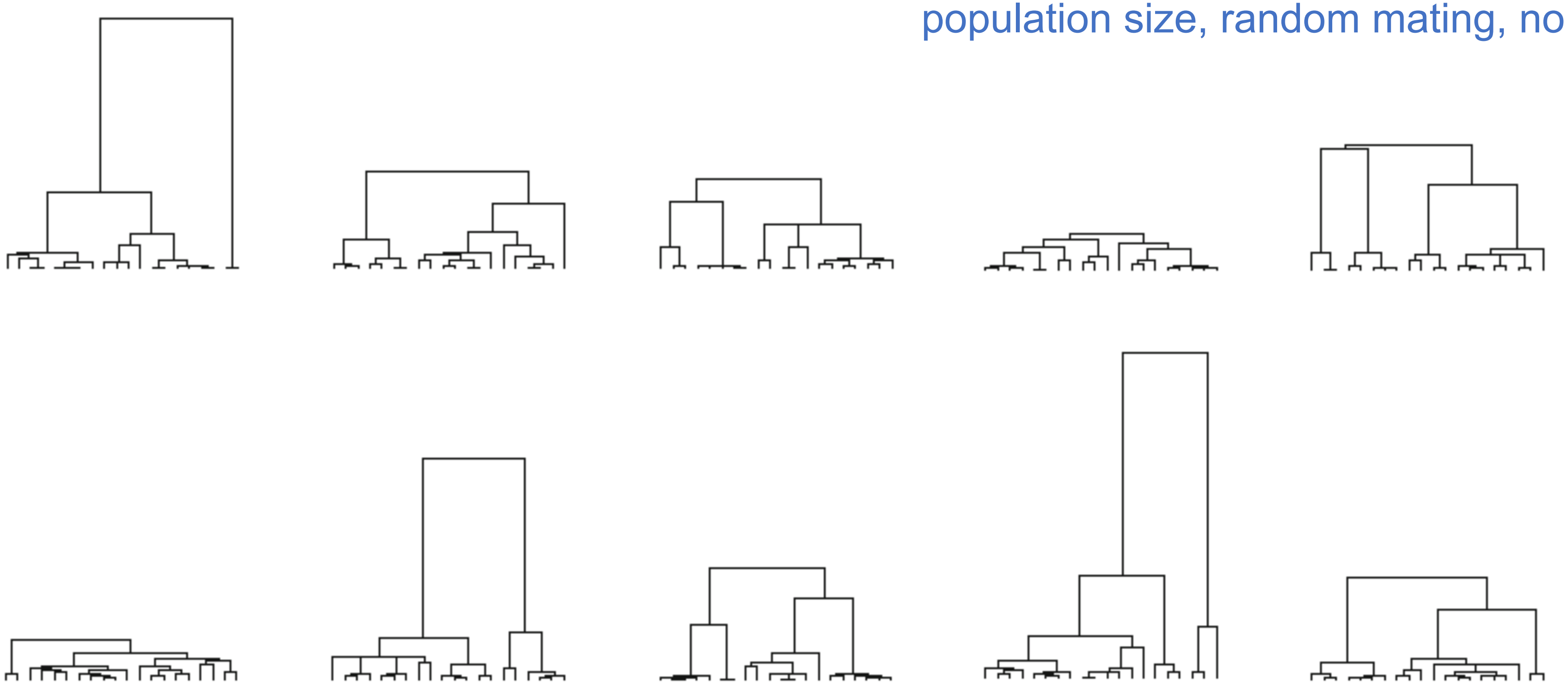
The Site Frequency Spectrum

Hancock

February 8, 2024

Ten independently generated gene genealogies (2N=20)



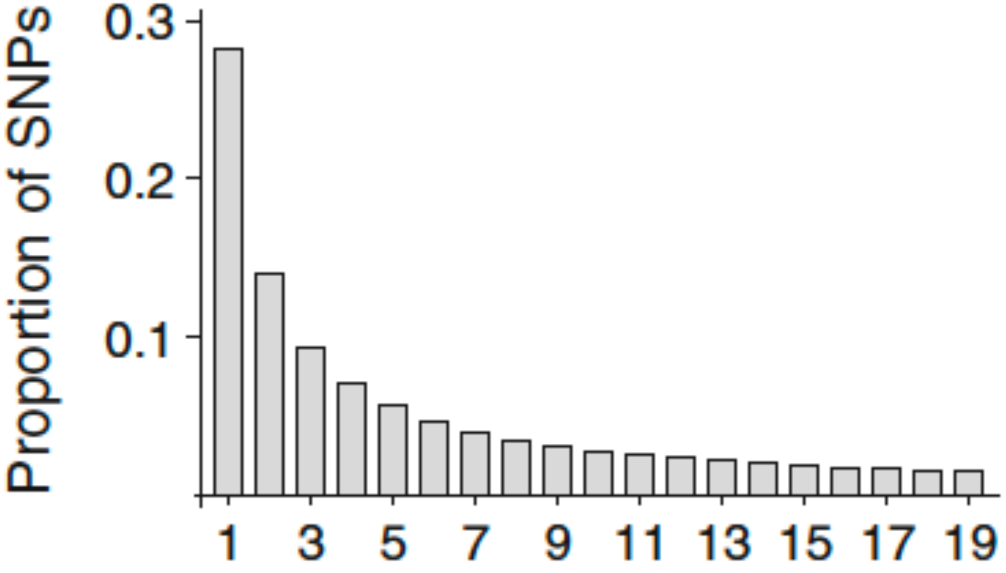
Genealogies generated under constant population size, random mating, no selection

Wide variation simply due to probabilistic nature of the timing of coalescence events

Coalescent Models, Wakeley, from Lohmueller and Nielsen

The site frequency spectrum (SFS)

The SFS is a histogram of allele counts



Note that: in different contexts, axes may be expressed as counts or as proportions (or probabilities)

Coalescent Models, Wakeley, from Lohmueller and Nielsen

A site's position in the spectrum depends on its position in the gene tree

```
----A------
           |
           |----B---
           |        |
-----------         |
                    |-----
                    |
                    |
------C-------------
```

Mutations A and C are singletons; B is a doubleton

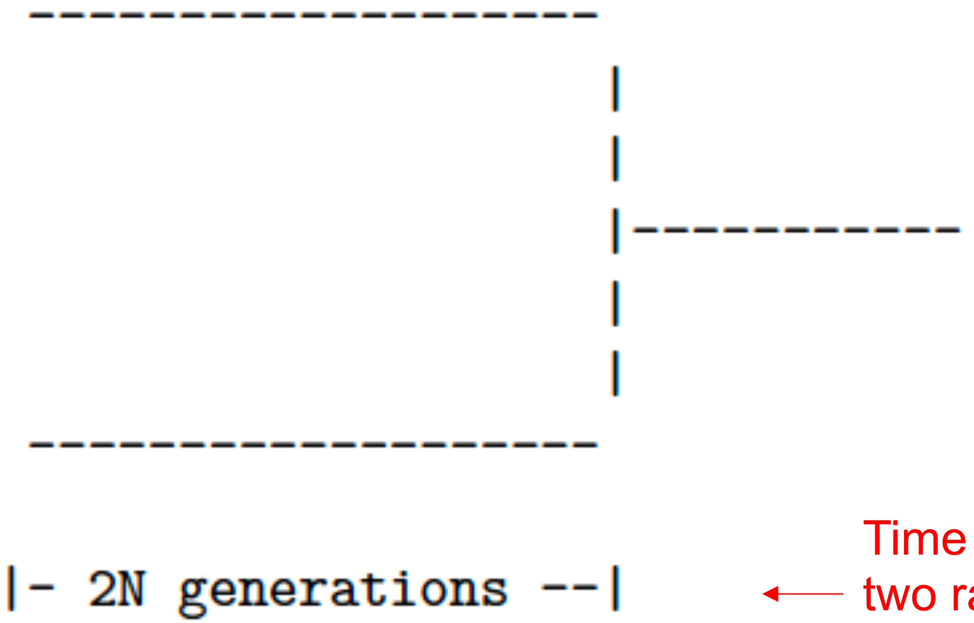
Most recent interval: singletons only

2nd most recent: singletons and doubletons

3rd most recent: singletons, doubletons, and tripletons

A.R. slides

A tree with 2 leaves has only singletons



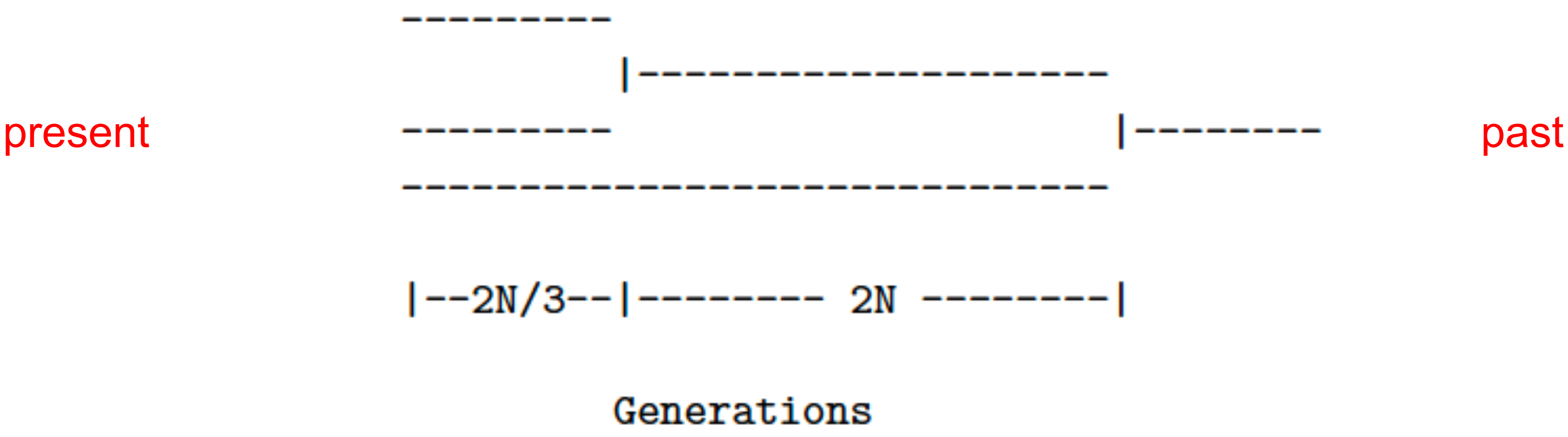
Time for coalescence to occur for two randomly sampled individuals (on average)

We expect $4Nu = \theta$ mutations, all singletons.

Number of branches x L x u = 2 * $2N$ * u = $4Nu$

A.R. slides

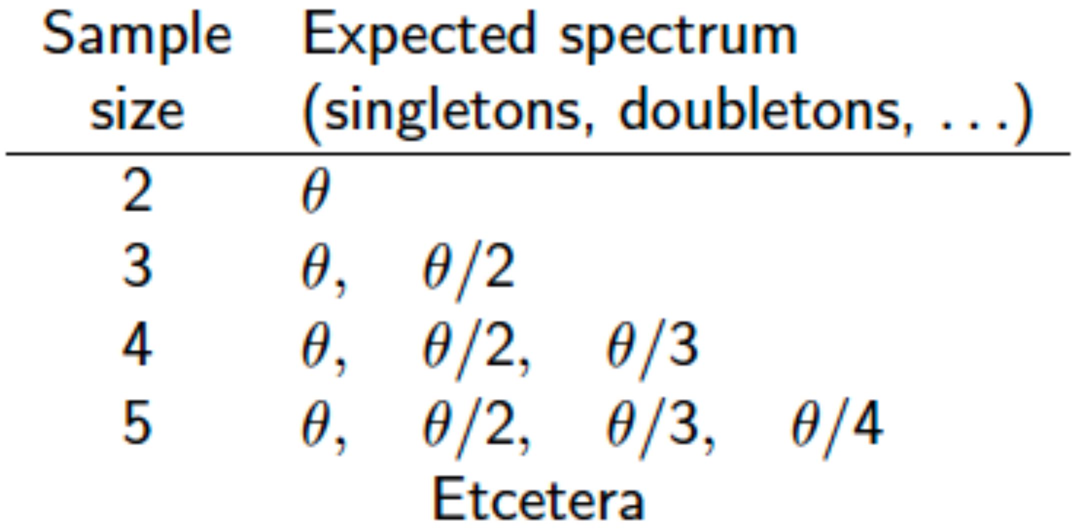
With 3 leaves, there are the same number of singletons but half as many doubletons



At time of coalescent event, $\theta/2$ singletons become doubletons.
New singletons in recent interval: $\underbrace{\frac{2N}{3} \times 3 \times u}_{L \text{ x } \# \text{ branches x } u} = 2Nu = \theta/2$.

A.R. slides

The expected spectrum in a population of constant size



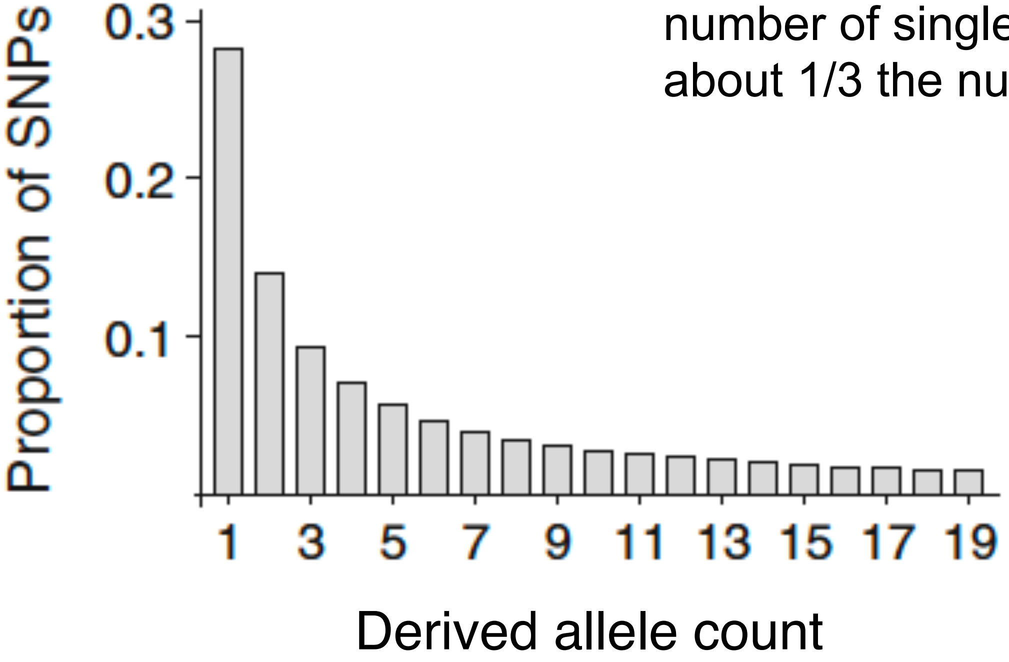
Sample size	Expected spectrum (singletons, doubletons, . . .)
2	θ
3	$\theta, \quad \theta/2$
4	$\theta, \quad \theta/2, \quad \theta/3$
5	$\theta, \quad \theta/2, \quad \theta/3, \quad \theta/4$
	Etcetera

Note that as we increase the sample size, the expected number of mutants in each category stays the same

A.R. slides

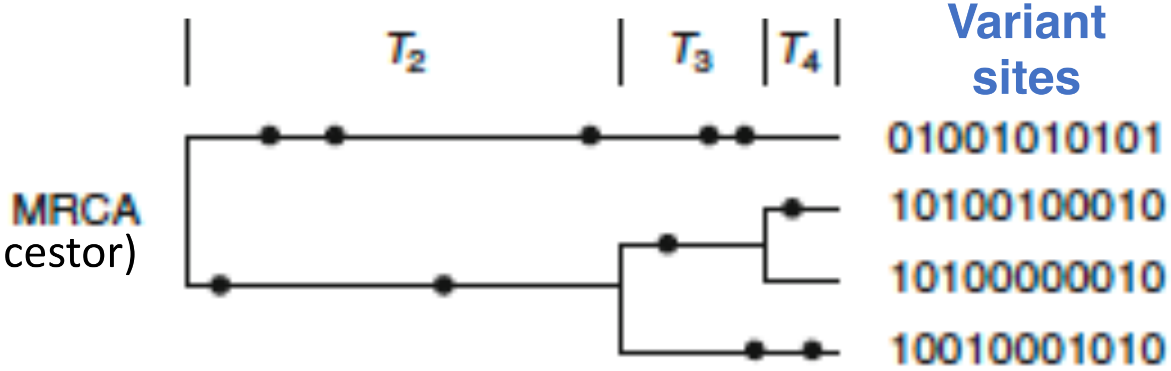
A neutral site frequency spectrum

So, a neutral (unfolded) SFS looks something like this, where the number of doubletons is about half the number of singletons, and the number of tripletons is about 1/3 the number of singletons, …



Coalescent Models, Wakeley, from Lohmueller and Nielsen

A coalescent genealogy with variant sites



Variant sites

(Most Recent Common Ancestor)

0: Ancestral
1: Derived

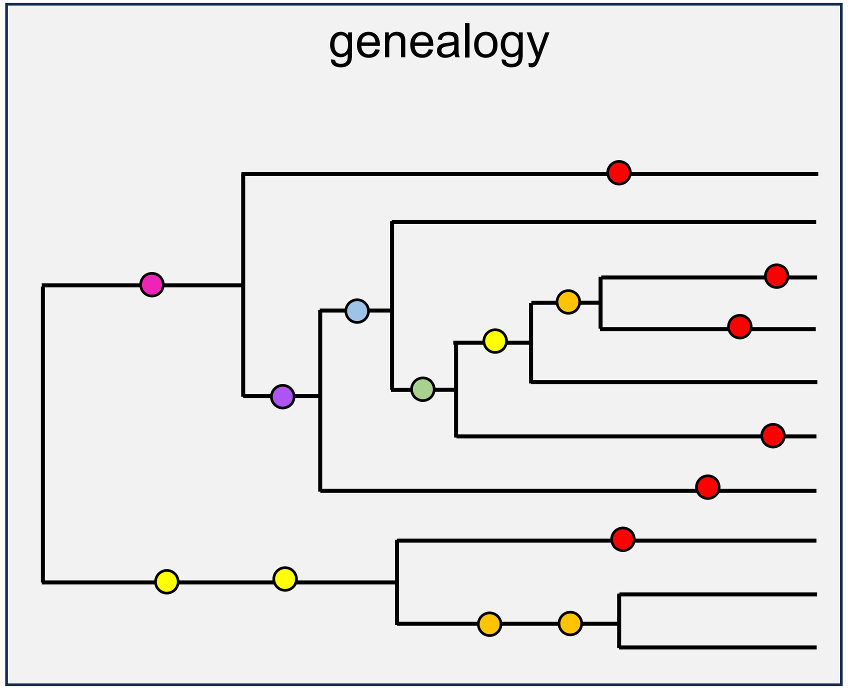
How many singletons?

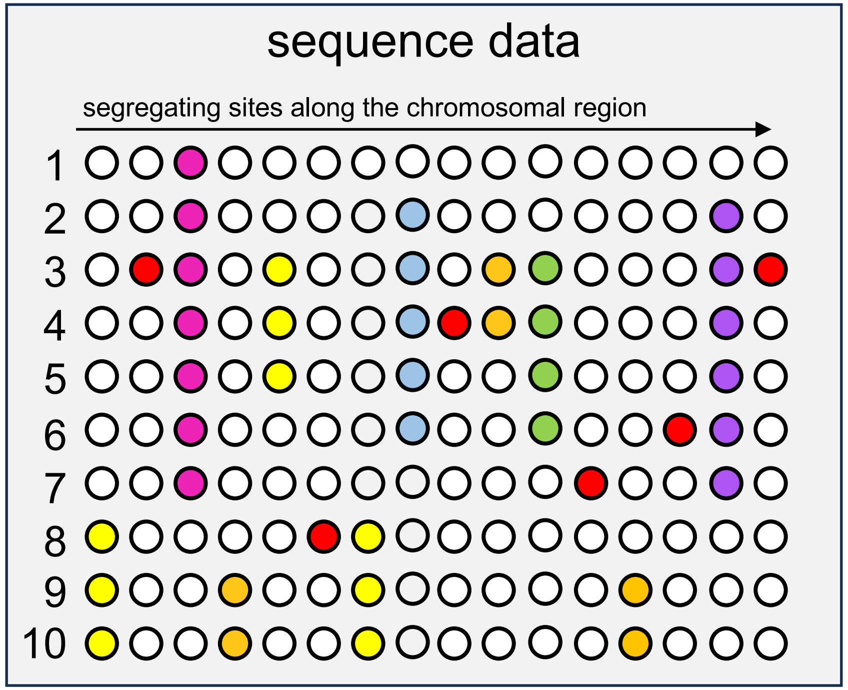
How many doubletons?

Coalescent Models, Wakeley, from Lohmueller and Nielsen

Relationship between a genealogy, sequence data, and the SFS

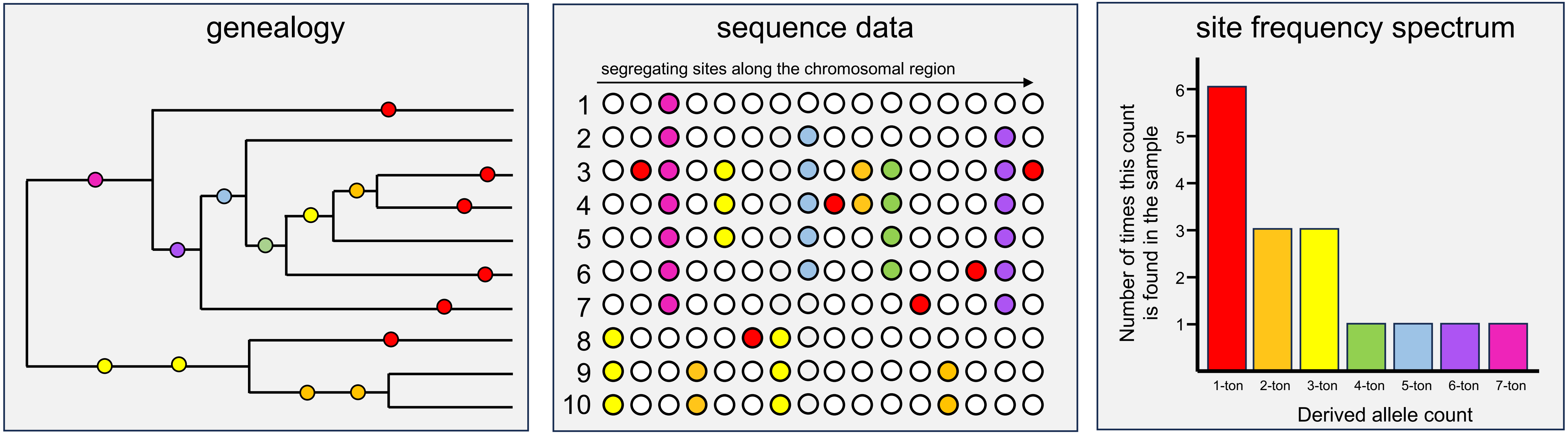
- 8 chromosomes ("genes") are sampled from the population
- This could be from 4 diploid individuals in a randomly mating population
- This is a region with no history of recombination



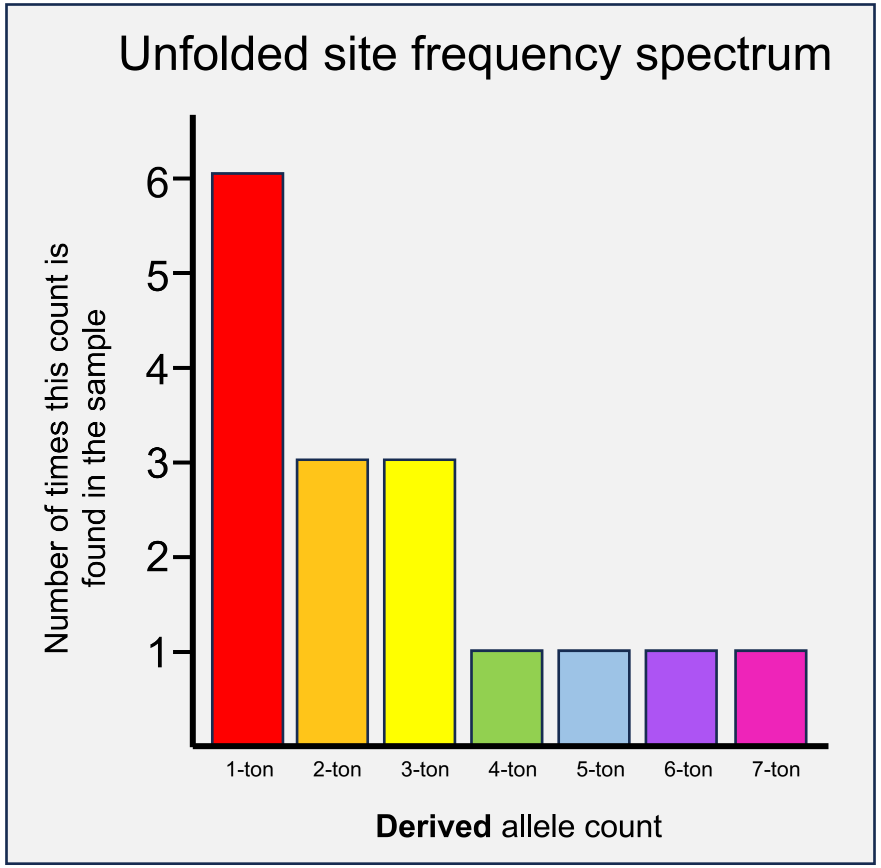


Let's build an SFS!

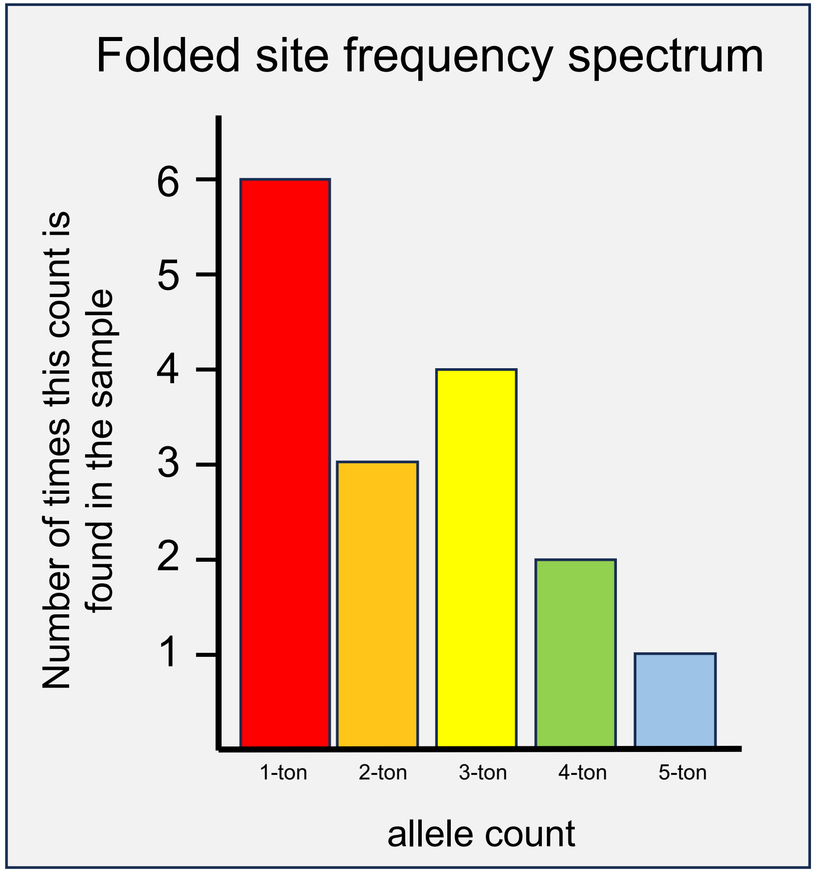
Relationship between a genealogy, sequence data, and the SFS



Convert an unfolded SFS to a folded SFS

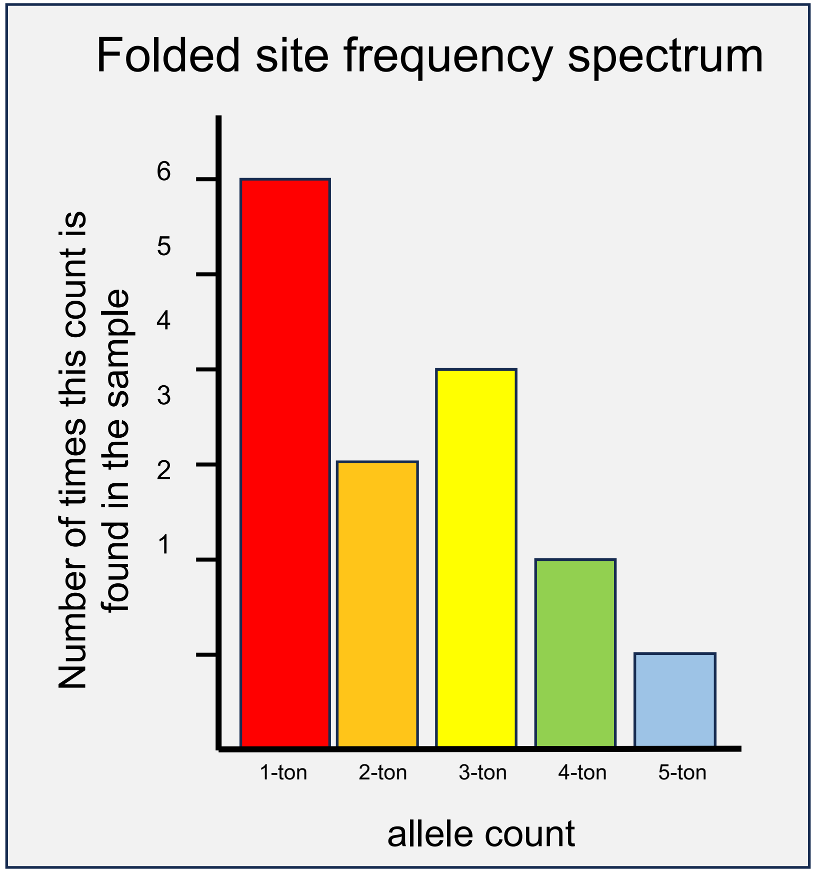


What changes are needed to get from the unfolded to folded SFS?

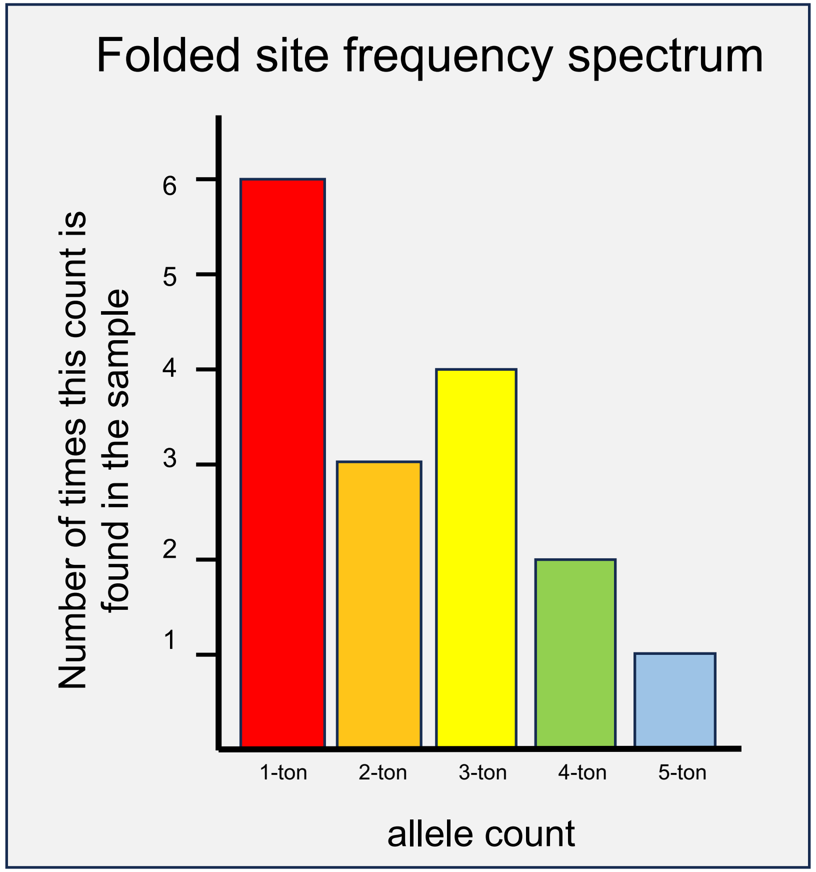


In the folded SFS, we only use information about which is the minor and which is the major allele, so the categories for the extremes are grouped (i.e., 1&9, 2&8, 3&7, 4&6 become 1, 2, 3, and 4)

How to calculate *S* from the folded SFS?



How to calculate *S* from the folded SFS?

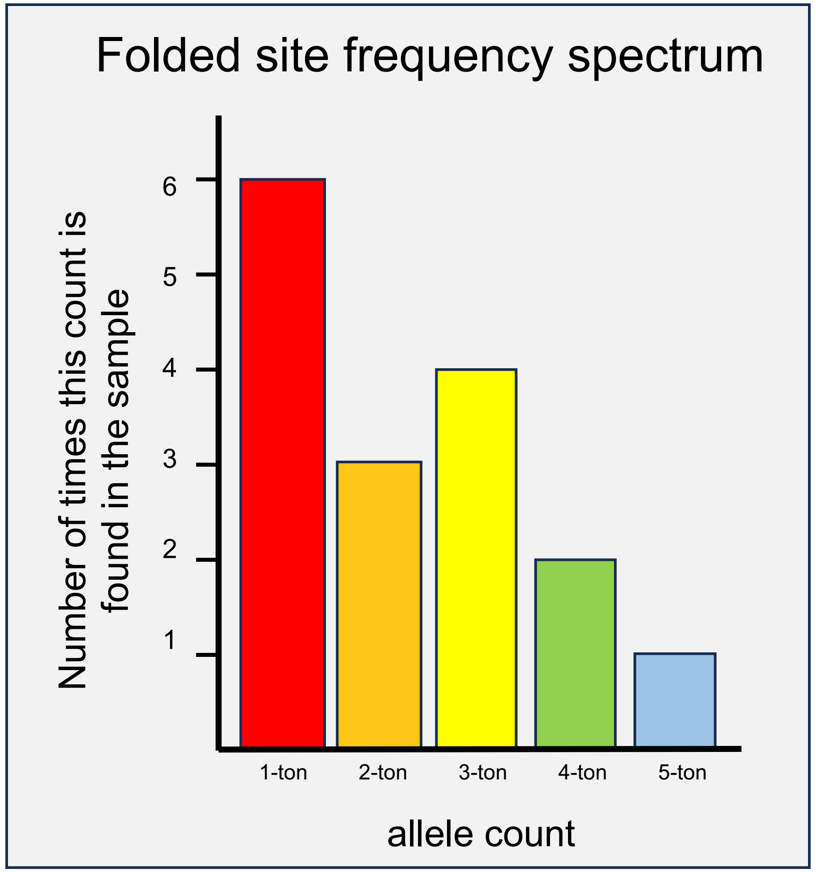


To get *S*, just add up the counts from each category

$S = 6 + 3 + 4 +2 + 1$

$= \ 16$

How to calculate (per sequence*) $\hat{\theta}_S$ from the folded SFS?



$$\hat{\theta}_S = \frac{S}{\sum_{i=1}^{K-1}\frac{1}{i}}$$

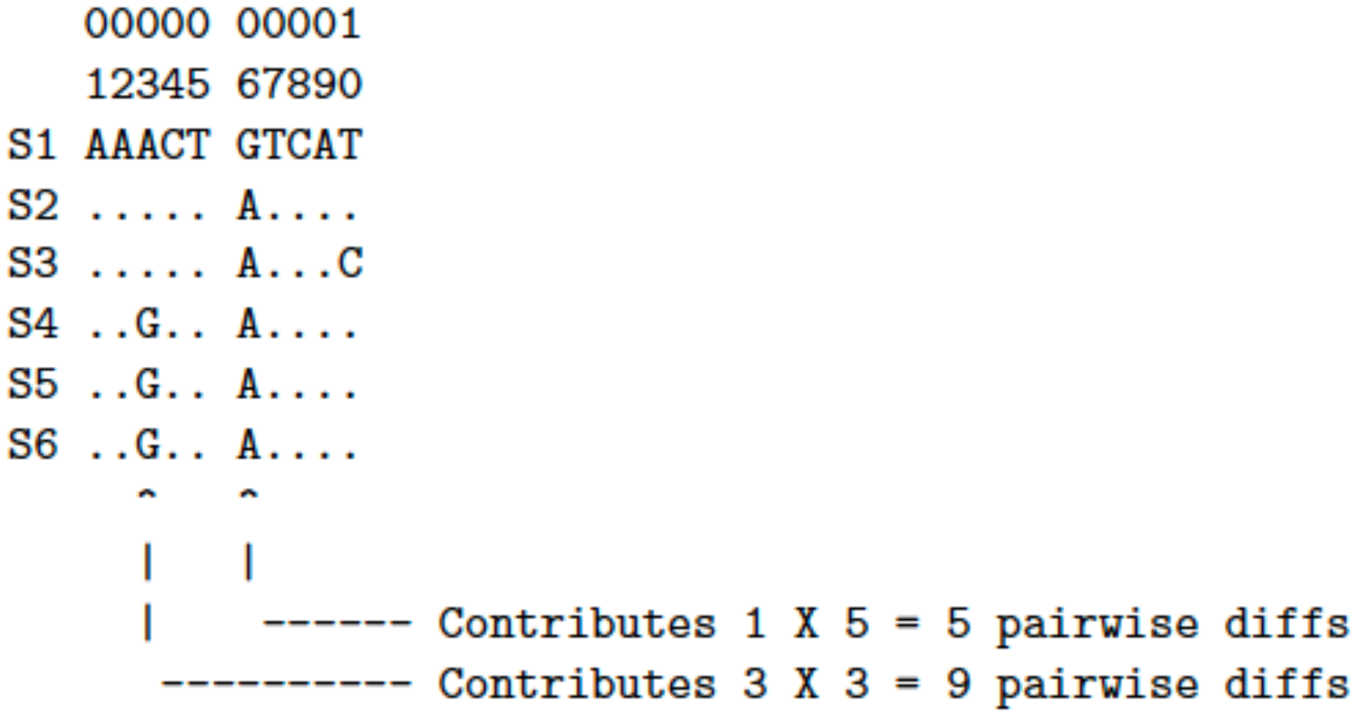
$$= \frac{16}{\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8}+\frac{1}{9}}$$

$$= \frac{16}{2.829} = 5.66$$

*recall: to calculate $\hat{\theta}_S$ per nucleotide, you would need to know the total number of assayed sites

How to calculate π from the folded SFS

Recall from Gene genealogies lecture:

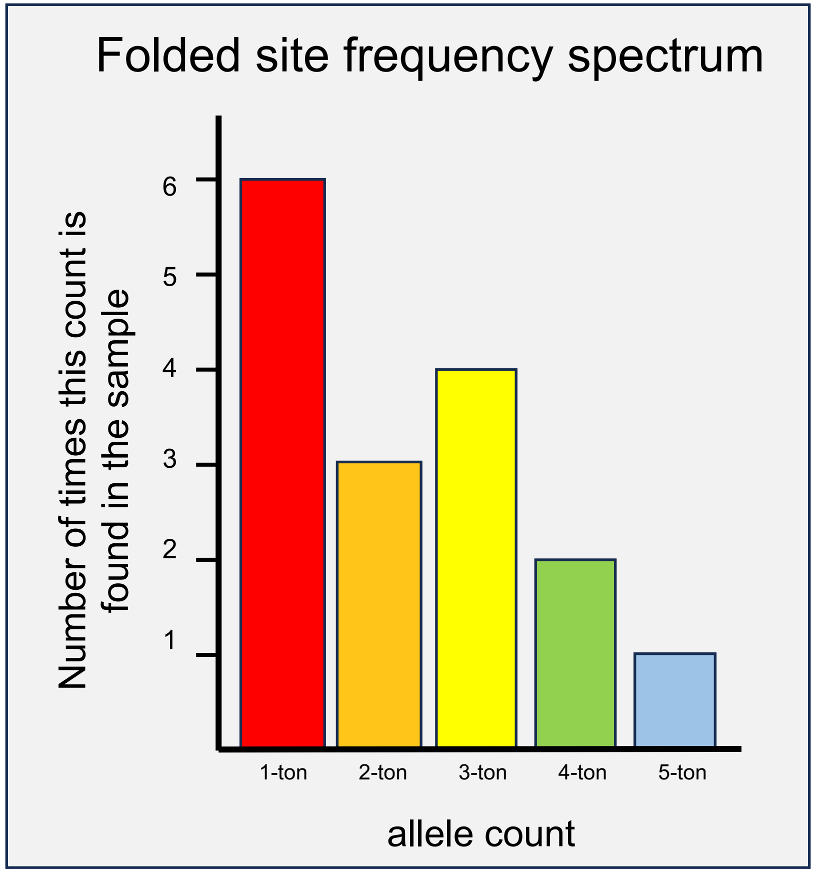


```
   00000 00001
   12345 67890
S1 AAACT GTCAT
S2 ..... A....
S3 ..... A...C
S4 ..G.. A....
S5 ..G.. A....
S6 ..G.. A....
     ^   ^
     |   |
     |    ------ Contributes 1 X 5 = 5 pairwise diffs
      ---------- Contributes 3 X 3 = 9 pairwise diffs
```

…Sum these up and divide by the number of pairwise comparisons

See gene genealogies lecture notes sec 1.3 for more info

How to calculate (per sequence*) $\widehat{\theta}_{\pi}$ from the folded SFS?



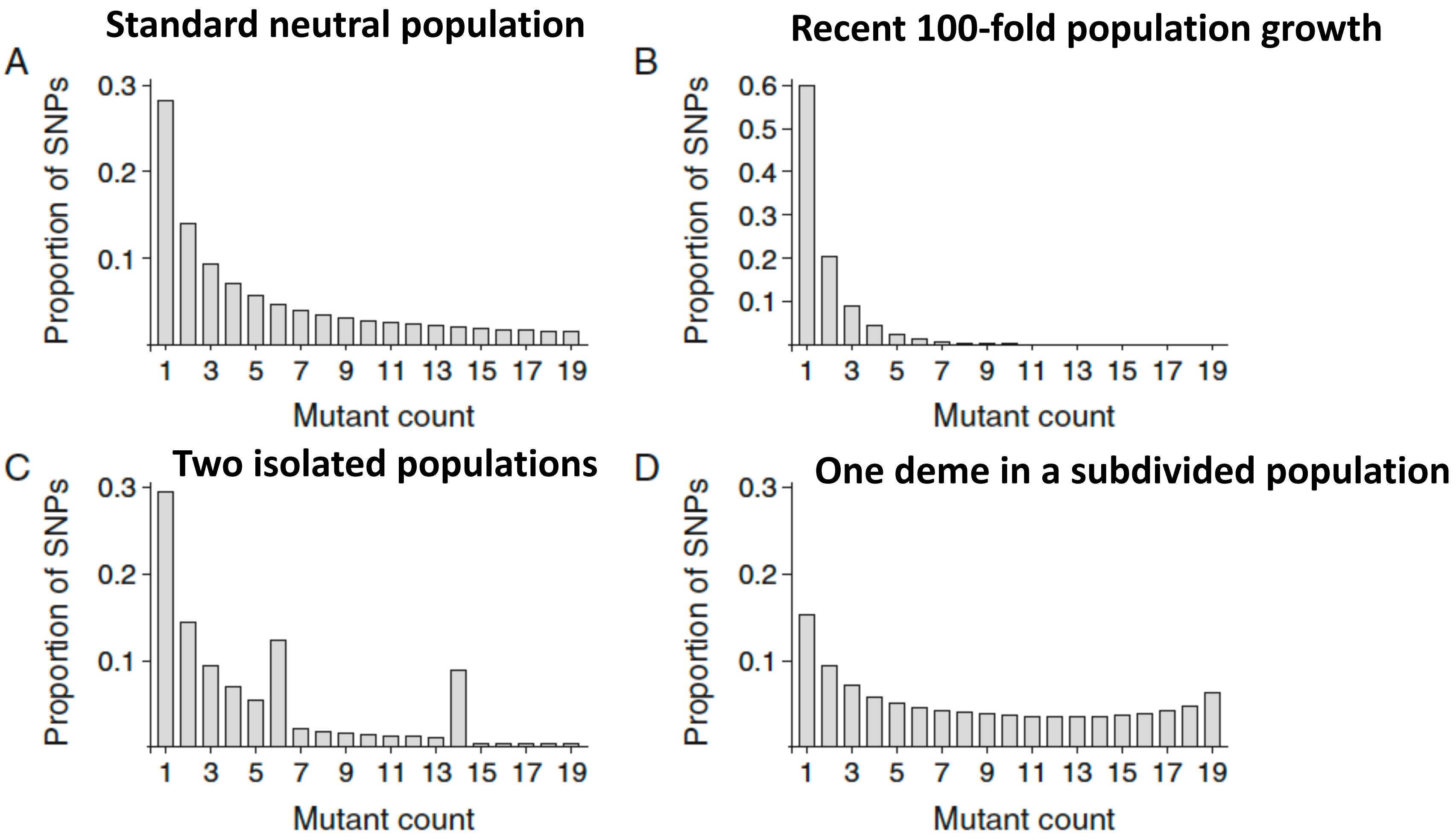
$$\hat{\theta}_{\pi} = \pi$$

$$= \frac{6(1 \text{ x } 9) + 3(2 \text{ x } 8) + 4(3 \text{ x } 7) + 2(4 \text{ x } 6) + 1(5 \text{ x } 5)}{45}$$

$$= \frac{54 + 48 + 84 + 48 + 25}{45} = 5.76$$

*recall: to calculate $\hat{\theta}_{\pi}$ per nucleotide, you would need to know the total number of assayed sites

Site frequency spectra under different population history models



Standard neutral population

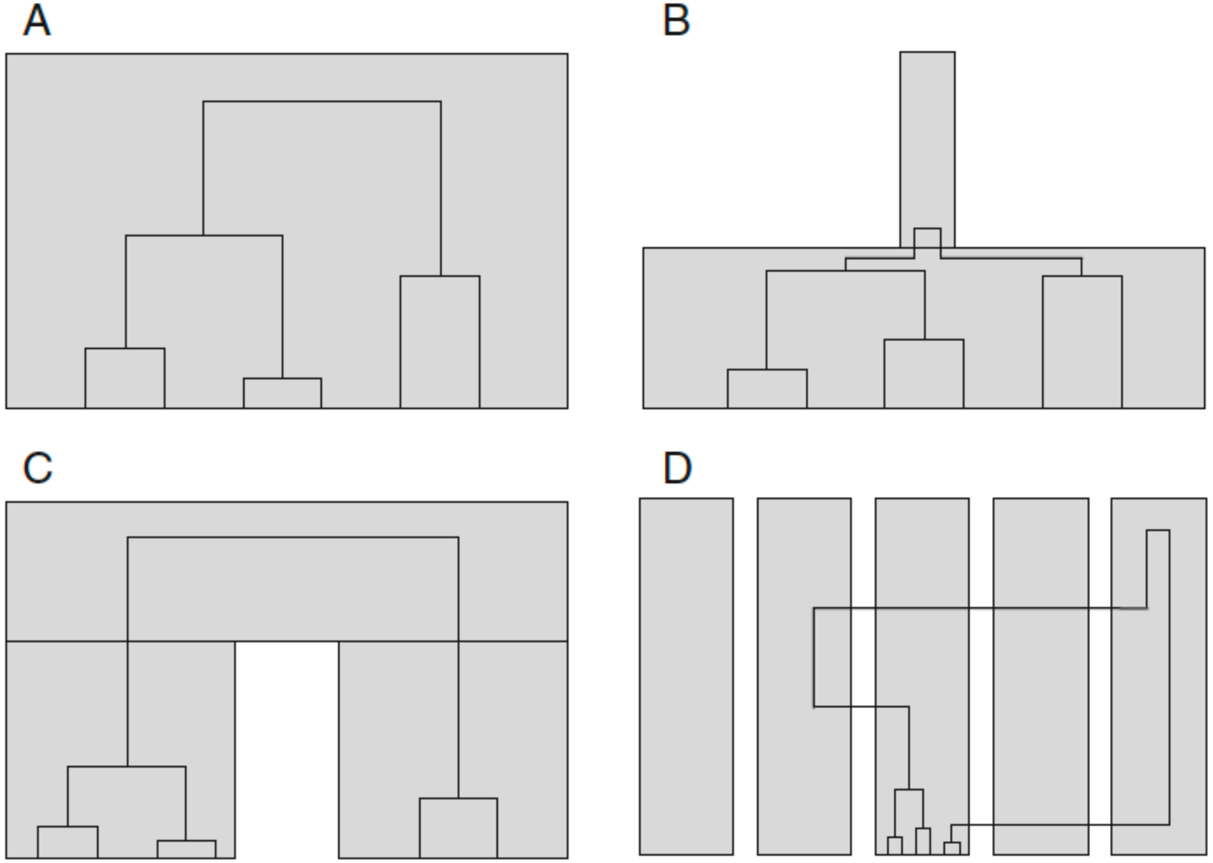
Recent 100-fold population growth

Two isolated populations

One deme in a subdivided population

Coalescent Models, Wakeley, from Lohmueller and Nielsen

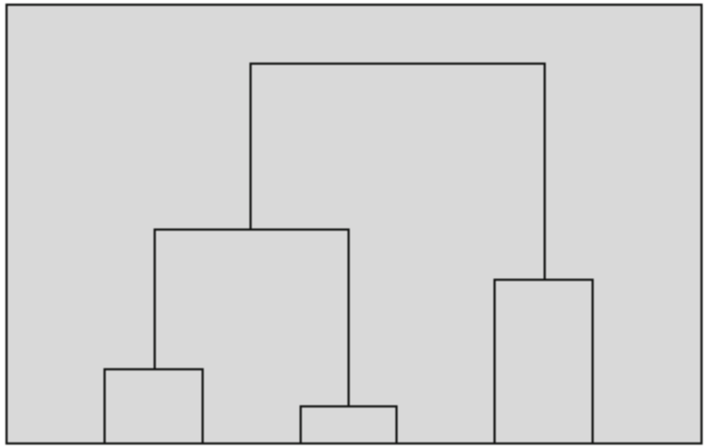
Cartoon depictions of genealogies from the four different population history models



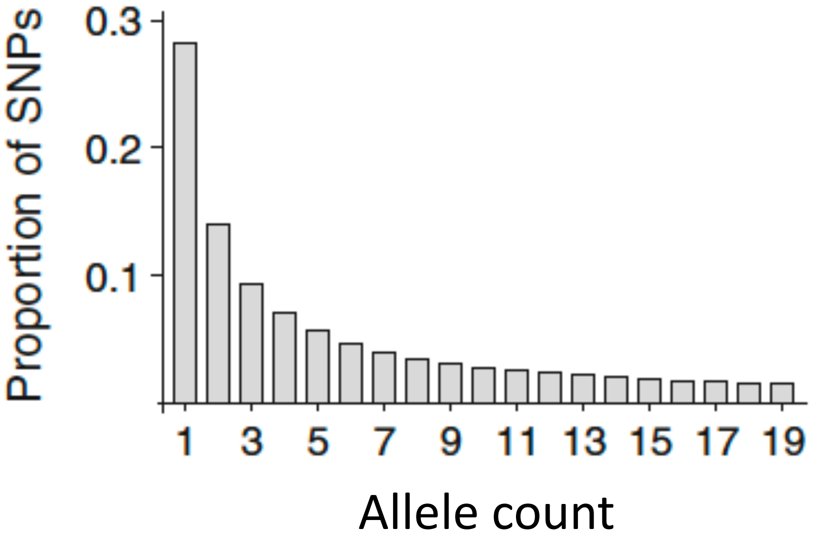
Coalescent Models, Wakeley, from Lohmueller and Nielsen

Standard neutral population

Schematic of coalescent history



Site frequency spectrum

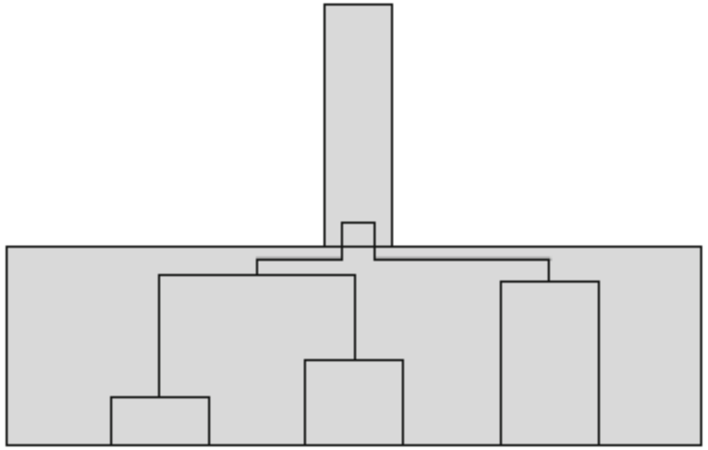


This SFS fits has roughly the expected distribution of frequencies (θ/i) for singletons, doubletons, tripletons, etc: θ, $\theta/2$, $\theta/3$, …

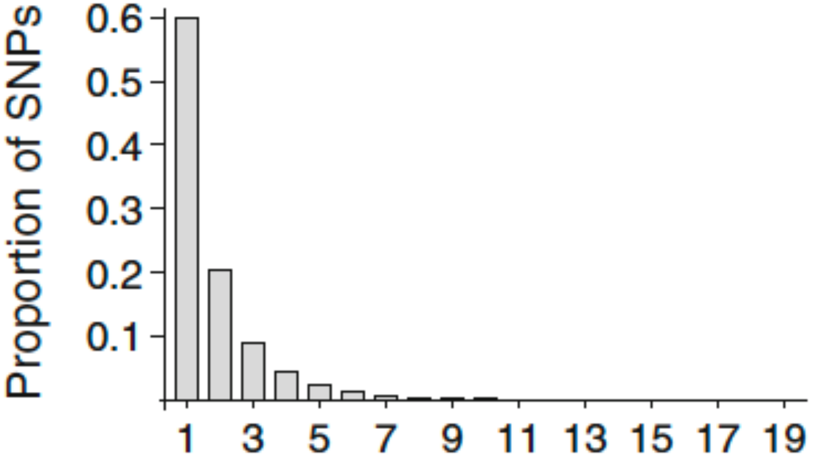
Modified from Coalescent Models, Wakeley, from Lohmueller and Nielsen

Recent 100-fold population growth

Schematic of coalescent history



Site frequency spectrum



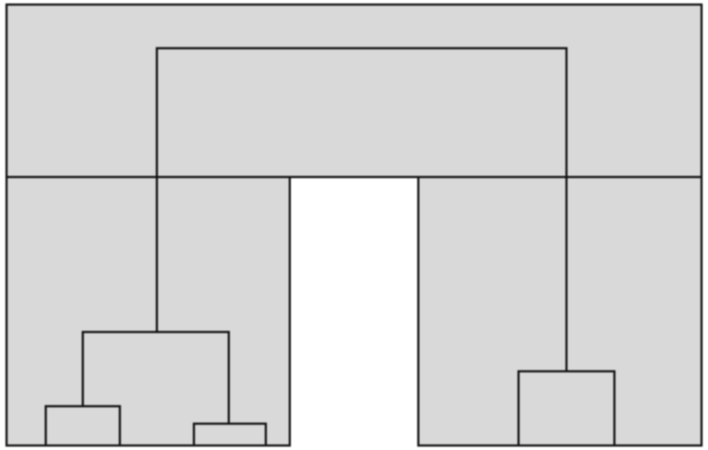
Allele count

An excess of low frequency variants in the SFS due to rapid population expansion

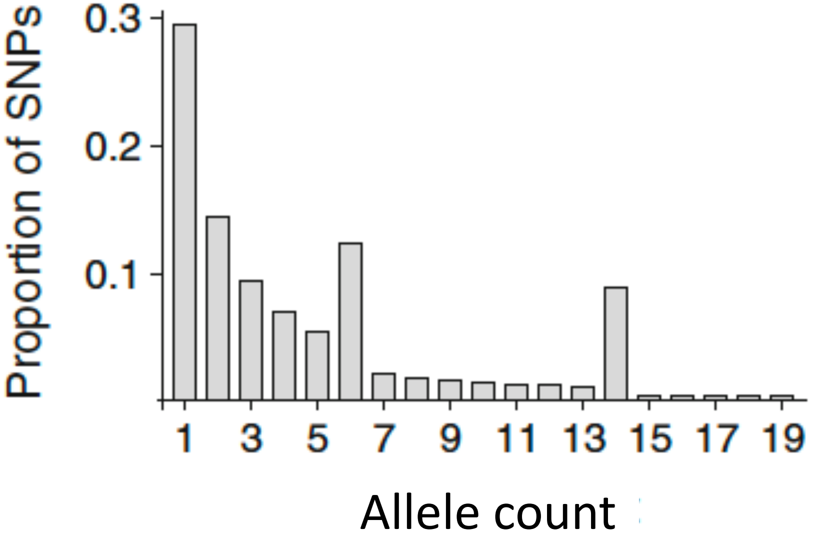
Modified from Coalescent Models, Wakeley, from Lohmueller and Nielsen

Two isolated populations

Schematic of coalescent history



Site frequency spectrum



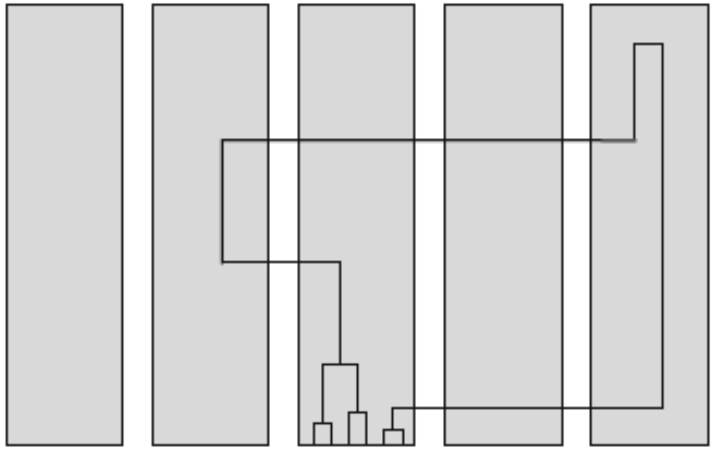
Population subdivision results in an excess of intermediate allele frequencies in the SFS because alleles often coalesce farther back in time

Modified from Coalescent Models, Wakeley, from Lohmueller and Nielsen

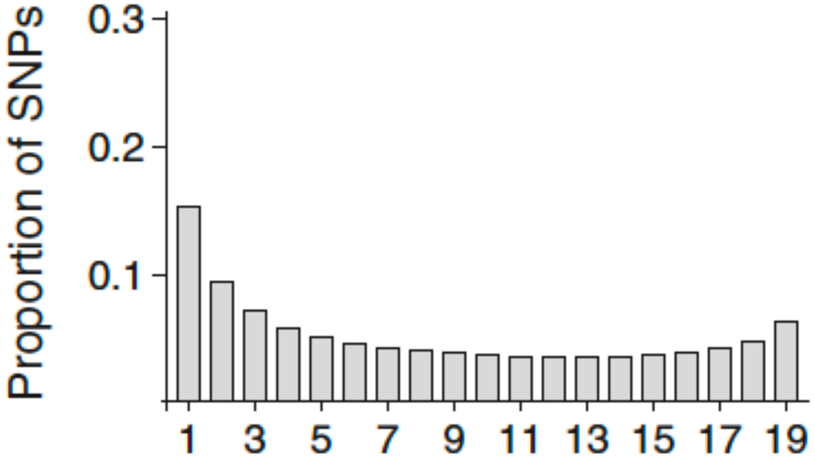
One deme in a subdivided population

A subdivided population in which migration can occur among five local populations

Schematic of coalescent history



Site frequency spectrum



Allele count

In this example, alleles tend to coalesce a long time ago and therefore to be at intermediate frequencies in the SFS

Modified from Coalescent Models, Wakeley, from Lohmueller and Nielsen

There are multiple ways to measure nucleotide diversity (θ)

The most popular estimates are:

- Waterson's θ_W, which is also called θ_S (based on S, the number of segregating sites)
- Tajima's θ_π(based on π, the number of pairwise differences)

There are multiple ways to measure nucleotide diversity (θ)

- Waterson's θ (based on S, the number of segregating sites)

$$\hat{\theta}_S = \frac{S}{\sum_{i=1}^{K-1} \frac{1}{i}}$$

- $\hat{\theta}_\pi$ (based the number of pairwise differences)

$$\hat{\theta}_\pi = \frac{\sum_{i<j} \pi(i,j)}{\binom{n}{2}}$$

$\pi(i,j)$ is the number of differences between two sequences
n is the number of sequences in the sample

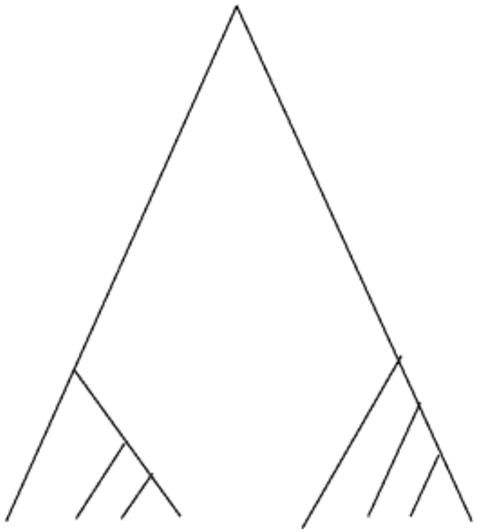
Comparing estimates of θ provides insights into the history of a population or locus

- Different theta estimates summarize different aspects of the site frequency spectrum (and different patterns of variation on the genealogy)
- By comparing these different estimates of theta, we can compare these different aspects of the site frequency spectrum (and the genealogy)
- There are several statistics that have been created to provide a way to summarize such comparisons. The most popular is called *Tajima's D*
- Events that occurred in the history of a ***population*** create ***genome-wide*** deviations from the expectation under random-mating
- In comparison, ***selective events affect single loci*** and create deviations in the statistics ***at a particular locus*** relative to the rest of the genome

Tajima’s D statistic

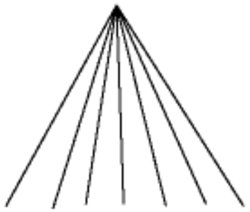
Compares estimates of *θ* based on the number of segregating sites (S) and *π* (the number of pairwise differences) in the sample

$$D = \frac{\widehat{\theta}_\pi - \widehat{\theta}_S}{\sqrt{\widehat{\mathbf{Var}}[\widehat{\theta}_\pi - \widehat{\theta}_S]}}$$



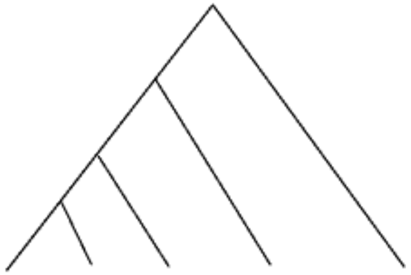
$\theta_\pi > \theta_W$

D positive



$\theta_\pi < \theta_W$

D negative

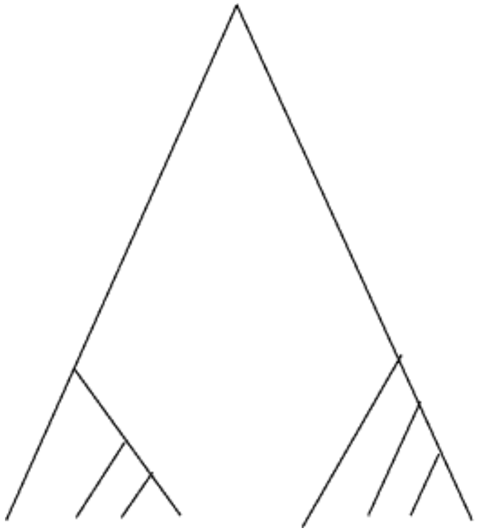


$\theta_\pi = \theta_W$

Genome-wide patterns of Tajima's D are impacted by population history

$\theta_\pi > \theta_W$

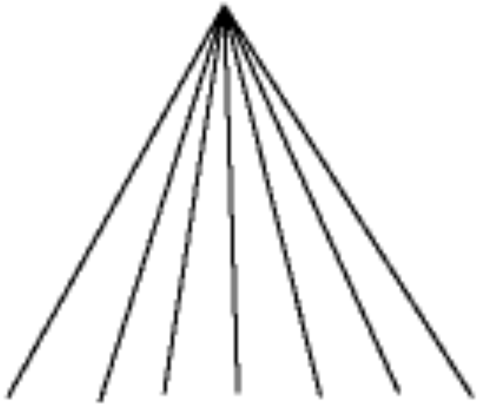
D positive



Can get this genome-wide from population subdivision or a bottleneck

$\theta_\pi < \theta_W$

D negative

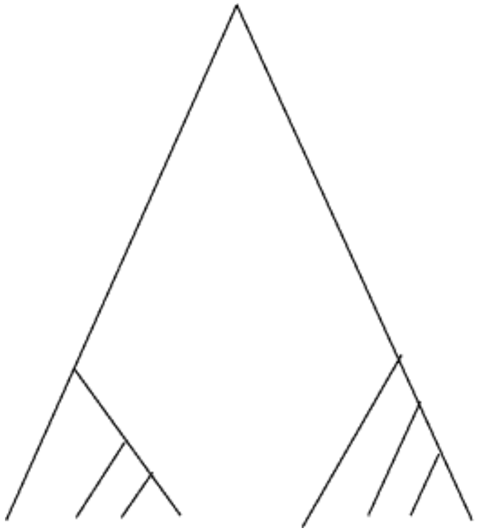


Can get this genome-wide from recent population growth

Single-locus patterns of Tajima's D are impacted by selection

$\theta_\pi > \theta_W$

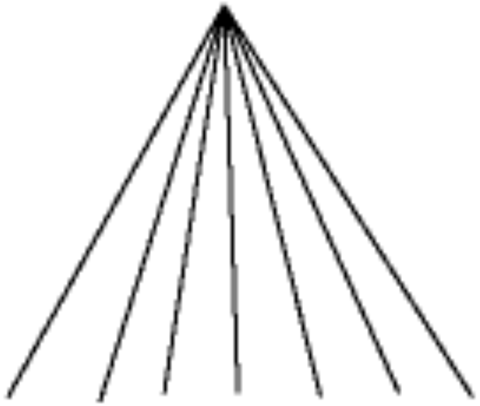
D positive



Can get this in a single region from balancing selection

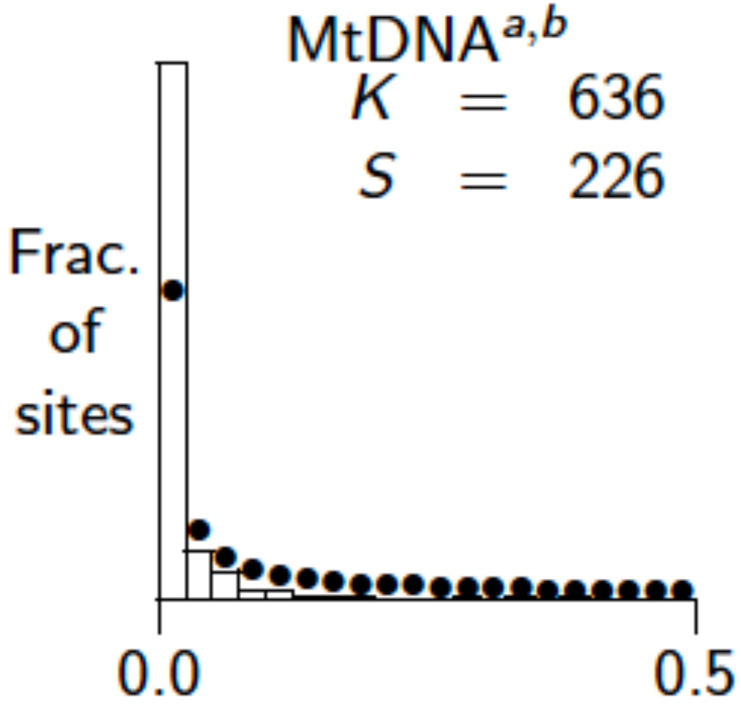
$\theta_\pi < \theta_W$

D negative



Can get this in a single region from a selective sweep

Real examples: human mitochondrial DNA



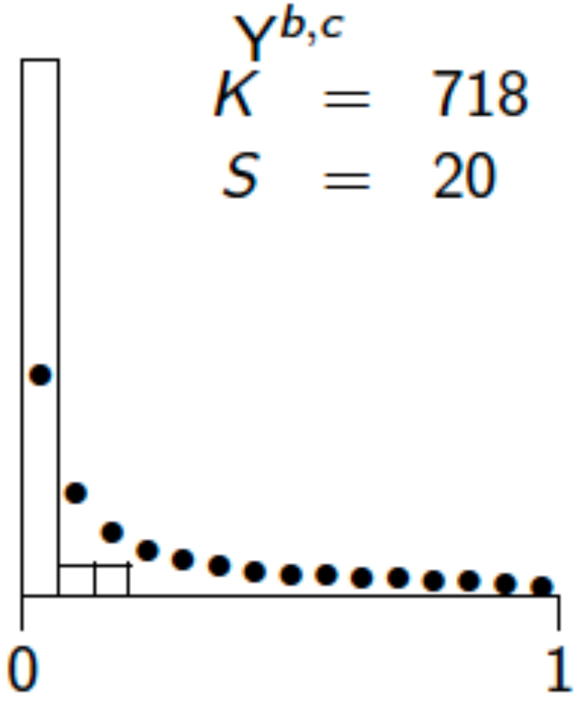
● Represents expected value

Bars represent observed values

In mtDNA, there is an excess of singletons relative to expected

from Gene genealogies lecture notes

Real examples: human Y chromosome



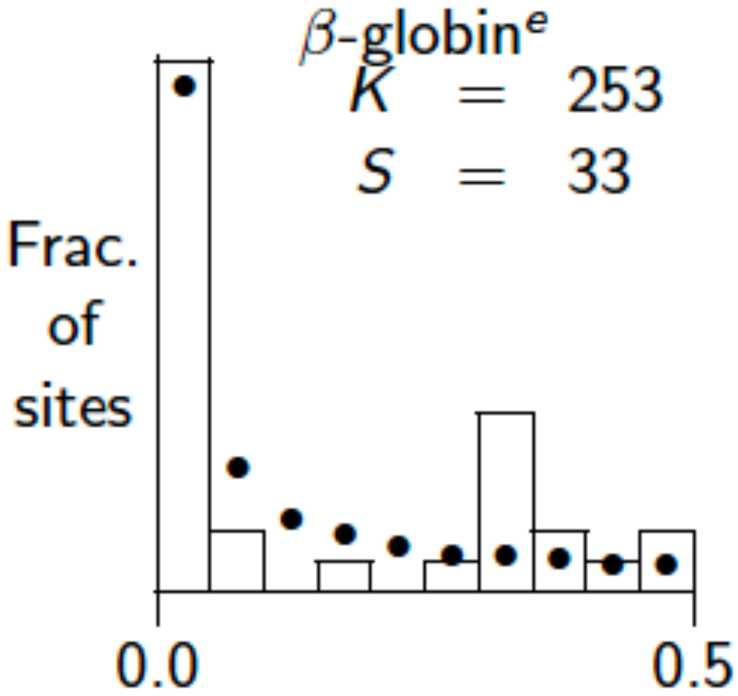
● Represents expected value

Bars represent observed values

On the Y chromosome, there is an excess of singletons relative to expected

from Gene genealogies lecture notes

Real examples: human beta globin



● Represents expected value

Bars represent observed values

At the Beta Globin locus, there is an excess of intermediate frequency alleles relative to expected

But why?

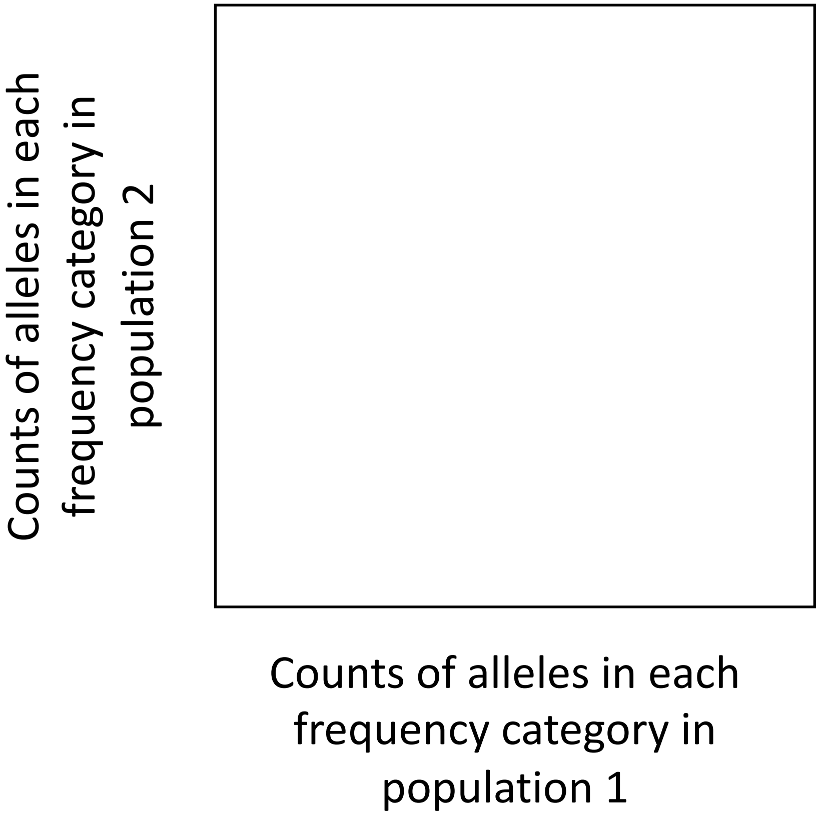
Some polymorphisms in *β-globin* are thought to protect against malaria but also to lead to sickle-cell anaemia and thalassemia, they are under balancing selection

from Gene genealogies lecture notes

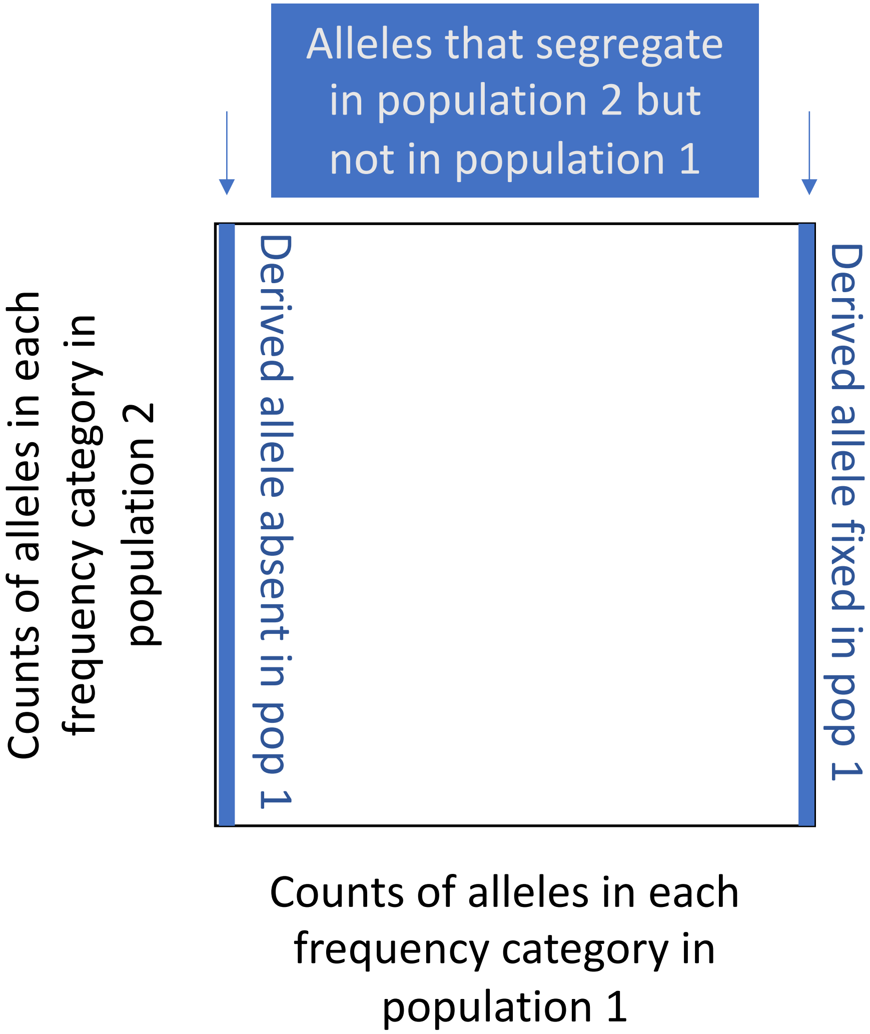
Comparing two populations using the joint site frequency spectrum (JSFS)

The joint site frequency spectrum includes information about sharing of alleles and their frequencies within and between populations

The joint site frequency spectrum (2D SFS)



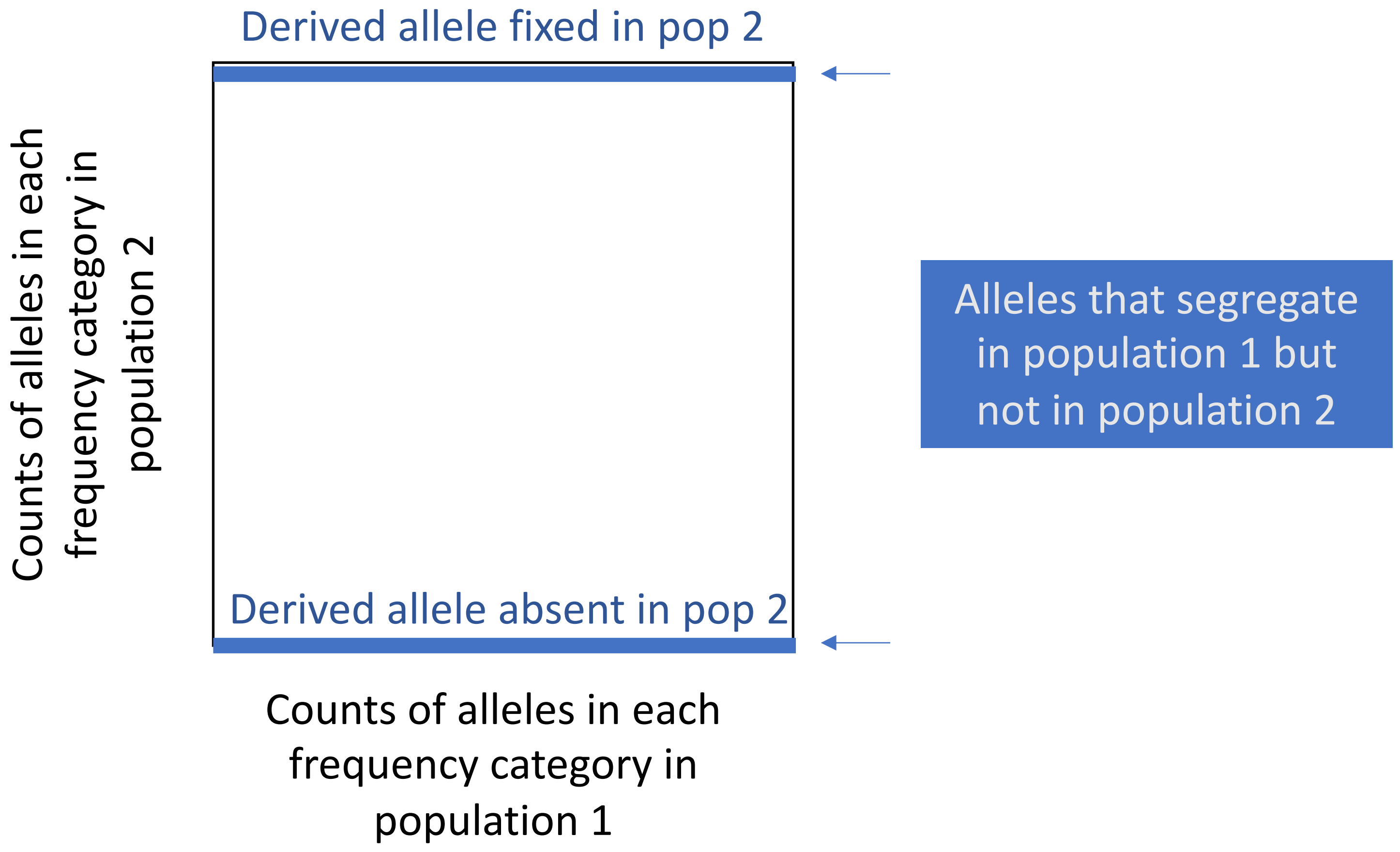
The joint site frequency spectrum (2D SFS)



Alleles that segregate in population 2 but not in population 1

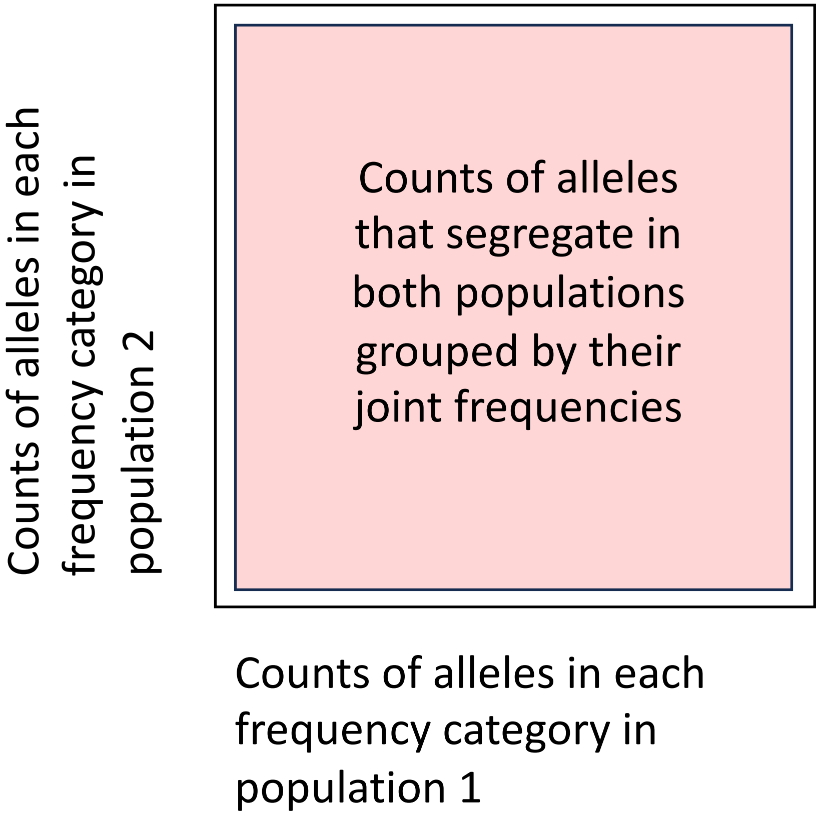
Derived allele absent in pop 1
Derived allele fixed in pop 1
Counts of alleles in each frequency category in population 2
Counts of alleles in each frequency category in population 1

The joint site frequency spectrum (2D SFS)



Derived allele fixed in pop 2
Counts of alleles in each frequency category in population 2
Alleles that segregate in population 1 but not in population 2
Derived allele absent in pop 2
Counts of alleles in each frequency category in population 1

The joint site frequency spectrum (2D SFS)

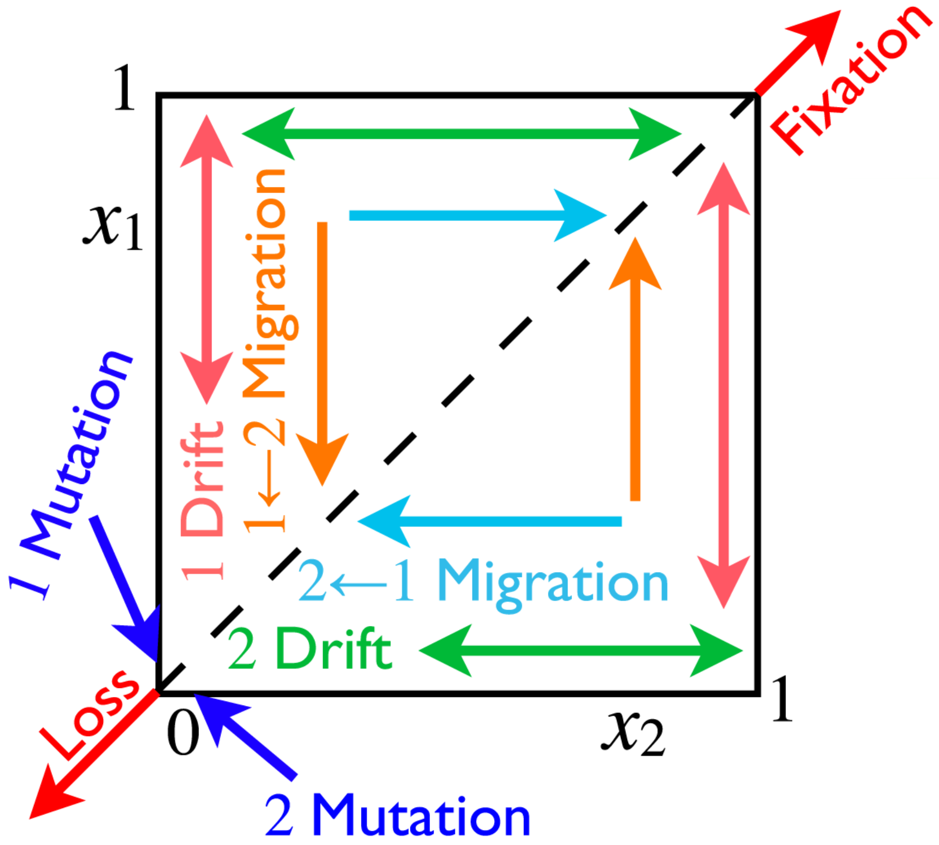


Counts of alleles that segregate in both populations grouped by their joint frequencies

Counts of alleles in each frequency category in population 2

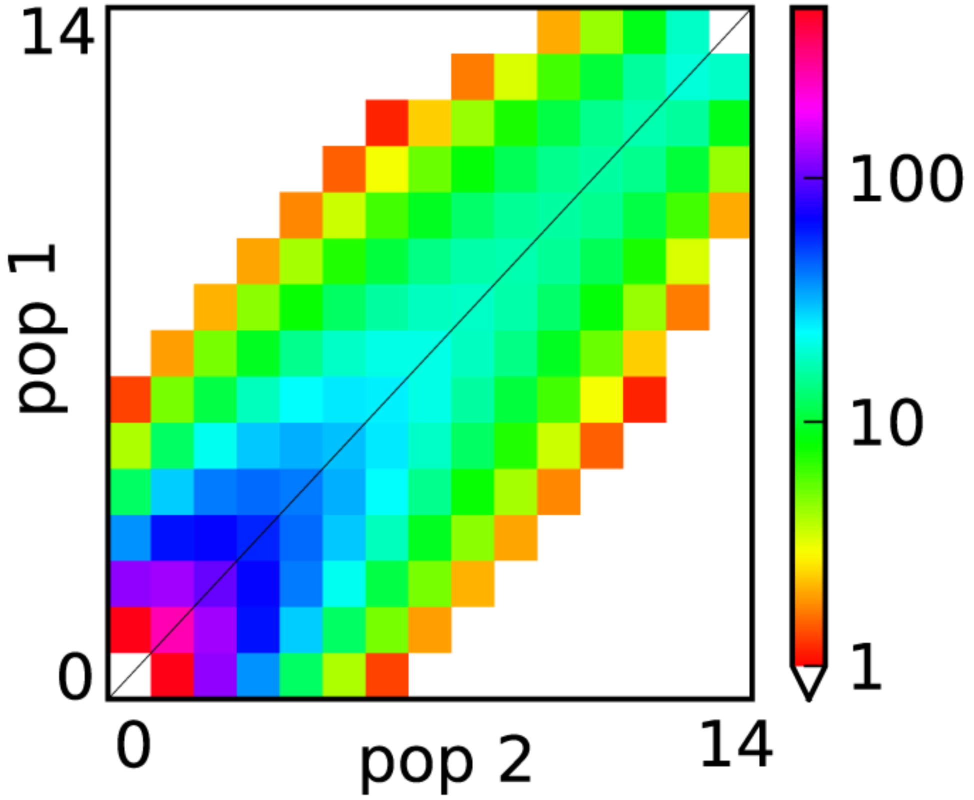
Counts of alleles in each frequency category in population 1

The joint site frequency spectrum (2D SFS)



Gutenkunst et al., 2009

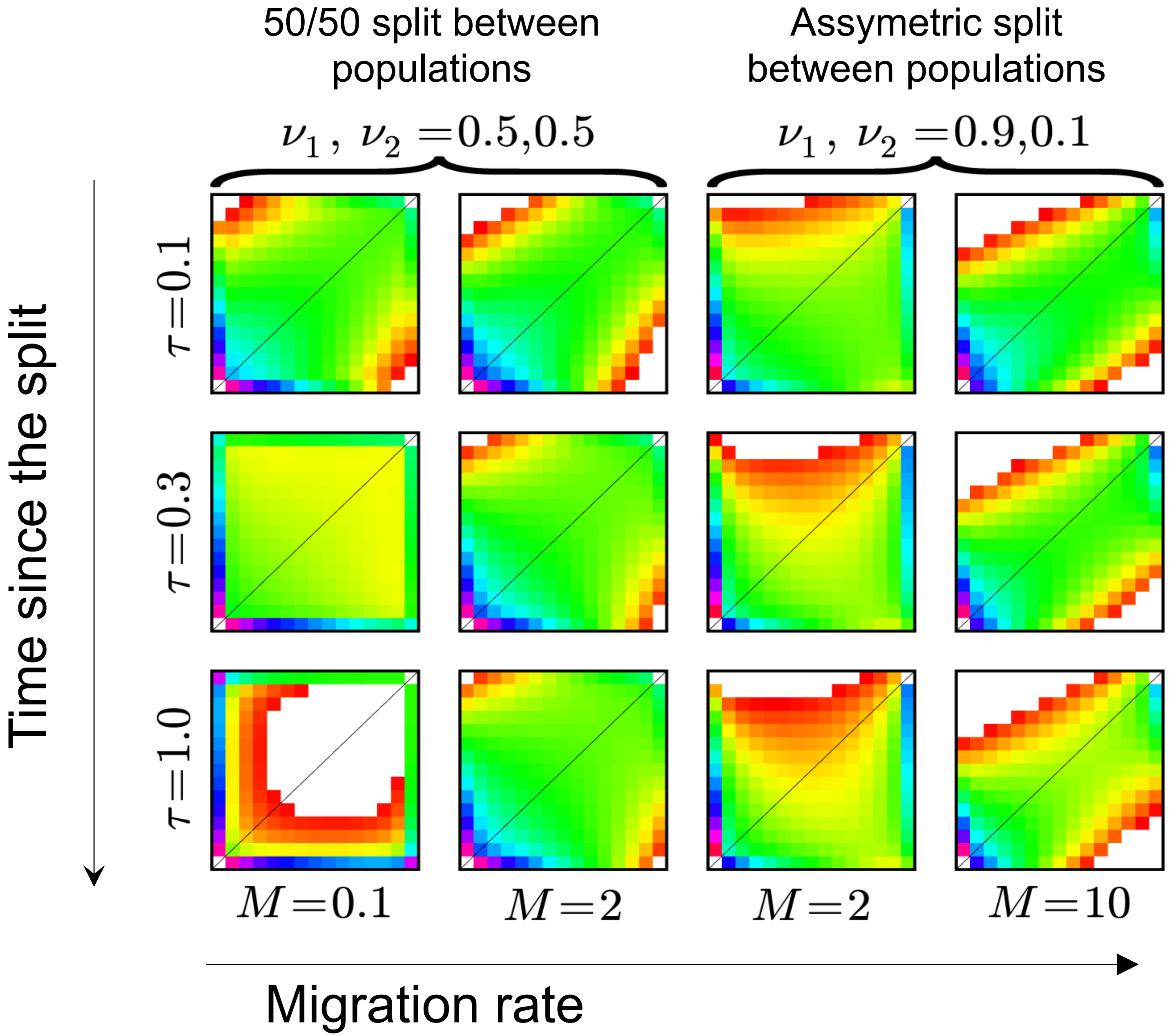
The joint site frequency spectrum (2D SFS)



Simulated a simple split, immediately after the split

Gutenkunst et al., 2009

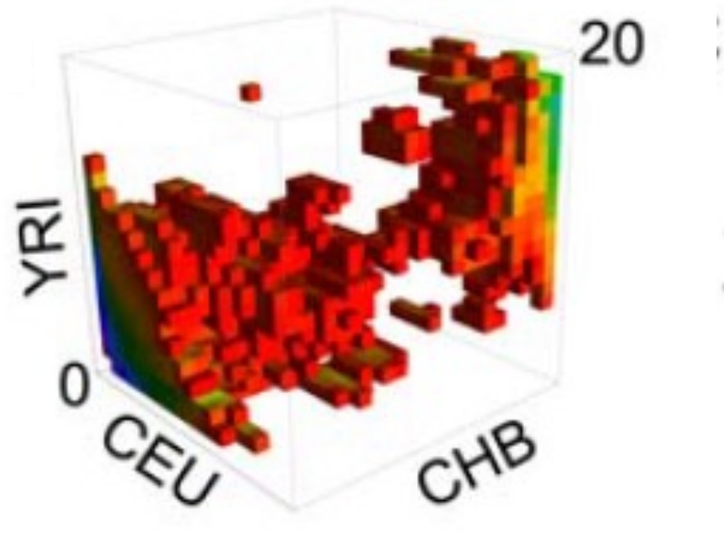
The impacts of demographic history on the JSFS



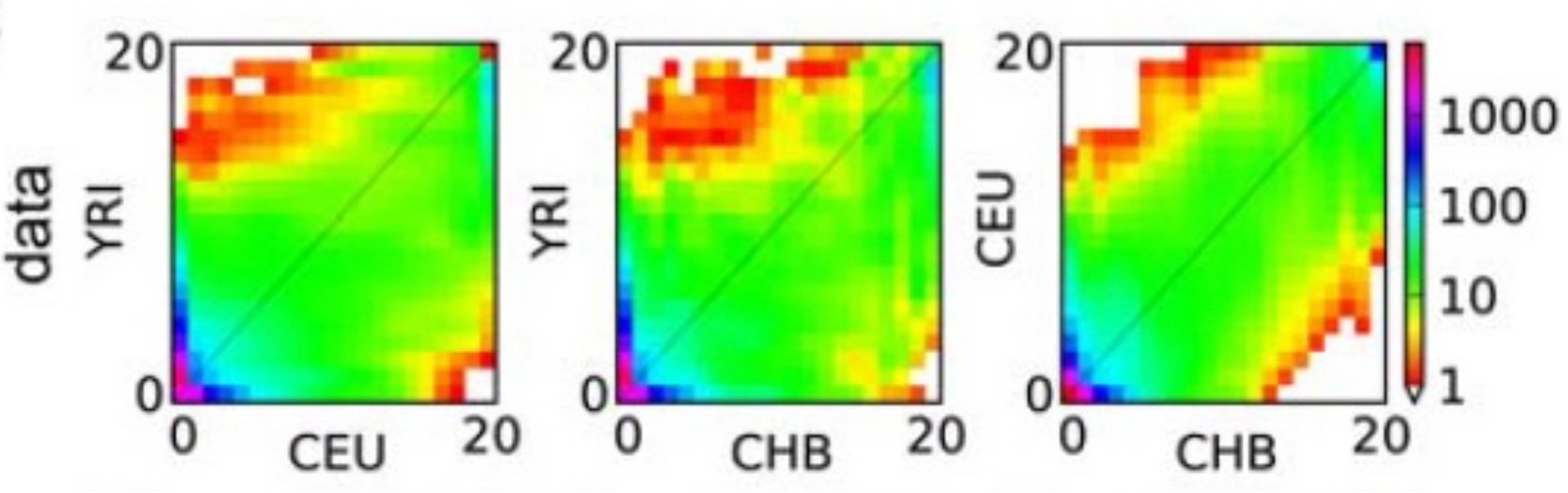
Gutenkunst et al., 2009

An example: human populations from Africa (YRI), Europe (CEU), and China (CHB)

3D SFS



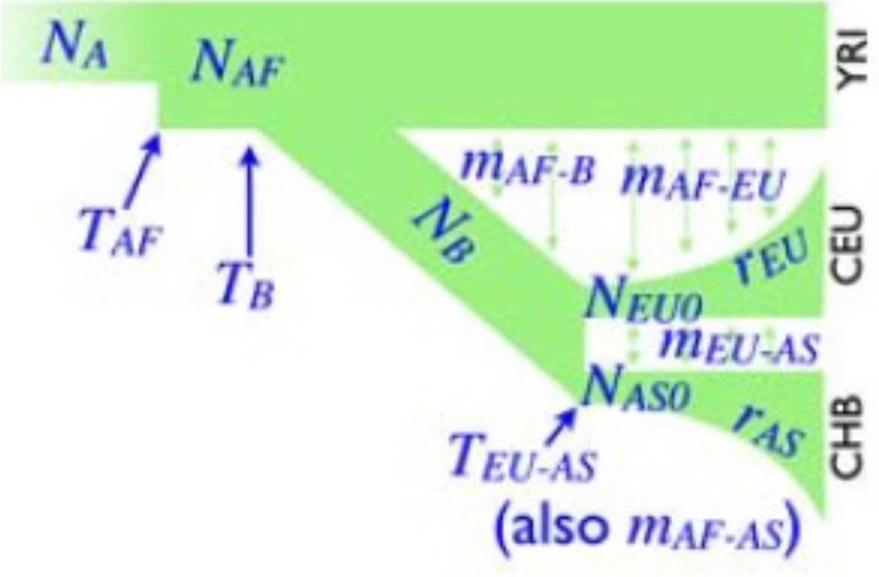
2D SFS

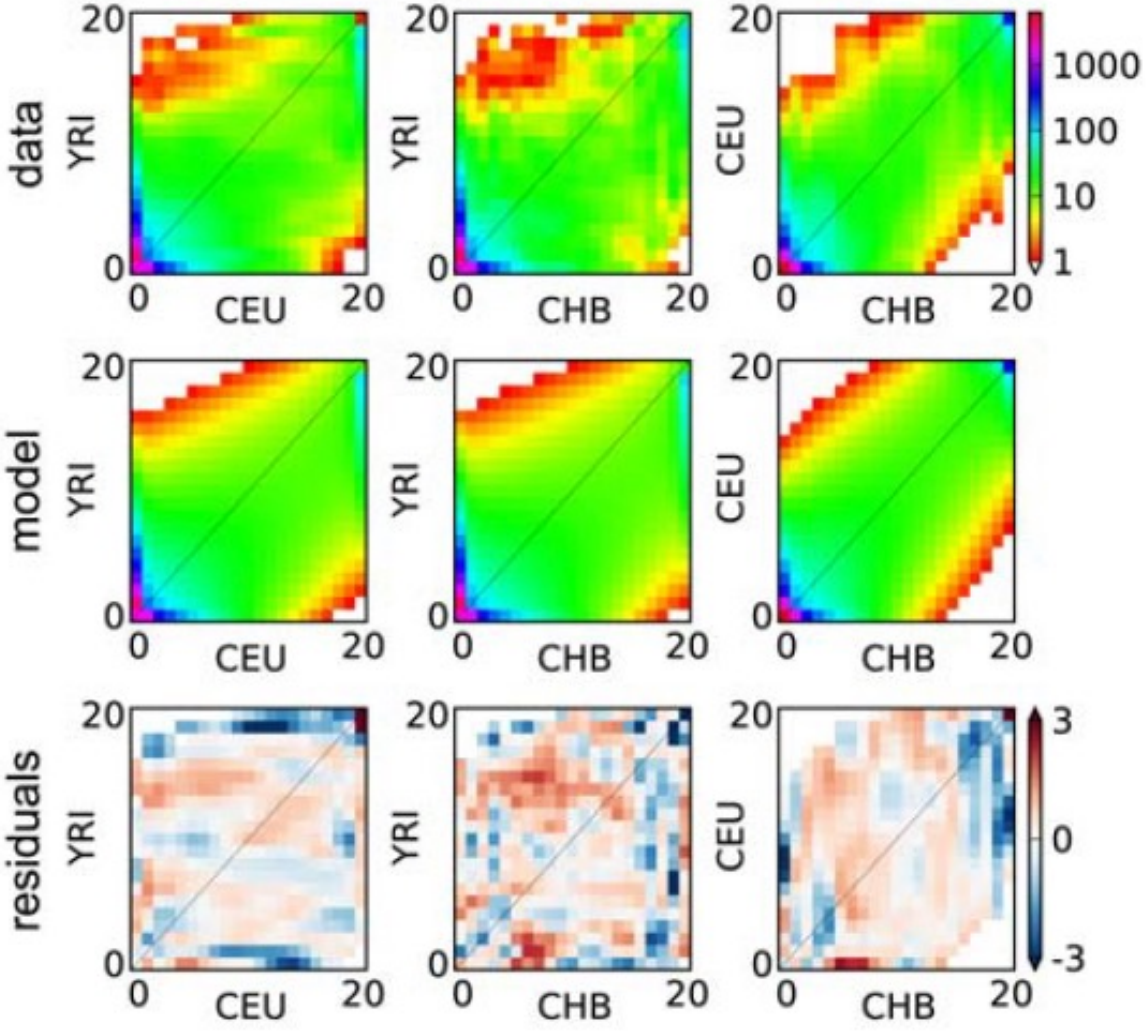


Gutenkunst et al., 2009

Fitting a demographic model to the JSFS

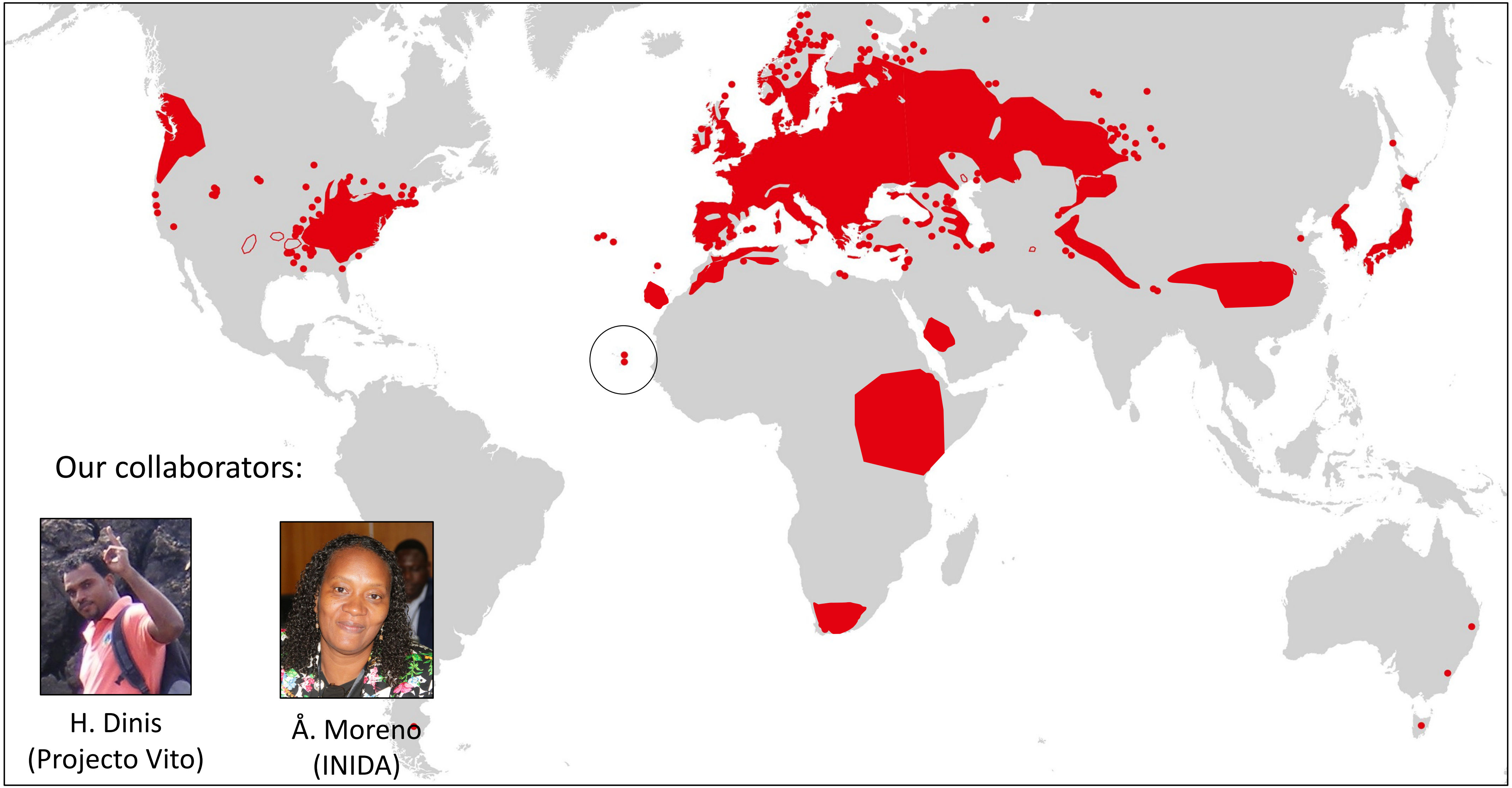
Out of Africa model





Gutenkunst et al., 2009

An example: *Arabidopsis thaliana* from Cape Verde



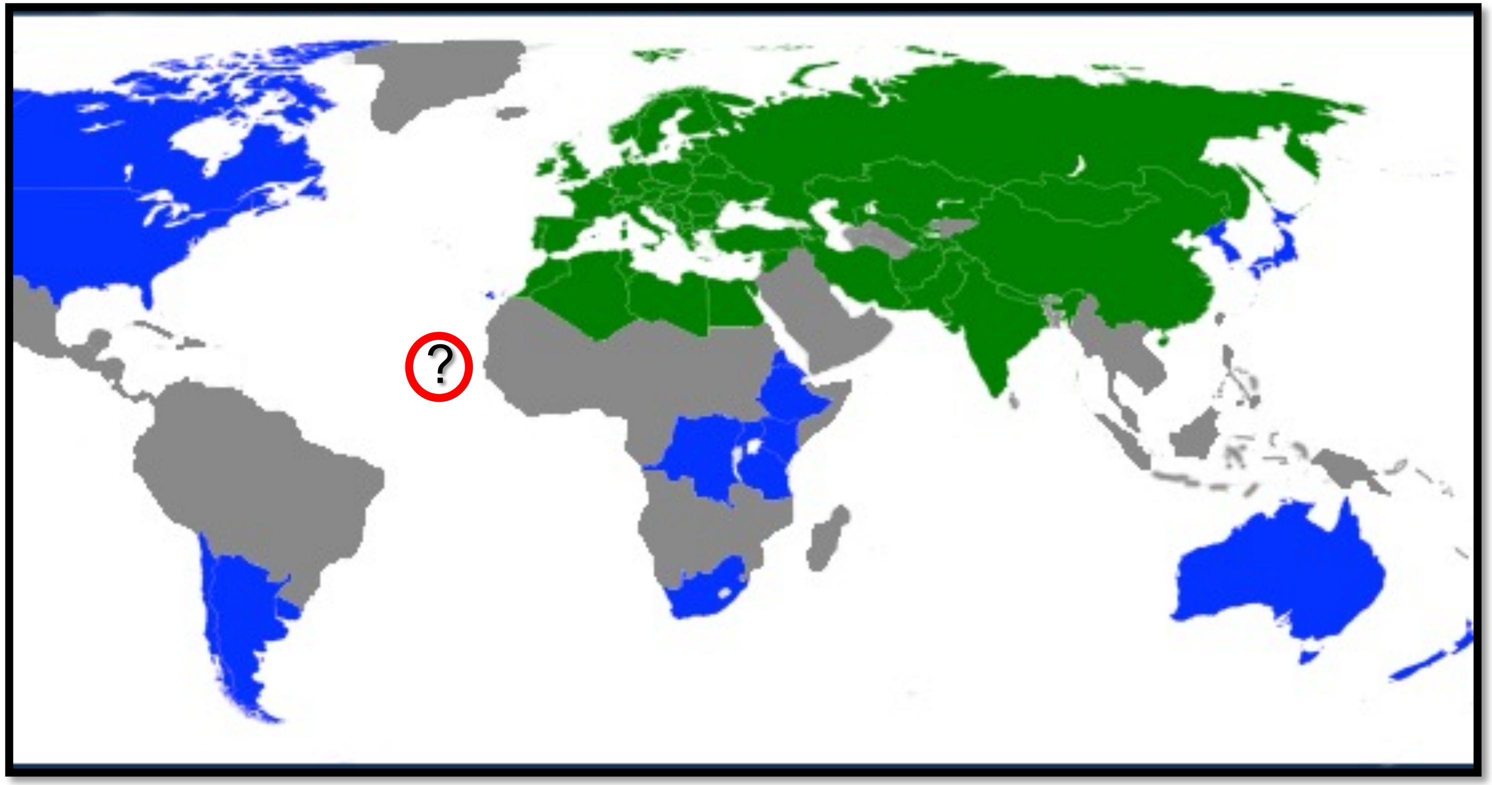
Our collaborators:

H. Dinis
(Projecto Vito)

Å. Moreno
(INIDA)

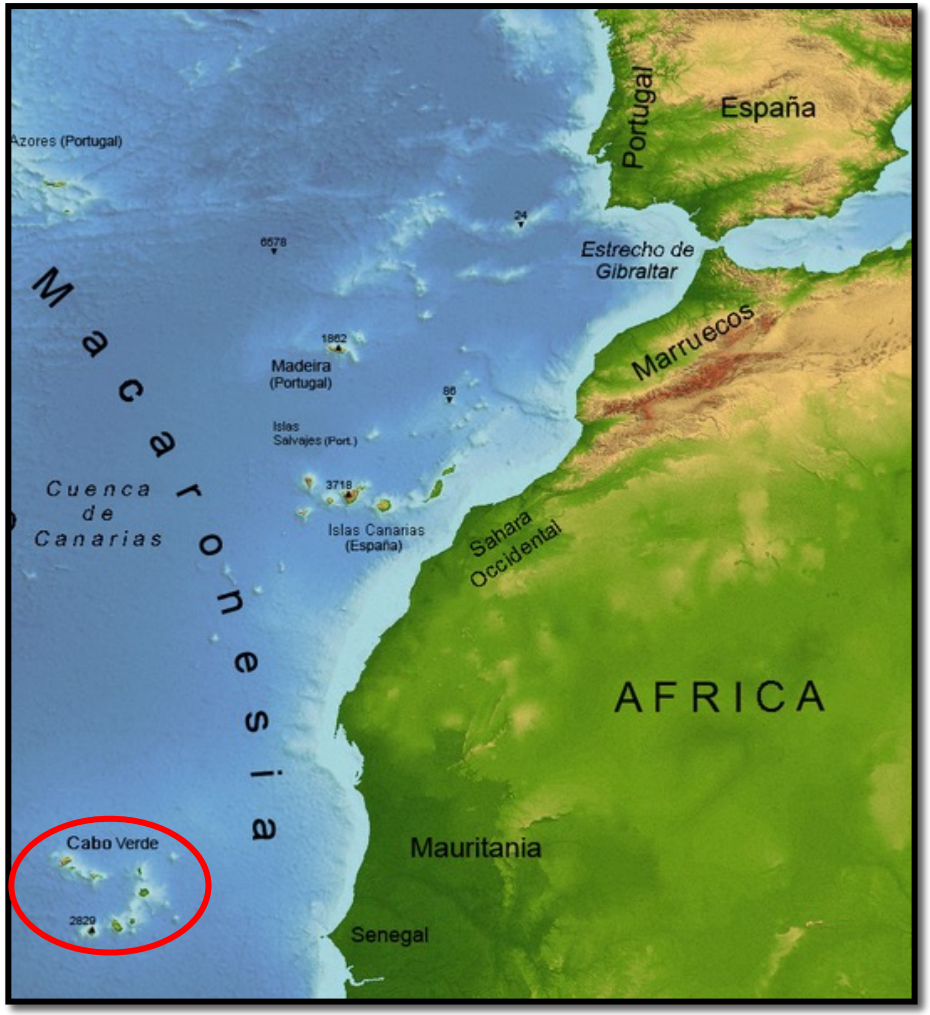
Fulgione, Neto et al., Nature Communications 2022
https://www.nature.com/articles/s41467-022-28800-z

Cvi-0: an enigmatic Arabidopsis accession

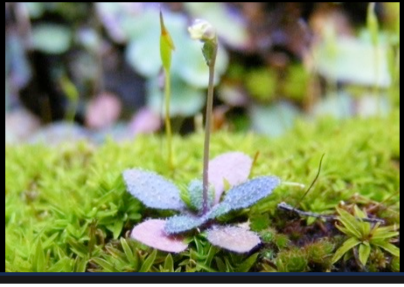


A single Arabidopsis plant (Cvi-0) was collected >35 years ago in the Cape Verde Islands, but it was not clear how it got there

History of the Cape Verde Islands

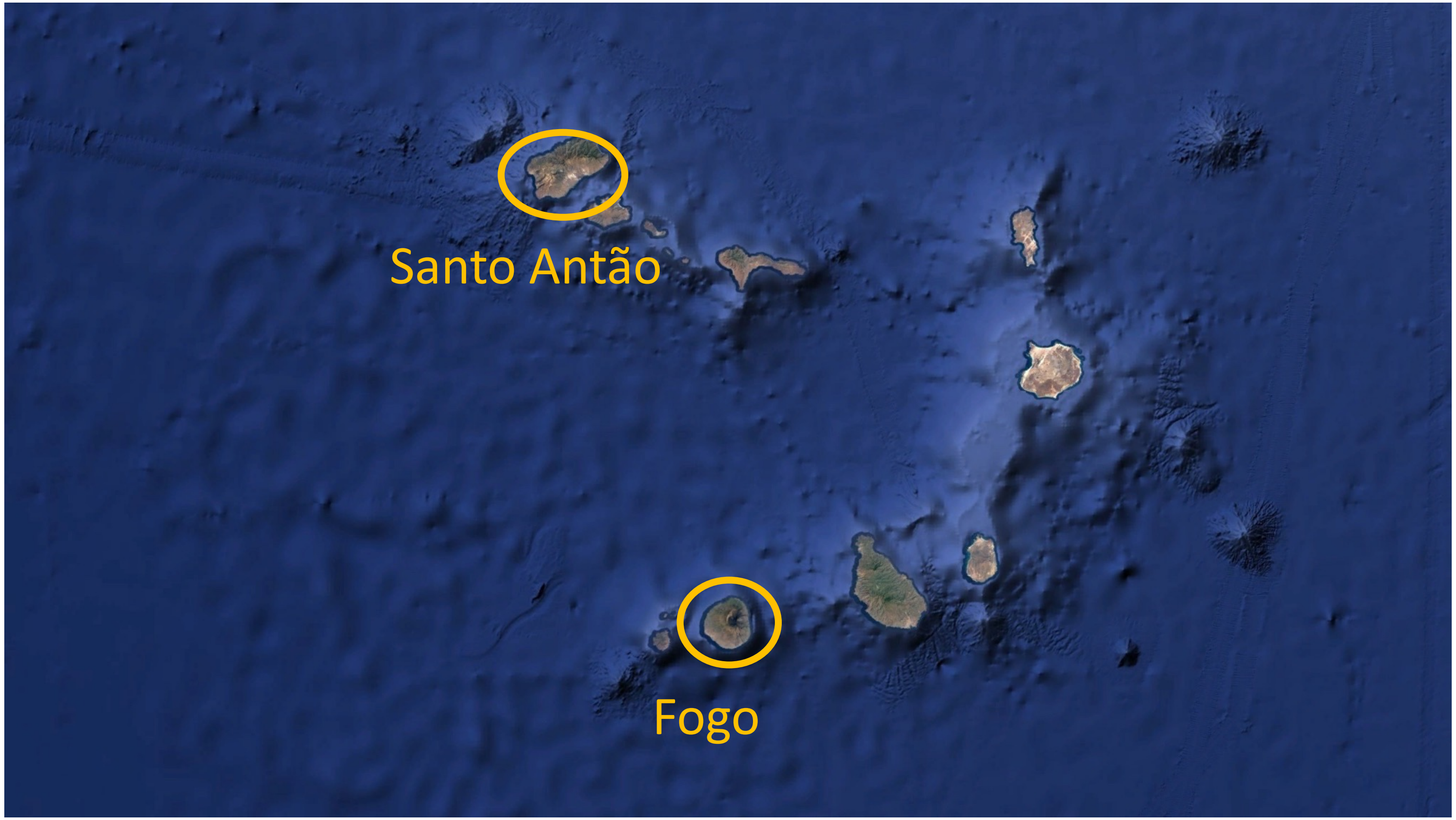


- Colonized by Portuguese in 1460
- Current flora is a mix of endemics and species introduced since colonization
- Main inputs of endemic flora derive from Africa and the Canary Islands

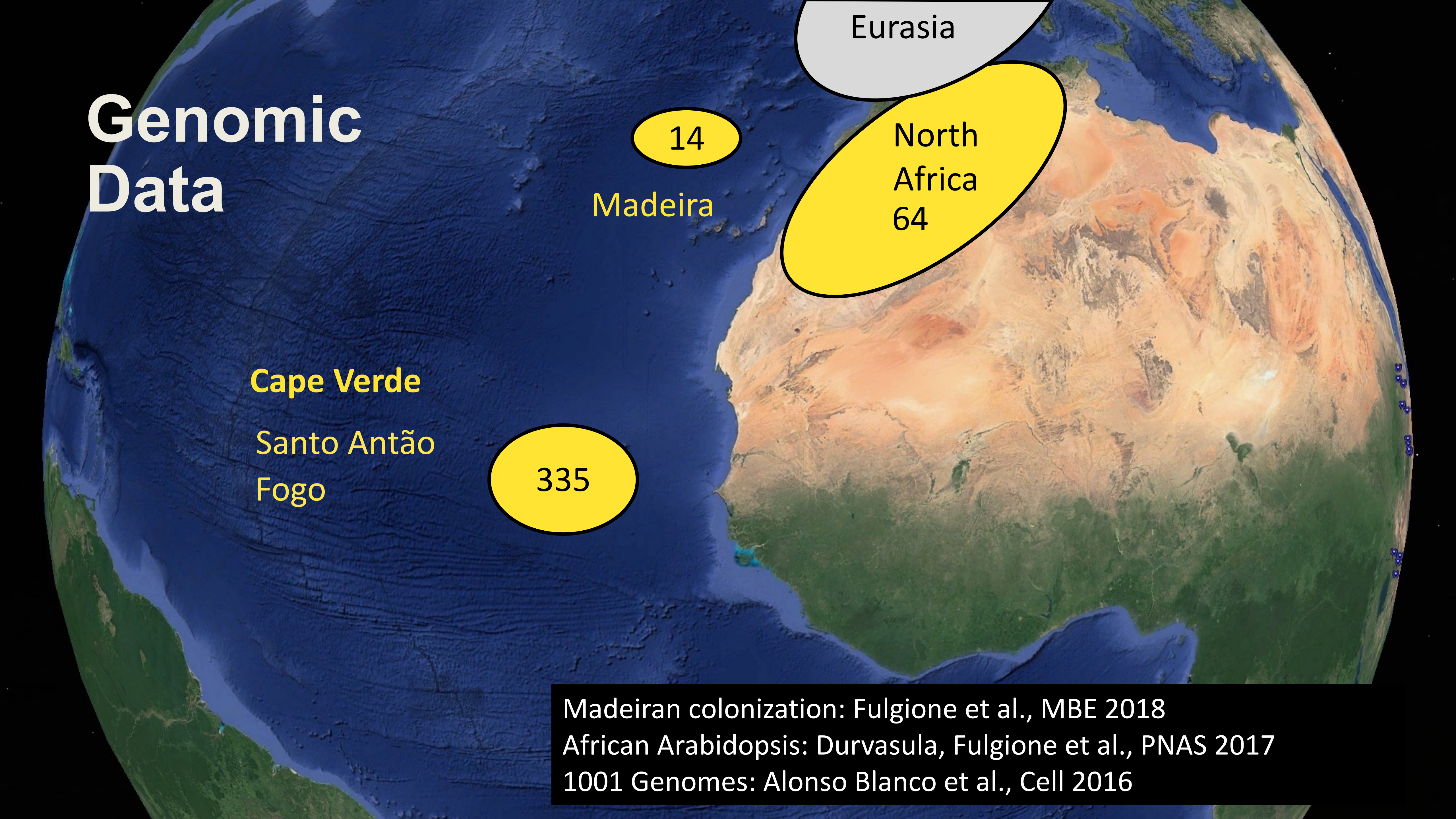


Arabidopsis in Cape Verde

Arabidopsis is present on two islands in Cape Verde



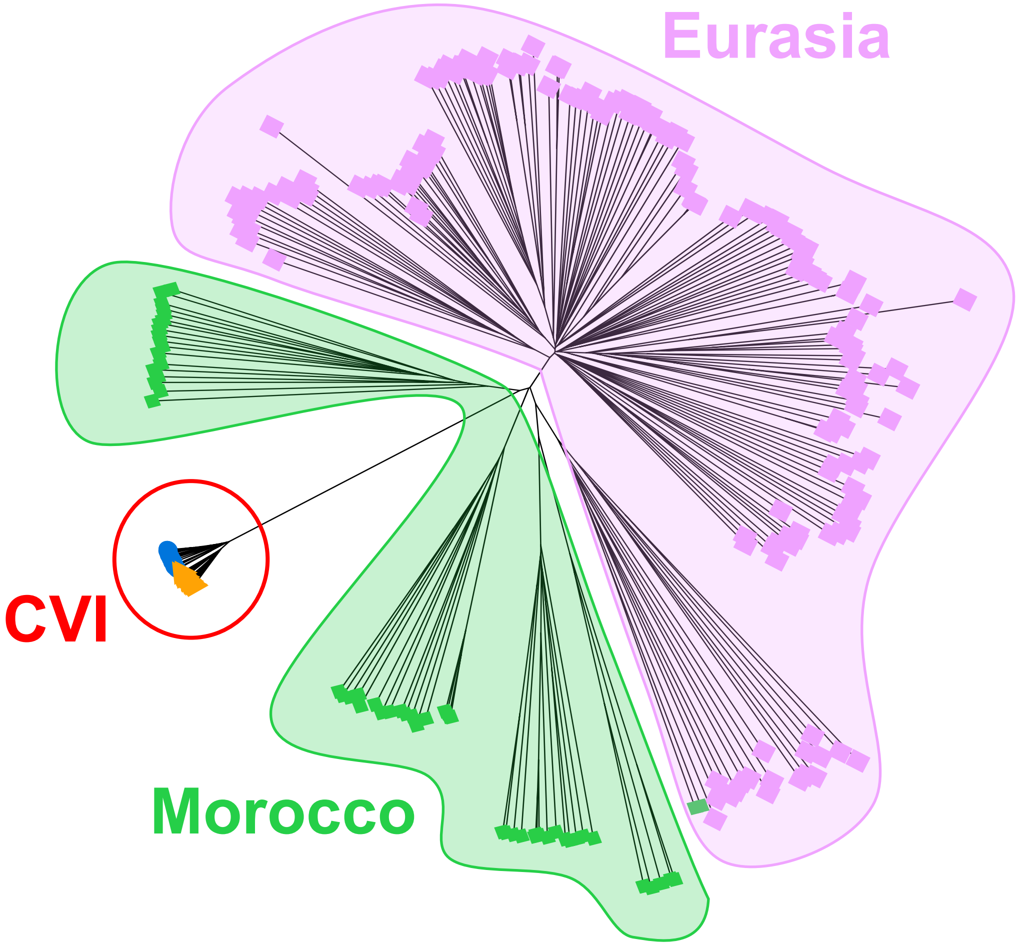
Genomic Data



Madeiran colonization: Fulgione et al., MBE 2018
African Arabidopsis: Durvasula, Fulgione et al., PNAS 2017
1001 Genomes: Alonso Blanco et al., Cell 2016

CVI populations represent a single migration from North Africa

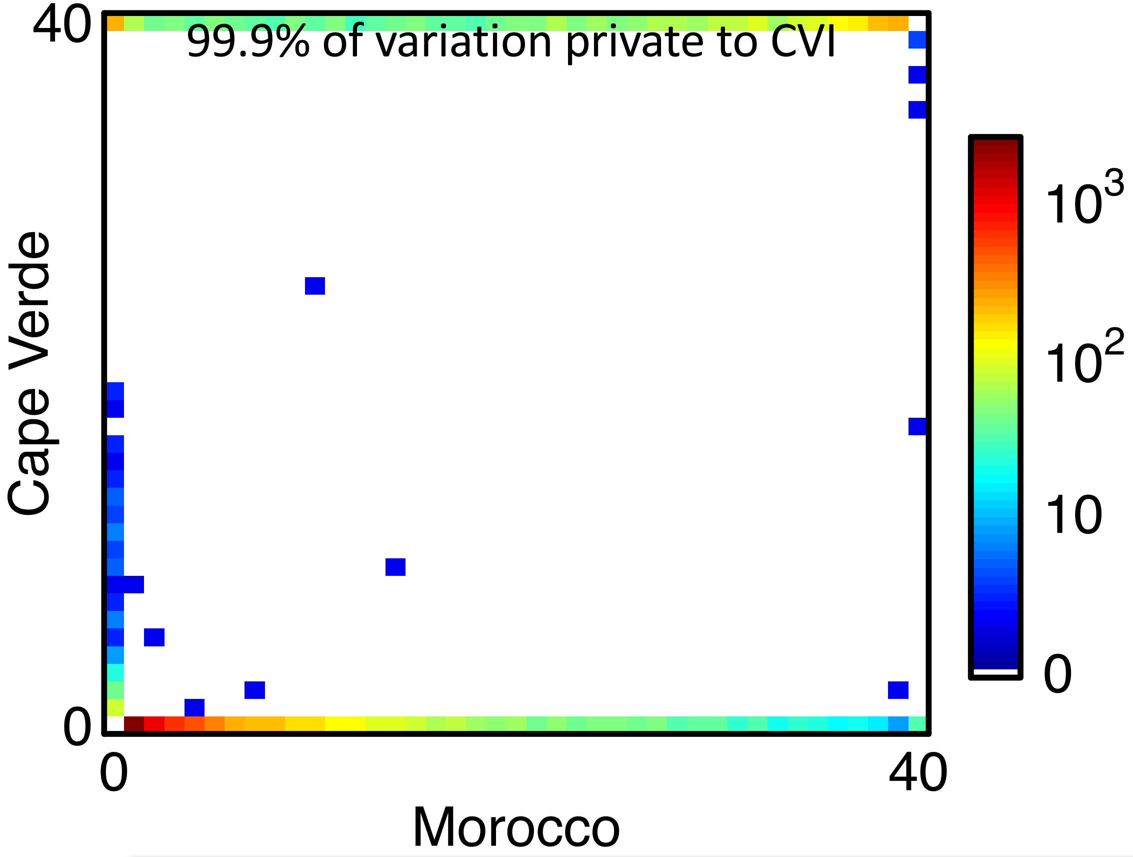
- CVI nested within Moroccan clade
- Divergence to Morocco is shared between islands
- Diversity in CVI is low
  - ○ Morocco θ_W = 5.56 x10^{-3}
  - ○ Santo Antao θ_W = 7.59x10^{-5}
  - ○ Fogo θ_W = 8.93x10^{-5}



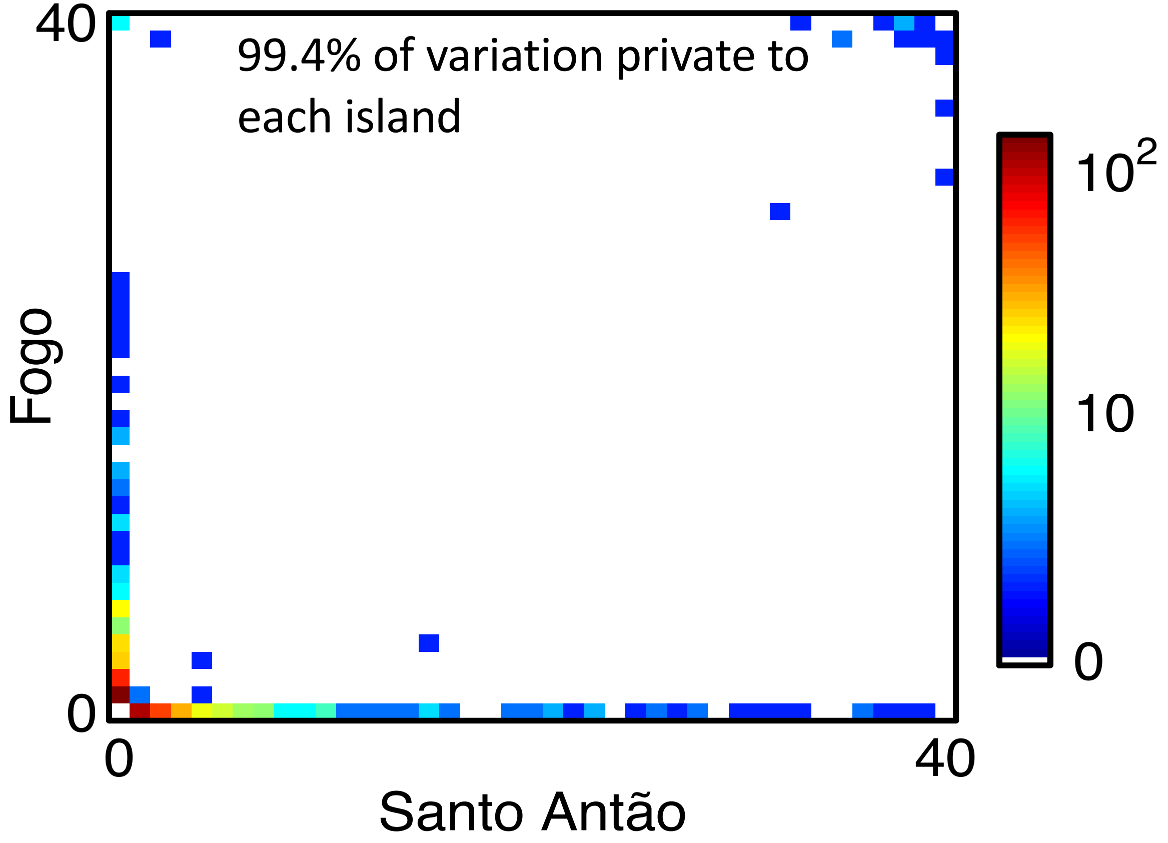
Andrea Fulgione

Fulgione, Neto et al., Nature Communications 2022
https://www.nature.com/articles/s41467-022-28800-z

CVI lineages are phylogenetically distinct



99.9% of variation private to CVI



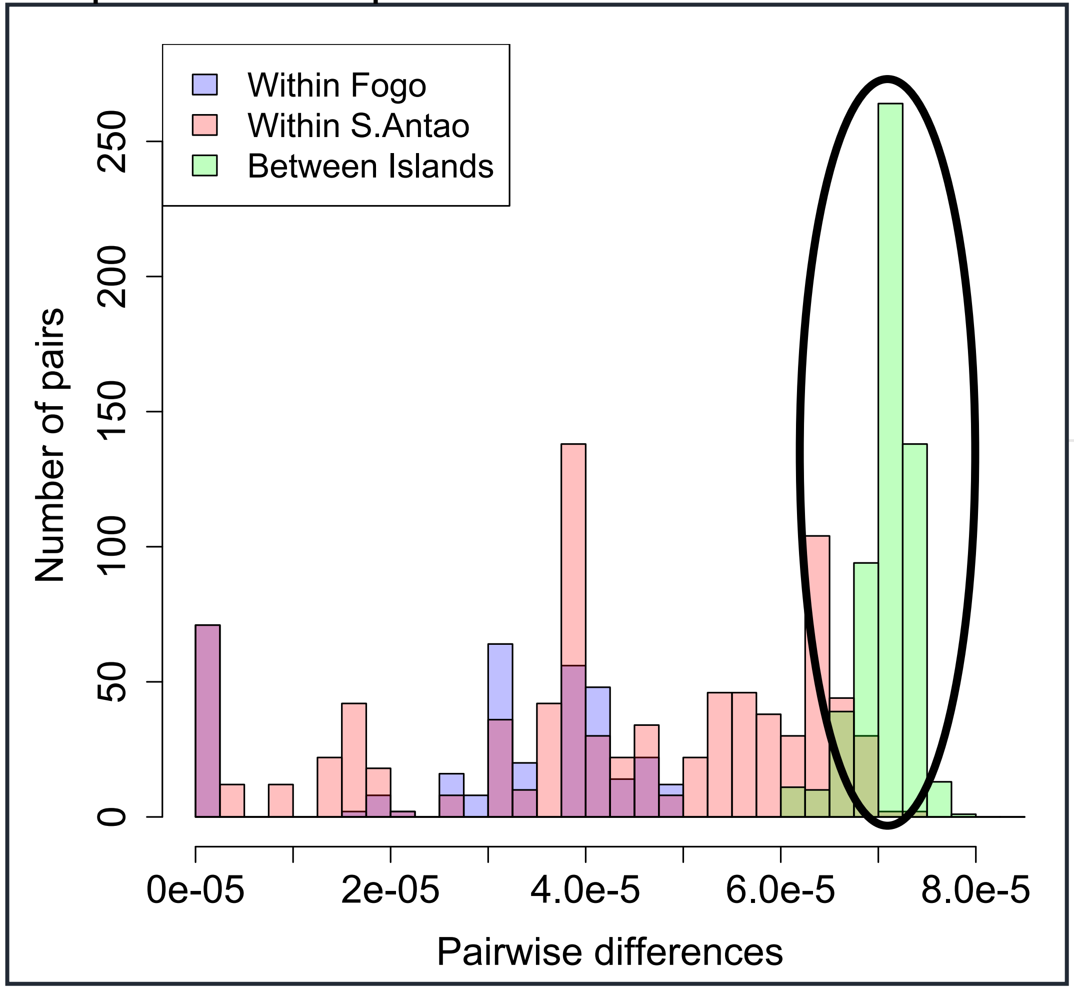
99.4% of variation private to each island

The patterns imply strong colonization bottlenecks with no subsequent migration

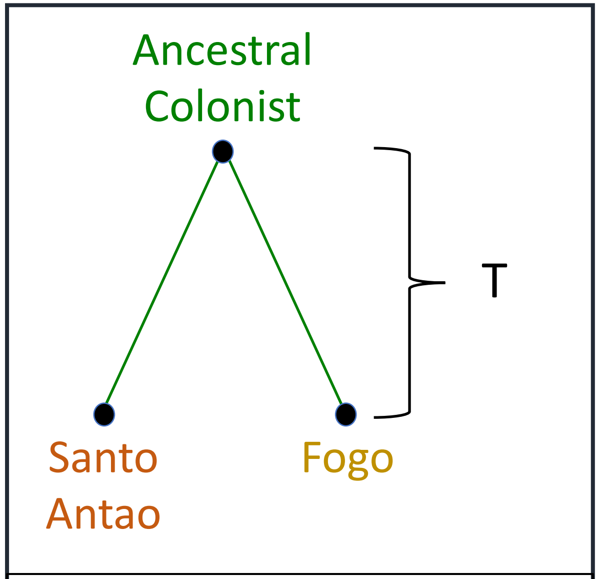
Fulgione, Neto et al., Nature Communications 2022
https://www.nature.com/articles/s41467-022-28800-z

Split time based on mean pairwise divergence across samples

Mismatch distributions:
pairwise comparisons between individuals



Based on a molecular clock (i.e., constant rate of mutation over time), we can estimate the split between islands

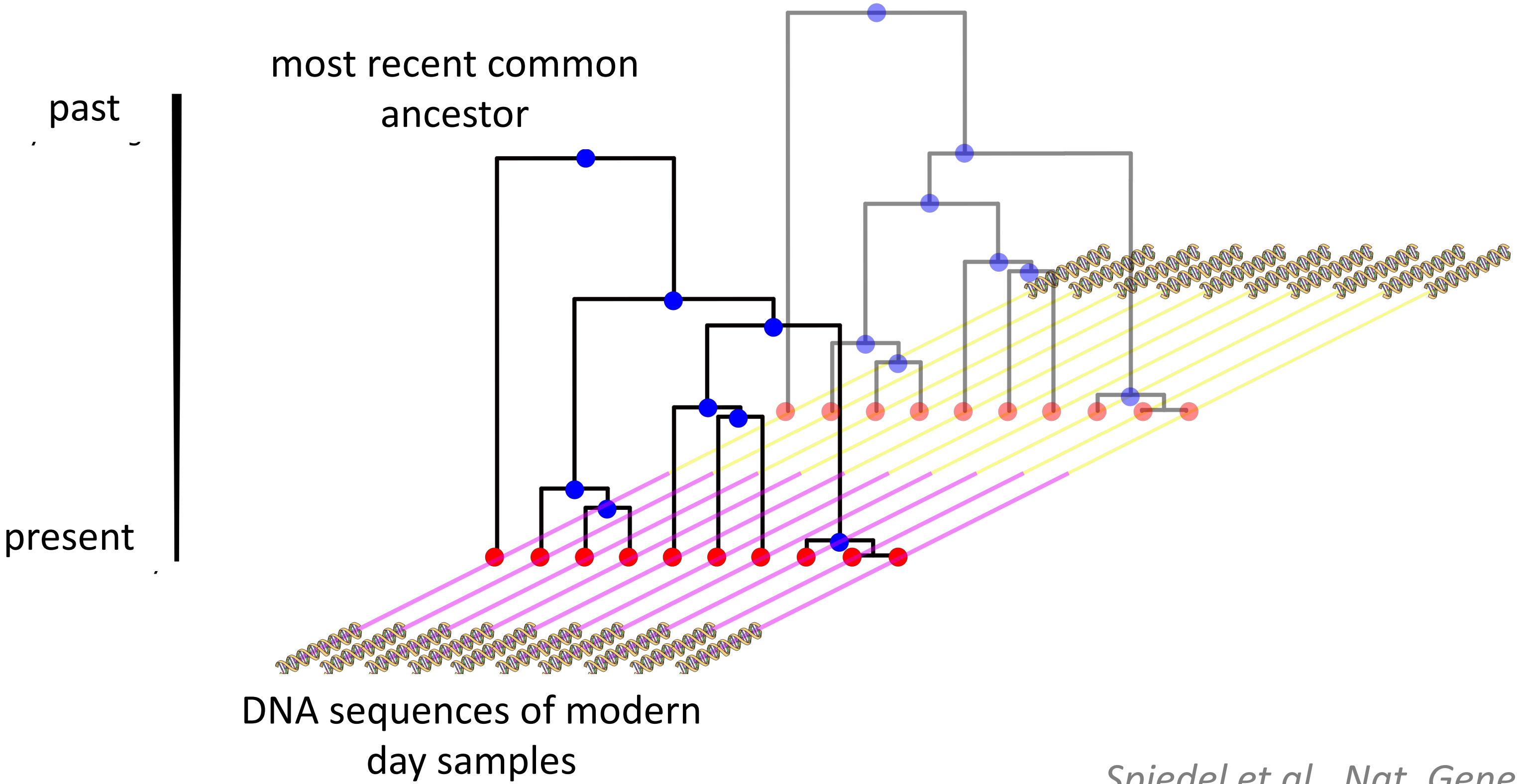


$D = 2T\mu$

$T \approx 5250$ gens

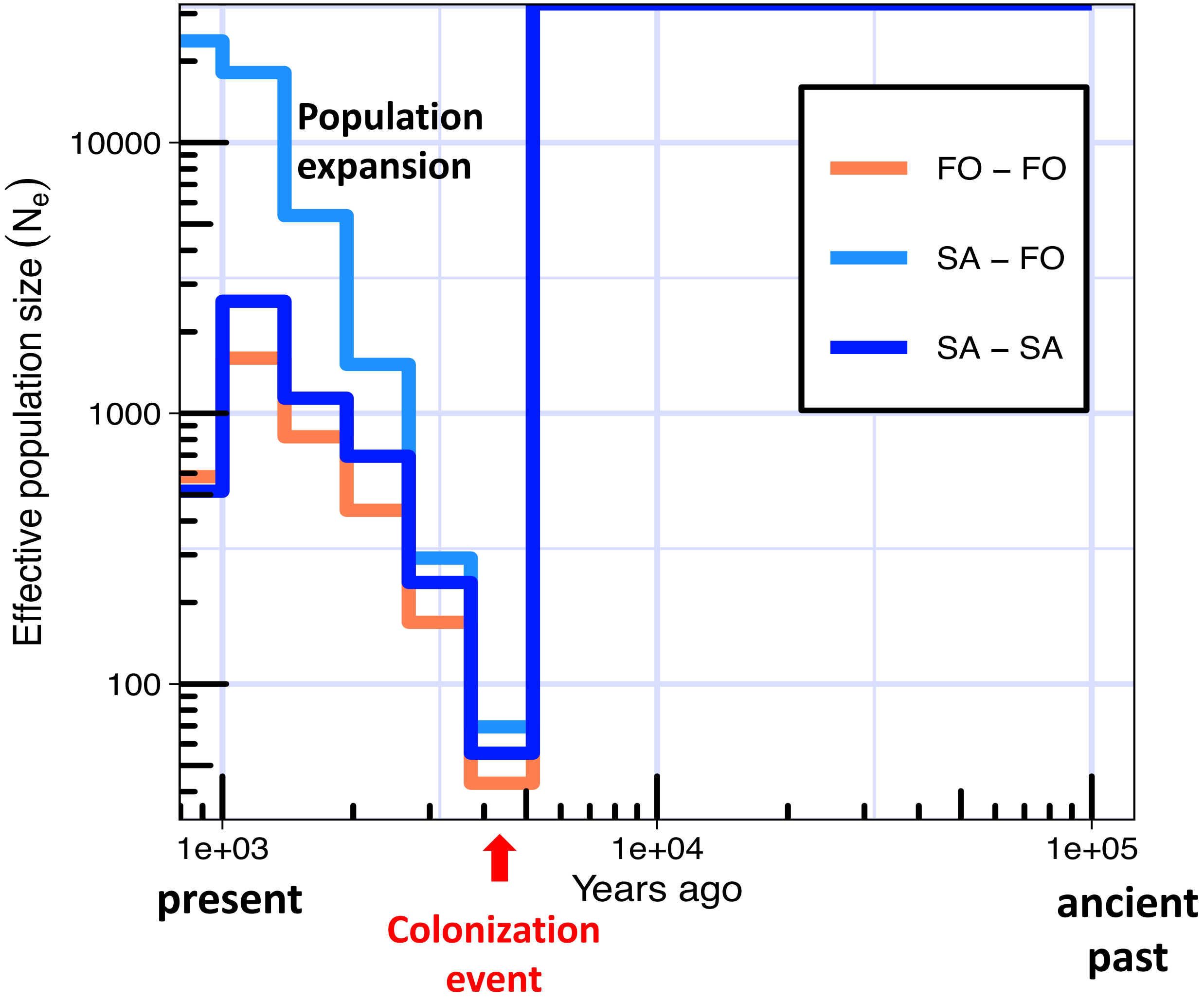
Inferring population history from sequence data

ARG-based methods use information from across the genome to infer coalescence times between chromosomes


most recent common ancestor
past
present
DNA sequences of modern day samples

Spiedel et al., Nat. Genet 2019

Split time based on the distributions of coalescence times across the genome



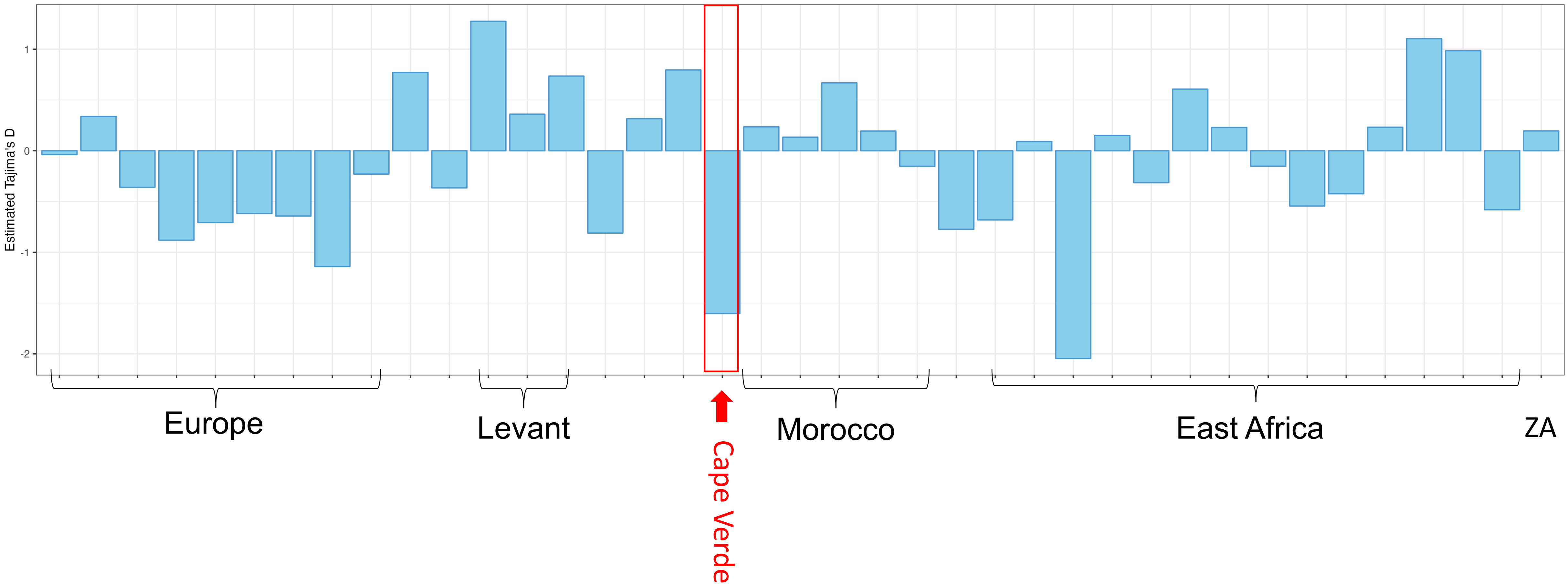
Given the rapid population expansion, would you expect Tajima's D to be positive or negative?

Ahmed Elfarargi

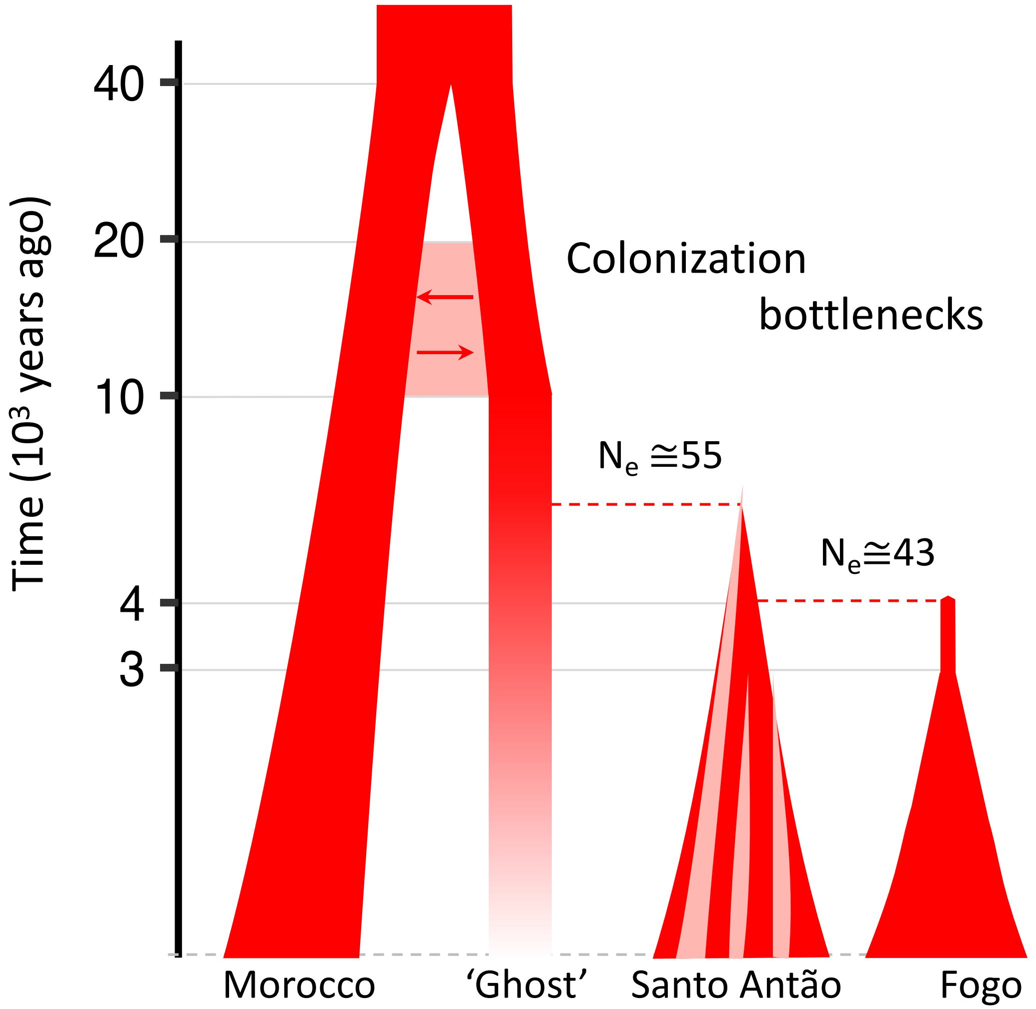
Fulgione, Neto* et al., Nat. Comm., 2022*

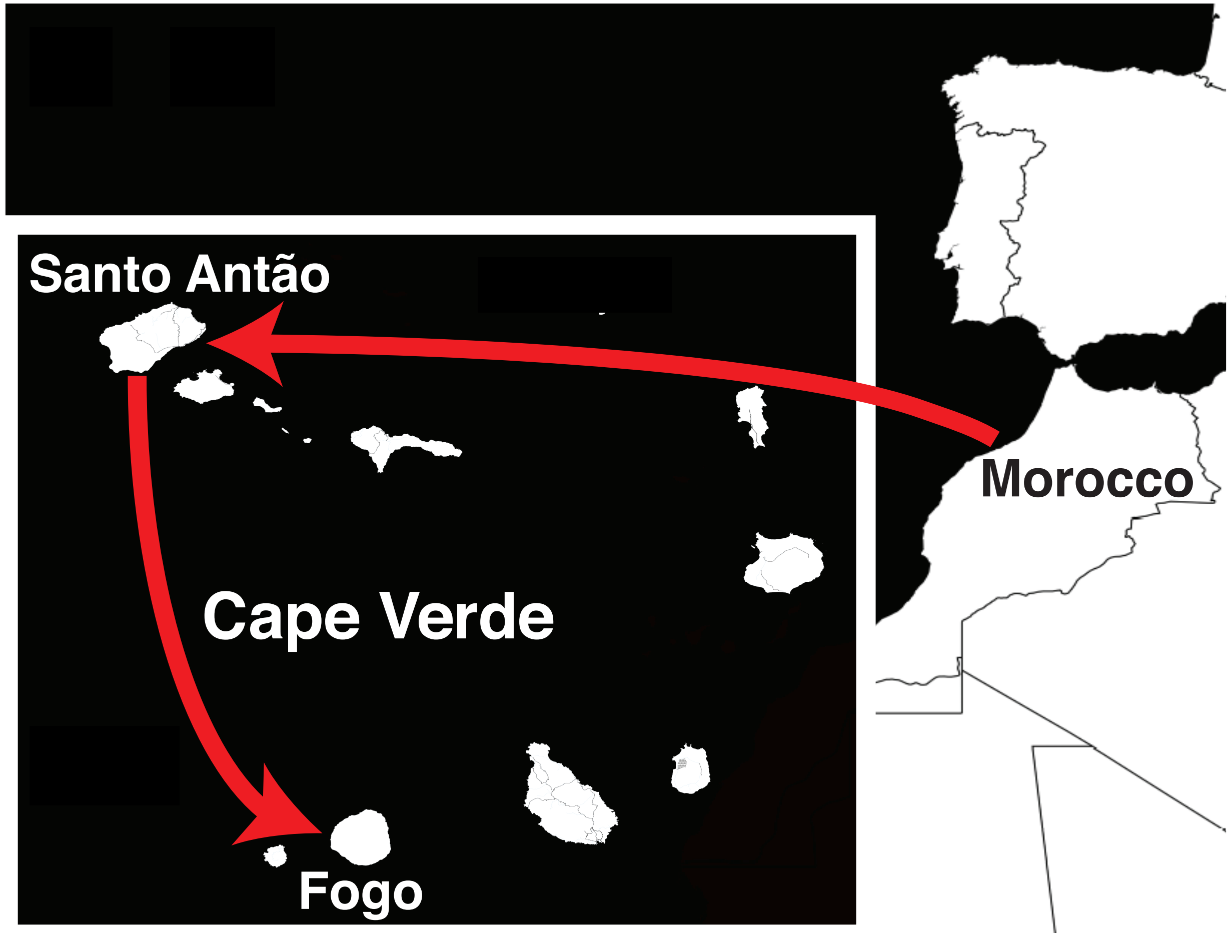
Tajima’s D across *Arabidopsis* populations

Cape Verde has a very negative Tajima’s D



Overall picture: CVI islands were colonized approximately 5 kya through a natural event





Fulgione, Neto* et al., Nat. Comm., 2022*

SFS Summary

- The site frequency spectrum (SFS) is a histogram of allele frequencies within a sample. It summarizes the count of alleles at each frequency in the sample.
- In a randomly mating population, under neutrality and constant population size, θ/i is the expected number of sites at which the derived allele is present in *i* copies. Note that this does *not* depend on sample size
- Genome-wide departures from this model imply something about the history of the population as a whole
- Locus-specific departures from this model imply selection specific to that locus
- Tajima's *D* is a statistic that allows you to compare different aspects of the frequency spectrum (different estimates of θ) to determine whether there is a departure from the model
- The Joint Site Frequency Spectrum (JSFS) summarizes the degree of sharing between two populations. It can be used to infer historical split times and migration rates.