Inference of ancestral recombination graphs for population genomics

Computational Methods in Evolution and Biodiversity

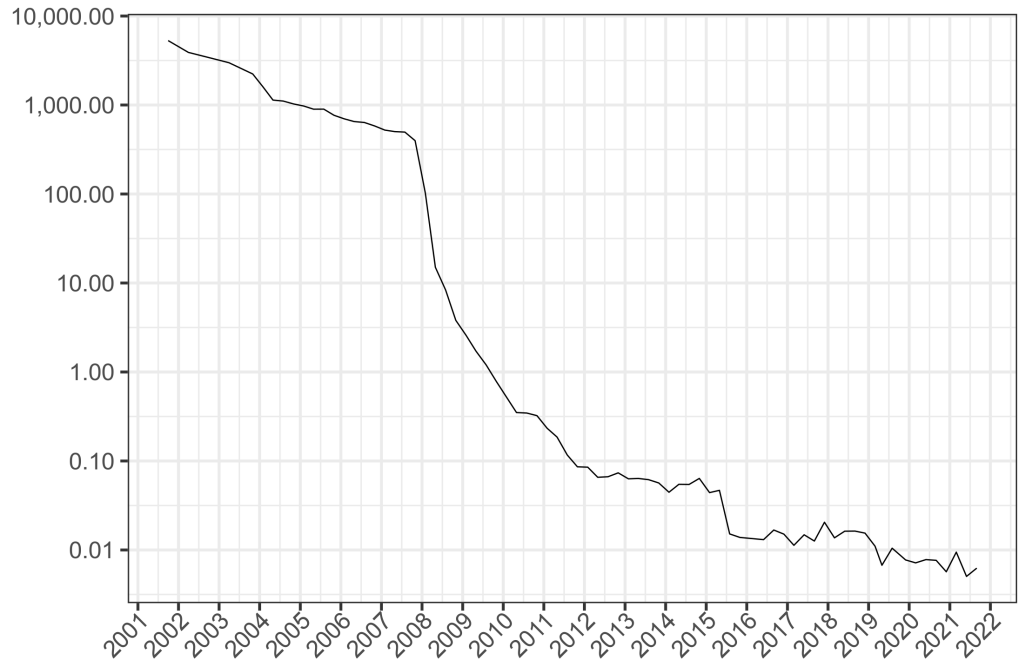
Per Unneberg

Based on slides by Yan Wong
https://github.com/hyanwong/genealogy_workshop

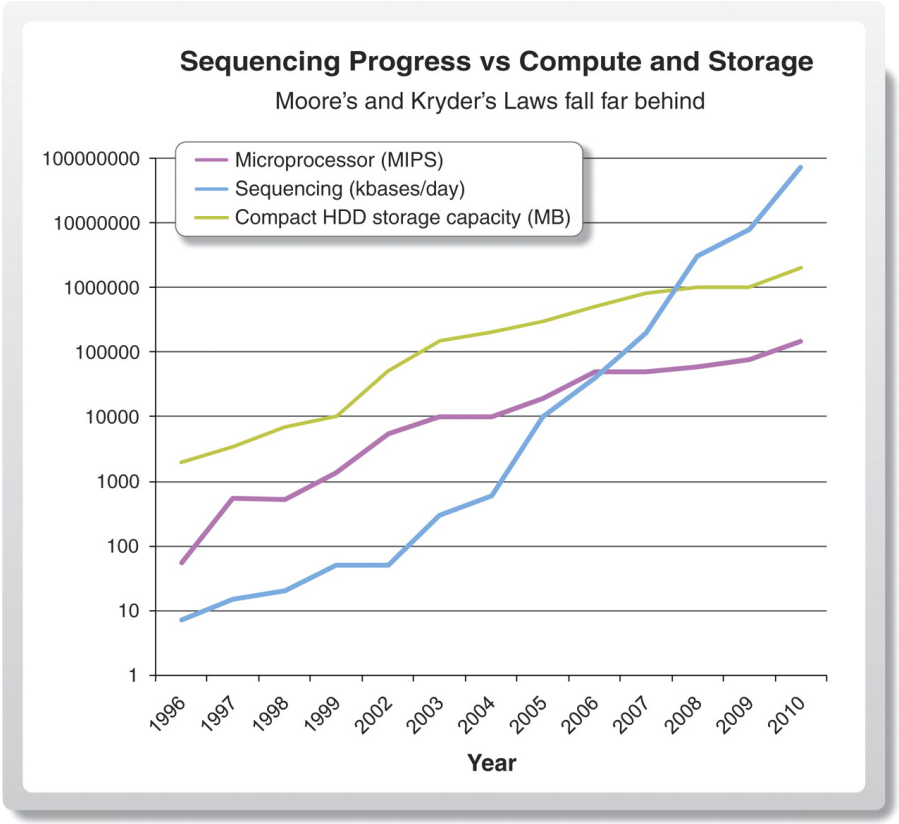
NBIS

20-Sep-2023

Genomics data is large



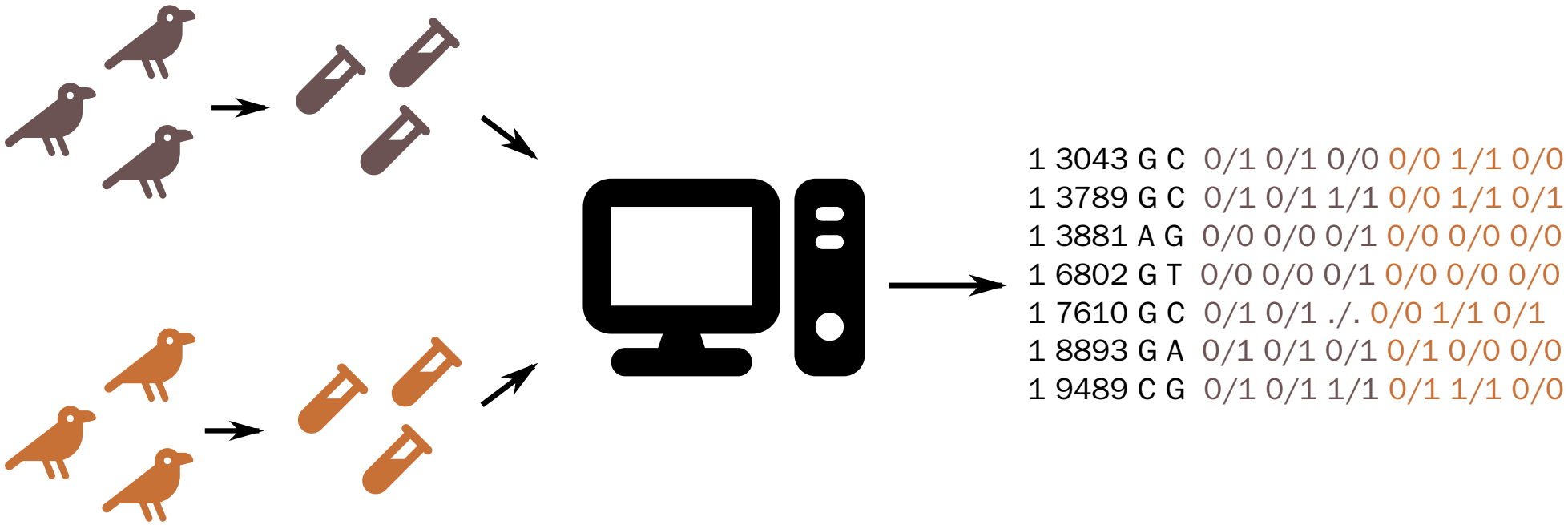
Sequencing cost ($) per megabase (Wetterstrand, KA, 2022)



On the future of genomic data (Kahn, 2011) (!)

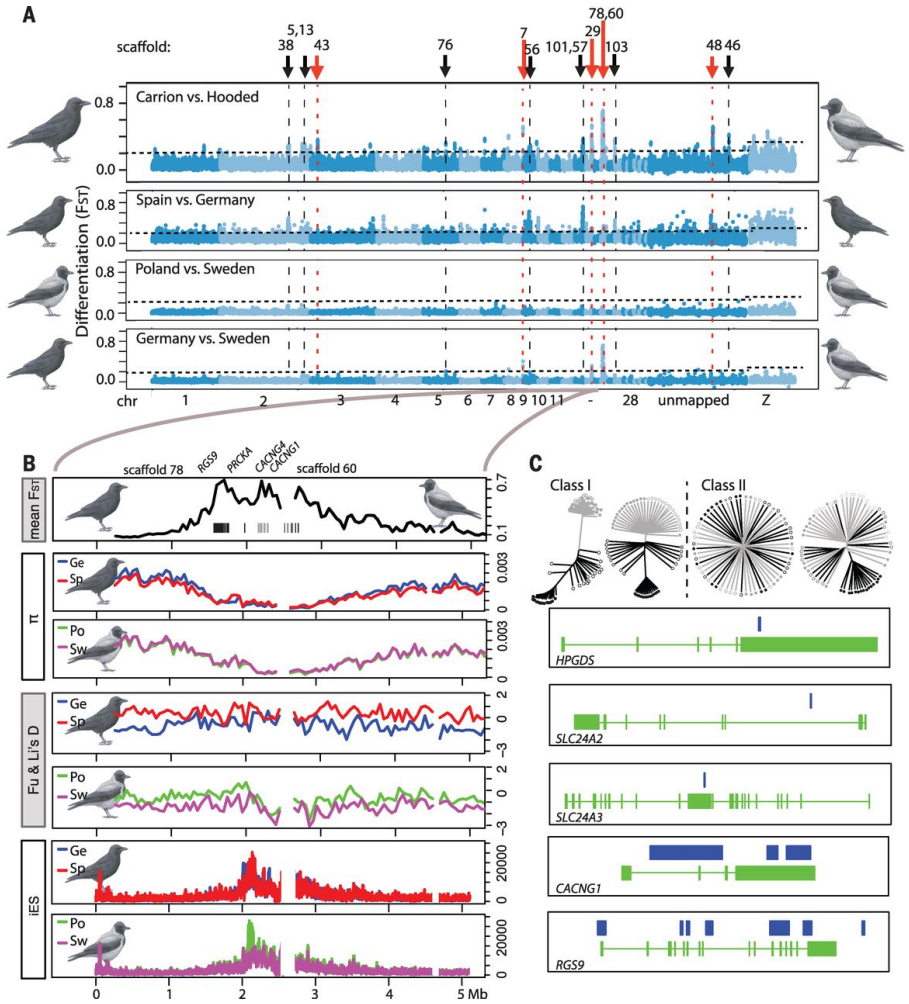
Population genomics

From sample collection and preparation through assembly, sequencing, variant calling and other data processing steps to genotype matrices



```
1 3043 G C  0/1 0/1 0/0 0/0 1/1 0/0
1 3789 G C  0/1 0/1 1/1 0/0 1/1 0/1
1 3881 A G  0/0 0/0 0/1 0/0 0/0 0/0
1 6802 G T  0/0 0/0 0/1 0/0 0/0 0/0
1 7610 G C  0/1 0/1 ./. 0/0 1/1 0/1
1 8893 G A  0/1 0/1 0/1 0/1 0/0 0/0
1 9489 C G  0/1 0/1 1/1 0/1 1/1 0/0
```

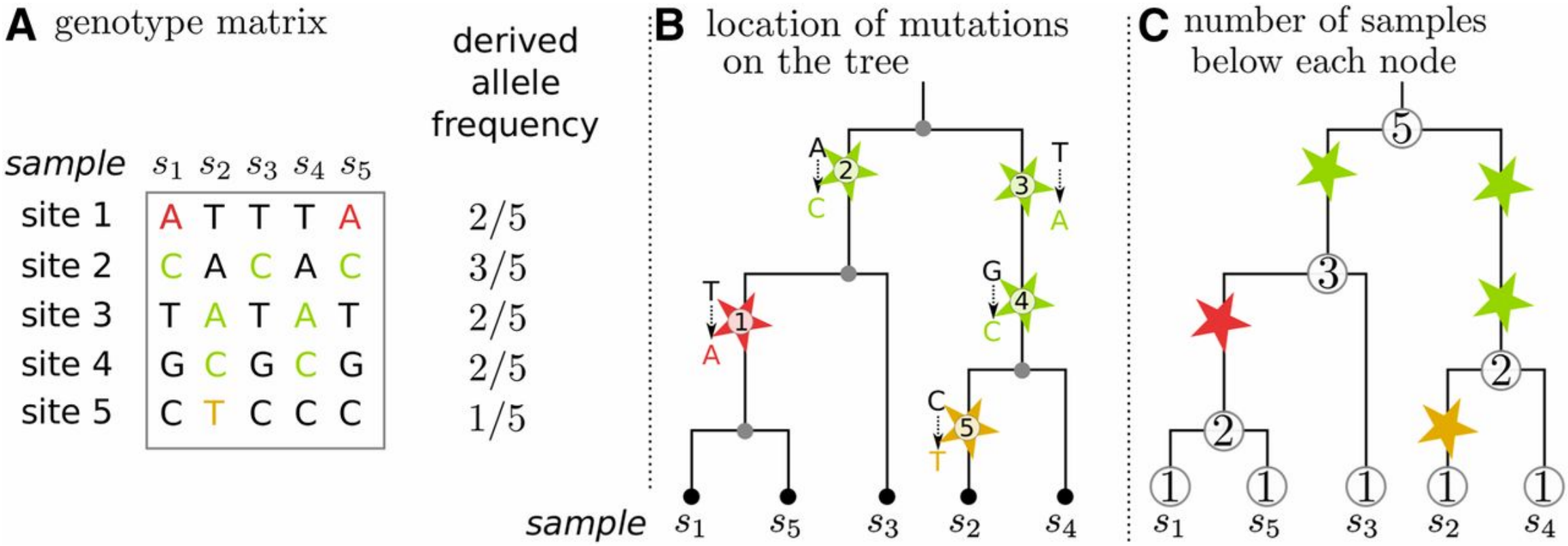
Making sense of variation data



The Genomic Landscape Underlying Phenotypic Integrity in the Face of Gene Flow in Crows
Poelstra et al. (2014)

Tree sequence inference

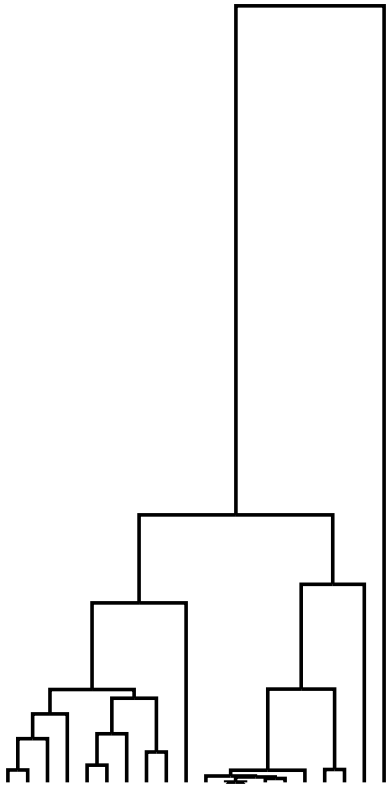
Genotype matrices and genealogical trees



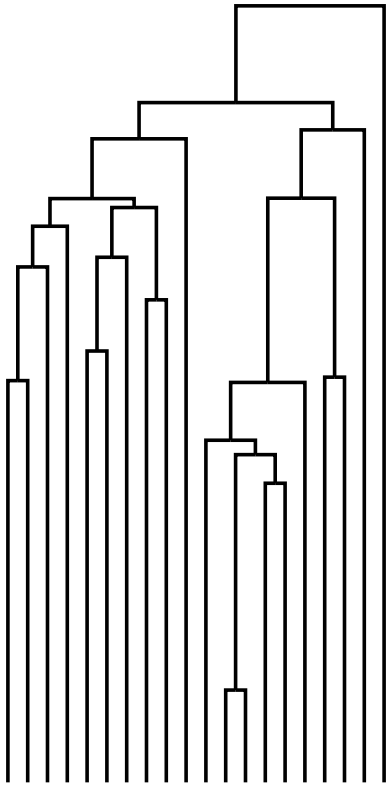
Efficiently Summarizing Relationships in Large Samples: A General Duality Between Statistics of Genealogies and Genomes. Ralph et al. (2020), Fig. 1

Trees capture biology

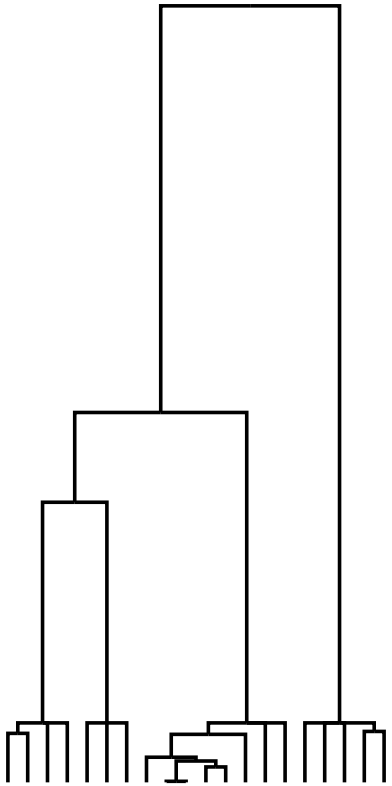
Neutral



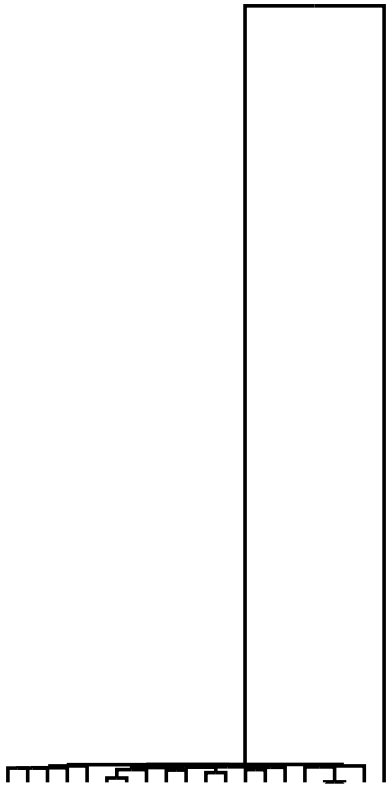
Expansion



Bottleneck

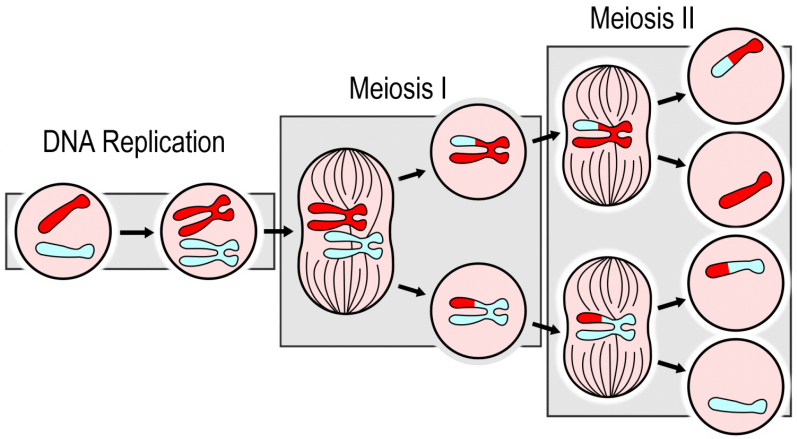


Selection

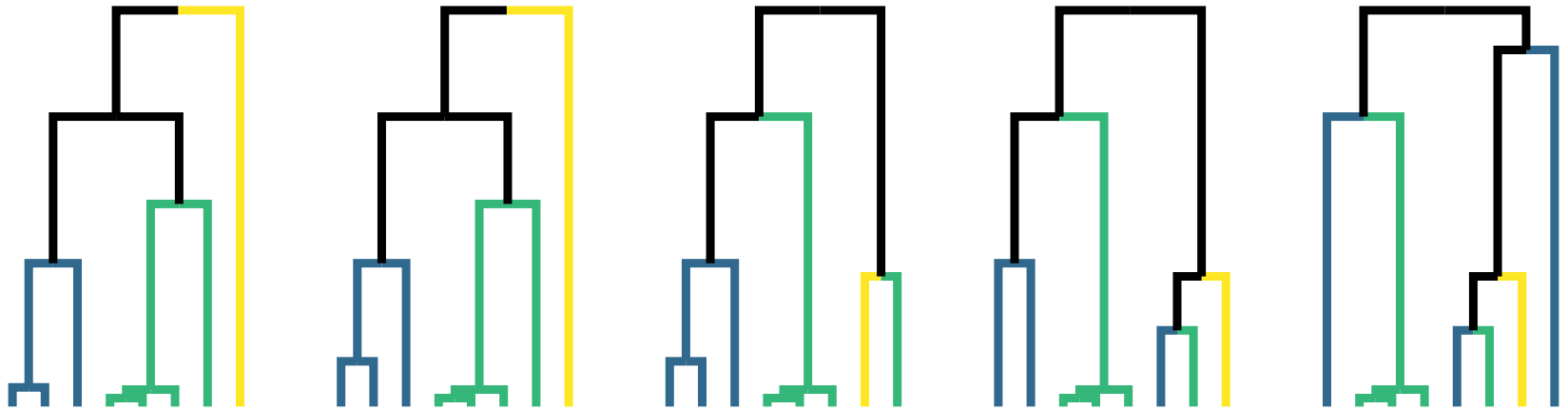


Recombination modifies gene genealogies

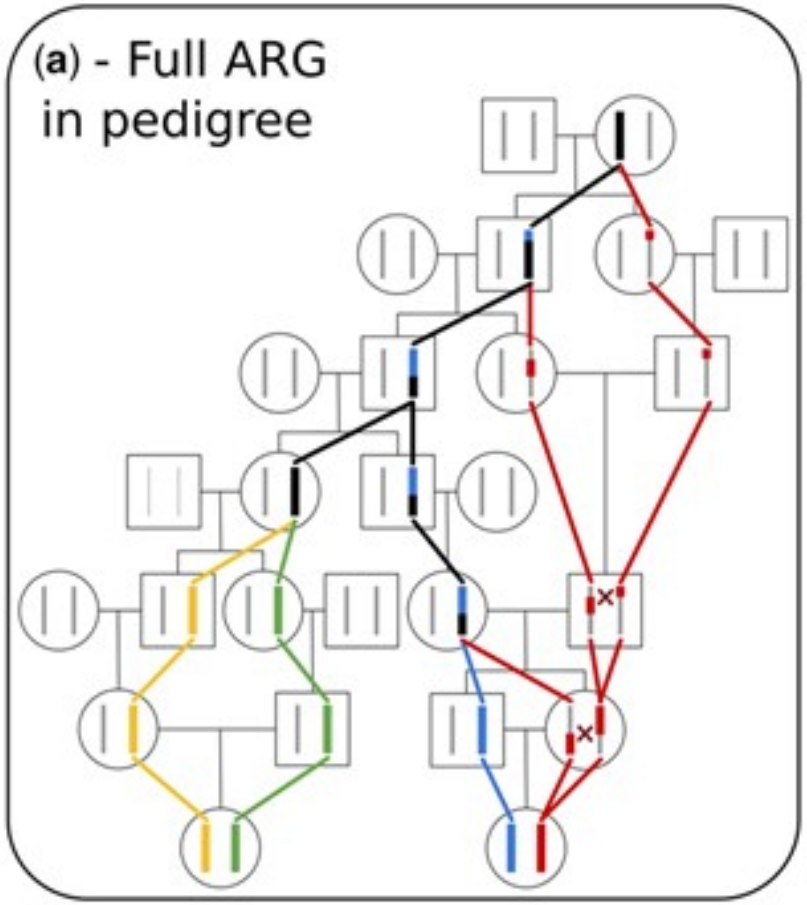
Most organisms recombine a lot!



Miller (2020), Fig. 5.12.3

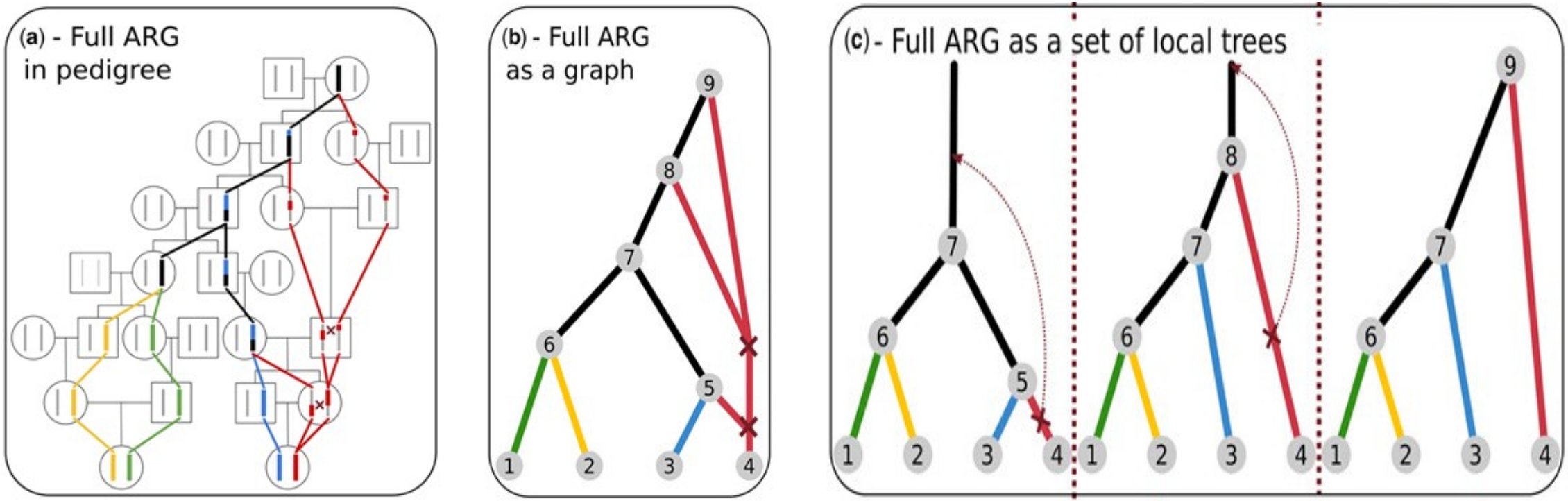


Ancestral recombination graphs (ARGs)



Y. C. Brandt et al. (2022, fig. 1a)

A “genetic genealogy” tracks genome-wide inheritance paths

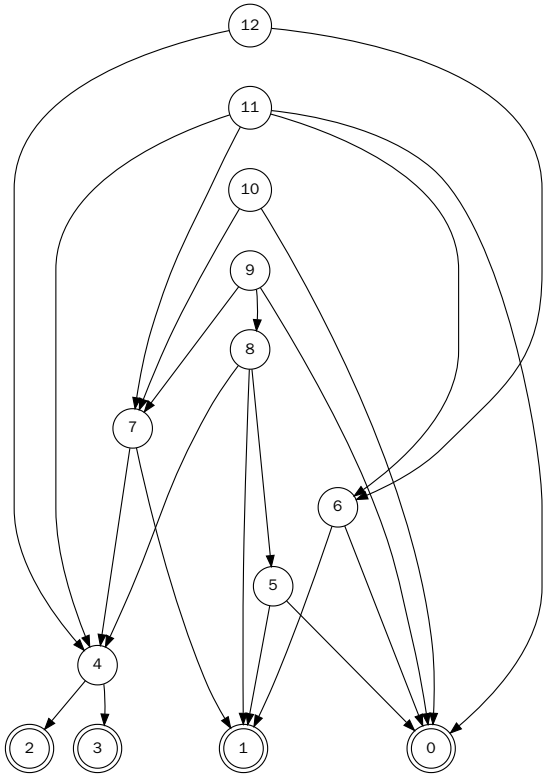


“Succinct tree sequences”

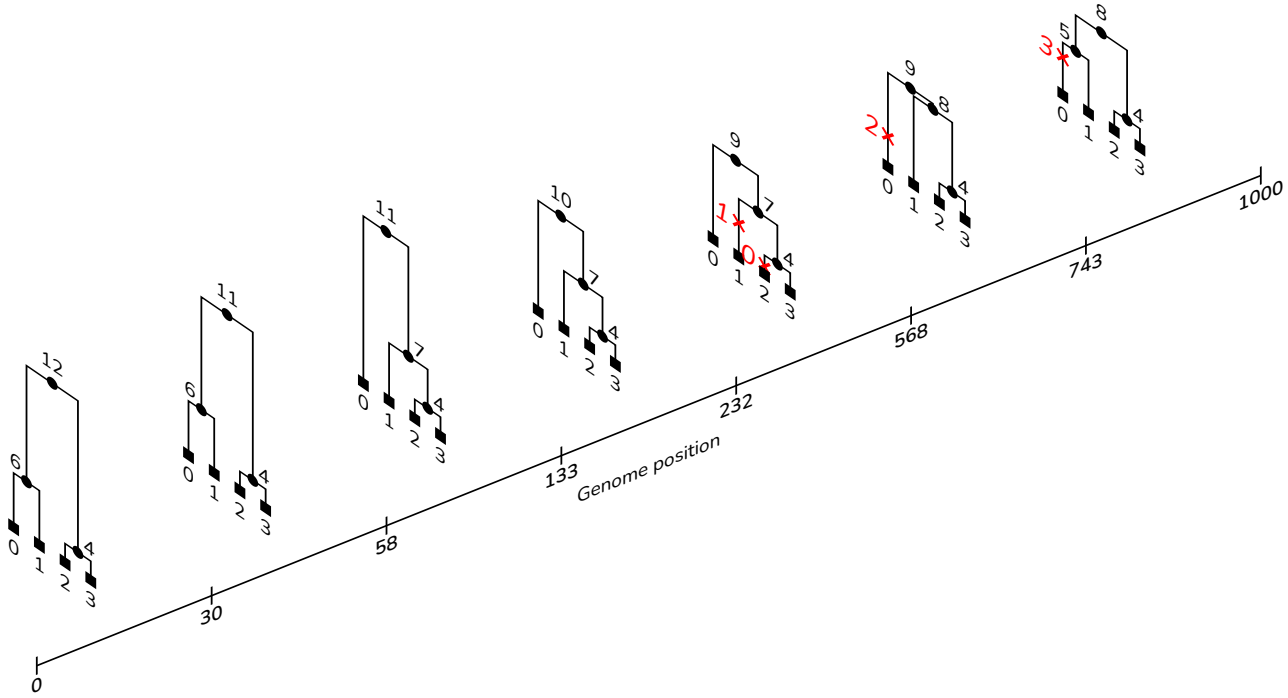
Evaluation of Methods for Estimating Coalescence Times Using Ancestral Recombination Graphs Y. C. Brandt et al. (2022, fig. 1)

msprime enables simulation of large chromosomes with recombination

Graph representation

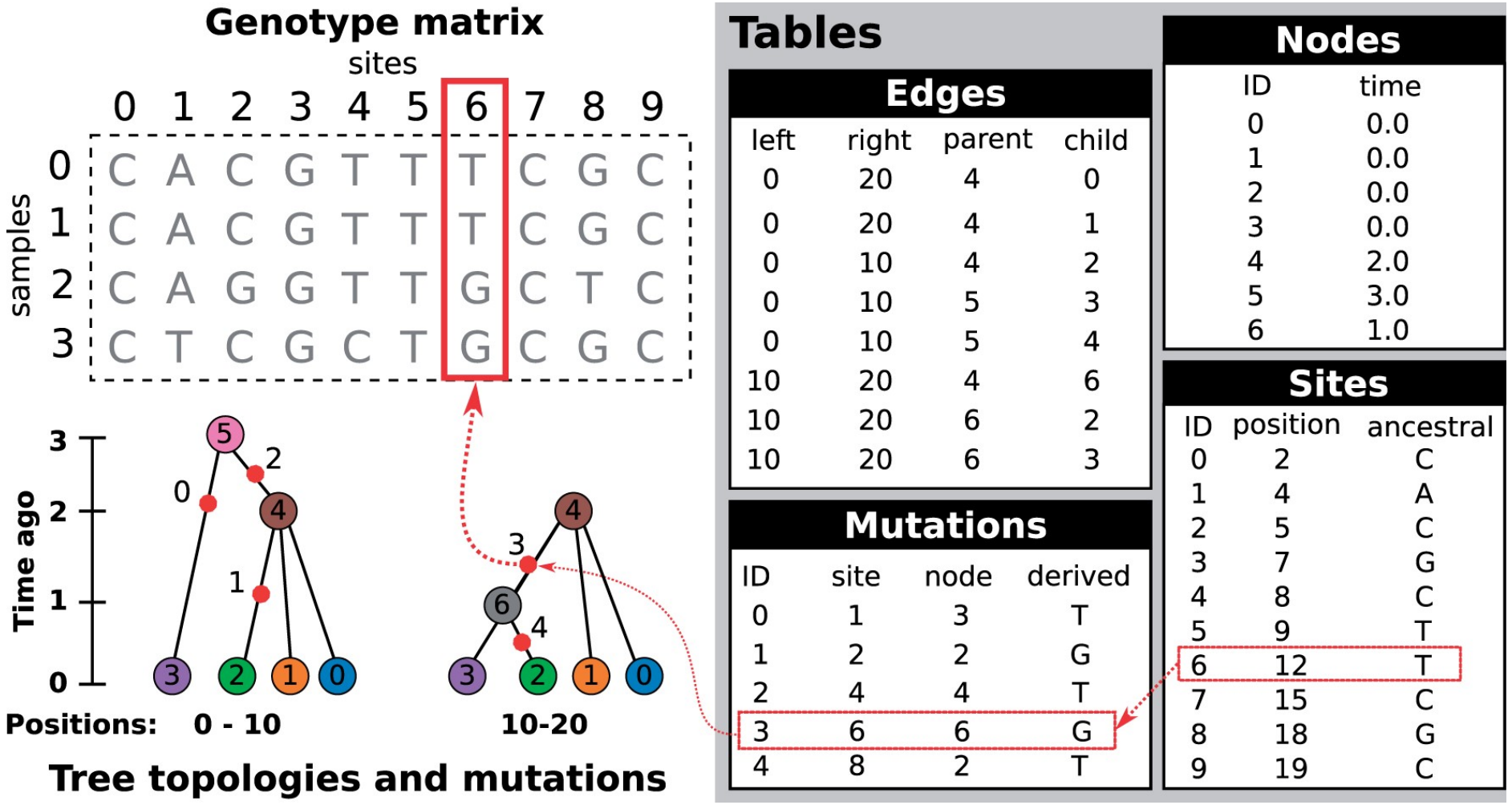


Local tree representation



cf https://github.com/tskit-dev/tutorials/issues/43

msprime stores data as succinct tree sequences

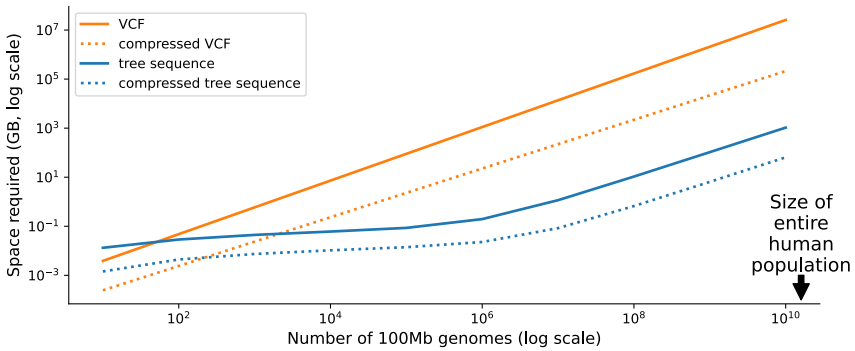


Tree sequences (Baumdicker et al., 2022, fig. 2)

Tree sequences compress data and speedup analyses

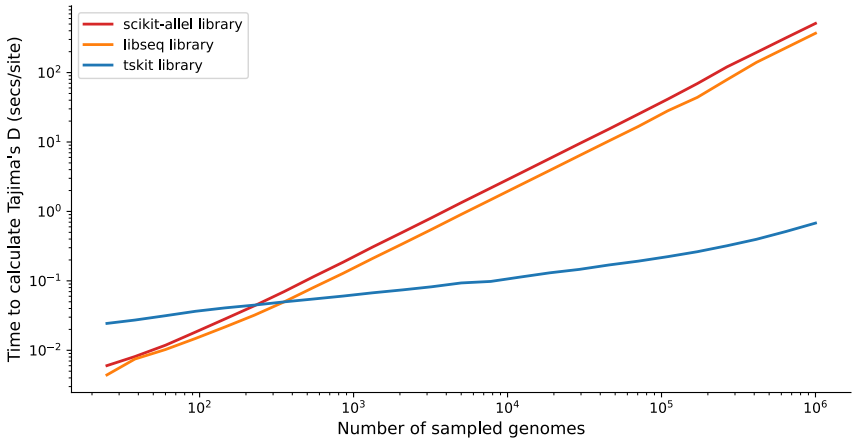
- Compact storage (“domain specific compression”)
- Fast, efficient analysis (a “succinct” structure)
- Well tested, open source (active dev community)

Data compression



- Built-in functionality (well documented: http://tskit.dev)

Speed

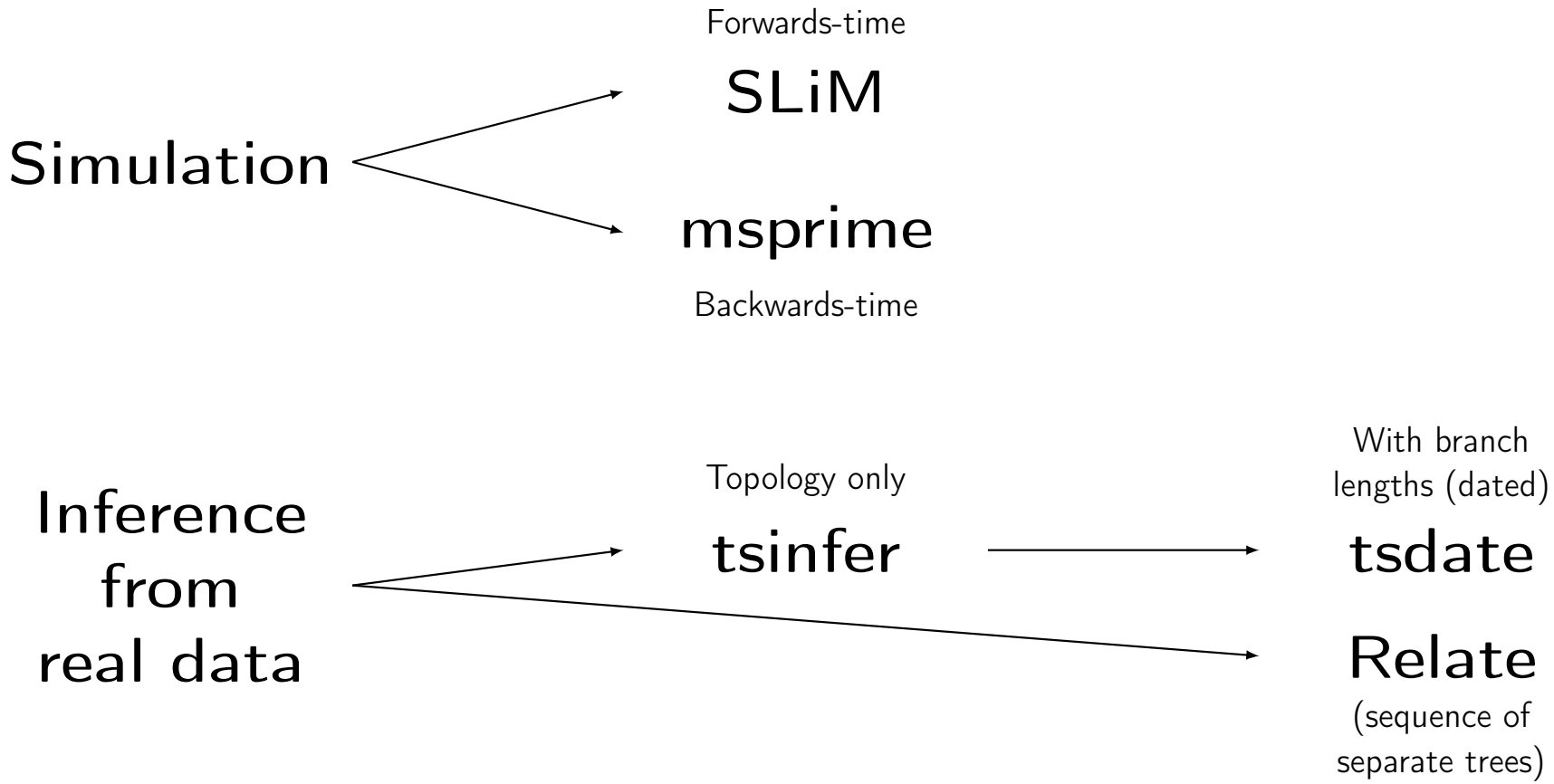


…but limited support for major genomic rearrangements (e.g. inversions, large indels): genomes should be (reasonably) aligned => current primary focus = population genetics

Tree sequence inference

Source: What is a tree sequence? (https://tskit.dev/tutorials/what_is.html)

Getting hold of tree sequences

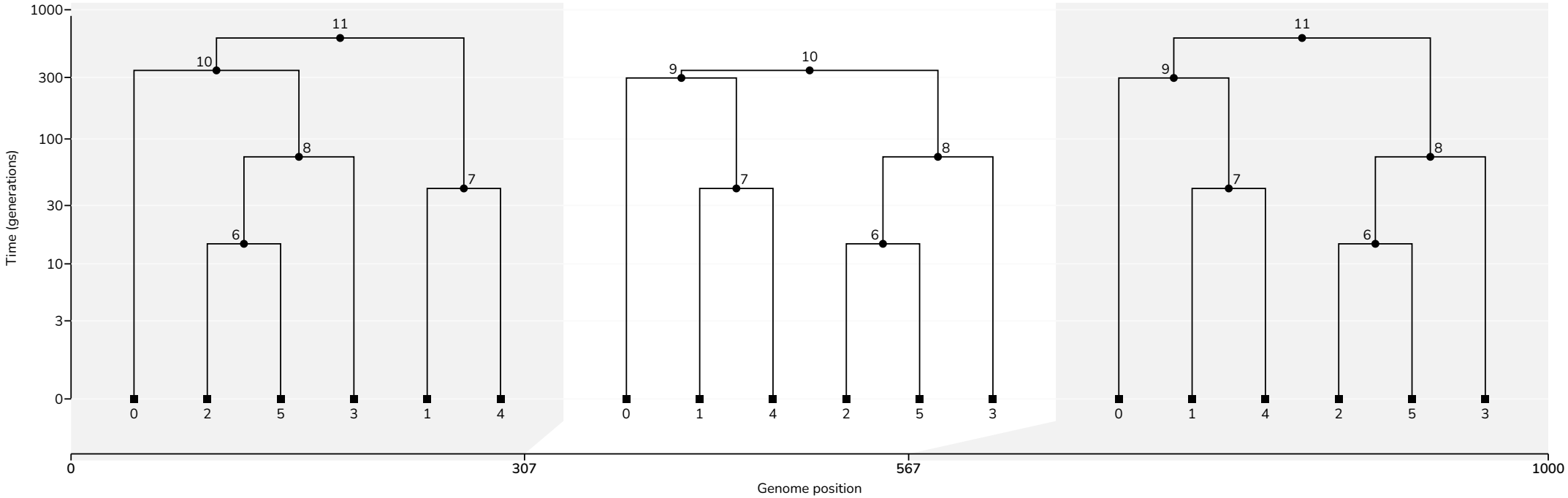


Other programs that don't output tree sequence format by default: *ARGweaver*, *Argneedle*

Analysing tree sequences

https://tskit.dev

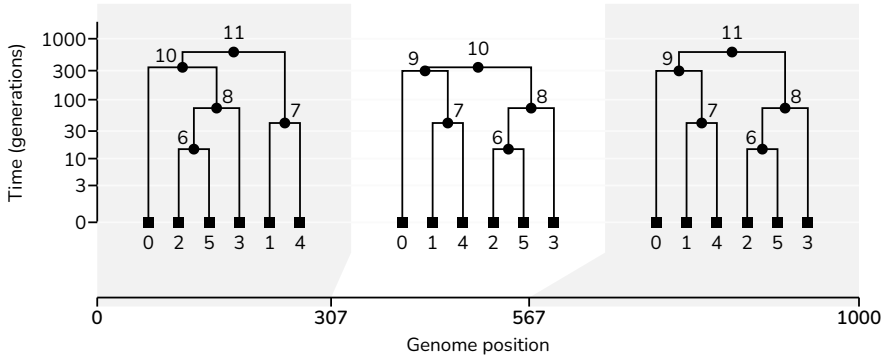
tskit terminology: the basics



- Multiple local trees exist along a genome of fixed length (by convention measured in base pairs)
- Genomes exist at specific times, and are represented by nodes (the same node can persist across many local trees)
- Some nodes are most recent common ancestors (MRCAs) of other nodes
- Entities are zero-based: the first node has id 0, the second id 1, …

Images from online tutorial “Terminology & concepts” https://tskit.dev/tutorials/terminology_and_concepts.html

tskit terminology: nodes and edges



Nodes (=genomes)

- exist at a specific time
- can be flagged as “samples”
- can belong to “individuals” (e.g., 2 nodes per individuals in humans) and, if useful, “populations”

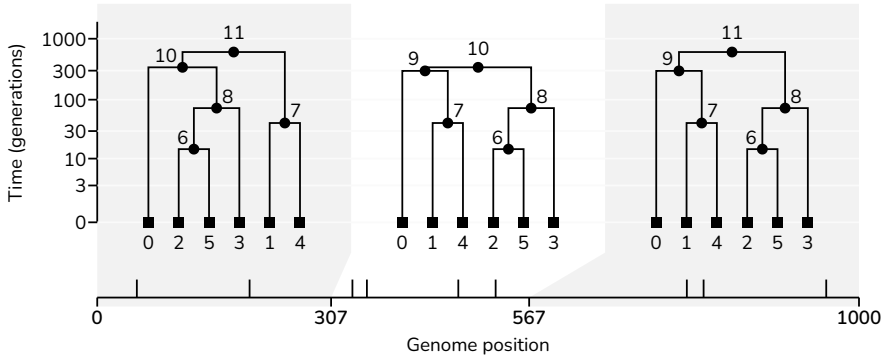
id	flags	population	individual	time	metadata
0	1	0	0	0.00000000	
1	1	0	0	0.00000000	
2	1	0	1	0.00000000	
3	1	0	1	0.00000000	
4	1	0	2	0.00000000	
5	1	0	2	0.00000000	
6	0	0	-1	14.70054184	
7	0	0	-1	40.95936939	
8	0	0	-1	72.52965866	
9	0	0	-1	297.22307150	
10	0	0	-1	340.15496436	
11	0	0	-1	605.35907657	

Edges

- Connect a parent & child
- Have a left & right genomic coordinate
- Usually span multiple trees (e.g., edges connecting nodes 1+7 and 4+7)

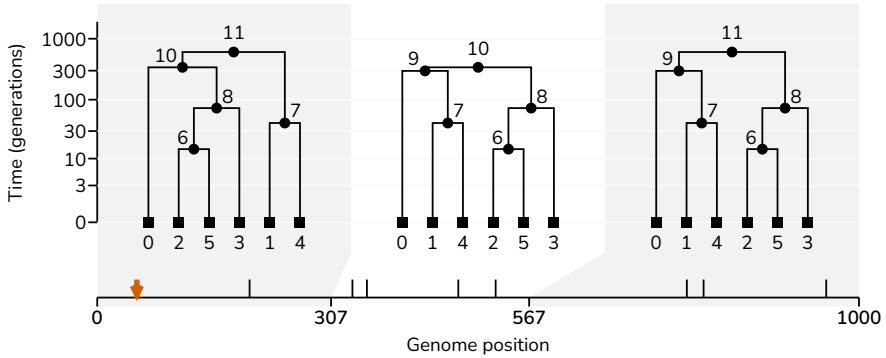
id	left	right	parent	child	metadata
0	0	1000	6	2	
1	0	1000	6	5	
2	0	1000	7	1	
3	0	1000	7	4	
4	0	1000	8	3	
5	0	1000	8	6	
6	307	1000	9	0	
7	307	1000	9	7	
8	0	307	10	0	
9	0	567	10	8	
10	307	567	10	9	
11	0	307	11	7	
12	567	1000	11	8	
13	567	1000	11	9	
14	0	307	11	10	

tskit terminology: sites and mutations



This is how we can encode genetic variation. Most genomic positions do not vary between genomes: usually we don't bother tracking these.

tskit terminology: sites and mutations

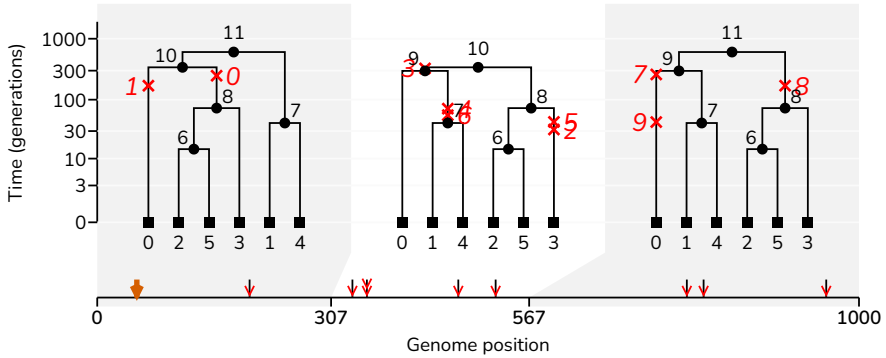


This is how we can encode genetic variation. Most genomic positions do not vary between genomes: usually we don't bother tracking these.

We can create a site at a given genomic position with a fixed ancestral state.

id	position	ancestral_state	metadata
0	52	C	
1	200	A	
2	335	A	
3	354	A	
4	474	G	
5	523	A	
6	774	C	
7	796	C	
8	957	A	

tskit terminology: sites and mutations



We can create a site at a given genomic position with a fixed ancestral state.

id	position	ancestral_state	metadata
0	52	C	
1	200	A	
2	335	A	
3	354	A	
4	474	G	
5	523	A	
6	774	C	
7	796	C	
8	957	A	

This is how we can encode genetic variation. Most genomic positions do not vary between genomes: usually we don't bother tracking these.

Normally, a site is created in order to place one or more mutations at that site

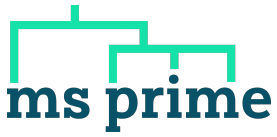
id	site	node	time	derived_state	parent	metadata
0	0	8	247.85988972	T	-1	
1	1	0	169.80687857	C	-1	
2	2	3	31.84262397	C	-1	
3	3	9	326.26095349	C	-1	
4	3	7	71.04212649	T	3	
5	4	3	42.72352948	C	-1	
6	5	7	55.44045835	T	-1	
7	6	0	259.82567754	T	-1	
8	7	8	169.87040769	G	-1	
9	8	0	42.47396523	C	-1	

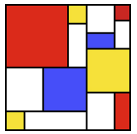
Using tskit

Docs and tutorials

https://tskit.dev/tskit/docs

https://tskit.dev/tutorials





tskit and biodiversity

tskit assumes

1. known ancestral state
2. phased genomes

and requires fairly large sample sizes to leverage power of data compression ($n > 1000$) and speedup of statistical analyses (n in the hundreds)

...conditions that are not always met for natural populations of non-model organisms

Reasons to use tskit ecosystem for evolution and biodiversity

future-proofing
: cheaper and longer read sequencing will require this sort of approach

simulation
: simulation software builds on tskit (msprime/SLiM/stdpopsim)

biology
: thinking in trees captures the “true” biology (unless structural variation)

statistical power
: trees capture genealogical history and variation and potentially have more statistical power than other methods, such as summary statistics

teaching
: biodiversity crowd very familiar with phylogenetic trees making the extension to tree sequences a short jump

modelling of complex histories
: complex histories with, e.g., hybridization / speciation, will have lots of ILS / conflicting trees which needs to be tackled somehow

alternatives to tsinfer
: tsinfer is only one way to infer genealogies but easy to introduce and demonstrate

Application: Evolutionary genomics of the *Motacilla alba* (white wagtails) radiation

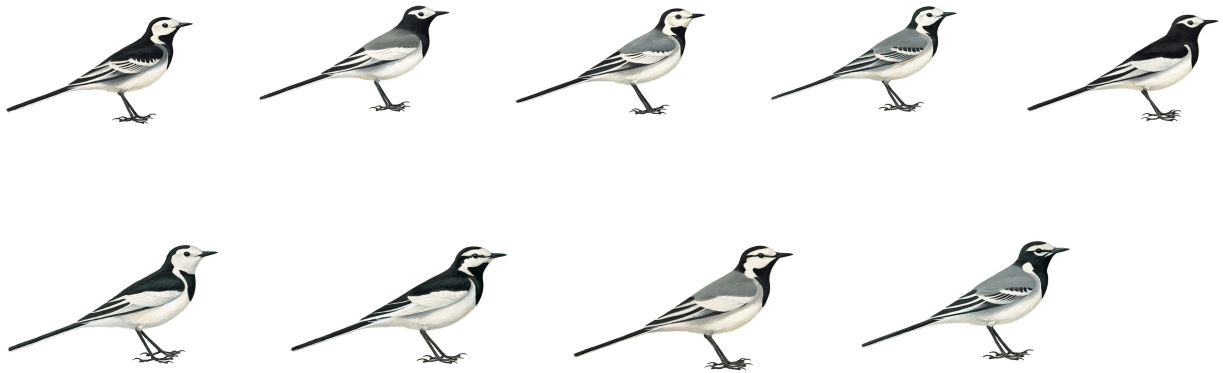
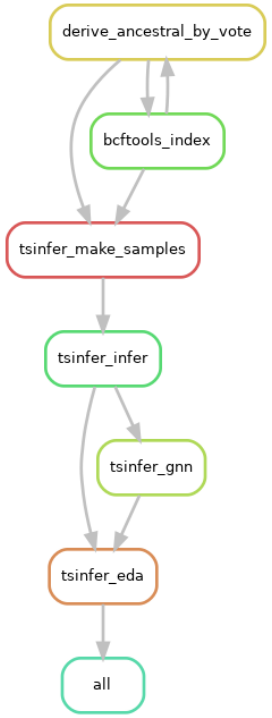


Figure 1: *Motacilla alba* subspecies; from top left *M. yarrellii*, *M. personata*, *M. baicalensis*, *M. alba*, *M. alboides*, *M. leucopsis*, *M. lugens*, *M. ocularis*, and *M. subpersonata*. Paintings by Bill Zetterström (from Alström & Mild (2003))

Together with Erik Enbody, Tom van der Valk, Leif Andersson, and Per Alström.

snakemake



Snakemake rulegraph

Source code at https://github.com/percyfal/tswf

Genealogical nearest neighbour chromosome plots

About

Collection of plots based on results from running tsinfer on phased bi-allelic snps from samples originating from multiple populations. tsinfer generates a tree sequence along a reference genome, which by algorithmic design guarantees that every position in the genome has a complete genealogy. Every position can therefore be represented as a genealogical tree, albeit with polytomies.

The plots have been created using bokeh. To the right of each plot, there is a toolbar with tools to pan, box zoom, box select, wheel zoom, save graphics, reset graphic, help, and hover tool, each of which can be turned on and off as required. The hover tool provides the mouse pointer with extra information when hovering the plot. The included plots are described in the following sections.

GNN clustering plot

Z-score normalised GNN proportions. Following Kelleher et al 2019, Fig 4 columns have first been Z-score normalised, followed by hierarchical clustering on rows. For each genomic position, genealogical nearest neighbours are assigned by looking at a focal node (where each node is a haplotype sample). The sister nodes (i.e. nodes sharing same parent) are the genealogical nearest neighbours. By labelling the nodes with the corresponding population of these sister samples, we can calculate the "population" proportion for the focal node. The rows correspond to the focal population, which is the average of all individuals in a population, and the columns are the corresponding GNN proportion assigned to the focal population. Note therefore that the matrix need not necessarily be symmetric; a focal node may predominantly be found together with a certain composition of neighbours, but a neighbour may predominantly be found in a completly different environment.

Mean chromosome Fst plot

Mean chromosome Fst values calculated from tree sequences.

GNN proportions plot

Plot showing average GNN proportions for all **individuals**.

Choropleth world map plot

The world map plot shows a choropleth overlayed with sampling sites. The plot is linked to the mean GNN clustering plot such that selections in one plot translate to selections in the other.

Average over all chromosomes

Genealogical nearest neighbour haplotype plots

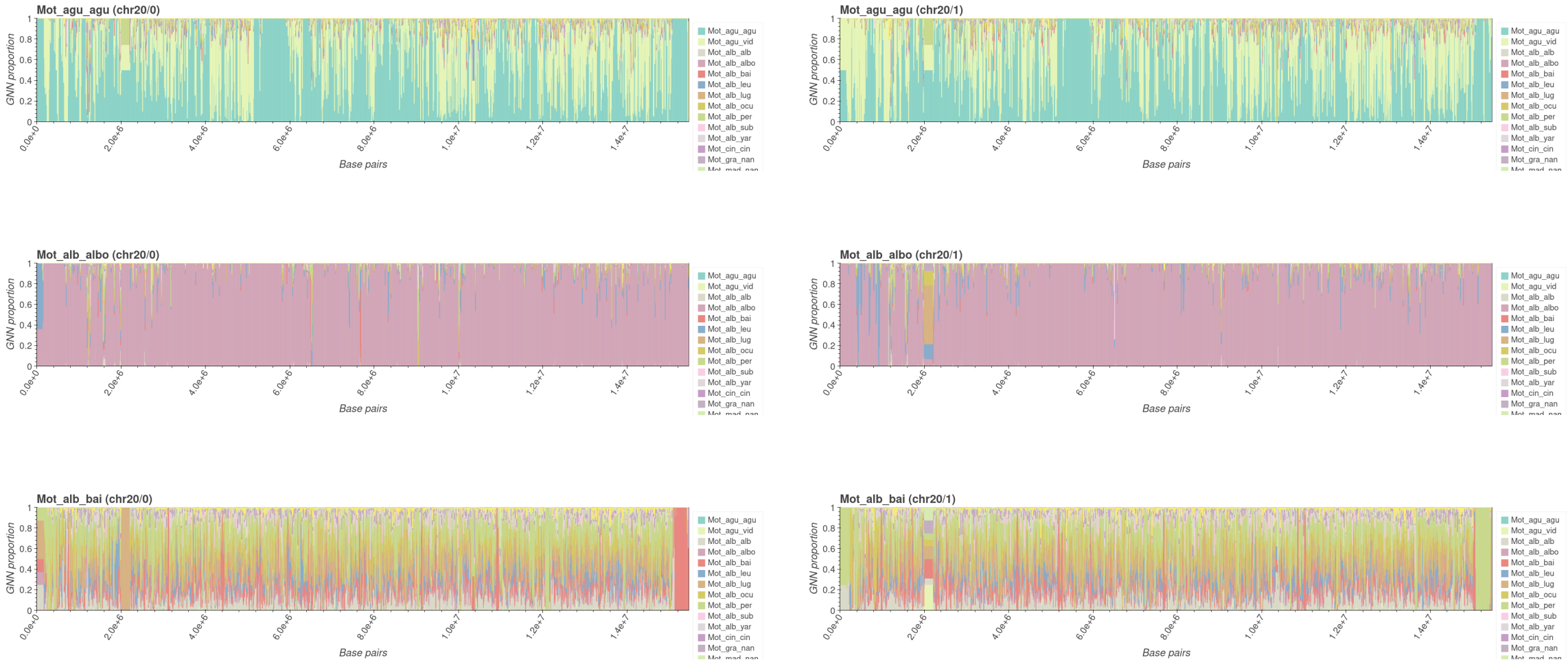


Figure 2: GNN proportions for chromosome 20 for selected individuals.

Acknowledgements

- Tom van der Valk
- Gabriel David

tskit development team

- Yan Wong
- Peter Ralph

Wagtail group

- Erik Enbody
- Leif Andersson
- Per Alström

DDLS Population Genomics in Practice

NBIS • Workshop

DDLS Population genomics in practice

Welcome to the DDLS Population Genomics in Practice homepage!

Important dates and information

See the canvas home page for the current course NBIS_POPGENIP_H23 for more information on how to apply.

Next round	06-Nov-2023 - 10-Nov-2023
Application opens	15-Aug-2023
Application deadline	30-Sep-2023

Bibliography

Alström, P., & Mild, K. (2003). *Pipits and Wagtails of Europe, Asia and North America*. A&C Black and Princeton University Press.

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., Zhu, S., Eldon, B., Ellerman, E. C., Galloway, J. G., Gladstein, A. L., Gorjanc, G., Guo, B., Jeffery, B., Kretzschumar, W. W., Lohse, K., Matschiner, M., Nelson, D., Pope, N. S., … Kelleher, J. (2022). Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, *220*(3), iyab229. https://doi.org/10.1093/genetics/iyab229

Hubisz, M., & Siepel, A. (2020). Inference of Ancestral Recombination Graphs Using ARGweaver. In J. Y. Dutheil (Ed.), *Statistical Population Genomics* (pp. 231–266). Springer US. https://doi.org/10.1007/978-1-0716-0199-0_10

Kahn, S. D. (2011). On the Future of Genomic Data. *Science*, *331*(6018), 728–729. https://doi.org/10.1126/science.1197891

Kelleher, J., Etheridge, A. M., & McVean, G. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Computational Biology*, *12*(5), e1004842. https://doi.org/10.1371/journal.pcbi.1004842

Miller, C. (2020). 5.12 Sexual Reproduction, Meiosis, and Gametogenesis. *Human Biology*. https://humanbiology.pressbooks.tru.ca/chapter/5-11-sexual-reproduction-meiosis-and-gametogenesis/

Poelstra, J. W., Vijay, N., Bossu, C. M., Lantz, H., Ryll, B., Müller, I., Baglione, V., Unneberg, P., Wikelski, M., Grabherr, M. G., & Wolf, J. B. W. (2014). The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science (New York, N.Y.)*, *344*(6190), 1410–1414. https://doi.org/10.1126/science.1253226

Ralph, P., Thornton, K., & Kelleher, J. (2020). Efficiently Summarizing Relationships in Large Samples: A General Duality Between Statistics of Genealogies and Genomes. *Genetics*, *215*(3), 779–797. https://doi.org/10.1534/genetics.120.303253

Wetterstrand, KA. (2022). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). In *Genome.gov*. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

Y. C. Brandt, D., Wei, X., Deng, Y., Vaughn, A. H., & Nielsen, R. (2022). Evaluation of methods for estimating coalescence times using ancestral recombination graphs. *Genetics*, iyac044. https://doi.org/10.1093/genetics/iyac044