Reconstructing the demographic history of populations

André E. R. Soares

Populations Genomics in Practice 2023

Annotated slides with additional notes

But why?

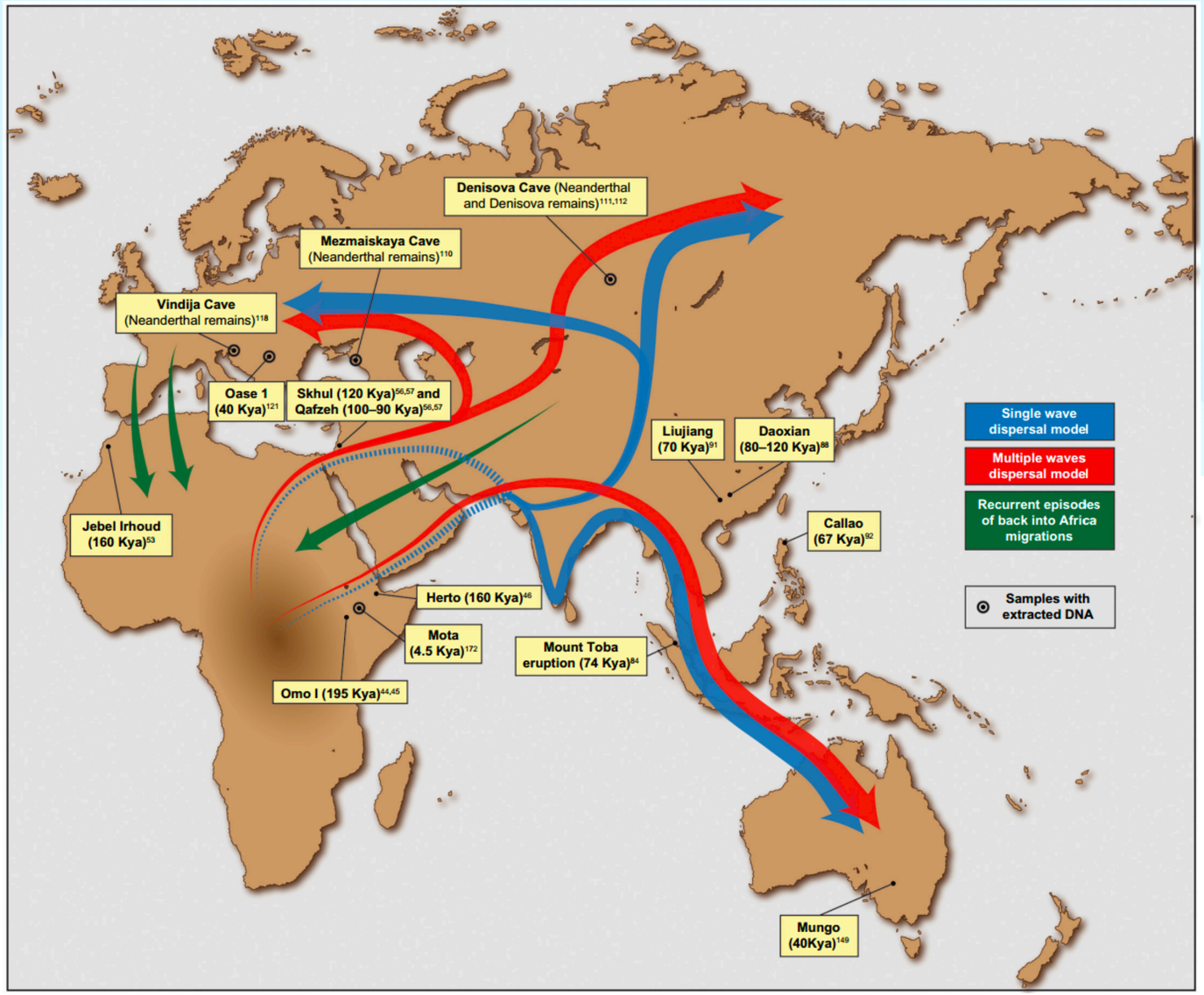
First things first

To understand the past

But why?

First things first

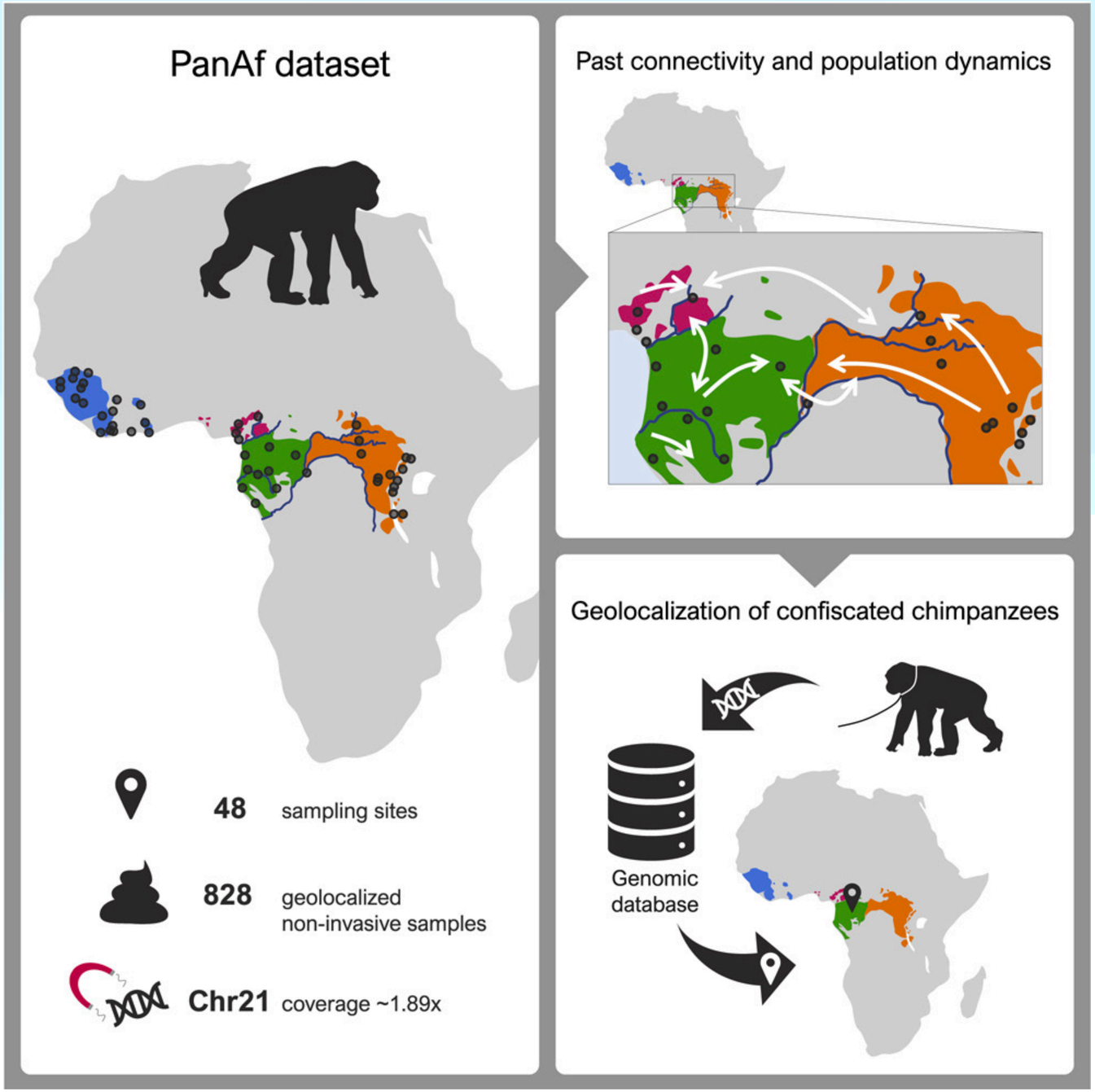
To understand where we came from.



But why?

First things first

To inform conservation projects and initiatives

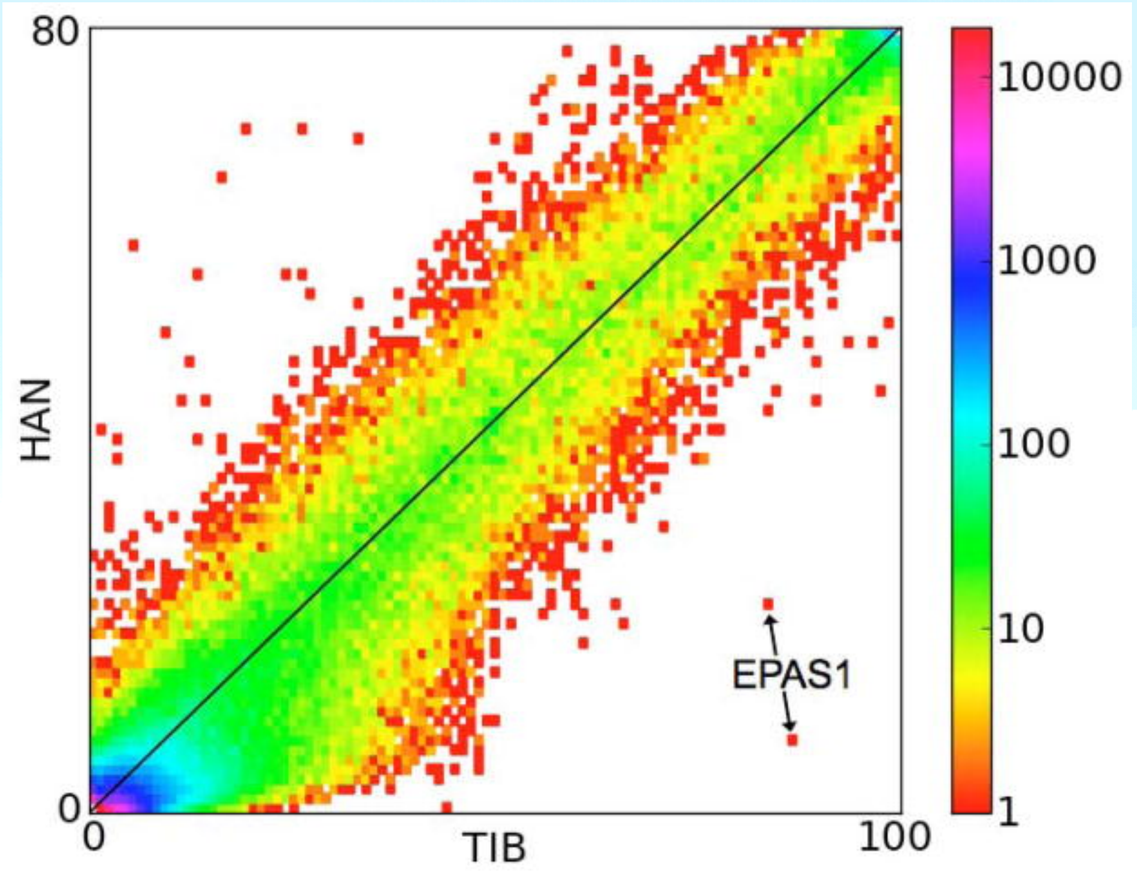


Fontsere *et al*. (2022) Cell Genomics

But why?

First things first

As a neutral background
for selection studies



Yu *et al*. (2010) Science

Population history

It can get messy

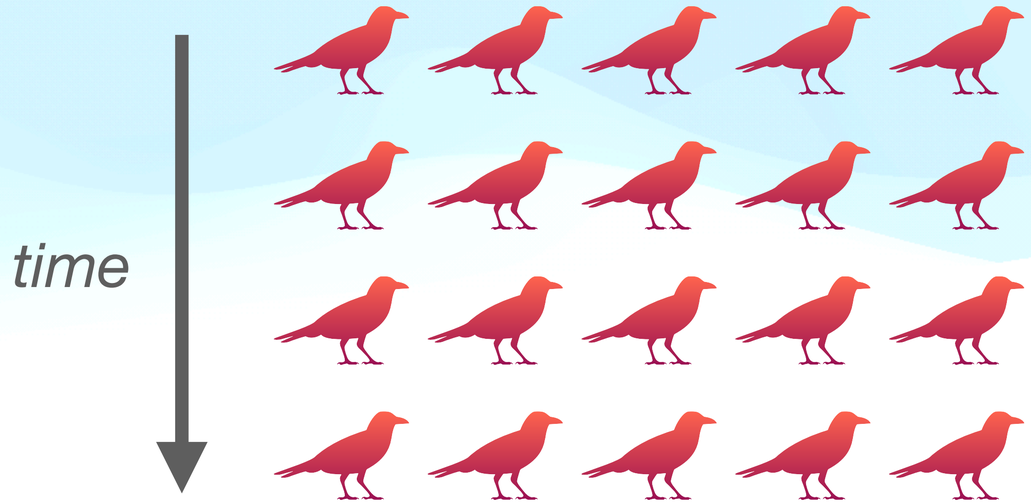
Demographic events:

- Population split

- Migration events

- Changes in effective population sizes

- Temporal changes in migration rates and effective sizes



Population history

It can get messy

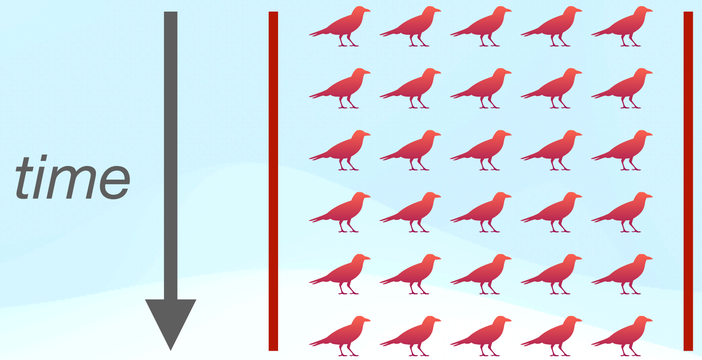
Demographic events:

- Population split

- Migration events

- Changes in effective population sizes

- Temporal changes in migration rates and effective sizes



Population history

It can get messy

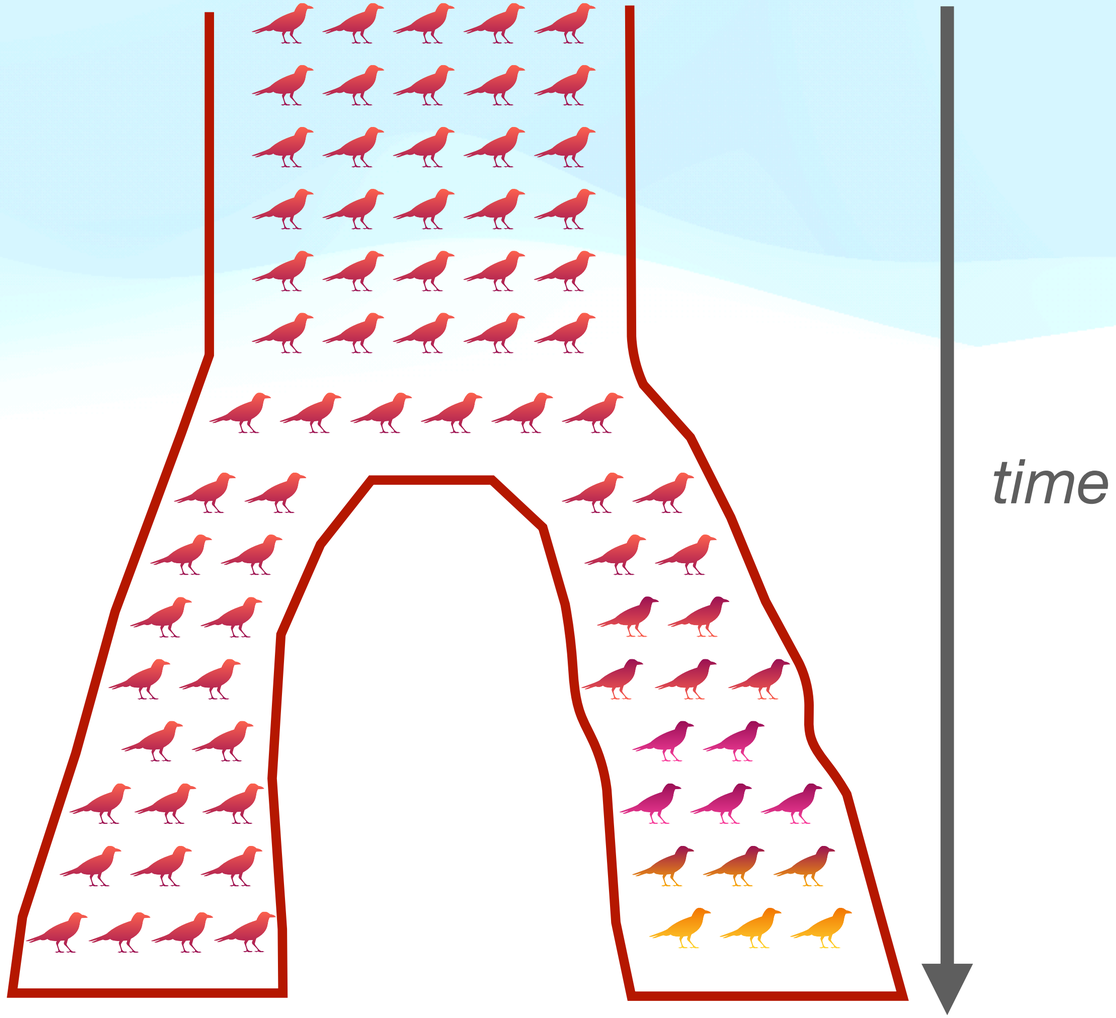
Demographic events:

- Population split

- Migration events

- Changes in effective population sizes

- Temporal changes in migration rates and effective sizes



Population history

It can get messy

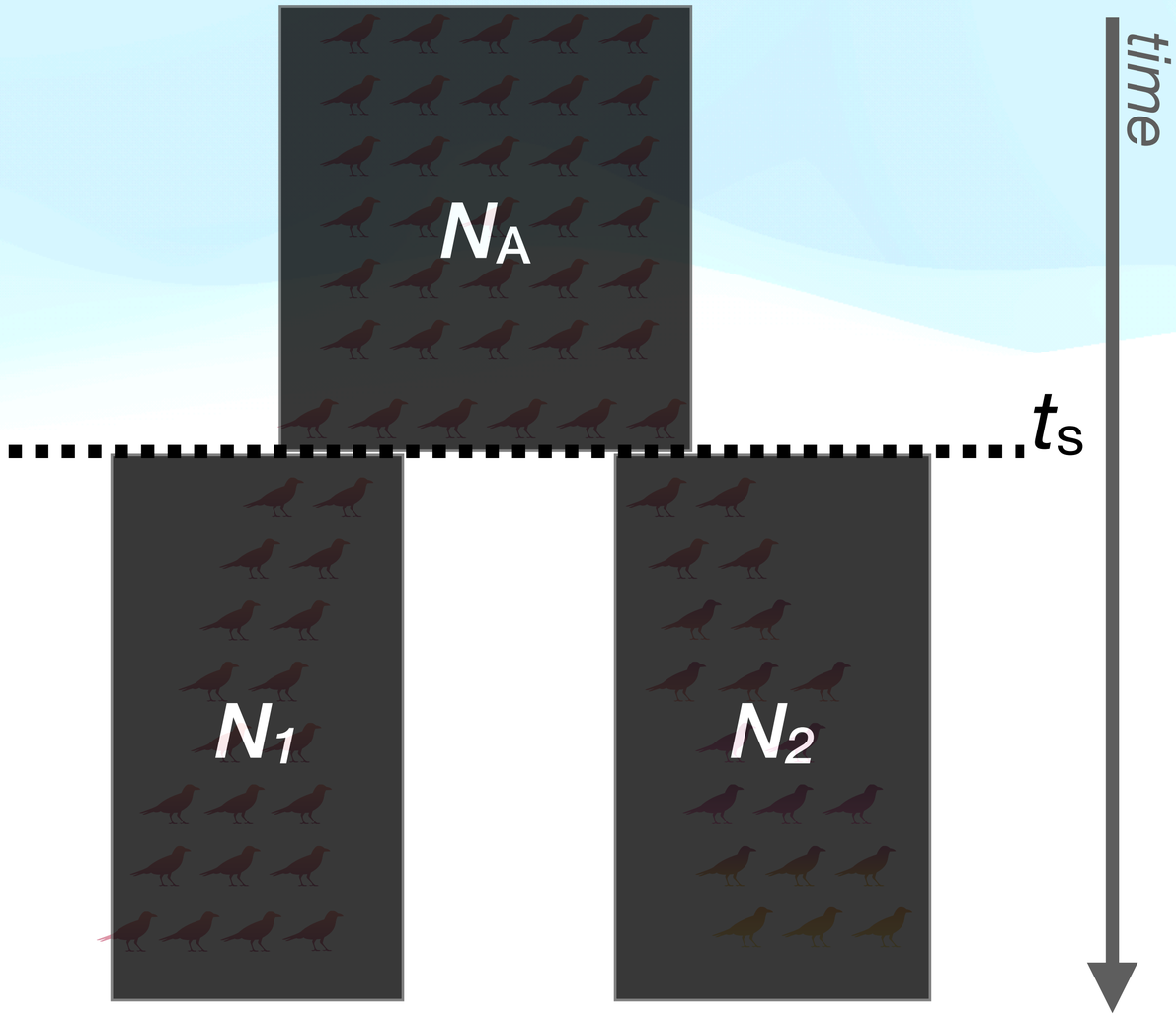
Demographic events:

- Population split

- Migration events

- Changes in effective population sizes

- Temporal changes in migration rates and effective sizes



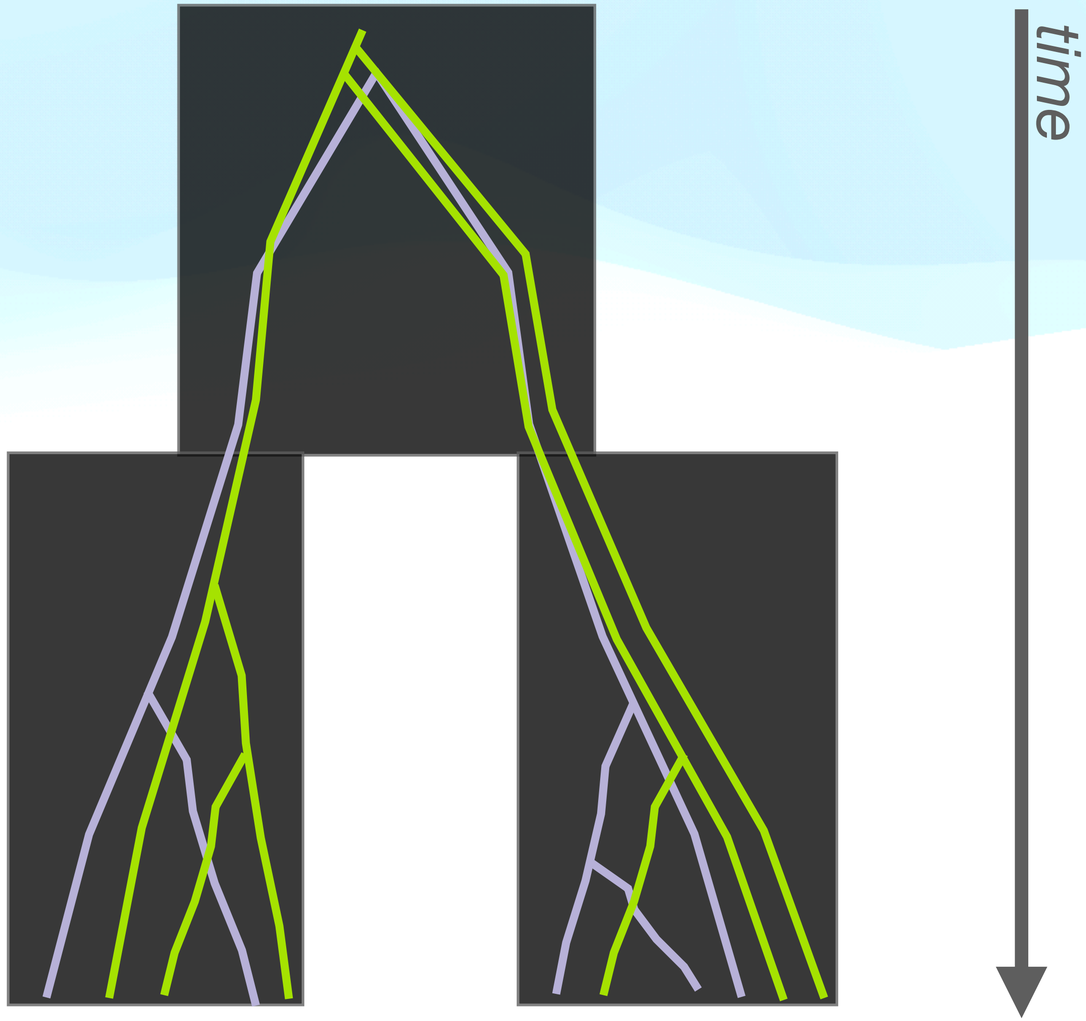
Genomes vs Demography

Demography wil affect the entire genome

- Recombination

- Natural selection acting on specific regions of the chromosome

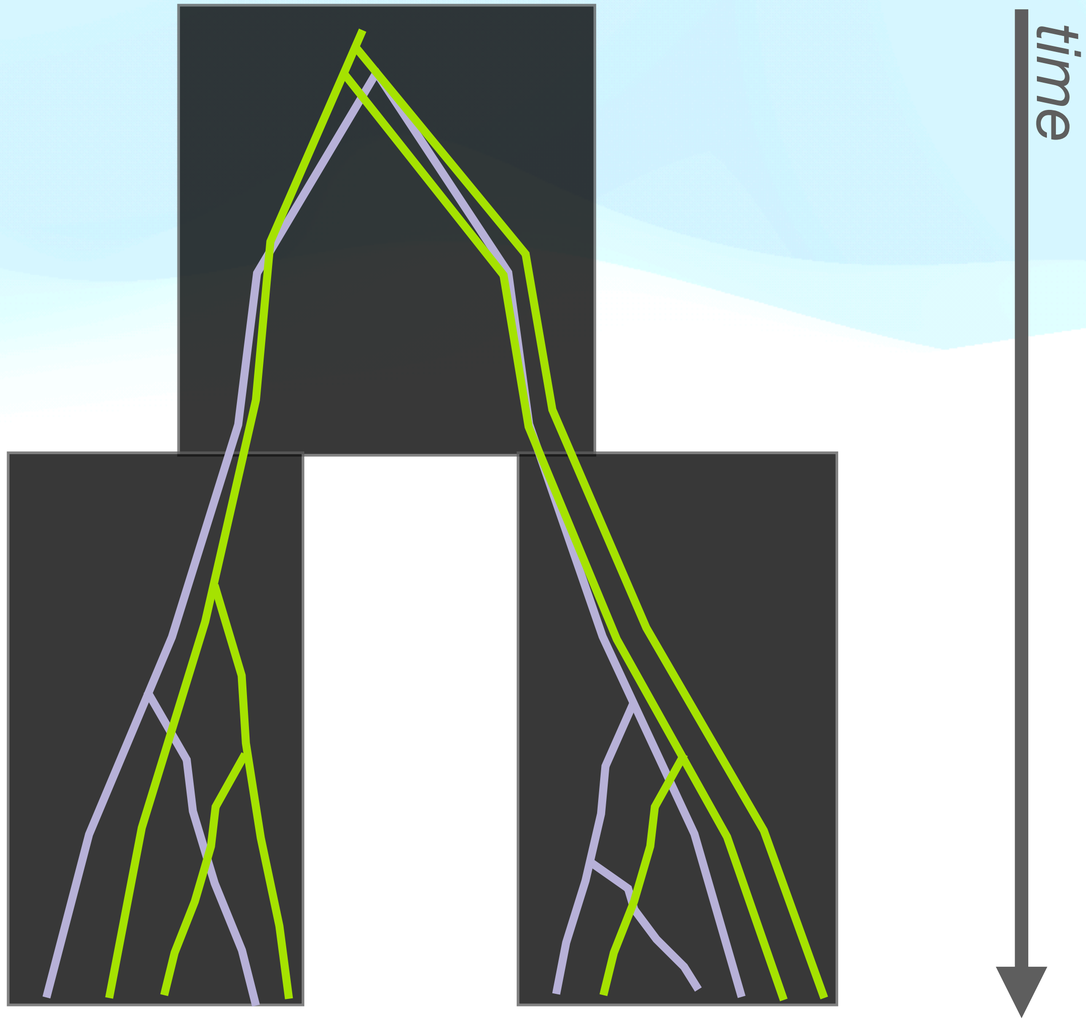
- The combination of all aspects will cause a difference between gene trees and population trees.



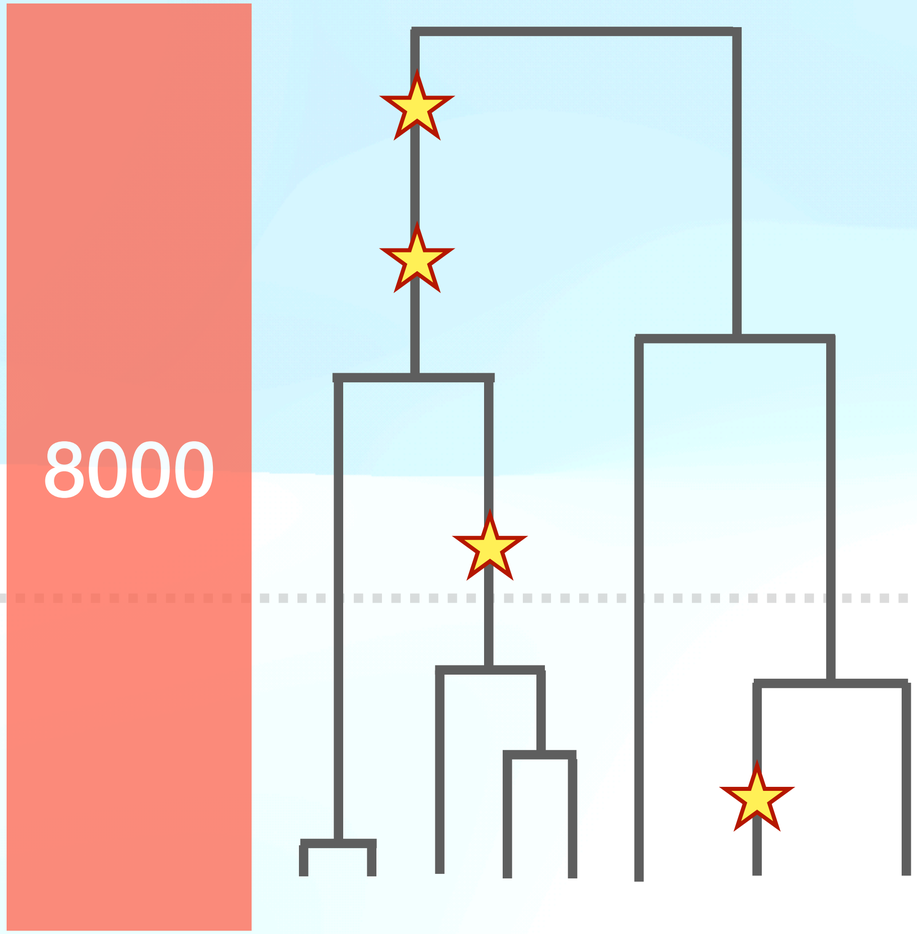
Neutral mutations

We assume all alleles have the same fitness. And the tree shape is determined by the demography of these populations, no mutations.

Mutations will accumulate as a Poisson process, so longer branches = more mutations.

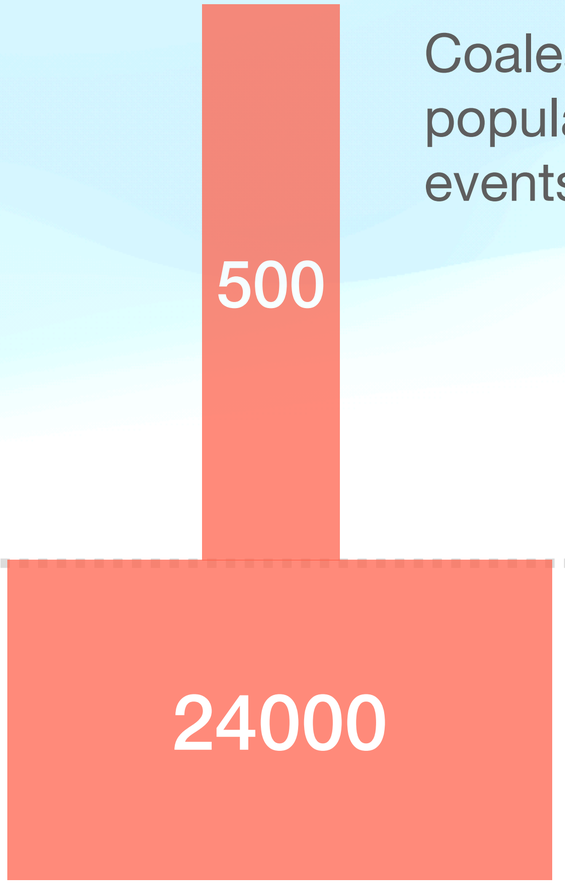


Gene trees *vs* growing populations

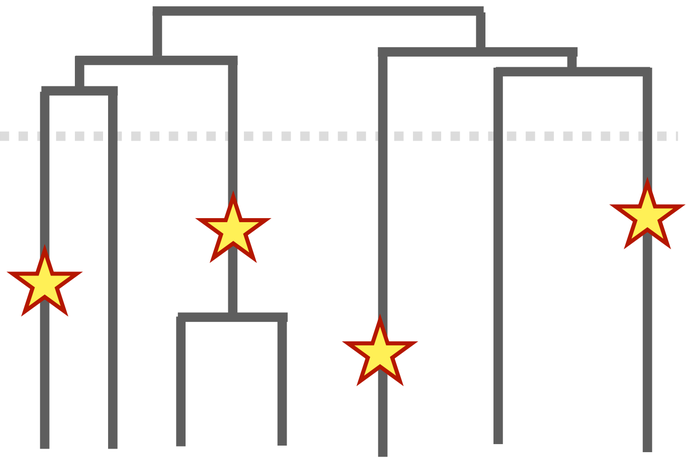


A mix of mutations shared by some lineages and singletons

time

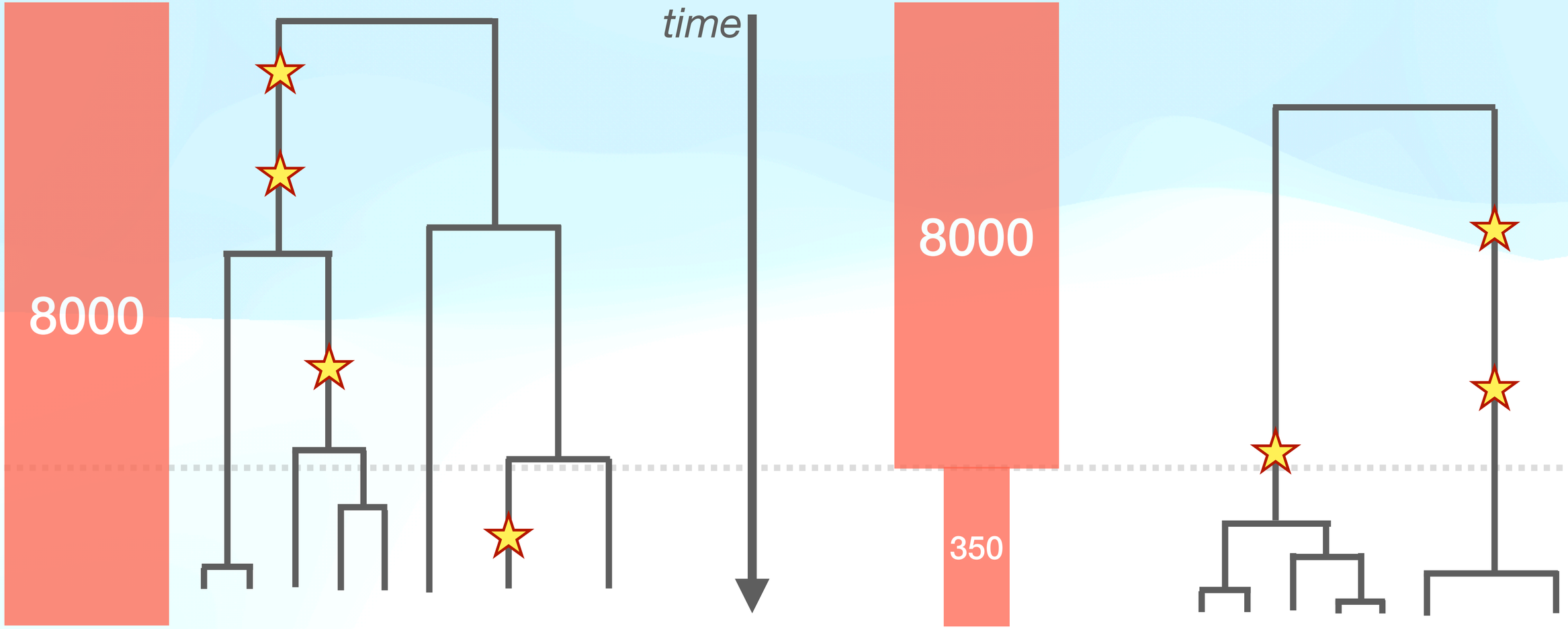


Coalescent rate is larger in smaller populations: smaller intervals, most events in ancestral population.



Most mutations are singletons, not shared between any lineages.

Gene trees *vs* bottlenecks



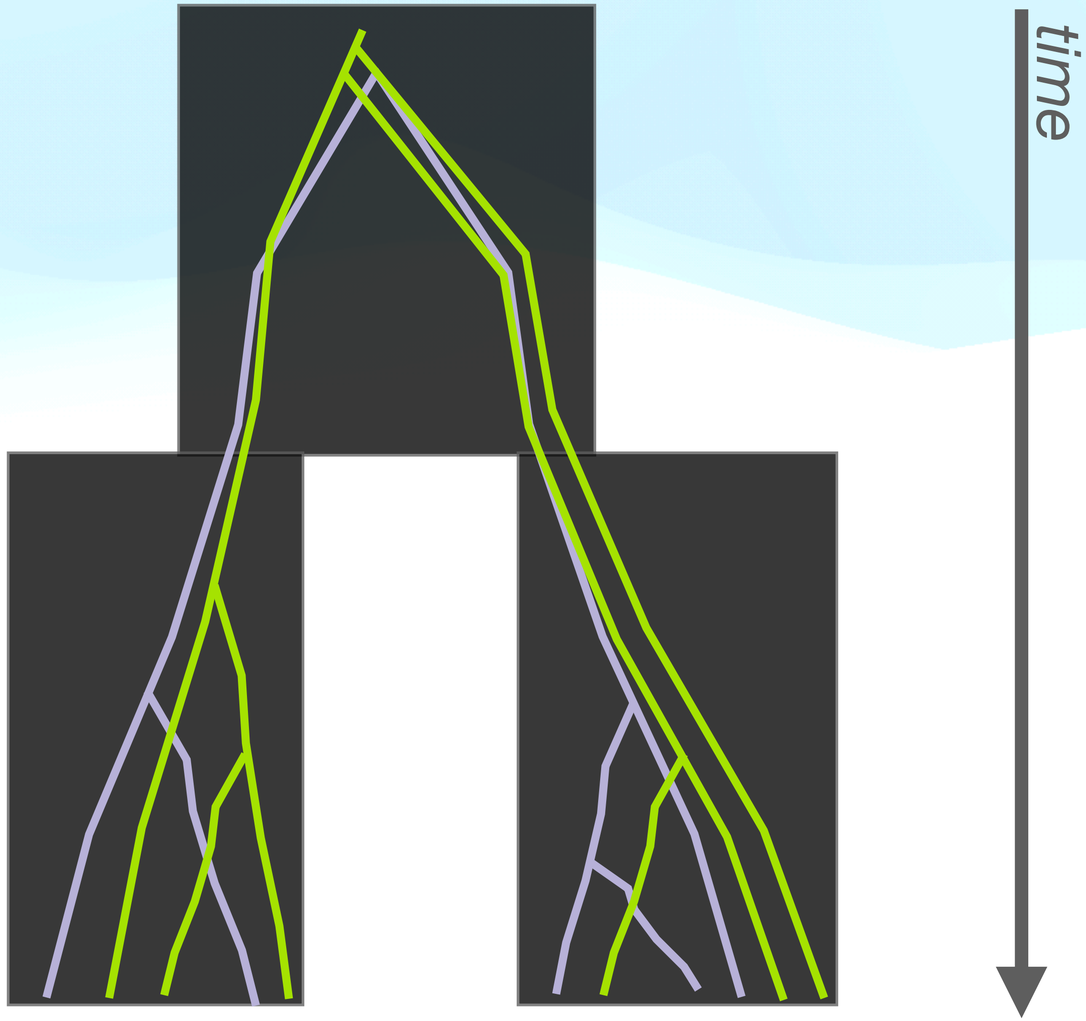
A mix of mutations shared by some lineages and singletons

Most lineages share the same mutations. Loss of diversity.

If only…

If we could observe all gene trees we could reconstruct the demographic history from them.

The next best thing is to observe mutations and allele frequencies.



Summarizing your genomic data

Observing allele frequencies

Booker *et al*. (2017) BMC Biology

Site frequency spectrum (SFS)

One way is to summarize your genomic data.

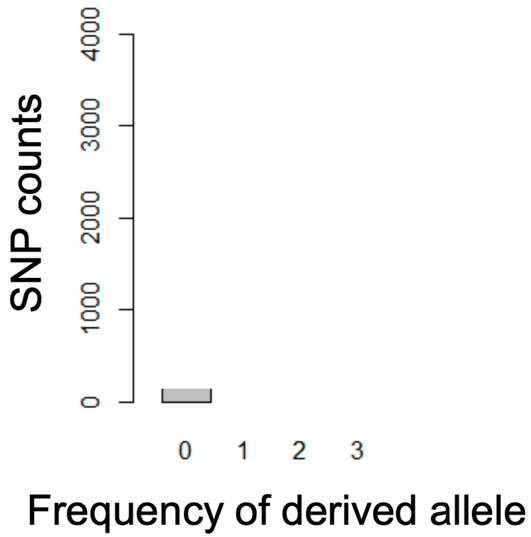
Data → **SFS**

Booker *et al*. (2017) BMC Biology

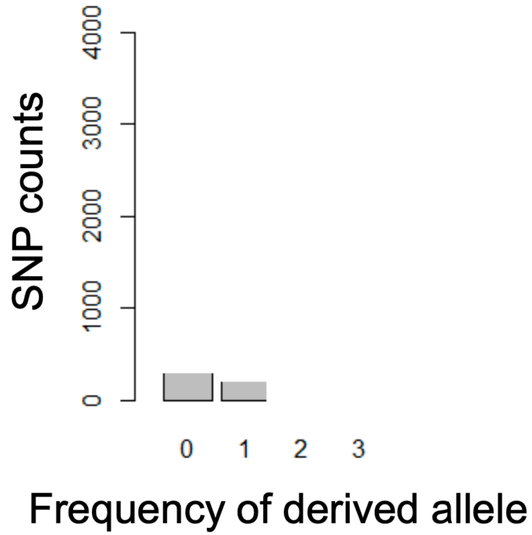
Site frequency spectrum (SFS)

One way is to summarize your genomic data.

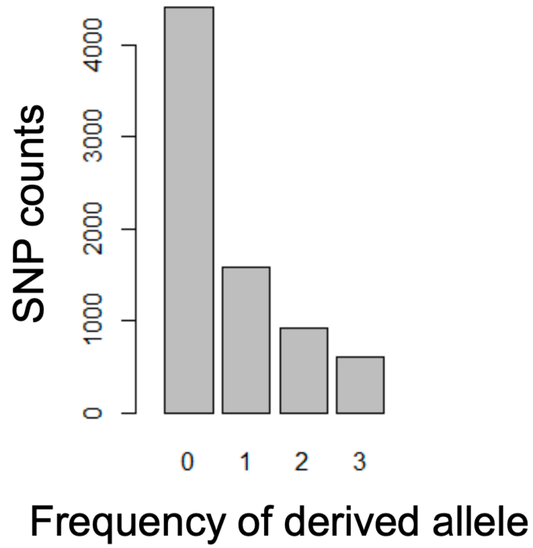
Outgroup ATACCG…
Individual 1 ATACCG…
Individual 2 ATTCGG…
Individual 3 ATACGG…



Outgroup ATACCG…
Individual 1 ATACCG…
Individual 2 ATTCGG…
Individual 3 ATACGG…



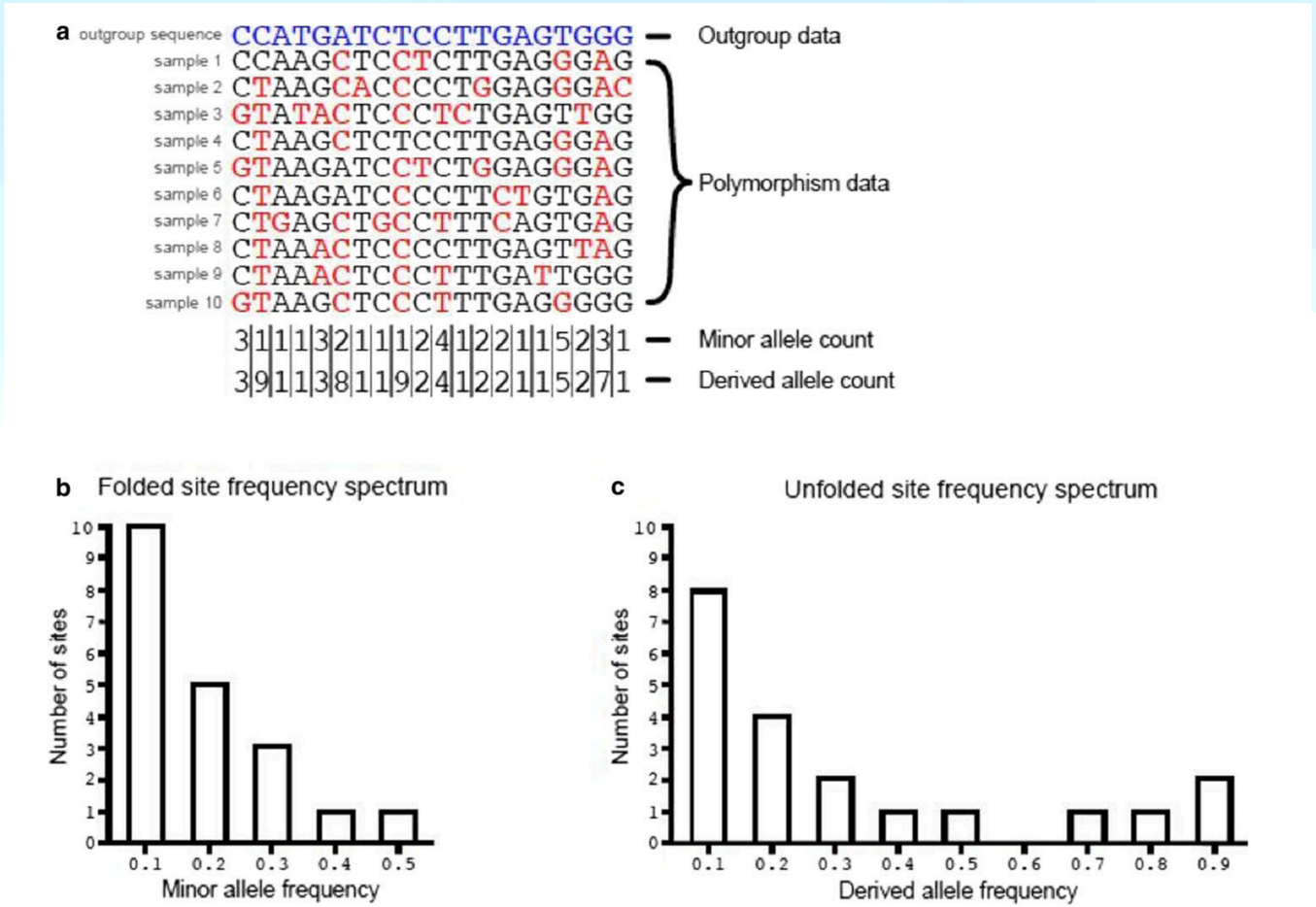
Outgroup ATACCG…
Individual 1 ATACCG…
Individual 2 ATTCGG..
Individual 3 ATACGG..



Site frequency spectrum (SFS)

Folded: We don’t have an outgroup, so we use the allele with higher frequency is treated as a reference.

Unfolded: We have an outgroup that helps us determine the ancestral state.



Additional suggested reading:
https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1010677

Booker *et al*. (2017) BMC Biology

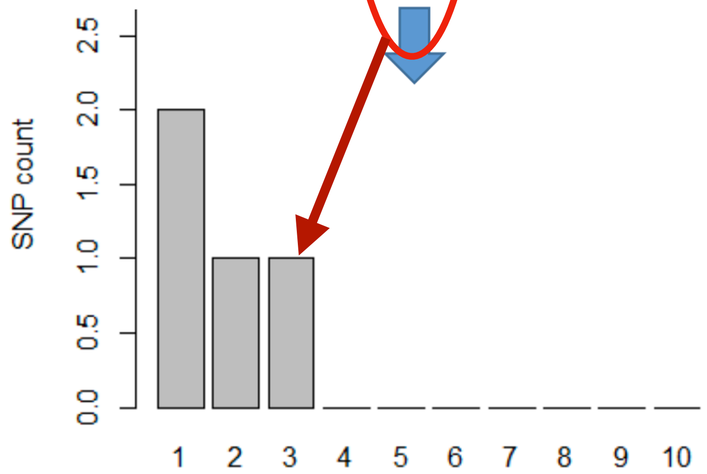
Site frequency spectrum (SFS)

It works with genotype call data

(You must have >10x coverage)

Low quality/low depth can inflate the number of singletons leadings to false inferences (like a false population expansion signal).

	SNP1	SNP2	SNP3	SNP4
Individual 1	0	2	0	1
Individual 2	0	0	1	0
Individual 3	1	0	0	0
Individual 4	0	1	0	0
Individual 5	0	0	1	0

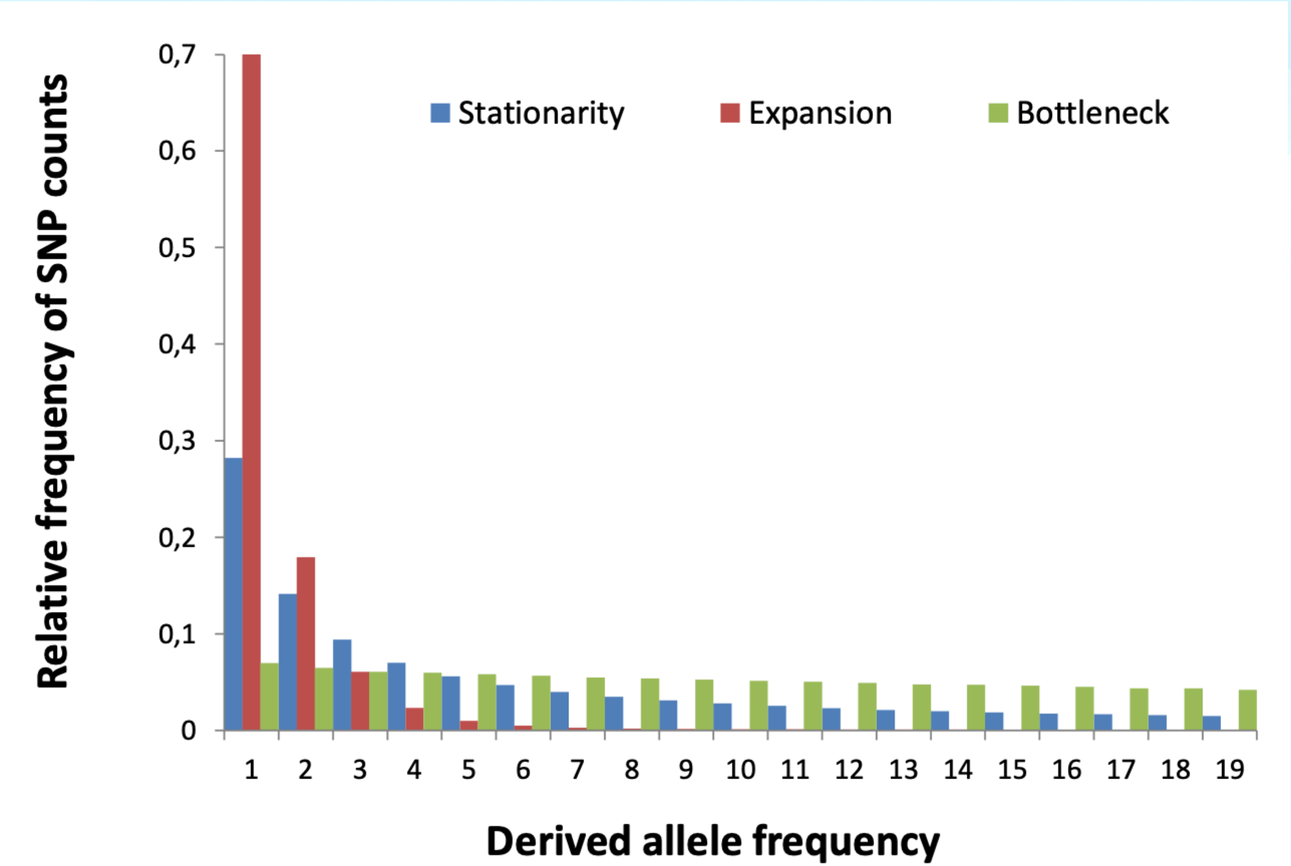


For the figure on the right. Think that 0 means homozygote for the reference allele, 1 is heterozygote and 2 is homozygote for the *alternative* allele. These are diploid individuals, so that's why you can have more categories than 5.

Booker *et al*. (2017) BMC Biology

Site frequency spectrum (SFS)

You can get certain initial insights from SFS…



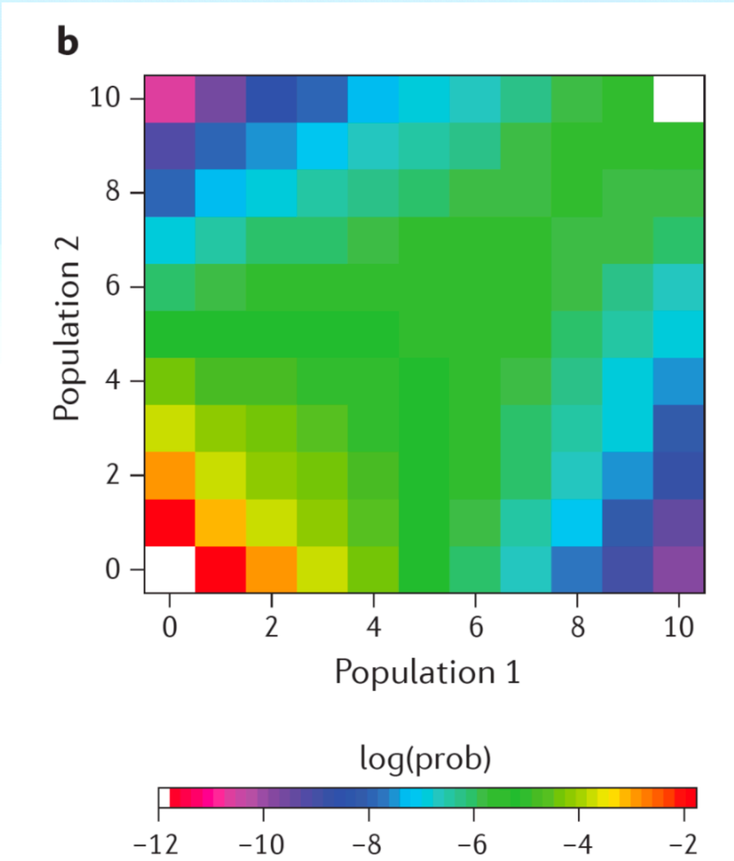
Booker *et al*. (2017) BMC Biology

Site frequency spectrum (SFS)

It works for 2 populations too…

But beware!

- It ignores linkage

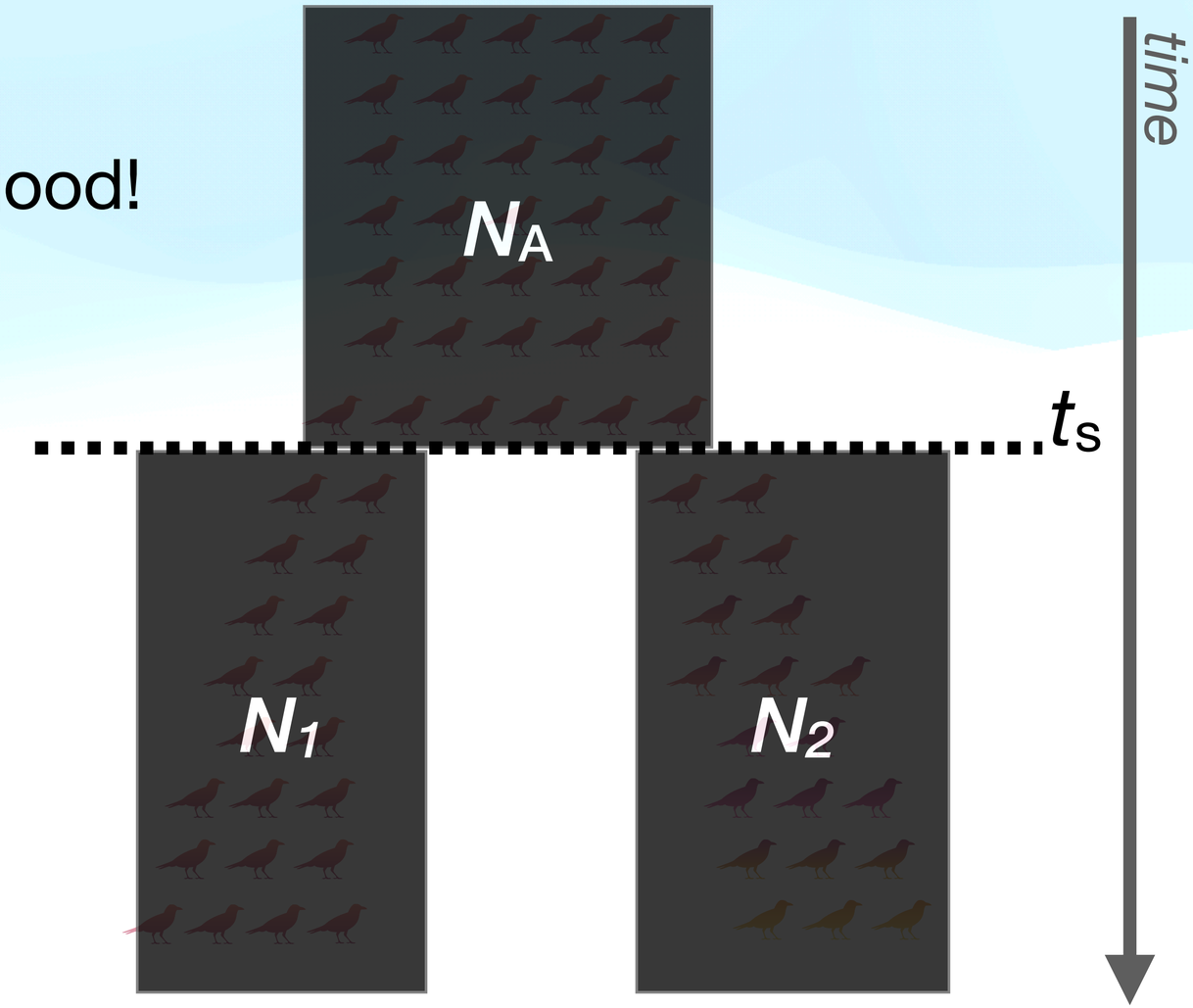


Booker *et al*. (2017) BMC Biology

What about that *modeling* thing?

We can also *model* the evolutionary history

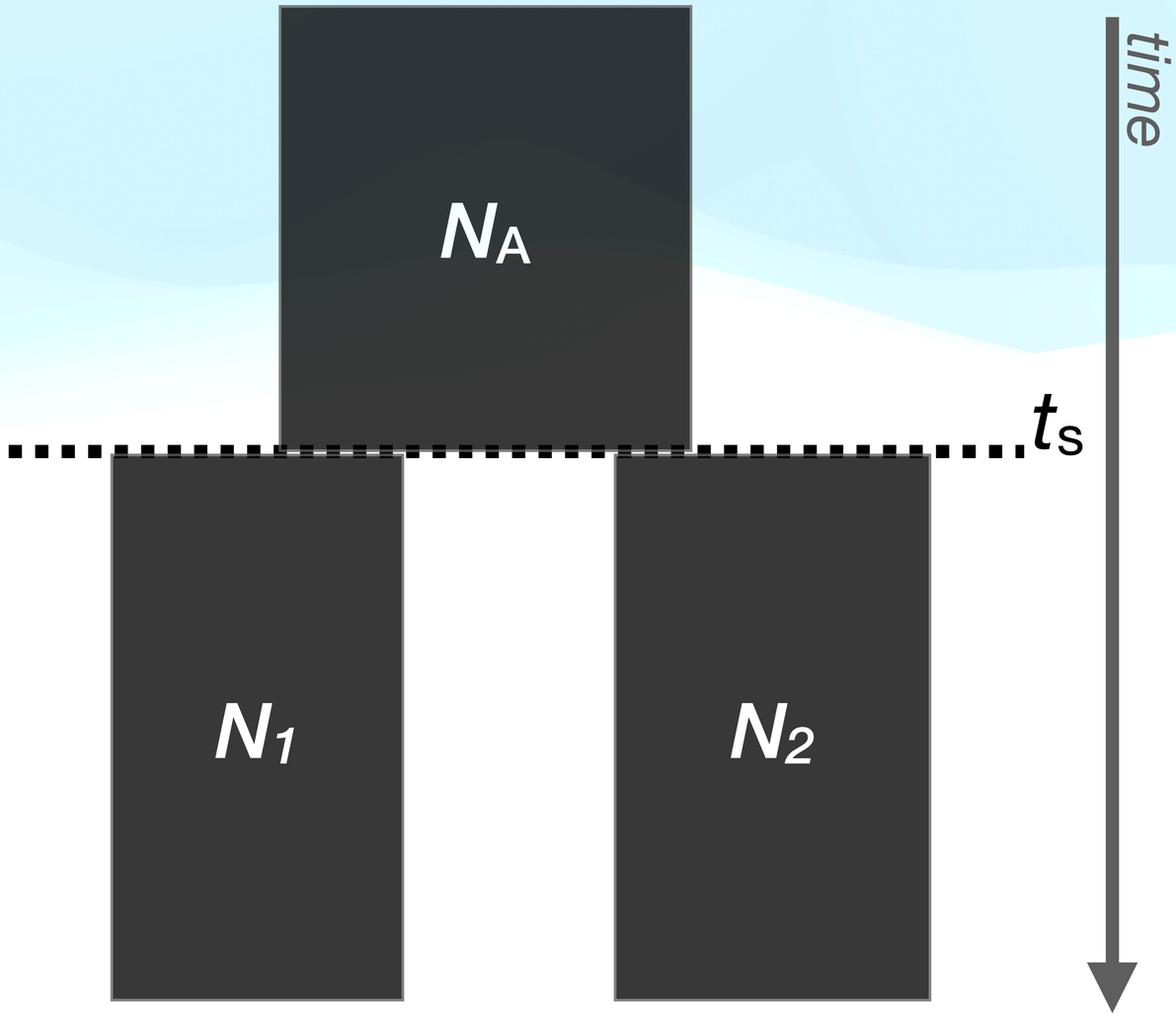
Remember it assumes your data is good!



We can also *model* the evolutionary history

You can incorporate inumerous parameters…

Because populations can have tricky histories.

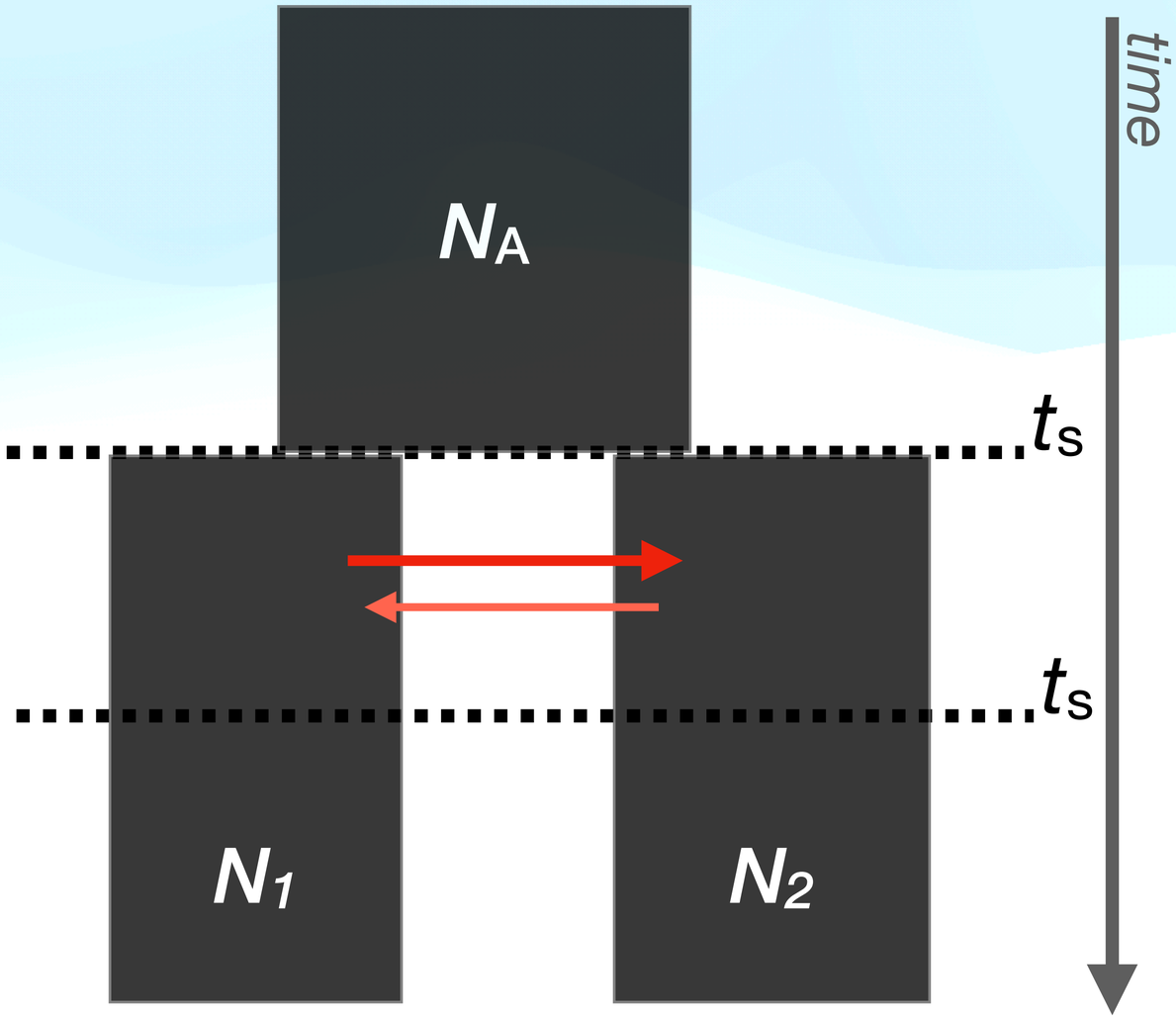


Isolation

We can also *model* the evolutionary history

You can incorporate inumerous parameters…

Because populations can have tricky histories.

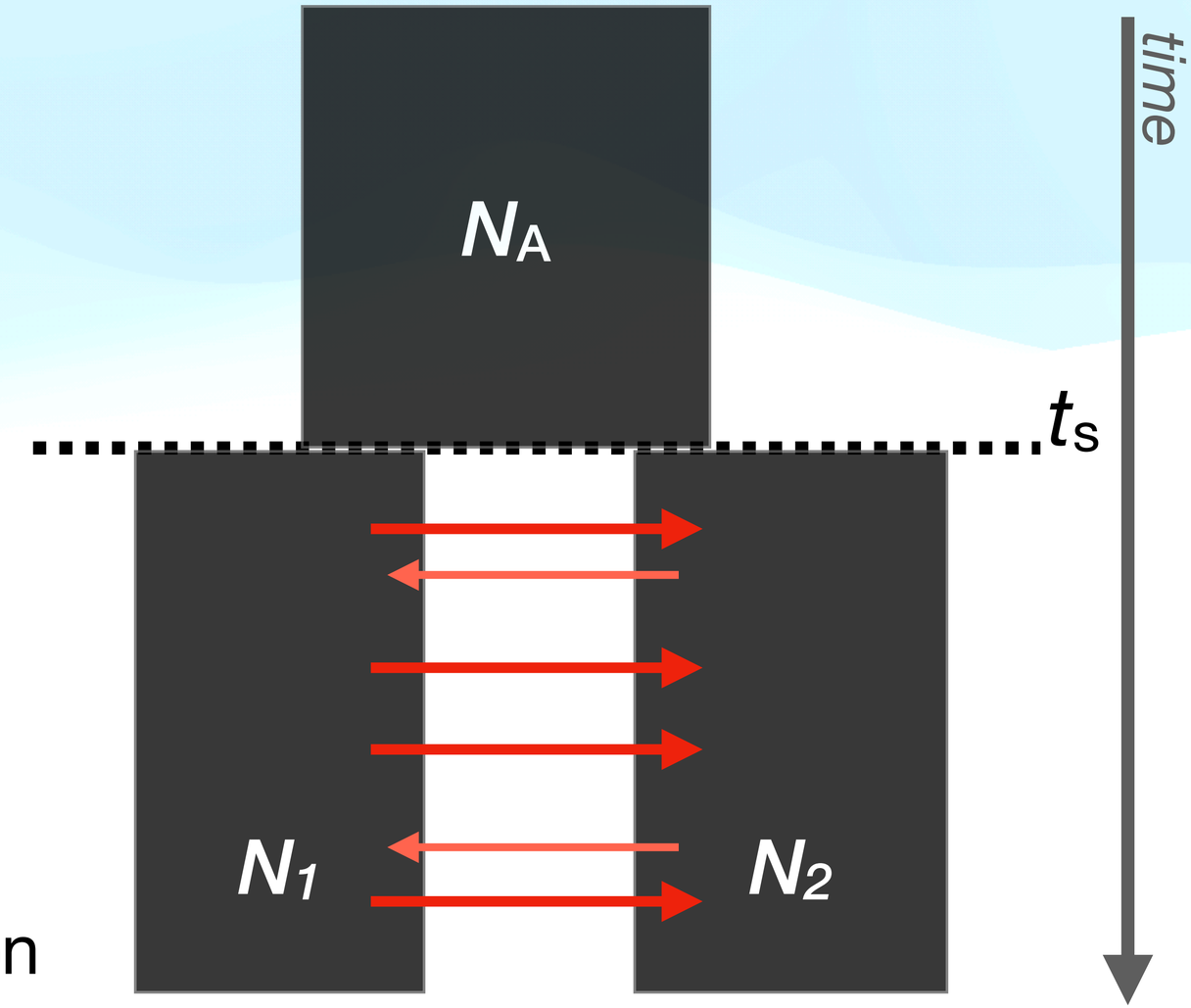


Isolation after migration

We can also *model* the evolutionary history

You can incorporate inumerous parameters…

Because populations can have tricky histories.

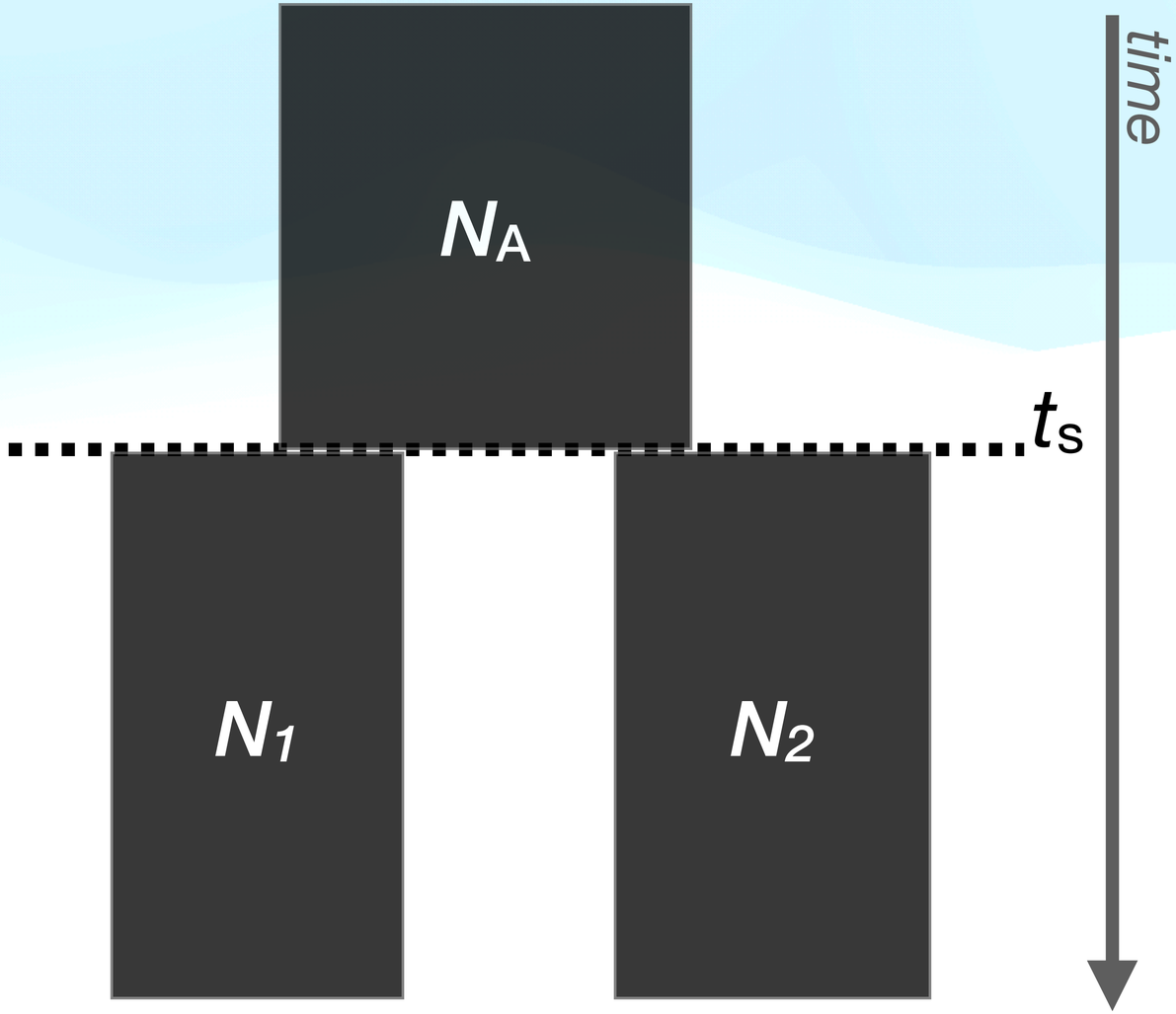


Isolation with migration

We can also *model* the evolutionary history

- *N*e (effective population size)
- The split time (t_s)
- Migration rates

- Selection
- Mutation rate
- Recombination rate



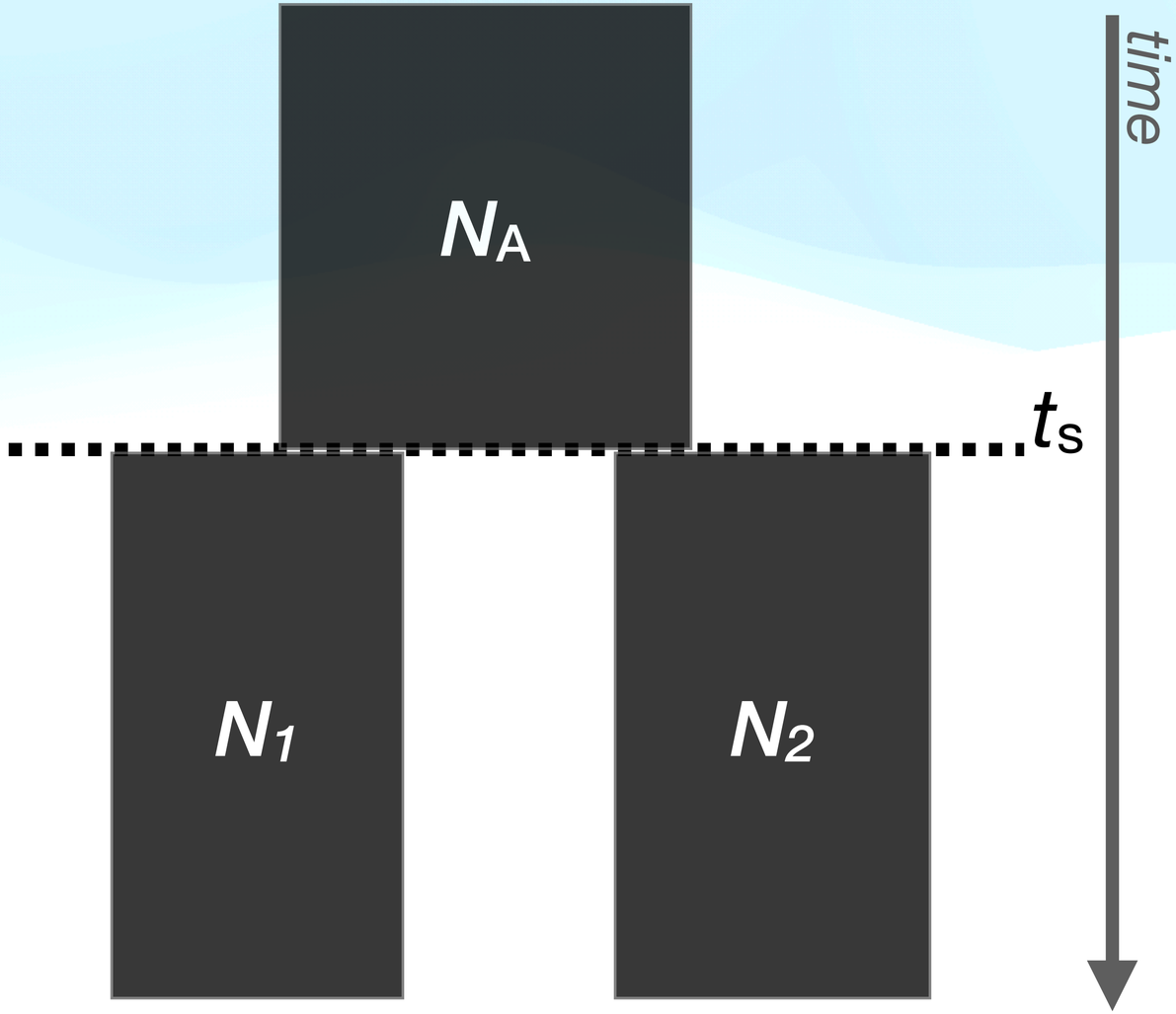
We can also *model* the evolutionary history

***Ne* (effective population size)**

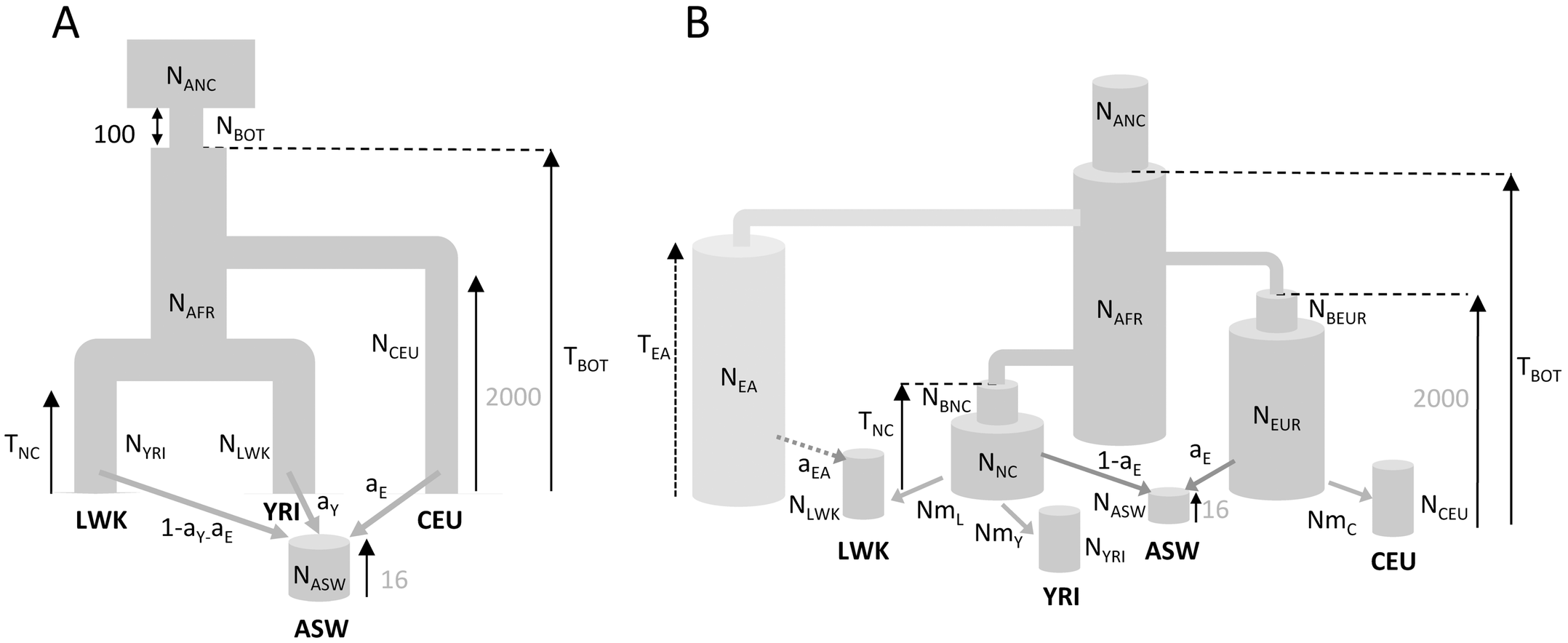
The size of the population that would give you the same behavior as the population of interest.

It's not the census size!

Sometimes more affected by selection than drift.

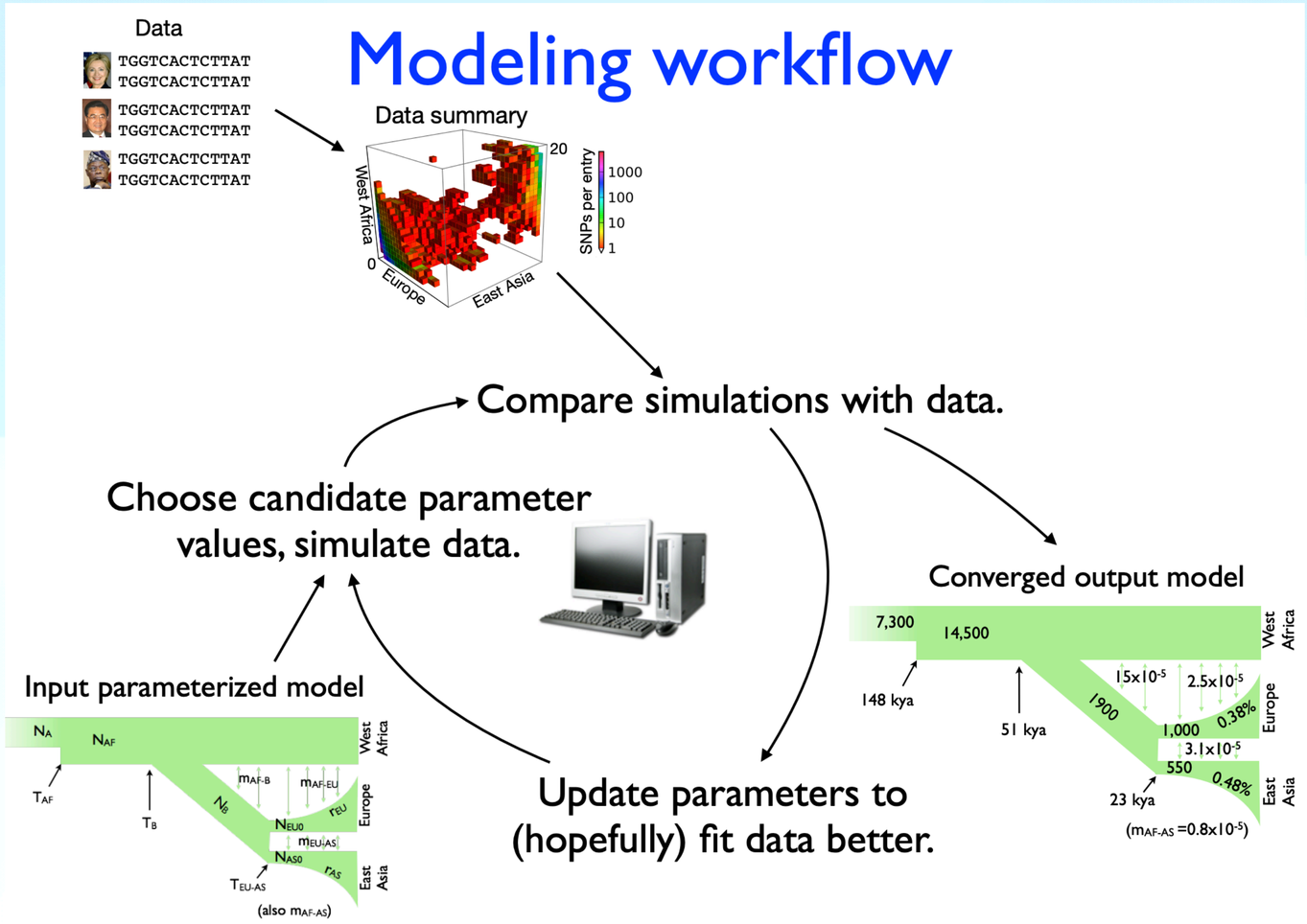


We can also *model* the evolutionary history



Excoffier *et al*. (2013) PloS Genetics

We can also *model* the evolutionary history



Many ways to simulate

Via Coalescent

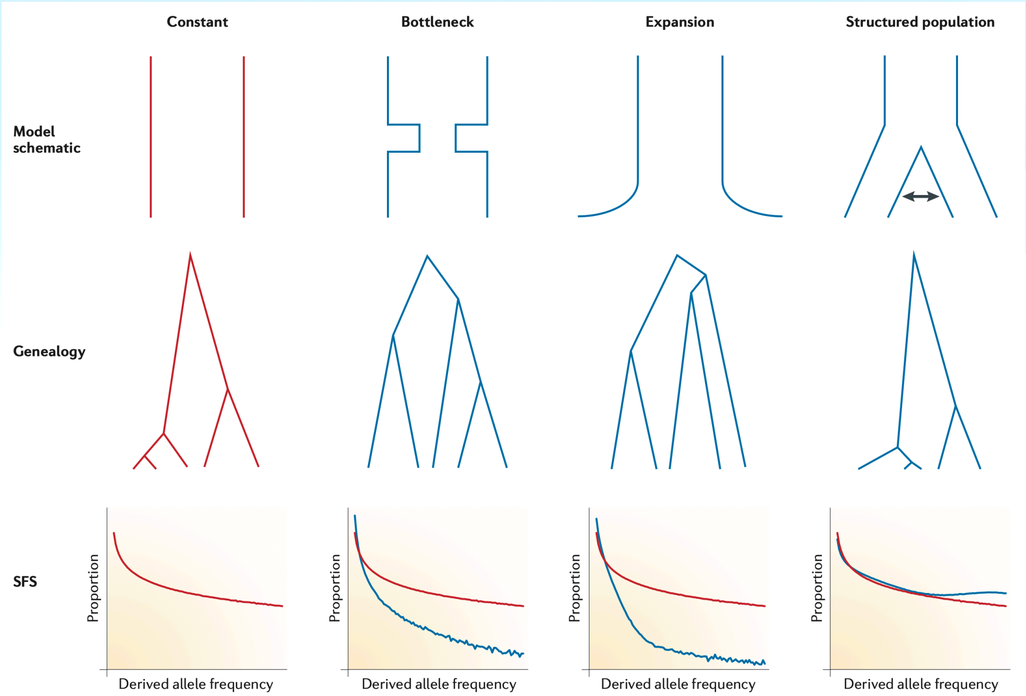
It aims to model the genealogy of sampled sequences;

The rate is proportional do 1/Ne;

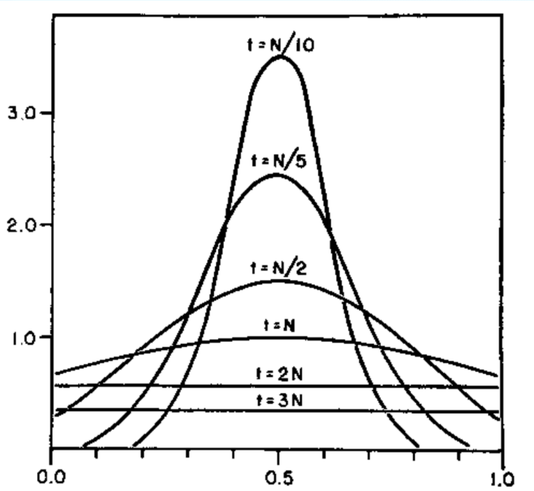
It can model recombination;

Mutations are added via a Poisson process;

Selection :(



Many ways to simulate



Kimura
(1964) *J Applied Prob*

Original slides says "Diffusion"

Forward-in-time

It aims model the distribution of allele frequencies in the population(s)

Simulation of selection is straightforward

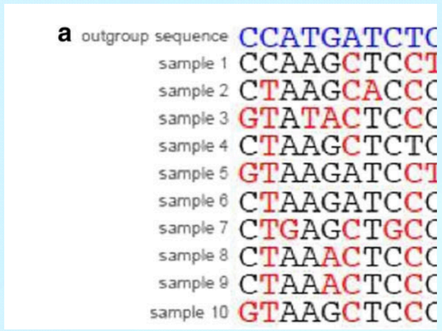
Linkage is very challenging

If you're curious about it, read the first paper on this:

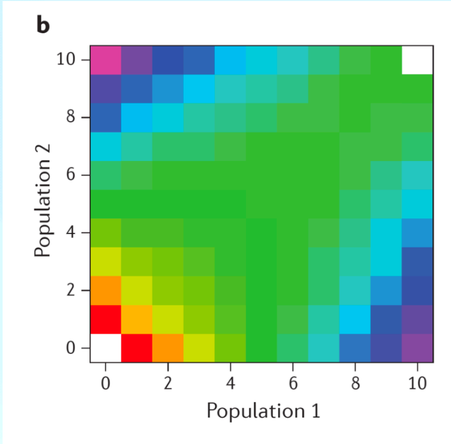
Diffusion models in population genetics by Motoo Kimura

https://www.well.ox.ac.uk/~gerton/Gulbenkian/kimura-diffusion.pdf

Inferring demographic history from SFS

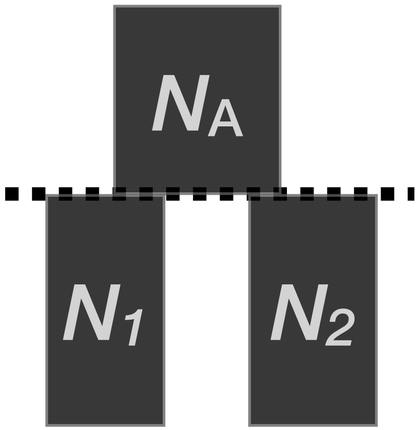


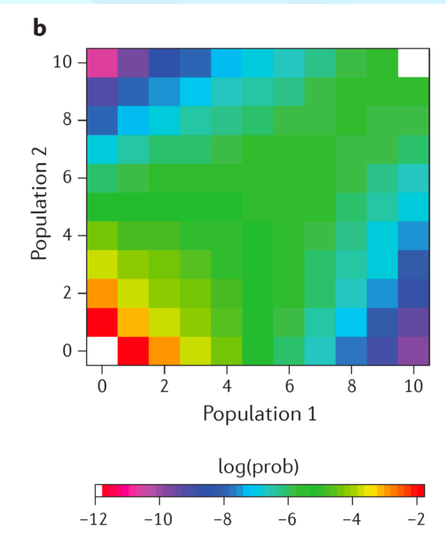
Genomic data



SFS (observed)

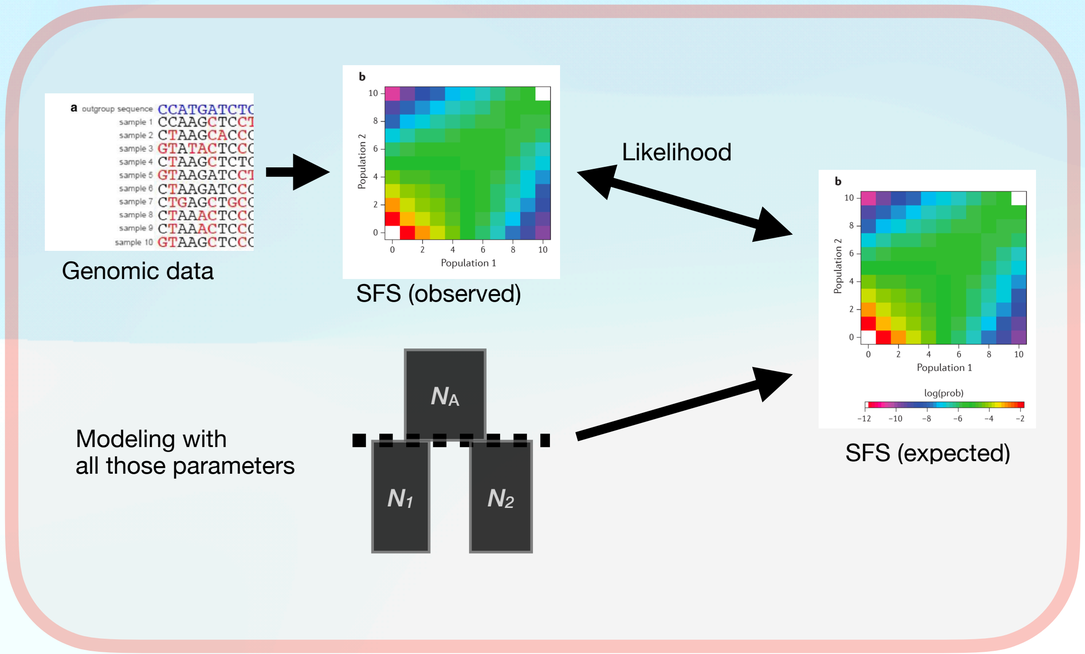
Modeling with
all those parameters





SFS (expected)

Ways to compare model and data



Frequentist

Likelihood

Bayesian

Ways to compare model and data

Likelihood

Probability of the data given the model

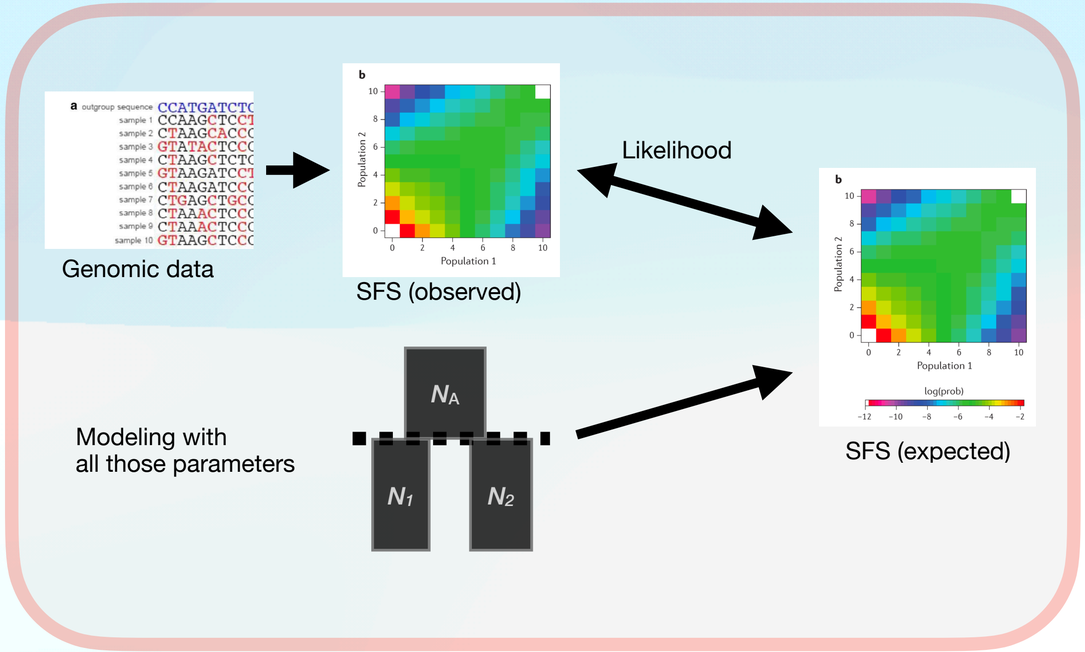
Frequentist

Maximize function to find best-fit parameters

Bayesian

Sample the posterior distribution of parameters based on a likelihood function

Inferring demographic history from SFS



Many programs do this:

(Coalescence)

Fastsimcoal2 (Excoffier et al. 2013)

Momi 1 and 2 (Kamm et al. 2015)

Rarecoal (Schiffels et al. 2016)

Original slides says "Diffusion"

(Forward-in-time)

ðaði (Gutenkunst et al. 2009)

Multipop (Lukic and Hey 2012)

Go to the last slide and get a more comprehensive list of programs for this!!

The table

It’s basically mandatory at this point

COMPUTATIONAL TOOLS

Methods and models for unravelling human evolutionary history

Joshua G. Schraiber and Joshua M. Akey

https://www.nature.com/articles/nrg4005

Nature Reviews Genetics, 2015; doi:10.1038/nrg4005

Table 1 | **Software for demographic inferences**

Name	Data type	Inference	Notes	Refs
STRUCTURE	Unlinked multi-allelic genotypes	Population structure, admixture	User-friendly GUI; can be computationally demanding	32
FRAPPE	Unlinked bi-allelic SNVs	Population structure, admixture	Alexander *et al.*[41] argue that convergence is not guaranteed	40
ADMIXTURE	Unlinked bi-allelic SNVs	Population structure, admixture	Estimates the number of populations via cross-validation error	41
fastSTRUCTURE	Unlinked bi-allelic SNVs	Population structure, admixture	Obtains variational Bayesian estimates of posterior probability distribution	42
Structurama	Unlinked multi-allelic genotypes	Population structure, admixture	Uses a Dirichlet process to estimate the number of populations	43
HAPMIX	Phased haplotypes; reference panel	Chromosome painting	Requires populations to be specified a priori	48
fineSTRUCTURE	Phased haplotypes	Population structure, admixture, chromosome painting	Can be used to identify the number and identity of populations	49
GLOBETROTTER	Phased haplotypes	Population structure, admixture, chromosome painting	Extends the fineSTRUCTURE approach to estimate unsampled ancestral populations and admixture times	7
LAMP	Phased haplotypes; reference panel	Chromosome painting	Identifies local ancestry in windows, rather than using an HMM, so is more discrete than other approaches	52
PCAdmix	Phased haplotypes	Chromosome painting, population structure	Uses PCA in small chunks followed by an HMM to estimate local ancestry	53
dadi	Frequency spectrum of unlinked bi-allelic SNVs	Demographic history	Requires some Python-coding skills; applicable to up to three populations	60
Fastsimcoal2	Frequency spectrum of unlinked bi-allelic SNVs	Demographic history	Can also be used to simulate data under the SMC	62,63
Treemix	Frequencies of unlinked bi-allelic SNVs	Admixture graph	Highly multimodal likelihood surface and heuristic search; redo inference from many starting points	64
fastNeutrino	Frequency spectrum of unlinked bi-allelic SNVs	Demographic history	Applicable only to a single population; designed specifically for extremely large sample sizes	65
DoRIS	Lengths of IBD blocks between pairs of individuals	Demographic history	IBD must be inferred (for example, using Beagle or GERMLINE); specification of lower cut-off minimizes false-negative IBD tracts	71,72
IBS tract inference	Lengths of IBS blocks between pairs of individuals	Demographic	IBS can easily be confounded by missing data and/or sequencing errors	76
PSMC	Diploid genotypes from one individual	Demographic history	Best used in MSMC’s PSMC mode, which uses the SMC to more accurately model recombination than the original PSMC; applicable to a single population	78
MSMC	Whole genome, phased haplotypes	Demographic history	Requires large amounts of RAM; cross-coalescence rate should not be interpreted as migration rate	82
CoalHMM	Whole genome, phased haplotypes	Demographic history	Multiple applications, including inference of population sizes, migration rates and incomplete lineage sorting	83–87
diCal	Medium-length, phased haplotypes	Demographic history	Uses shorter sequences than MSMC, but can be applied to multiple individuals in complex demographic models; infers explicit population genetic parameters for migration rates	89,92
LAMARC	Short, phased haplotypes	Demographic history	Requires Monte Carlo sampling of coalescent genealogies; very flexible	93
BEAST	Short, phased haplotypes	Species trees, effective population sizes	Used mainly as a method of phylogenetic inference. Can also infer population size history	94
MCMCcoal	Short, phased haplotypes	Divergence times between populations	Now incorporated into the software BPP[131]	95
G-PhoCS	Short, (un)phased haplotypes	Demographic history	Incorporates migration into the MCMCcoal framework. Averages over unphased haplotypes	96
Exact likelihoods using generating functions	Short, phased haplotypes	Demographic history	Implemented in Mathematica; applicable only to specific classes of multi-population models	97,98

BEAST, Bayesian evolutionary analysis by sampling trees; BPP, Bayesian phylogenetics and phylogeography; CoalHMM, coalescent HMM; *dadi*, diffusion approximations for demographic inference; diCal, demographic inference using composite approximate likelihood; DoRIS, demographic reconstruction via IBD sharing; G-PhoCS, generalized phylogenetic coalescent sampler; GERMLINE, genetic error-tolerant regional matching with linear-time extension; GUI, graphical user interface; HMM, hidden Markov model; IBD, identity by descent; IBS, identity by state; LAMARC, likelihood analysis with metropolis algorithm using random coalescence; LAMP, local ancestry in admixed populations; MCMC, Markov chain Monte Carlo; MSMC, multiple SMC; PCA, principal components analysis; PSMC, pairwise SMC; RAM, random access memory; SMC, sequentially Markov coalescent; SNVs, single nucleotide variants.