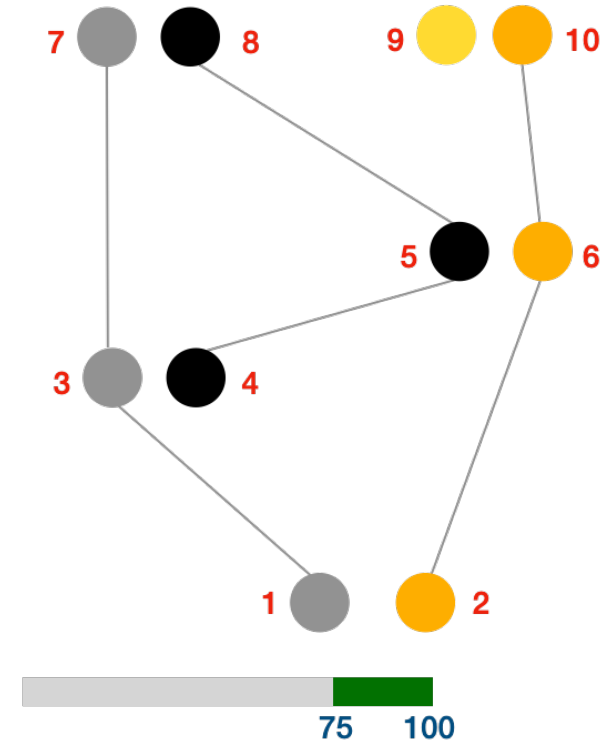
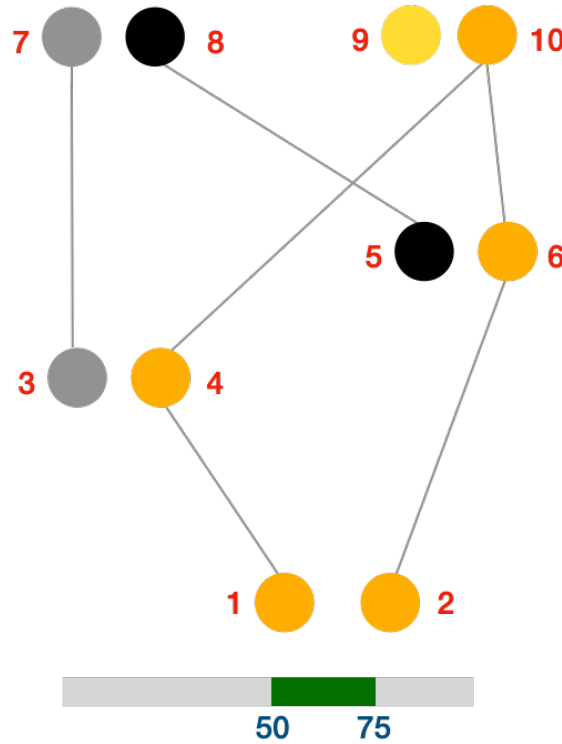
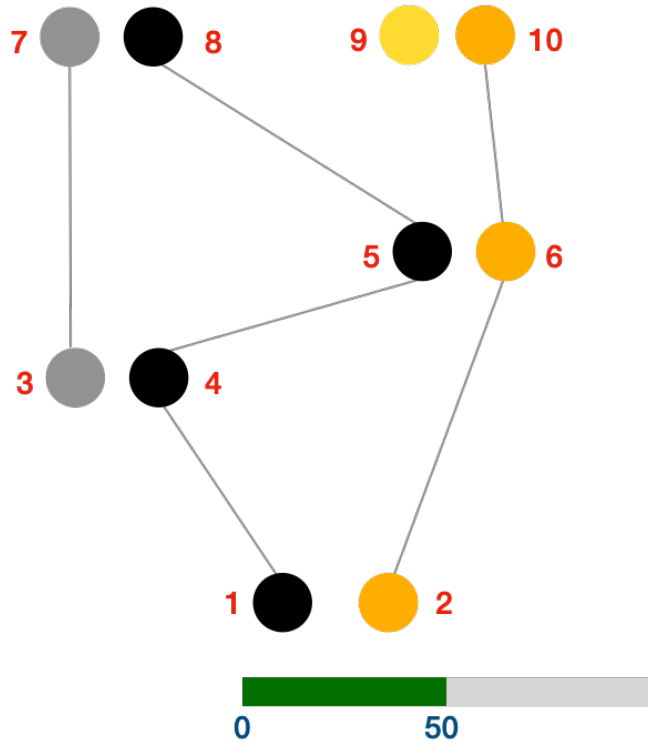
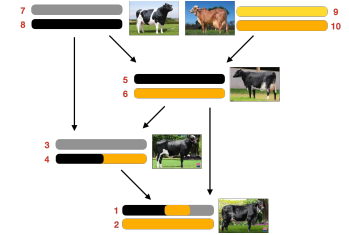
The genomic challenge



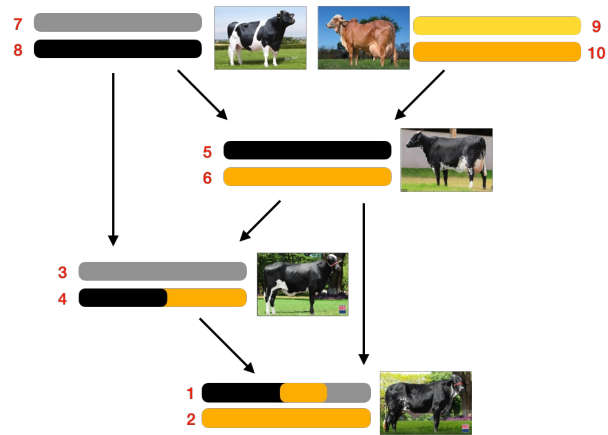
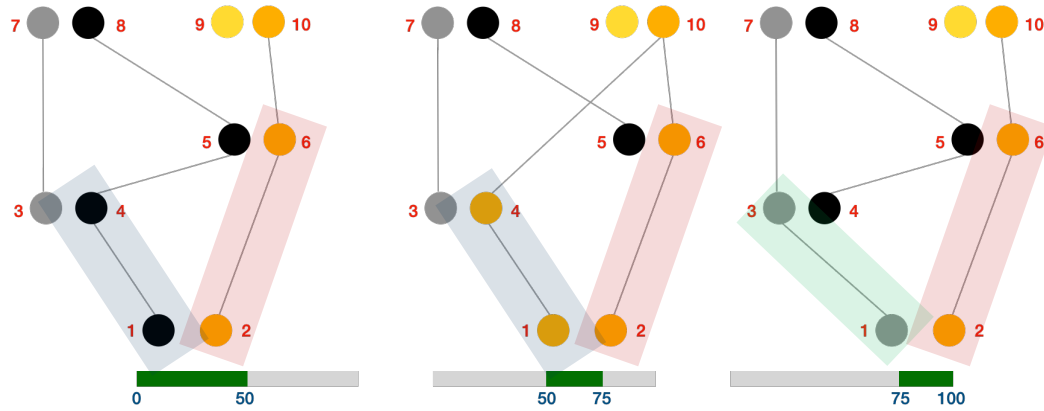
The genomic challenge



The genomic challenge



Tree sequence within a pedigree



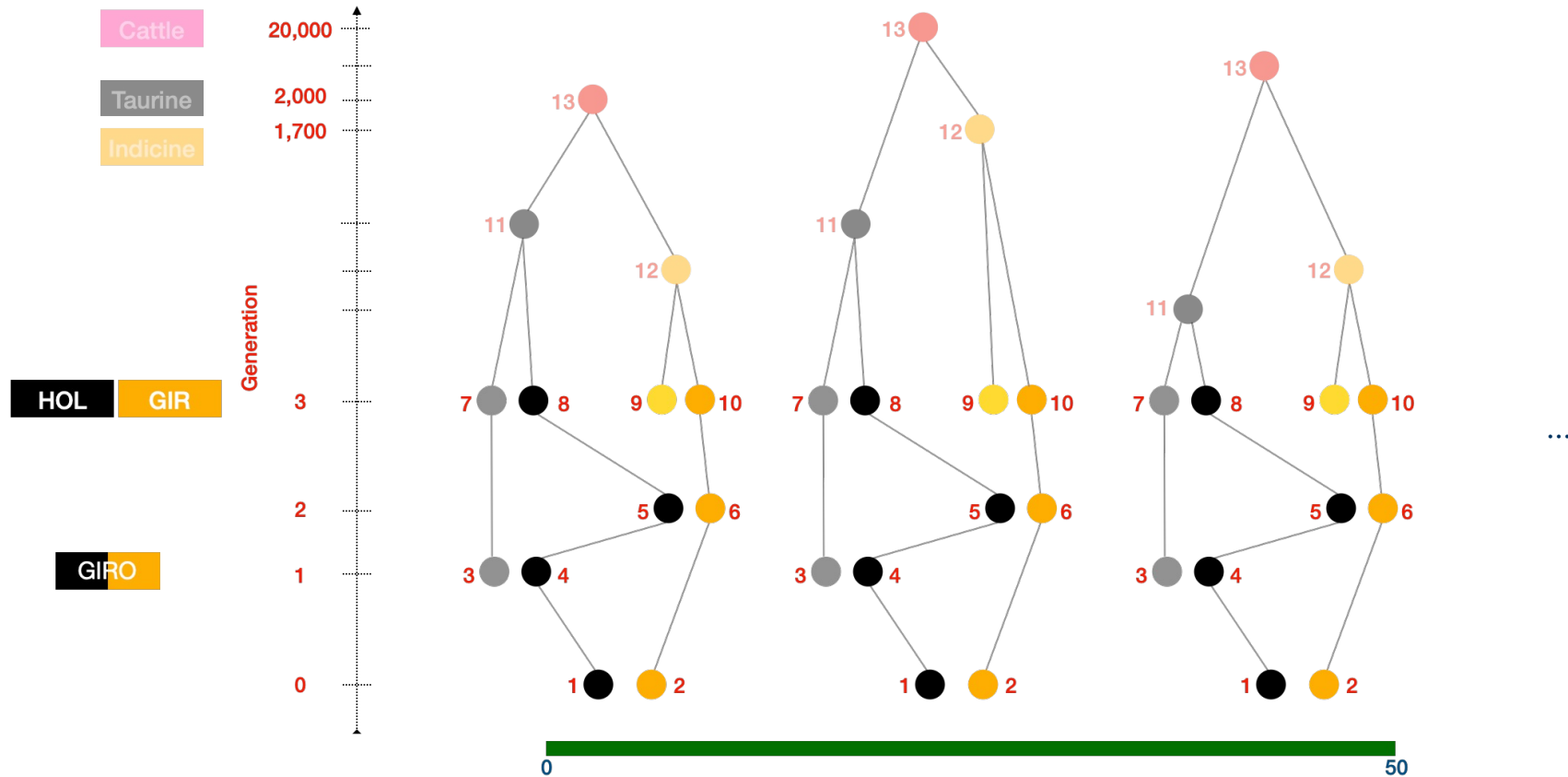
Nodes

Node	Time
10	3
9	3
8	3
7	3
6	2
5	2
4	1
3	1
2	0
1	0

Edges

Child	Parent	Left	Right
1	4	0	75
1	3	75	100
2	6	0	100
3	7	0	100
4	5	0	50
4	9	50	75
4	5	75	100
5	8	0	100
6	10	0	100

Tree sequence “recapitulation” (deep ancestry)

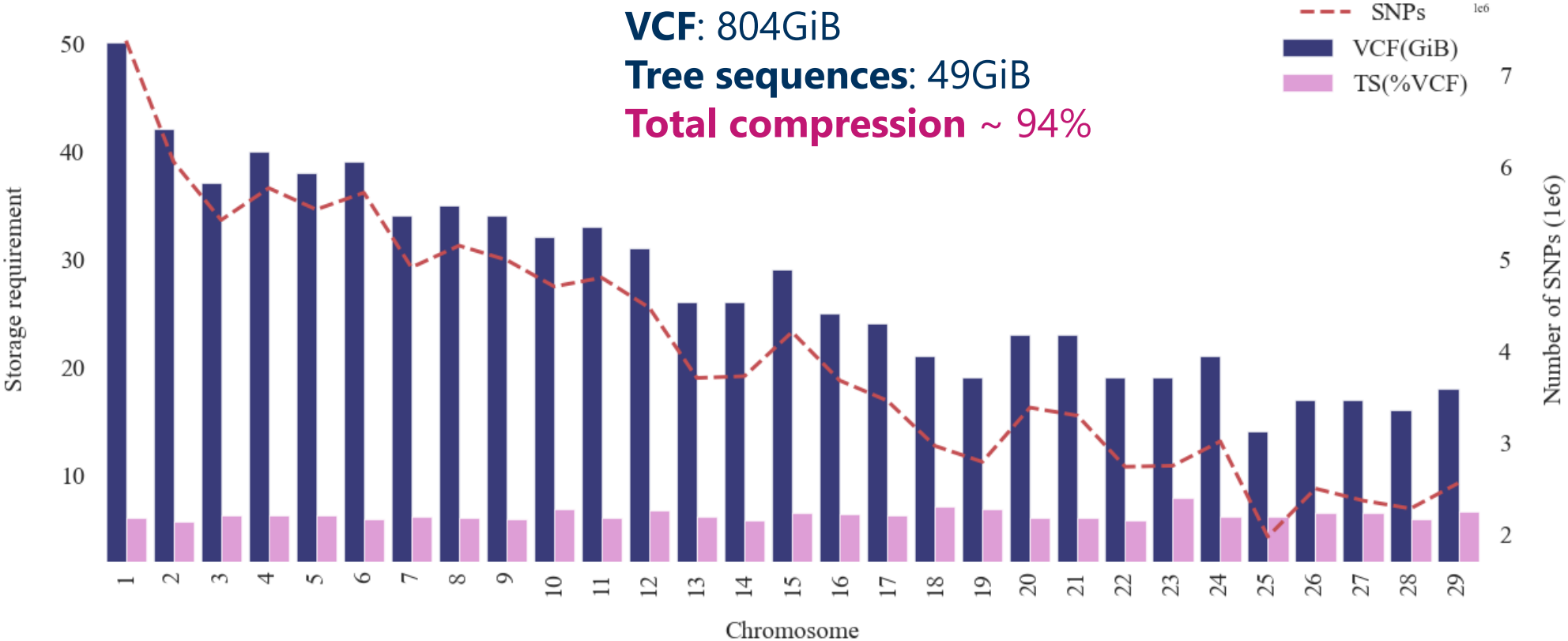


Tree sequence on 1000 Bull Genomes data

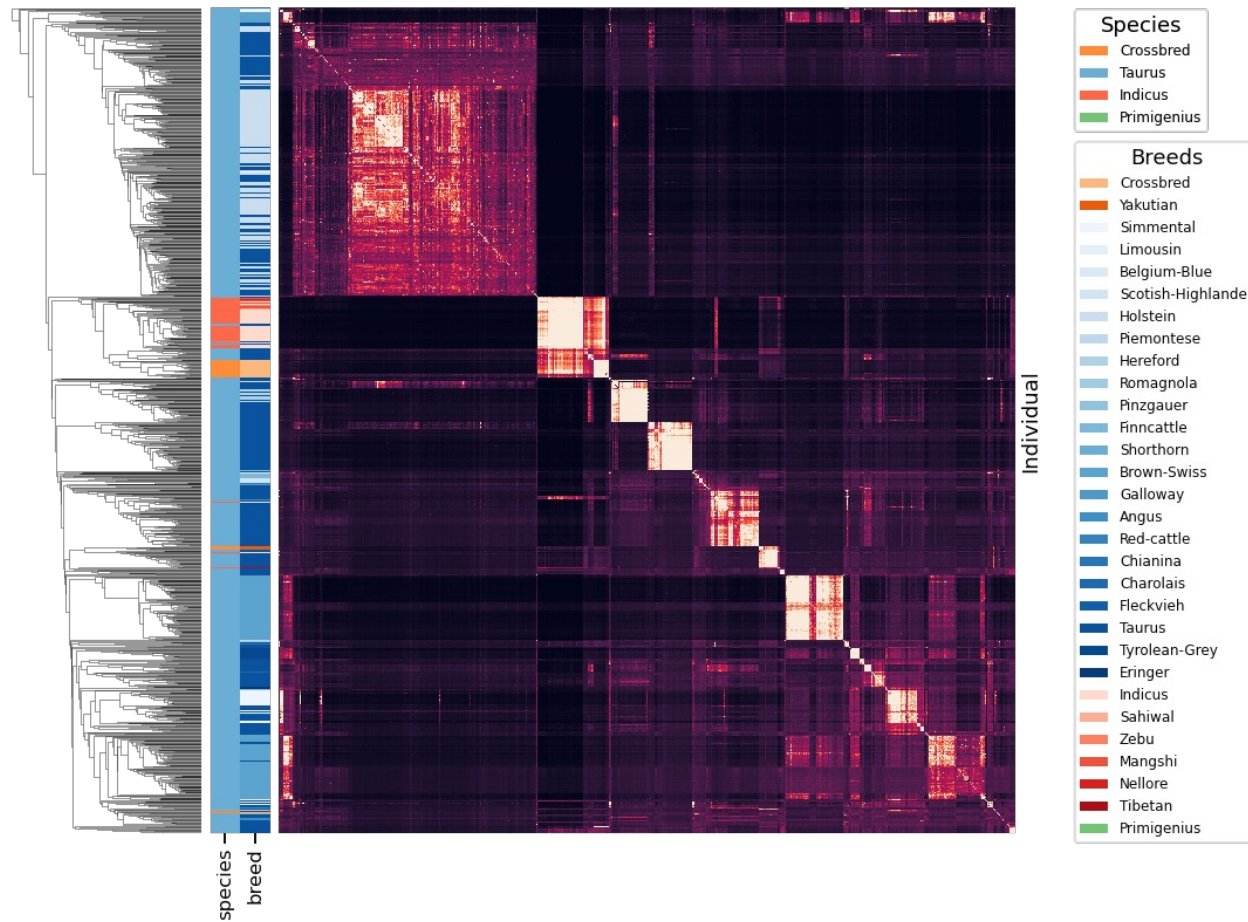
- 2,716 samples & 157 groups
 - Bos taurus
 - Bos indicus
 - Crossbred
 - African
 - Bos taurus coreanae (nat. Korean)
 - Bos primigenius (auroch)
 - Bos grunniens (yak)
- 29 autosomal chromosomes with ~116M variants
- Pipeline
 - Shapelt (phase)
 - tsinfer (infer tree sequence)
 - tskit (analyse)
- Files
 - VCFs: 804 GB
 - Tree sequences: 49 GB
 - "Compression": 94%
- Tree sequence analysis: FAST;)

Tree sequence naturally represents genomic data

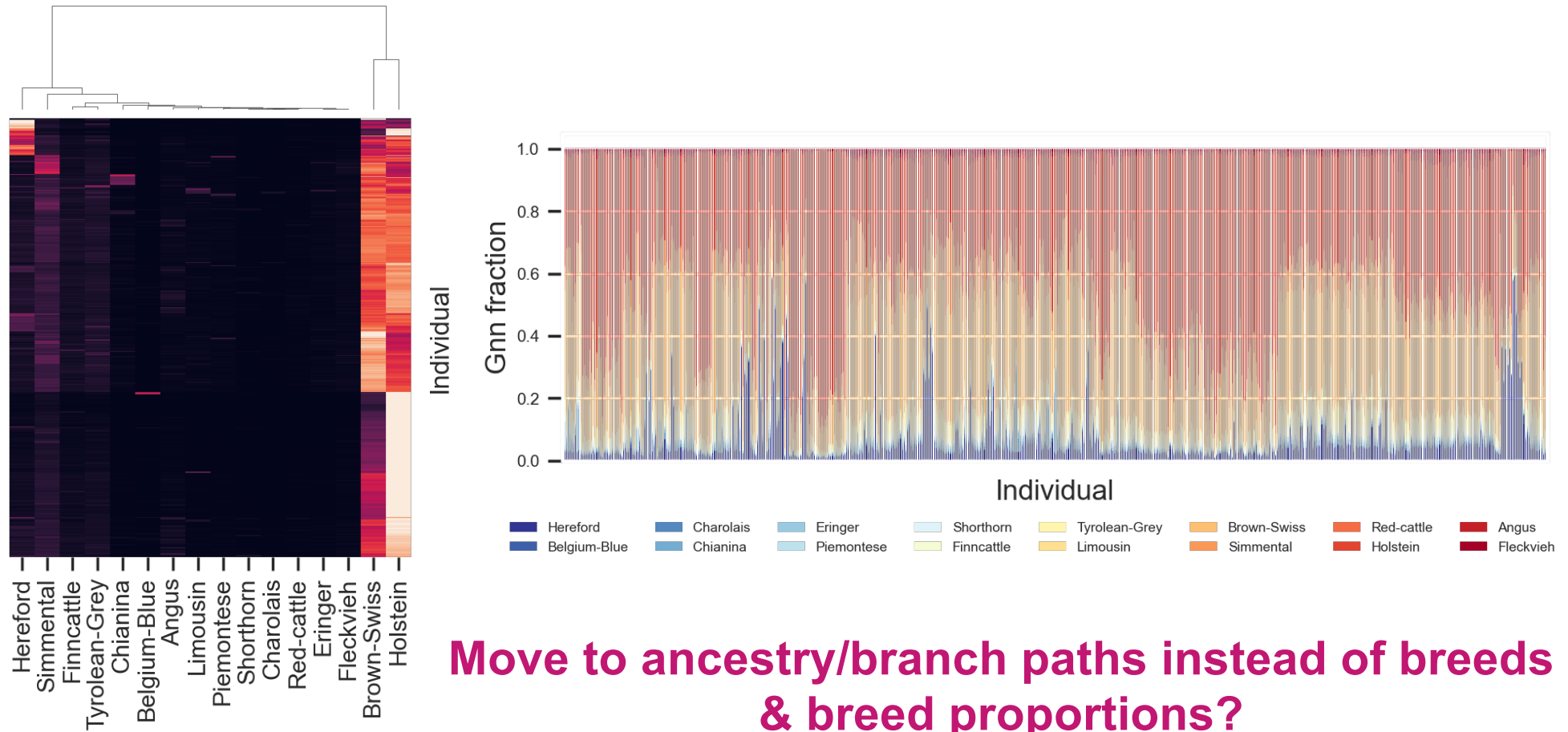
VCF: 804GiB
Tree sequences: 49GiB
Total compression ~ 94%



Population structure summary (Genealogical Nearest Neighbours)



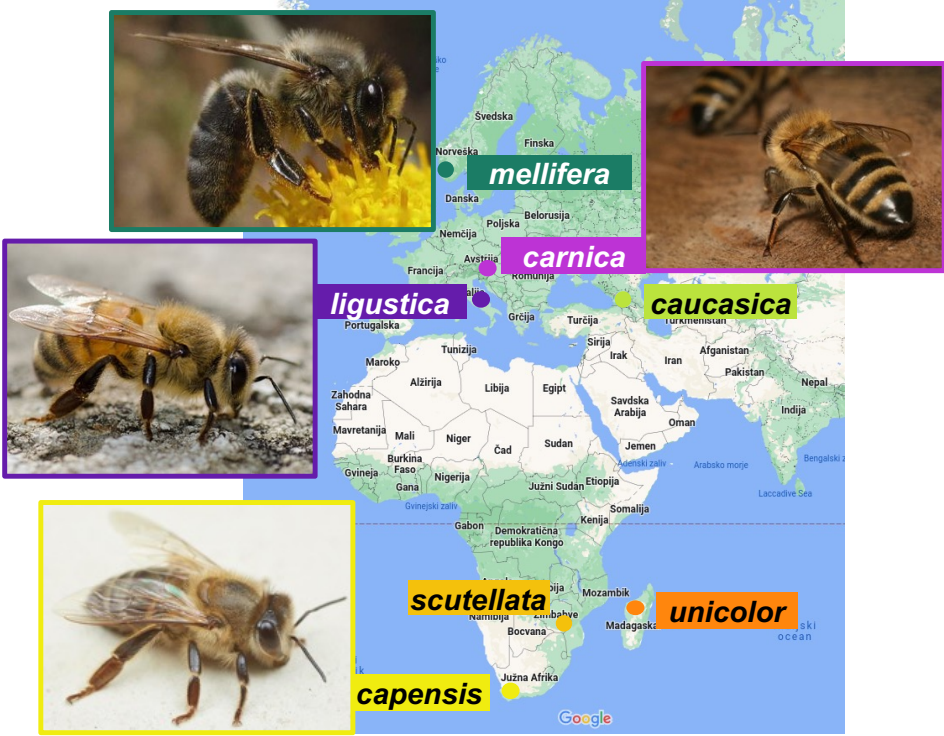
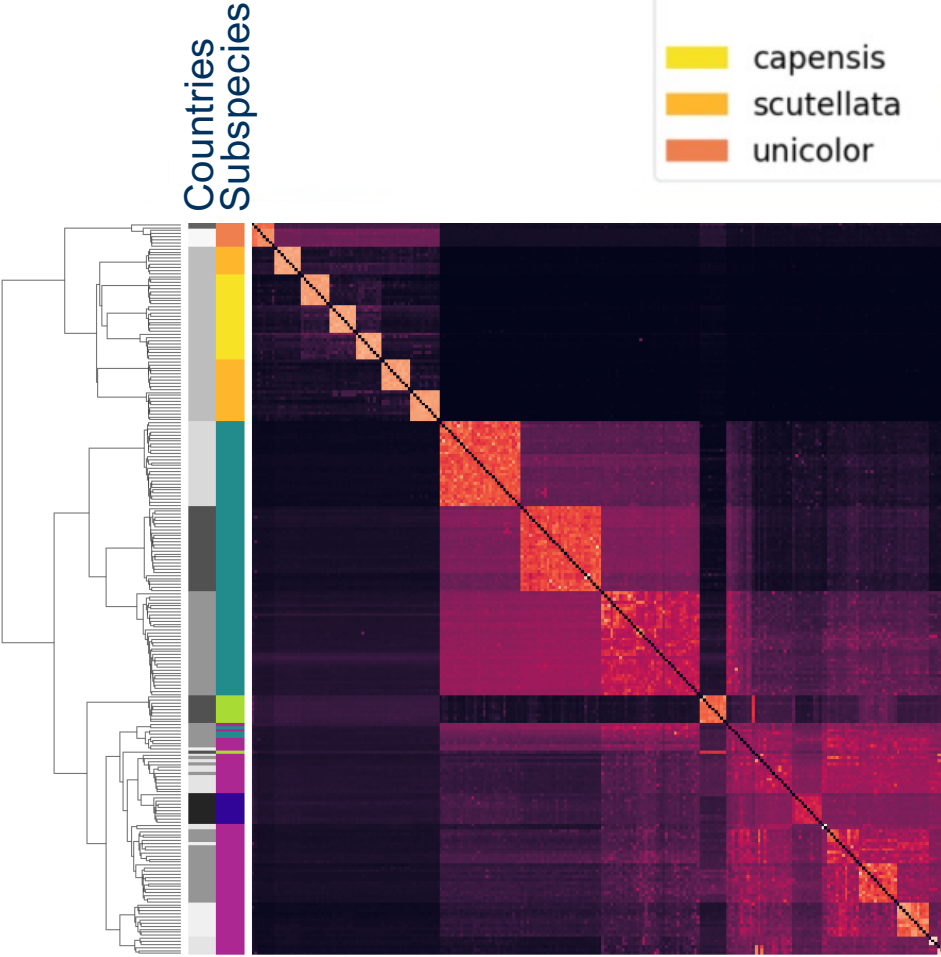
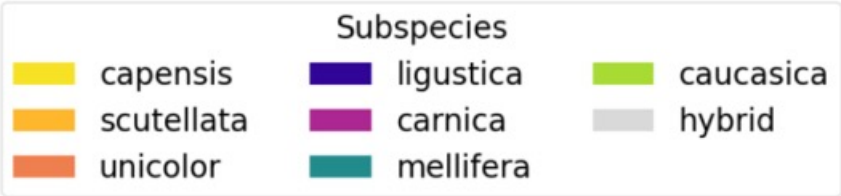
Inferring breed proportions with Genealogical Nearest Neighbours



Bigger picture

- Girolando is a “simple” crossing system
- More complex admixture → African setting
- Crossing systems in the Global North
 - beef-on-dairy
 - beef crosses
- Quantitative genetics within and across breeds & subspecies

Honey bee work



Where have we been & where are we going?

- Started career with pedigree-based mixed models
- Rode on the excitement of introducing genomic selection
 - 2010s $\sim 10^3$ individuals
 - 2015s $\sim 10^5$ individuals
 - 2020s $\sim 10^6$ individuals
 - 2030s $\sim 10^9$ individuals???
- Contributed to the whole-genome sequencing “craze”
 - 2015s $\sim 10^3$ individuals
 - 2020s $\sim 10^{4-6}$ individuals
 - 2030s $\sim 10^{6-9}$ individuals???

SNP array genotypes
(~50K markers)

Whole-genome sequences
(with & without imputation,
~20M+ variants)

Where are we going in ag genetics?

- Started career with pedigree-based mixed models
 - 2010s
 - 2020s
- Rode on genomic selection
 - 2010s
 - 2020s
- Core challenge: **MEGA-SCALE!!!!**
 - 2010s
 - 2020s
 - 2030s ~ 10^9 animals???

Individuals (nInd)

Sites (nSites)

MEGA-SCALE!!!!

**Genotype matrix
(nInd x nSites)**

Roslin-Genus/PIC ~1 million pig genomes project

- Pedigree & SNP array data
 - 9 lines with a total of ~450K pigs
 - ~15K-50K markers
- Whole-genome sequence
 - ~8K pigs (a mix of ~1x and ~30x)
 - ~46M variants passed quality control across lines
- Accurate imputation of whole-genomes
 - ~450K diploid pigs * 2 = **~900,000 haploid genomes**
 - ~450K pigs * ~46M sites * 8 bytes / 2^{40} = ~152 TiB of memory ☹️
(2-bit storage → ~5TiB of memory)



Roslin-Genus/PIC ~1 million pig genomes project

- How will we analyse this DATA BEAST ?
- Let's create a better SNP array
 - preselect SNPs in one way or another
 - 0.03-0.04 increase in genomic prediction accuracy
 - disappointment?!

Tree sequence to the rescue!

- Back-of-the-envelope calculations!
- 46M variants
 - Sites table
 - 46M sites * 3 columns * 8 B / 2³⁰ = 1 GB
 - Mutations table
 - Assuming 1 mutation / site
 - 46M mutations * 4 columns * 8 B / 2³⁰ = 1.4 GB
- 450K pigs
 - Nodes table
 - Total number of whole-chromosome haplotypes = nNodes
 - 450K * 2 ploidy * 19 chromosomes = 171M haplotypes
 - 171M haplotypes * 2 columns * 8 B / 2³⁰ = 2.5 GB
 - Edge table
 - ?

Tables				Nodes	
Edges				ID	time
left	right	parent	child	0	0.0
0	20	4	0	1	0.0
0	20	4	1	2	0.0
0	10	4	2	3	0.0
0	10	5	3	4	2.0
0	10	5	4	5	3.0
10	20	4	6	6	1.0
10	20	6	2		
10	20	6	3		
Mutations				Sites	
ID	site	node	derived	ID	position
0	1	3	T	0	2
1	2	2	G	1	4
2	4	4	T	2	5
3	6	6	G	3	7
4	8	2	T	4	8
				5	9
				6	12
				7	15
				8	18
				9	19
					ancestral
					C
					A
					C
					C
					G
					C
					T
					C
					C
					G
					C

Tree sequence to the rescue!

- Back-of-the-envelope calculations!
- ...
- 450K pigs
 - Nodes table
 - Total number of whole-chromosome haplotypes = nNodes
 - 450K pigs * 2 ploidy * 19 chromosomes = 171M haplotypes
 - 171M haplotypes * 2 columns * 8 B / 2³⁰ = 2.5 GB
 - Edge table
 - in 9 lines, so, 50K pigs / line, say, 10 years → 5K pigs / line / year
 - 5K pigs from 1000 sows (5 progeny) and 50 boars (100 progeny)
 - Assume 1 recombination / whole-chromosome haplotype → 2 edges / progeny haplotype
 - Assume 9 lines * (1000 sows + 50 boars) = ~10K founders – no edges for them (not yet)
 - 440K pigs * 2 ploidy * 19 chromosomes = 167M progeny haplotypes
 - 167M haplotypes * 2 edges = 334M edges
 - 334M edges * 4 columns * 8 B / 2³⁰ = ~10 GB

Tables				Nodes		
Edges				ID	time	
left	right	parent	child	0	0.0	
0	20	4	0	1	0.0	
0	20	4	1	2	0.0	
0	10	4	2	3	0.0	
0	10	5	3	4	2.0	
0	10	5	4	5	3.0	
10	20	4	6	6	1.0	
10	20	6	2			
10	20	6	3			
Mutations				Sites		
ID	site	node	derived	ID	position	ancestral
0	1	3	T	0	2	C
1	2	2	G	1	4	A
2	4	4	T	2	5	C
3	6	6	G	3	7	G
4	8	2	T	4	8	C
				5	9	T
				6	12	T
				7	15	C
				8	18	G
				9	19	C

Tree sequence to the rescue!

- **Memory requirements for the genotype matrix?**
 - 440,610 pigs and 46,344,624 sites
→ $n_{\text{Ind}} * n_{\text{Site}} * 8 / 2^{30} = 152,140 \text{ GB} = \sim 152 \text{ TB}$
- **Memory requirements for the tree sequence?**
 - Sites table 46M sites * 3 columns → $46\text{M} * 3 \text{ columns} * 8 \text{ B} / 2^{30} = 1 \text{ GB}$
 - Mutations table 46M sites * 4 columns → $46\text{M} * 4 \text{ columns} * 8 \text{ B} / 2^{30} = 1.4 \text{ GB}$
 - Nodes table 171M haplotypes * 2 columns * 8 B / $2^{30} = 2.5 \text{ GB}$
 - Edge table 334M edges * 4 columns * 8 B / $2^{30} = \sim 10 \text{ GB}$
 - Total $1 + 1.4 + 2.5 + 10 = \sim 15 \text{ GB}$
- 15 GB out of 152 TB is $\sim 99.9\%$ “compression”!

UK dairy herd

- ~2M COWS (1.85M <https://www.statista.com/statistics/616188/dairy-cow-numbers-united-kingdom-uk>)
- 50K SNP array $\rightarrow 2M * 50K * 8 / 2^{30} = 745 \text{ GB}$
 - Sites table
 - 50K sites * 3 columns * 8 B / $2^{20} = 1 \text{ MB}$
 - Mutations table
 - Assuming 1 mutation / site
 - 50K sites * 3 columns * 8 B / $2^{20} = 1.5 \text{ MB}$
 - Assuming 25%/year replacement we have 800K 1st parity cows / year
 - Assume 100 sires used every year to sire the 800K
 - Nodes table
 - 800K cows * 2 ploidy * 30 chromosomes = 48M haplotypes
 - 48M haplotypes * 2 columns * 8 B / $2^{30} = 0.7 \text{ GB}$
 - Edge table
 - Assume 1 recombination / whole-chromosome haplotype $\rightarrow 2$ edges / progeny haplotype
 - 800K cows * 2 ploidy * 30 chromosomes = 48M progeny haplotypes
 - 48M haplotypes * 2 edges = 96M edges
 - 96M edges * 4 columns * 8 B / $2^{30} = \sim 2.9 \text{ GB}$

Conclusion

- MEGA-SCALE genomic datasets are here & growing
- ARGs & Tree sequence data format to the rescue!?
 - PROS
 - Succinctly encodes the inheritance process
 - Combines pedigree, coalescent, phylogenetics (gene & species trees), segregation, recombination, gene conversion, mutations, IBS, IBD, ...
 - Significant storage reduction & fast analyses
 - Novel insights & modelling
 - CONS (on-going work)
 - Novel way of thinking
 - Ancestral alleles, inference from real data, and inputs still HUGE
 - Need to develop specialised algorithms

Learning objectives

- Motivate tree-thinking as a way forward in genomics
- Revisiting the fundamentals of DNA events and how to efficiently encode DNA variation
- Understand ARGs & tree sequence format
- Showcase ongoing agricultural applications

Questions?!



THE UNIVERSITY
of EDINBURGH



Ancestral recombination Graphs (ARGs)

Gregor Gorjanc, Chris Gaynor, Jon Bancic, Daniel Tolhurst

UNE, Armidale

2024-02-09

