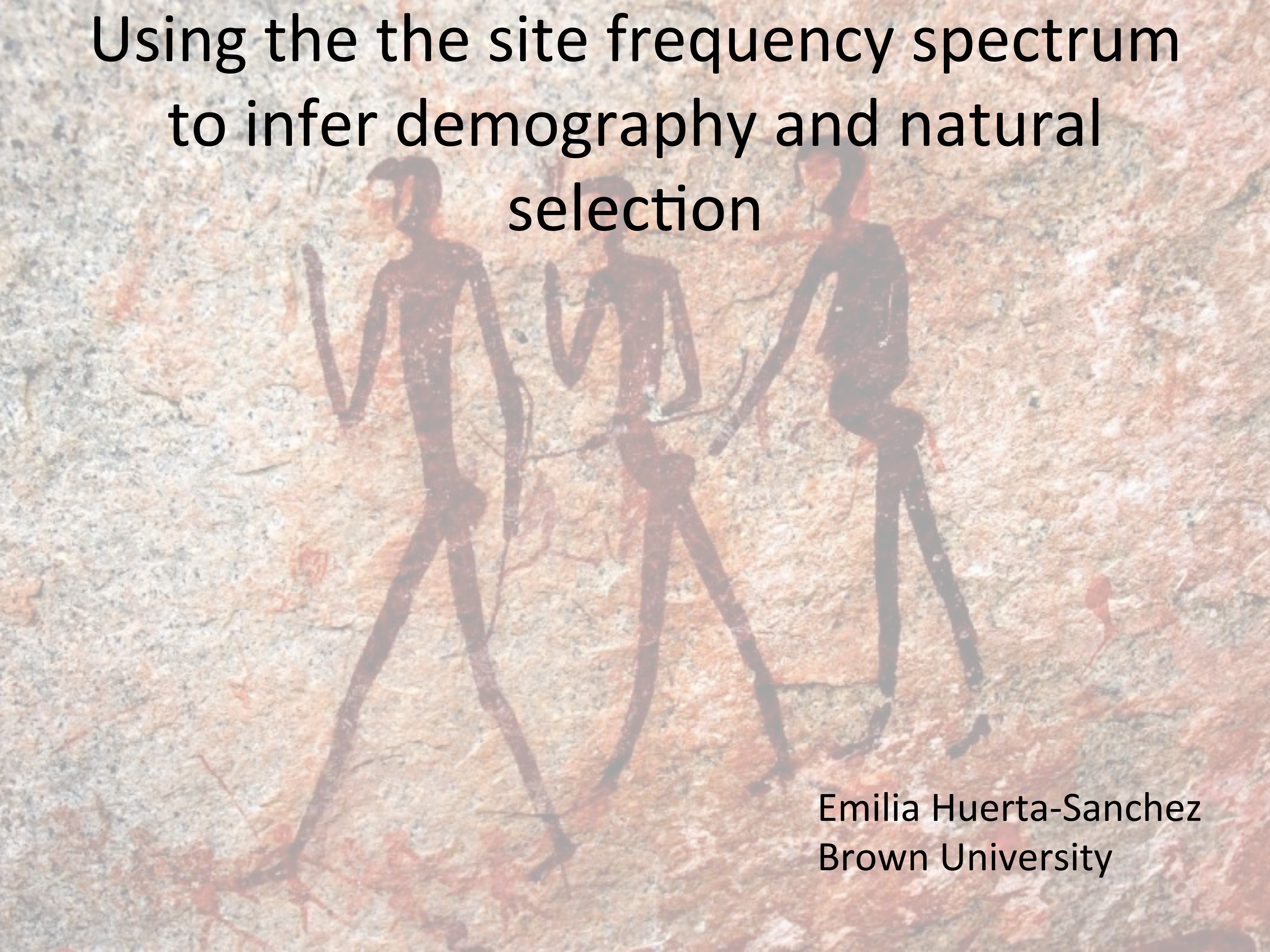
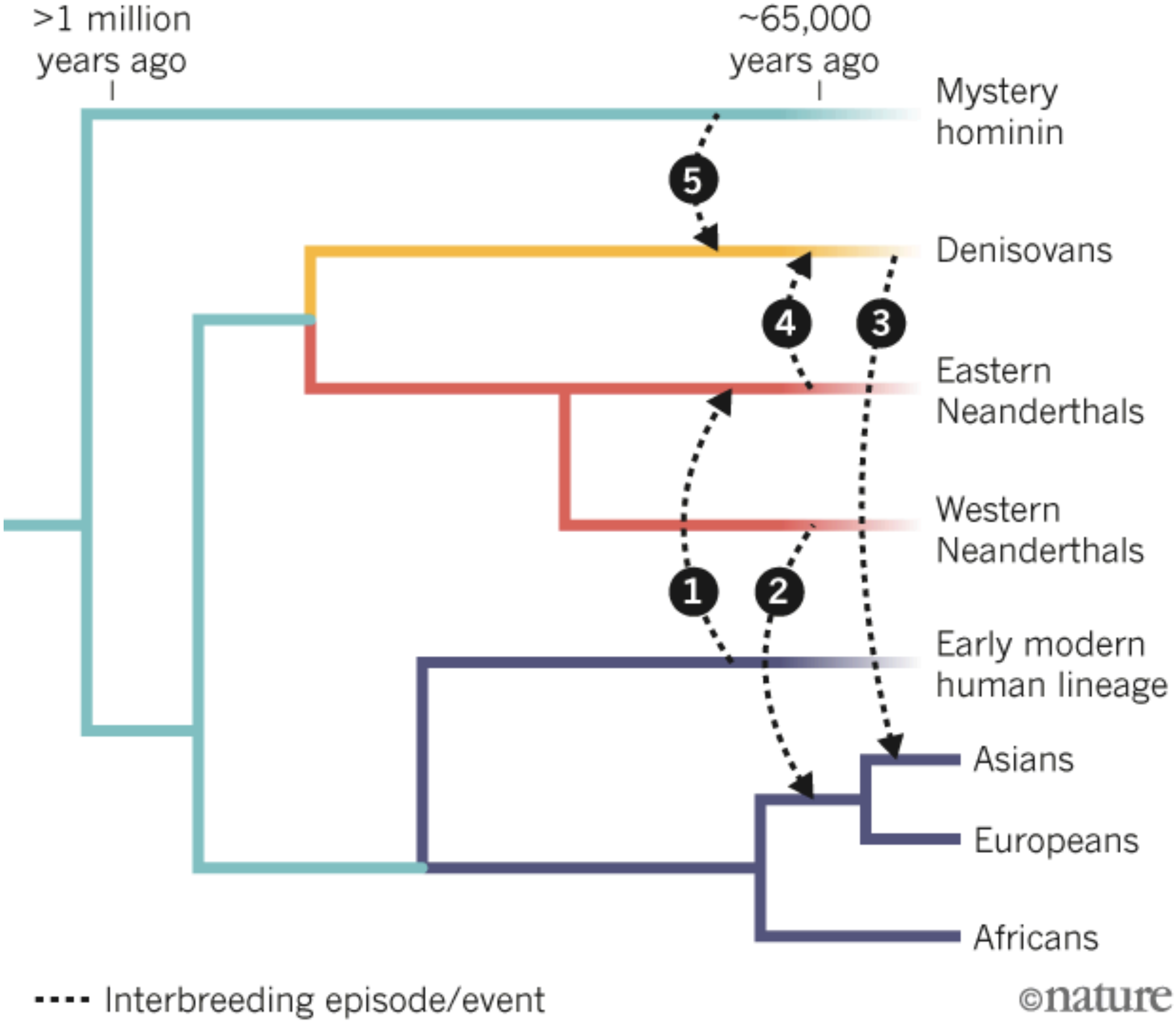
Using the the site frequency spectrum to infer demography and natural selection



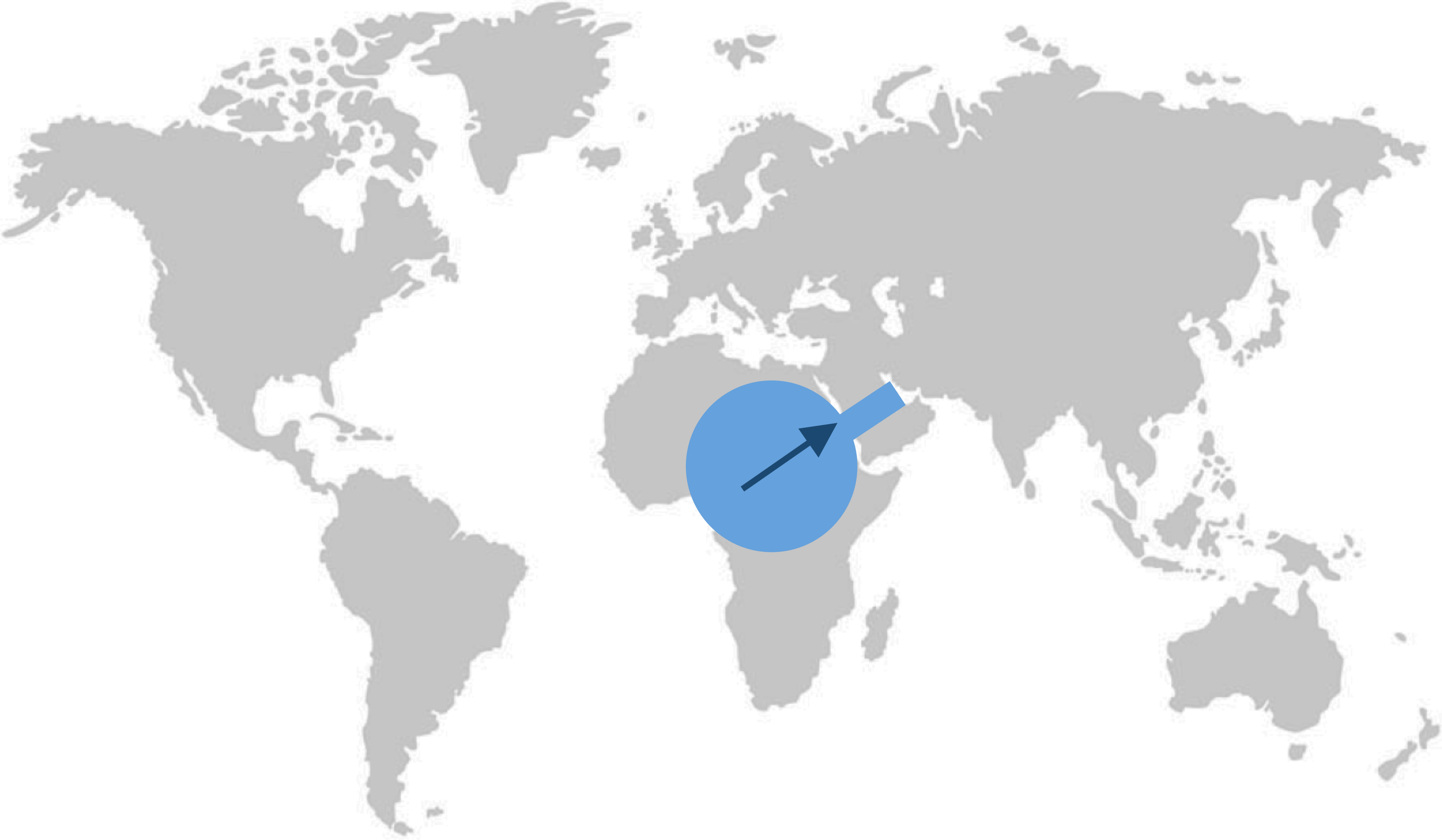
Emilia Huerta-Sanchez
Brown University

Recent time scale

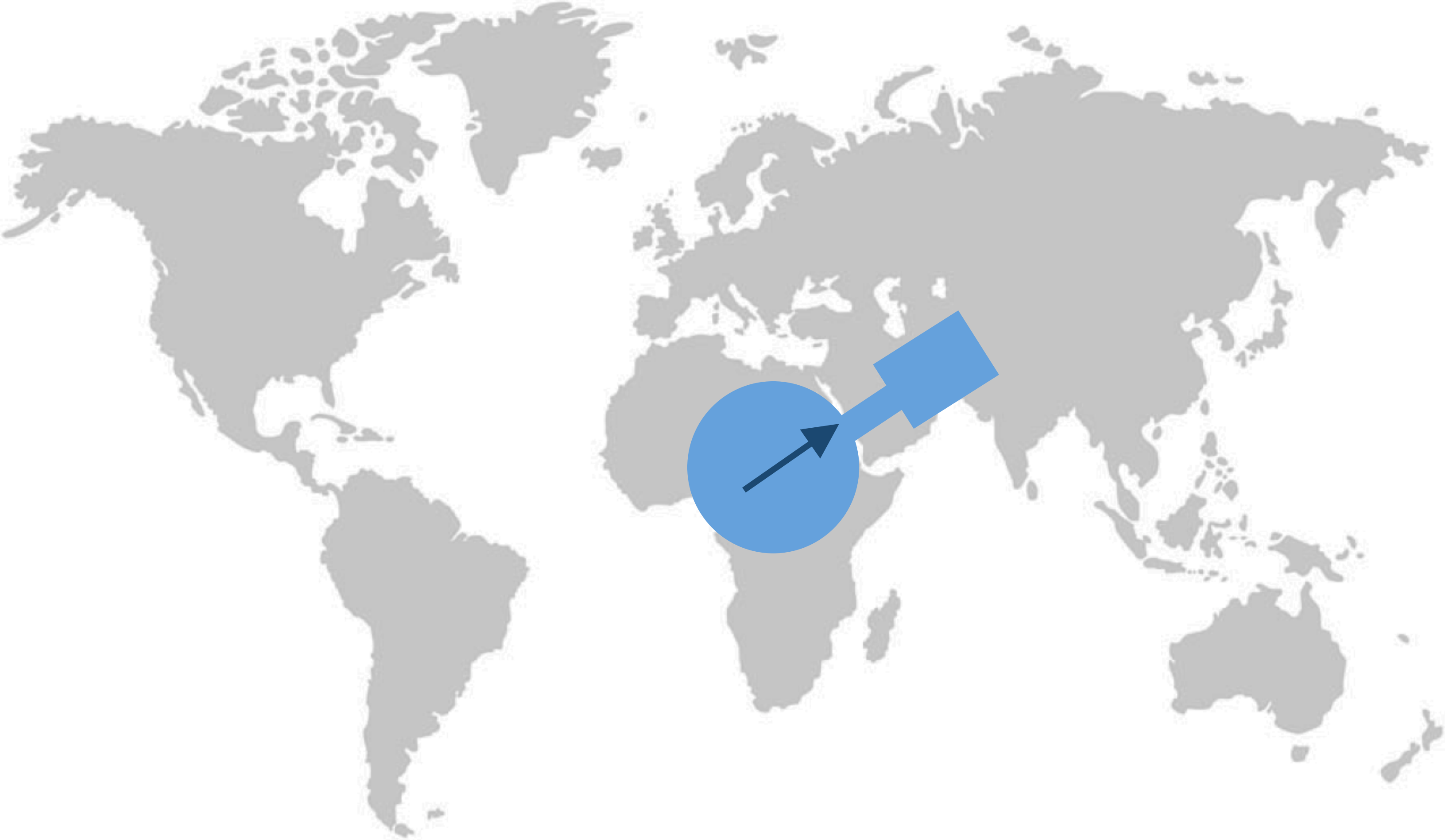


Population dispersal

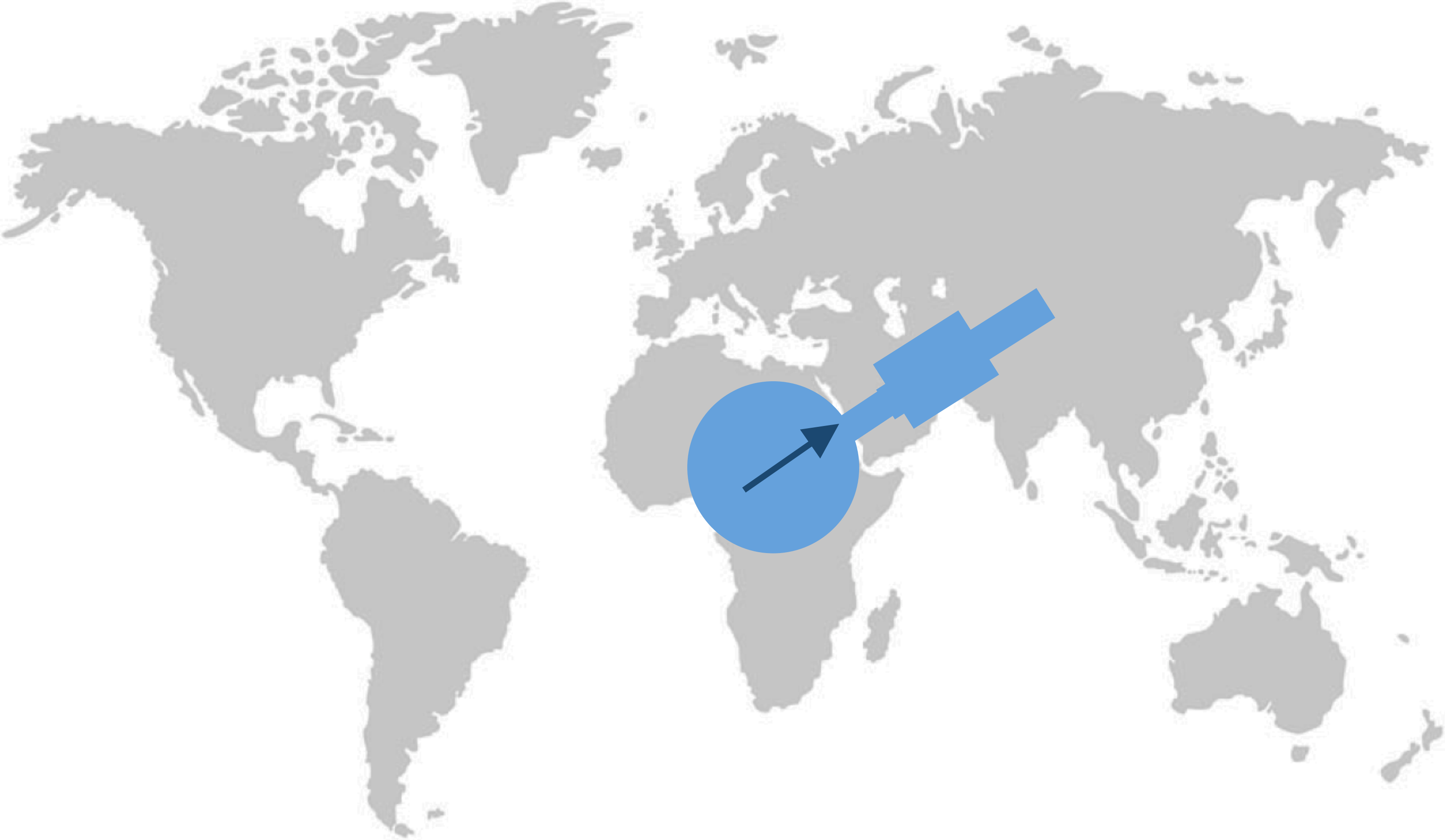
Population dispersal



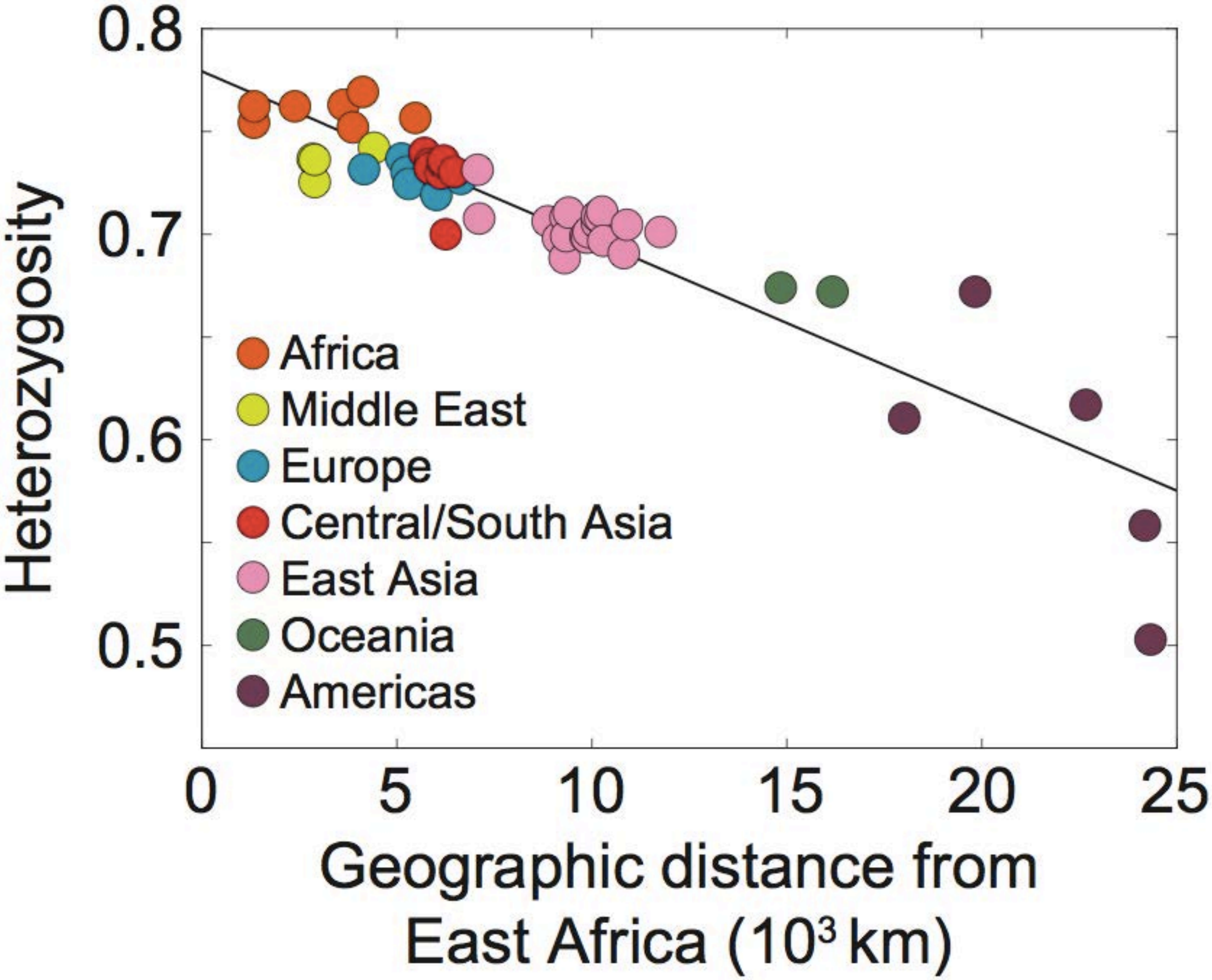
Population dispersal



Population dispersal

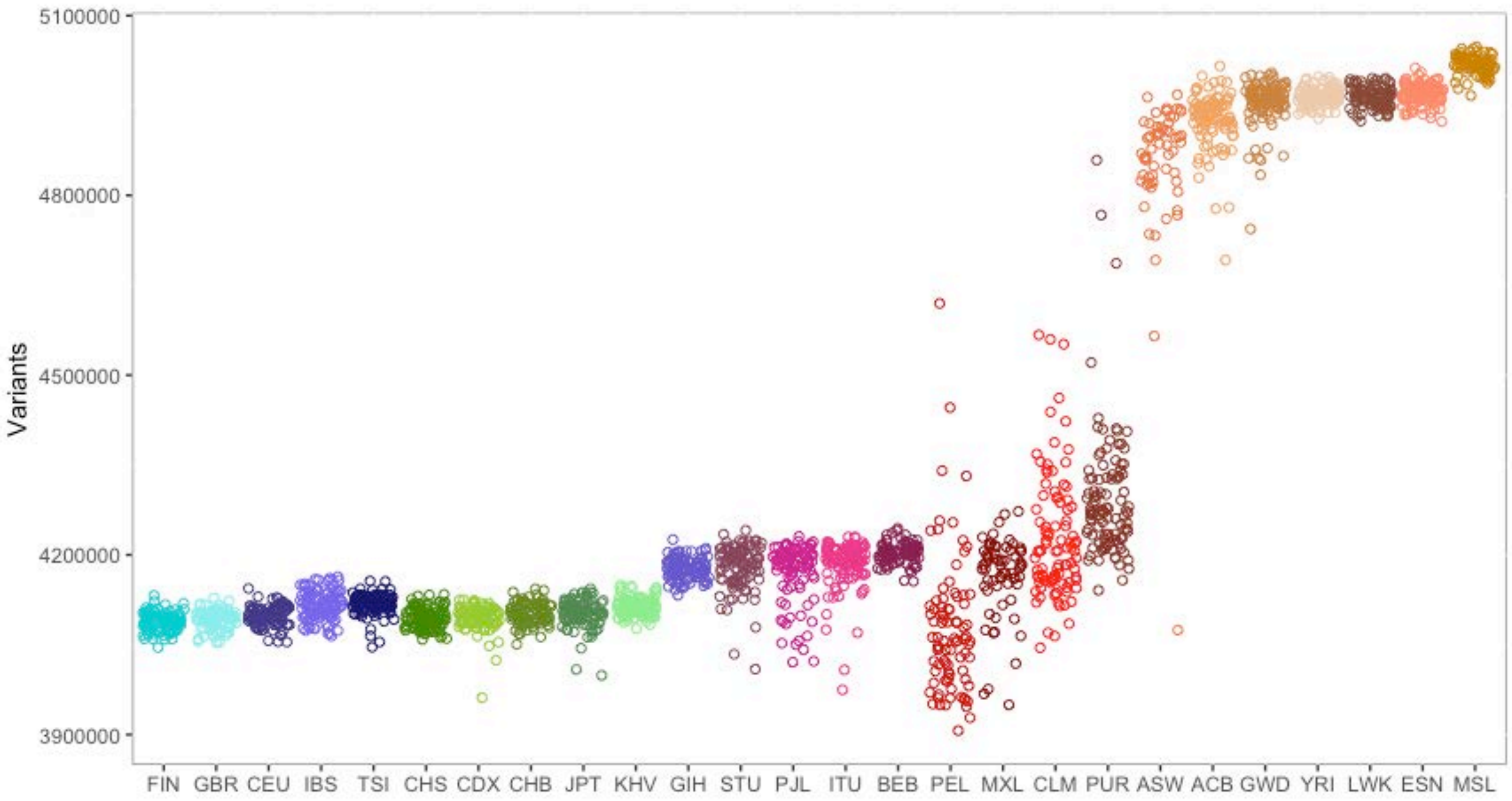


Heterozygosity



DeGiorgio et al. 2009

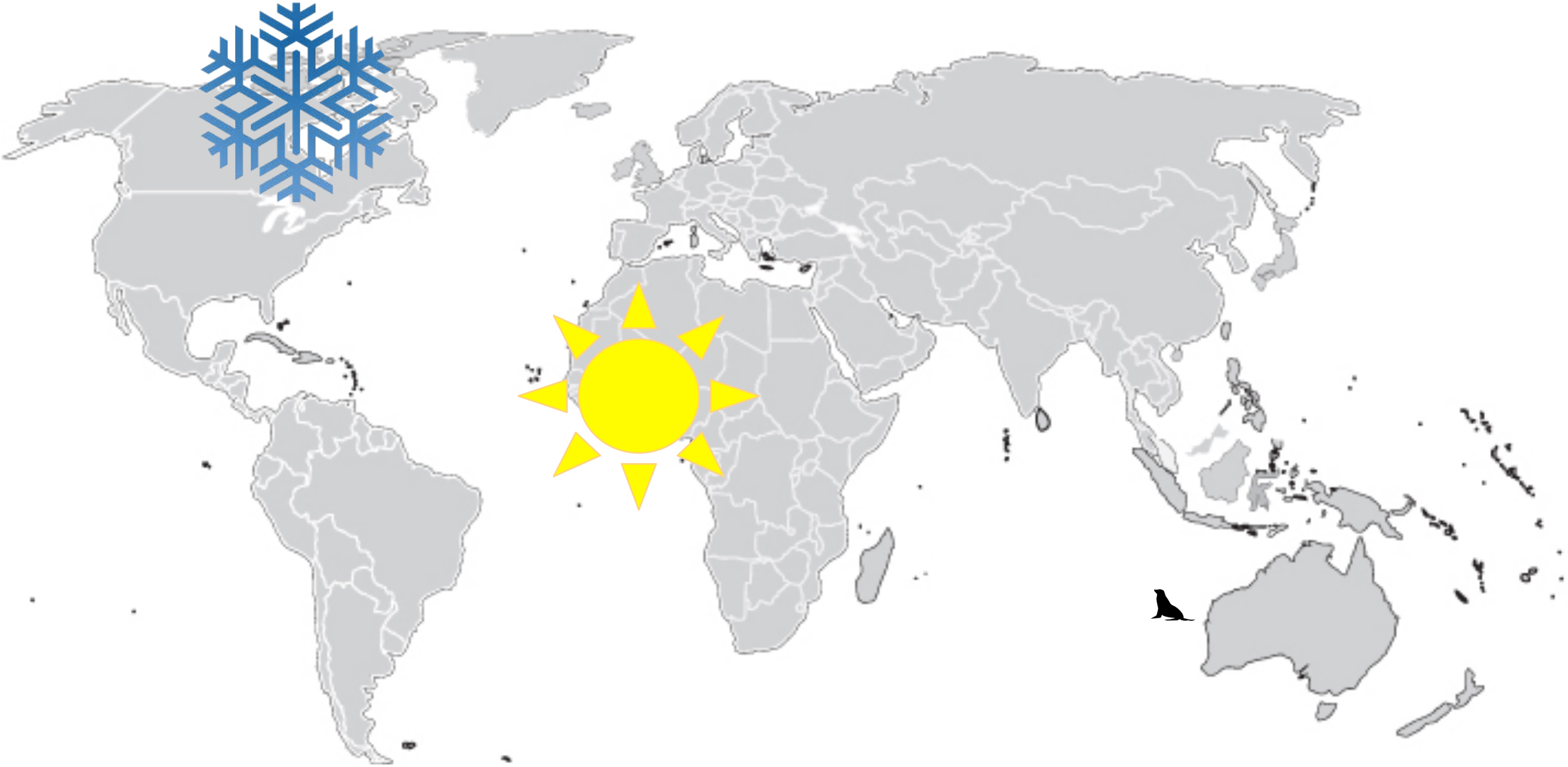
Number of variants per individual



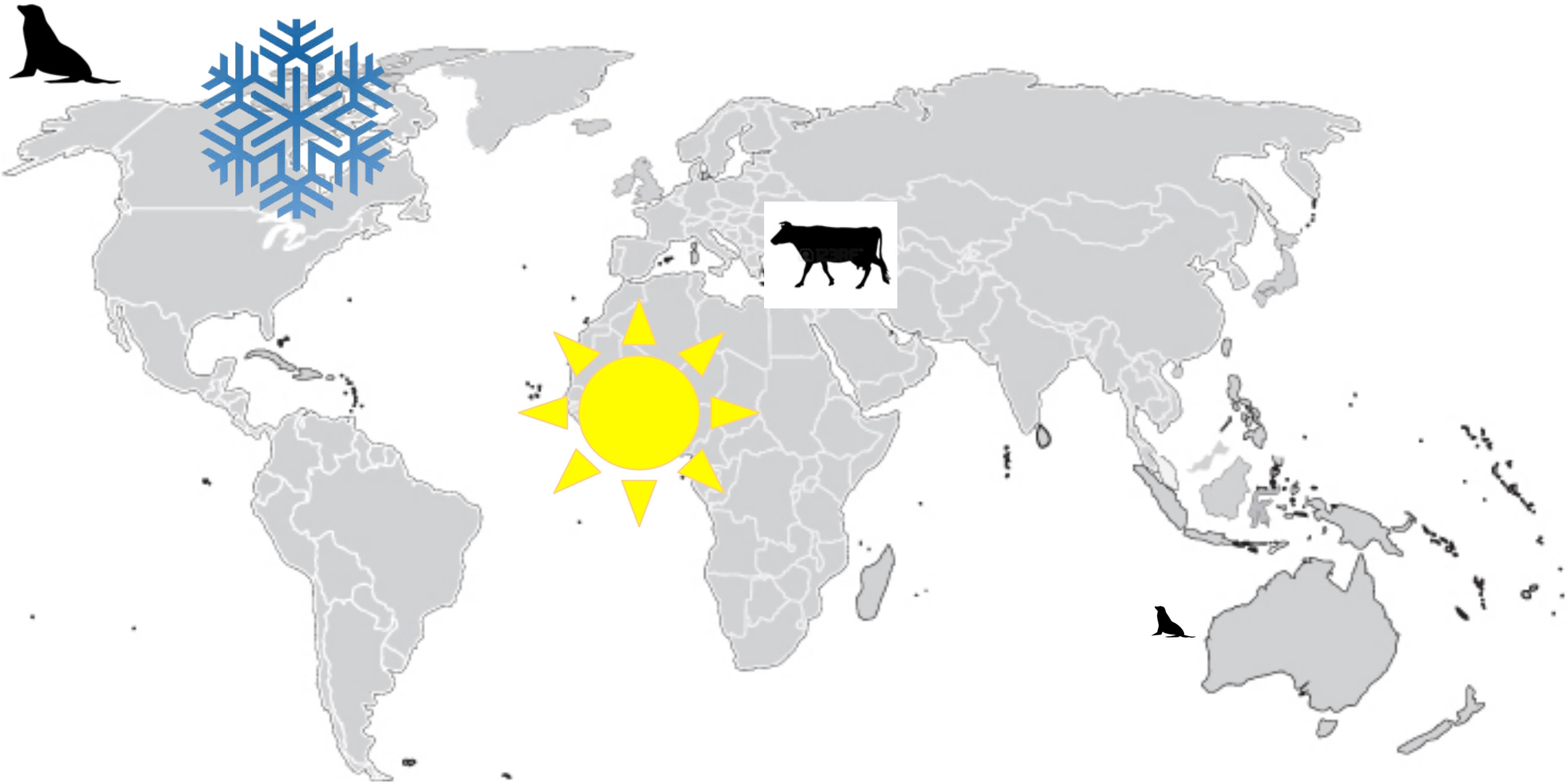
1000 Genomes Project, Nature (2015)

Positive natural selection

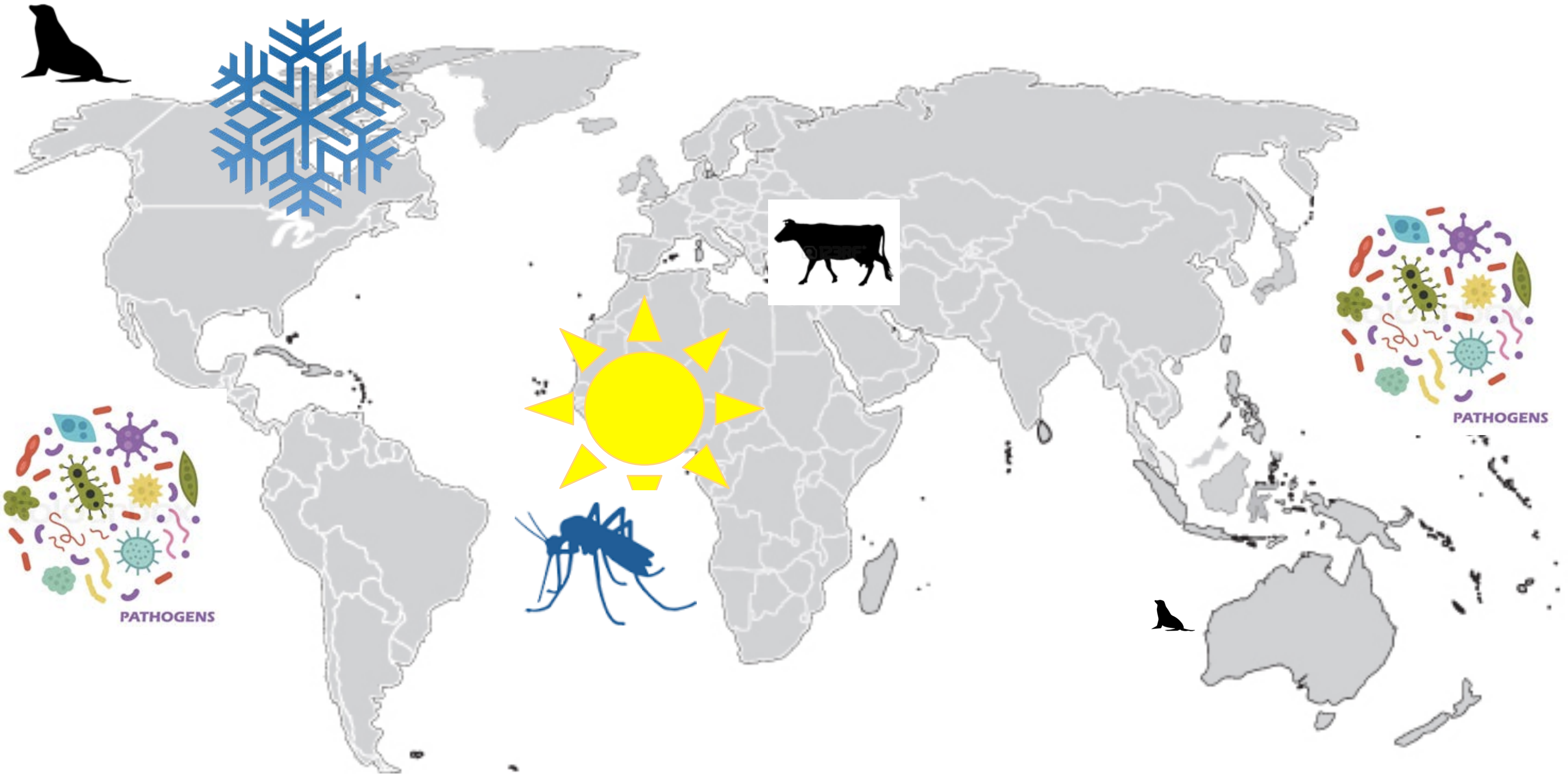
Positive natural selection



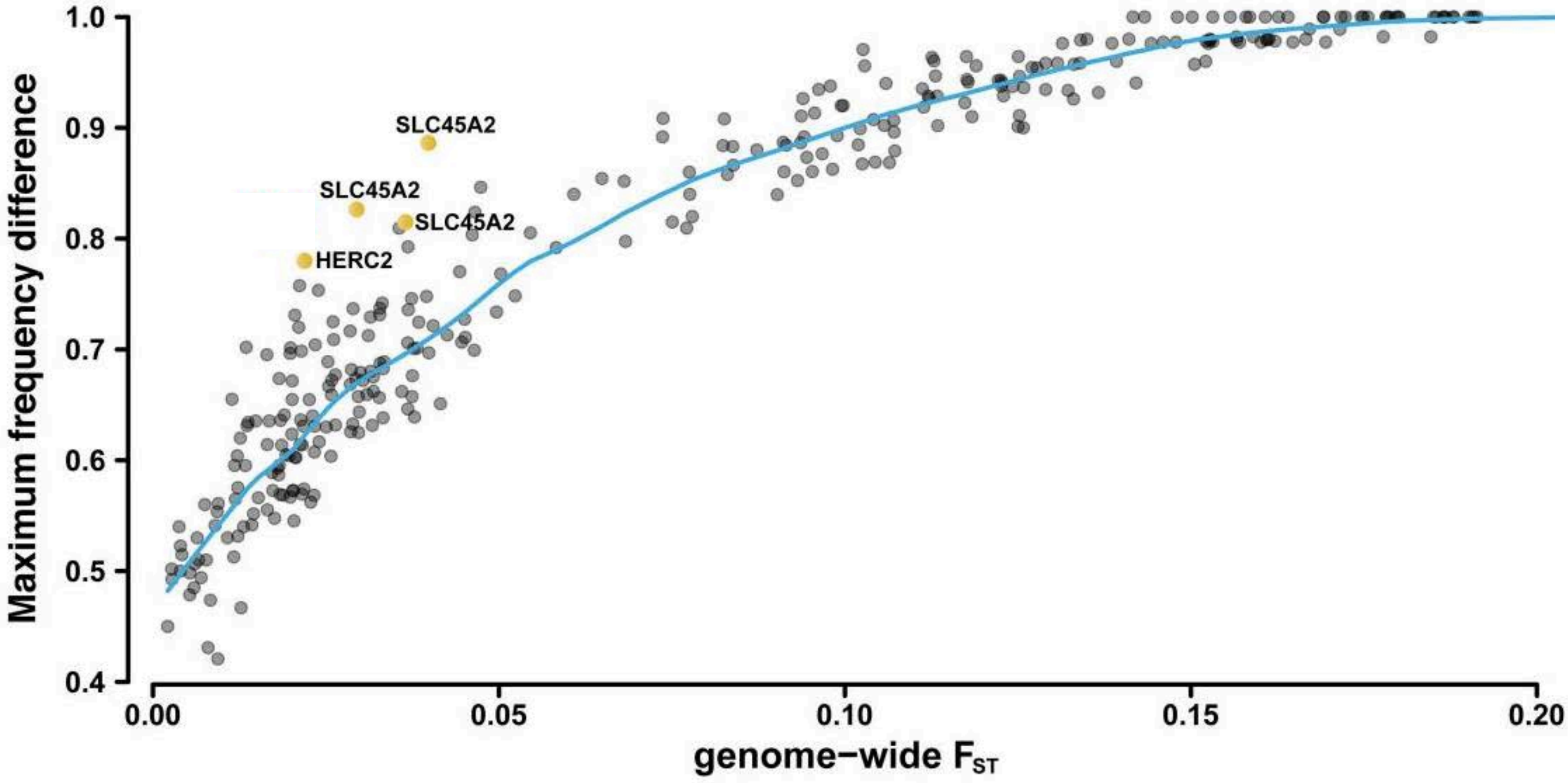
Positive natural selection



Positive natural selection

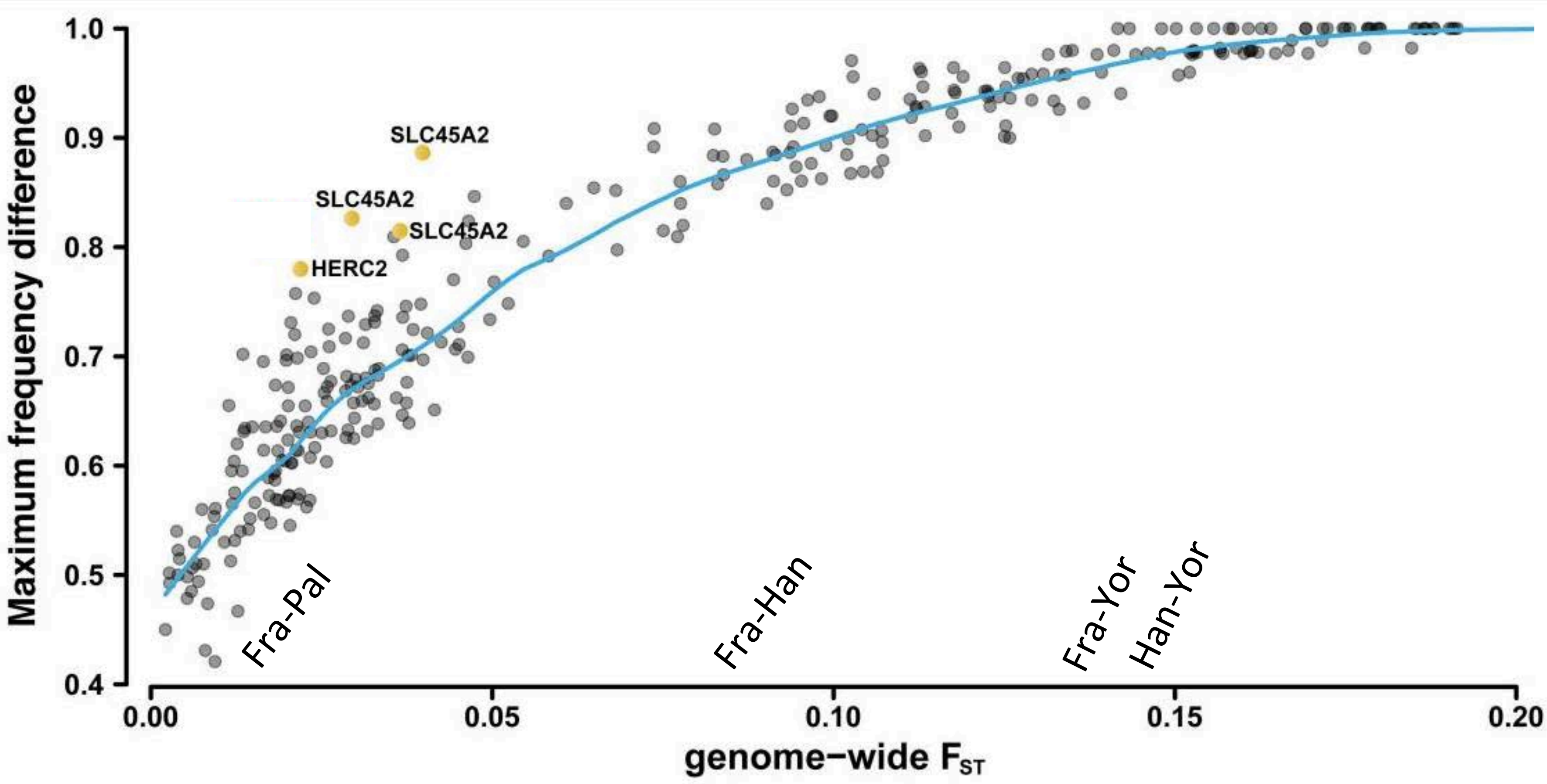


Genetic Differentiation



Redrawn from Coop et al. 2009

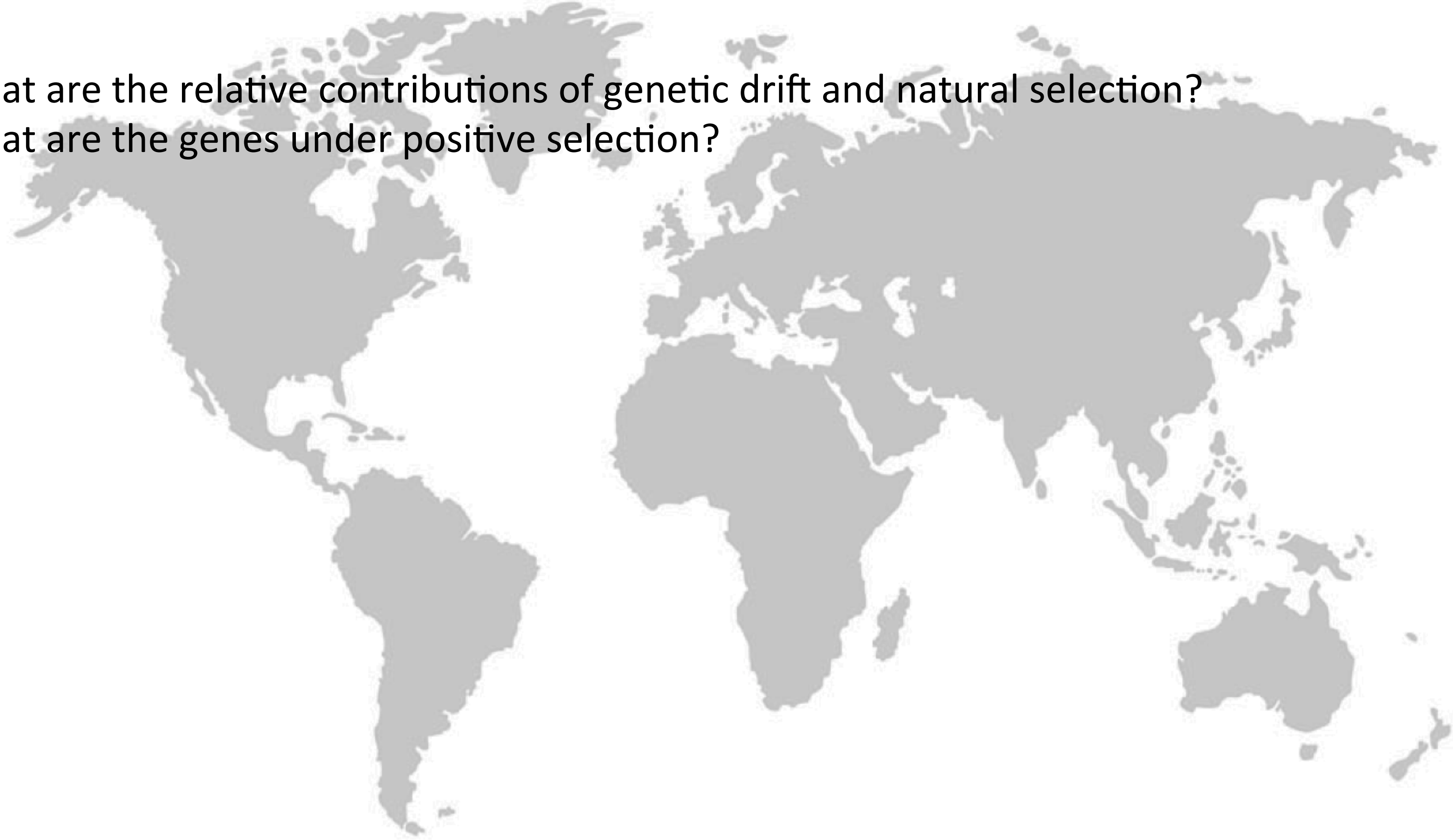
Genetic Differentiation



Redrawn from Coop et al. 2009

Tons of modern human DNA sequence data

What are the relative contributions of genetic drift and natural selection?
What are the genes under positive selection?



Tons of modern human DNA sequence data

What are the relative contributions of genetic drift and natural selection?
What are the genes under positive selection?

- Population size has not been constant in time
- The effect of demography can mimic the effect of natural selection

DNA sequence data

← Sites sequenced →

← individuals →

A T G A C C A G A C C T A G T A A C T T G T A G T C G T C A T A

A C G A C A A T A G C T A C C G A C T T C C T G A A G T C A T A

A C G A C A T T A G C A A C T A T G T A G C T G T A G T C A T G

A T G T C C T T A G T A A C T A T C T A G C T G A C G T C A T G

A C G A G A T T C G C A G C T A T C A T G T A G A C G T C A T A

A C T A C C A T A G C A A G T G T C A T G T A G A C G T C T T A

A C G A C C T T A G C A A C T G T C T A G C A G A A G T T T T A

A C G A C C T G A G C A A C T G T C T A G C A G A A G T T A T G

Chimp reference

DNA sequence data

← Sites sequenced →

← individuals →

```
A T G A C C A G A C C T A G T A A C T T G T A G T C G T C A T A
A C G A C A A T A G C T A C C G A C T T C C T G A A G T C A T A
A C G A C A T T A G C A A C T A T G T A G C T G T A G T C A T G
A T G T C C T T A G T A A C T A T C T A G C T G A C G T C A T G
A C G A G A T T C G C A G C T A T C A T G T A G A C G T C A T A
A C T A C C A T A G C A A G T G T C A T G T A G A C G T C T T A
A C G A C C T T A G C A A C T G T C T A G C A G A A G T T T T A

0 2 1 1 1 3 3 6 1 1 1 2 1 2 1 4 2 1 2 4 1 3 3 0 2 4 0 0 6 2 0 5
```

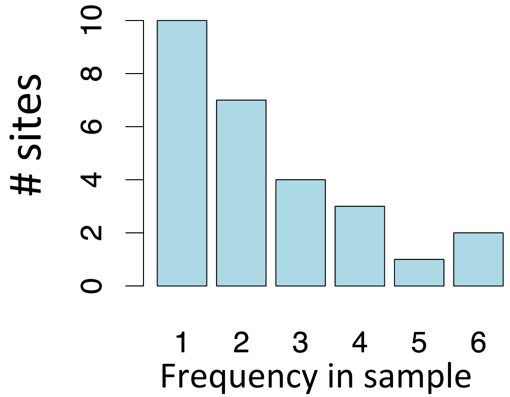
DNA sequence data

← Sites sequenced →

← individuals →

```
A T G A C C A G A C C T A G T A A C T T G T A G T C G T C A T A
A C G A C A A T A G C T A C C G A C T T C C T G A A G T C A T A
A C G A C A T T A G C A A C T A T G T A G C T G T A G T C A T G
A T G T C C T T A G T A A C T A T C T A G C T G A C G T C A T G
A C G A G A T T C G C A G C T A T C A T G T A G A C G T C A T A
A C T A C C A T A G C A A G T G T C A T G T A G A C G T C T T A
A C G A C C T T A G C A A C T G T C T A G C A G A A G T T T T A

0 2 1 1 1 3 3 6 1 1 1 2 1 2 1 4 2 1 2 4 1 3 3 0 2 4 0 0 6 2 0 5
```



DNA sequence data

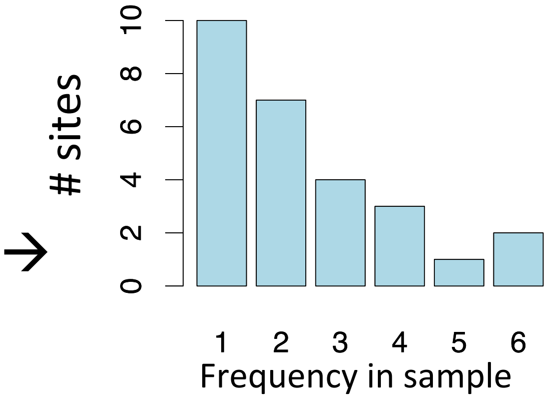
← Sites sequenced →

← individuals →

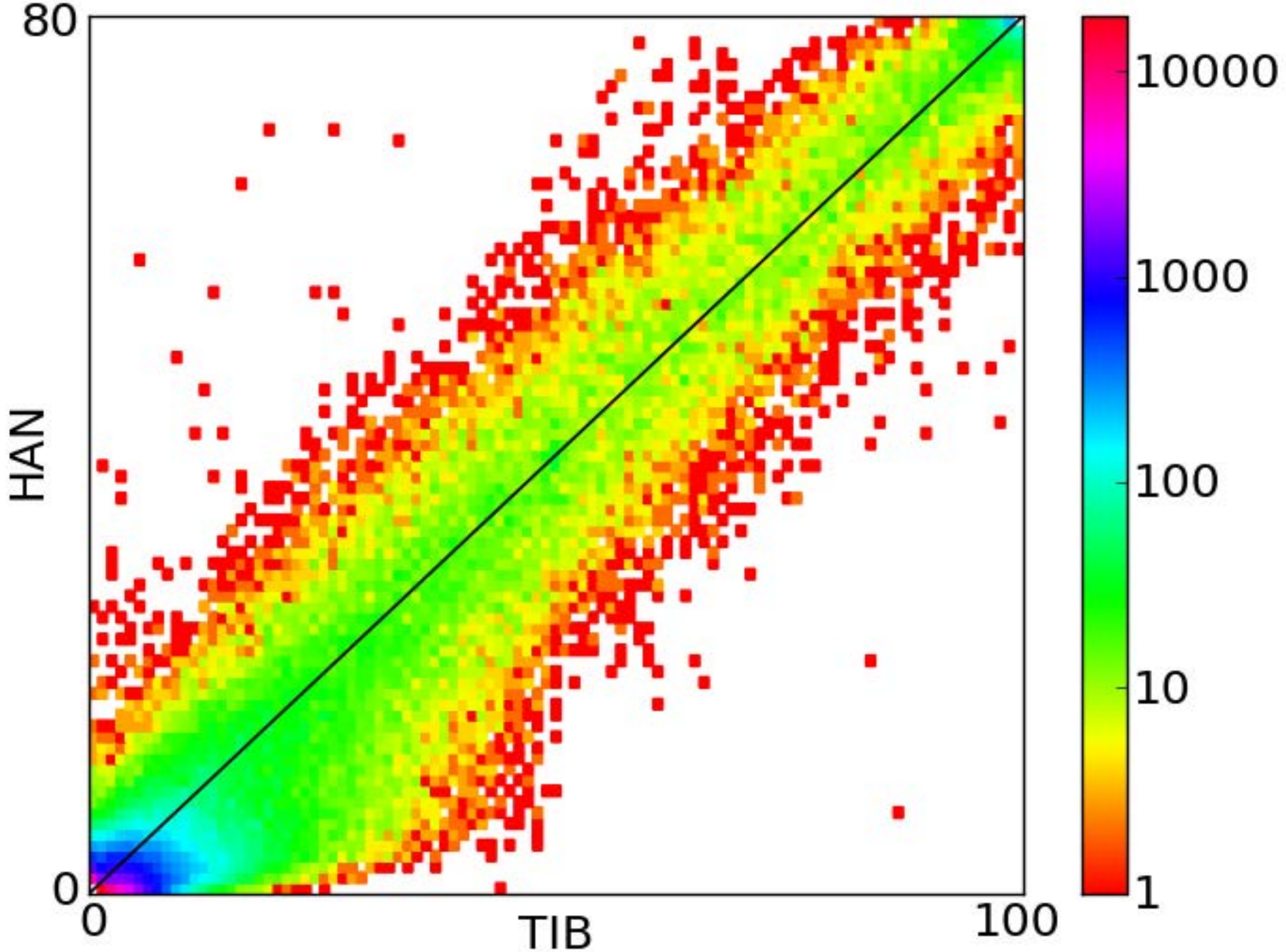
A T G A C C A G A C C T A G T A A C T T G T A G T C G T C A T A
A C G A C A A T A G C T A C C G A C T T C C T G A A G T C A T A
A C G A C A T T A G C A A C T A T G T A G C T G T A G T C A T G
A T G T C C T T A G T A A C T A T C T A G C T G A C G T C A T G
A C G A G A T T C G C A G C T A T C A T G T A G A C G T C A T A
A C T A C C A T A G C A A G T G T C A T G T A G A C G T C T T A
A C G A C C T T A G C A A C T G T C T A G C A G A A G T T T T A

0 2 1 1 1 3 3 6 1 1 1 2 1 2 1 4 2 1 2 4 1 3 3 0 2 4 0 0 6 2 0 5

The site frequency spectrum →

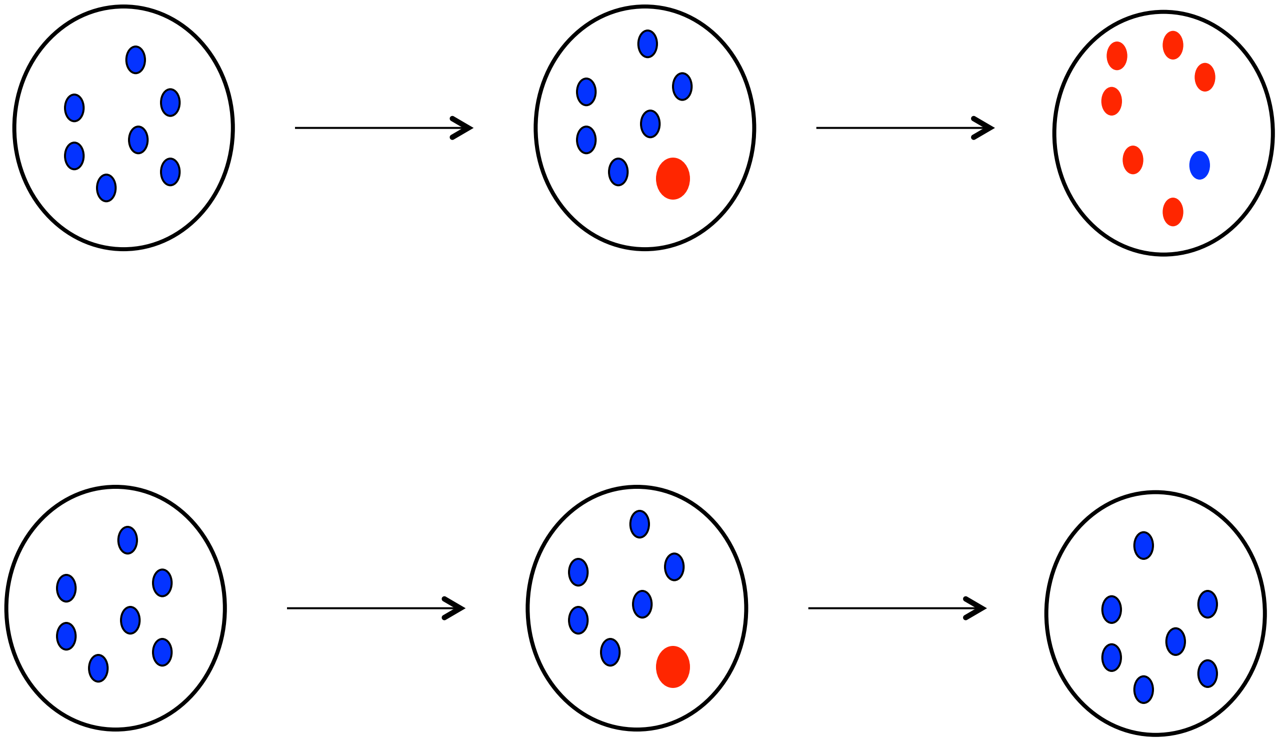


2D SFS



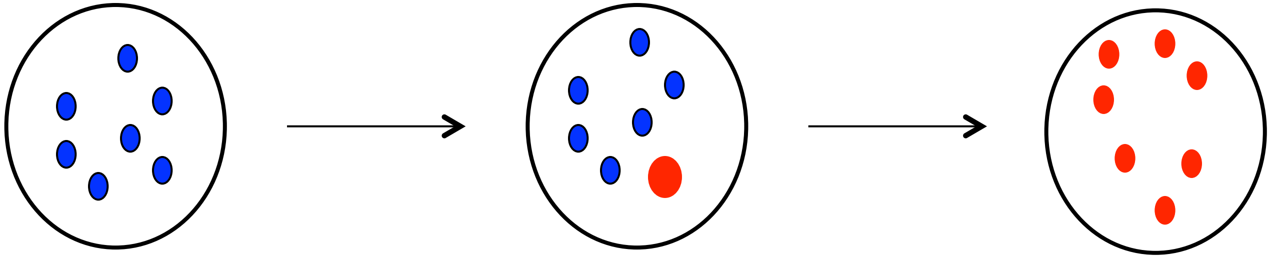
Neutrality

By chance …



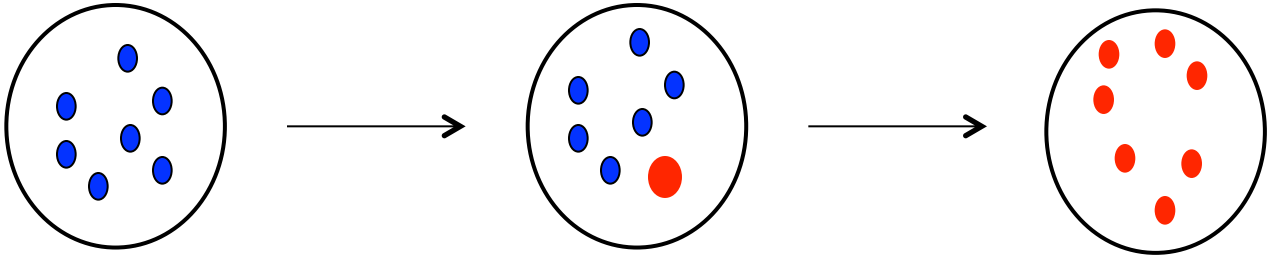
Natural selection

Positive Selection

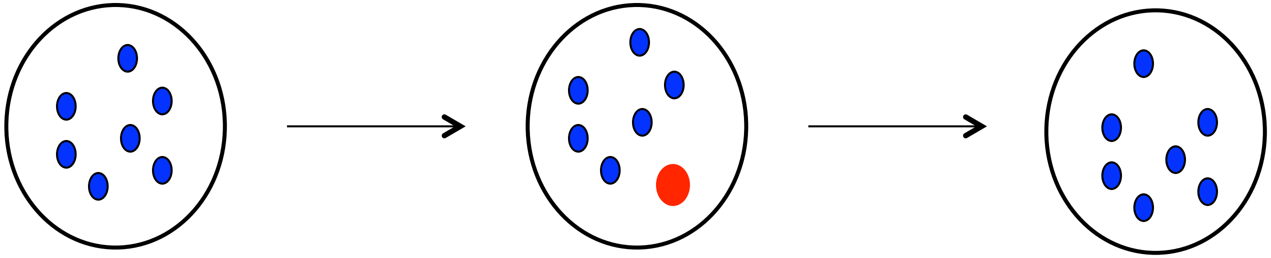


Natural selection

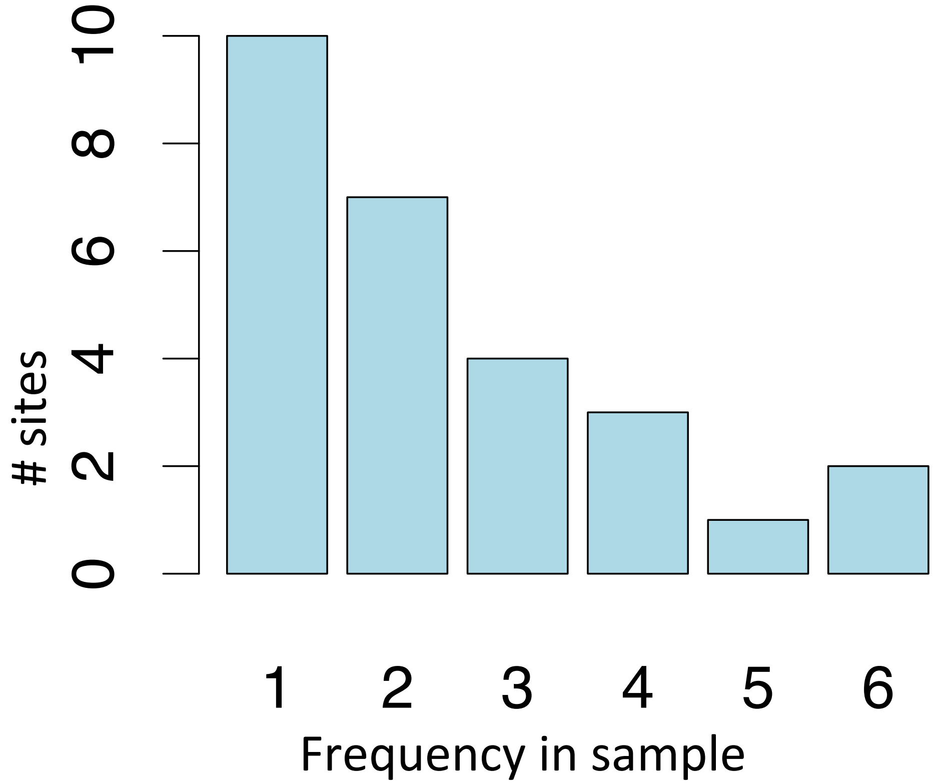
Positive Selection



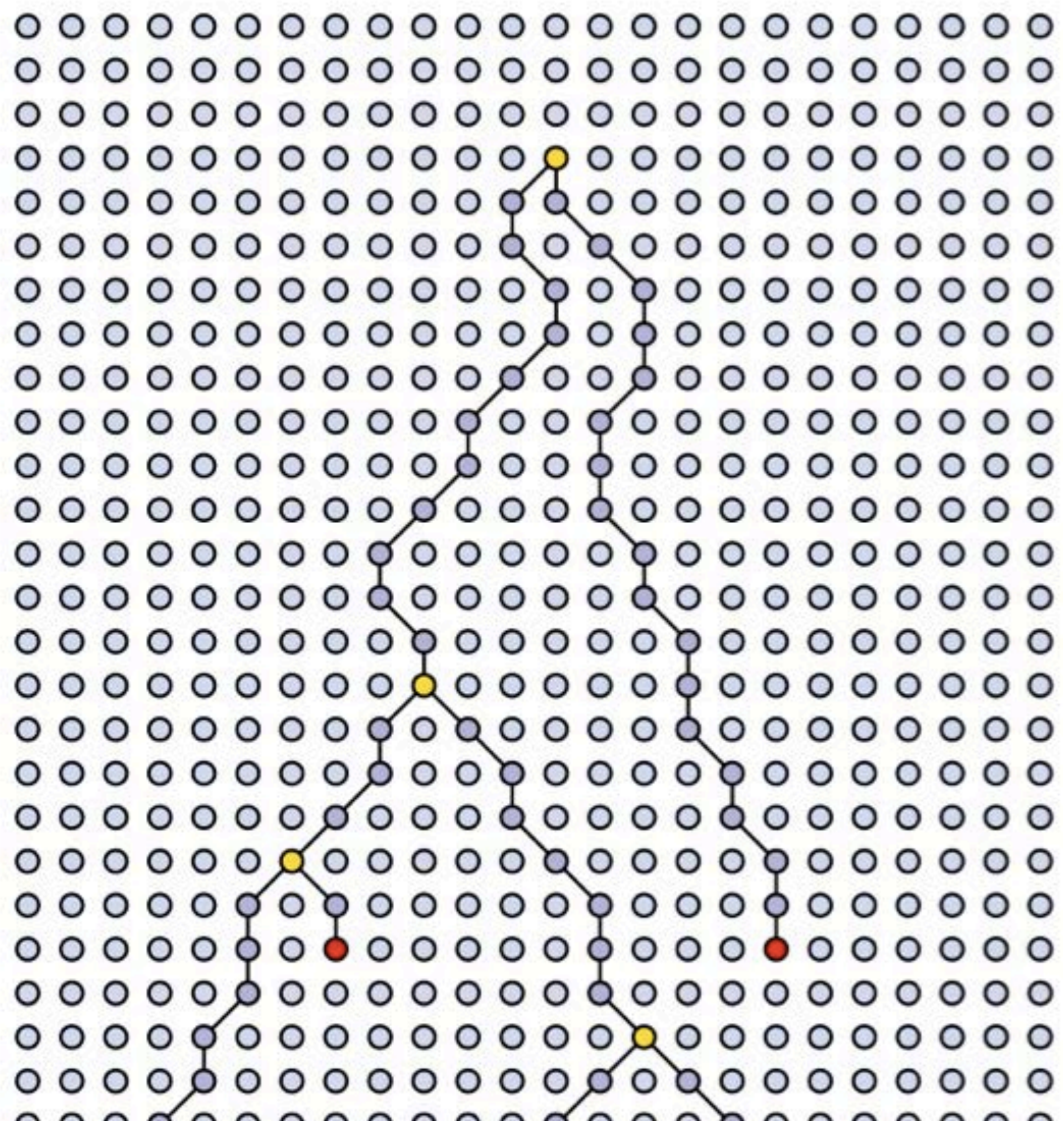
Negative Selection



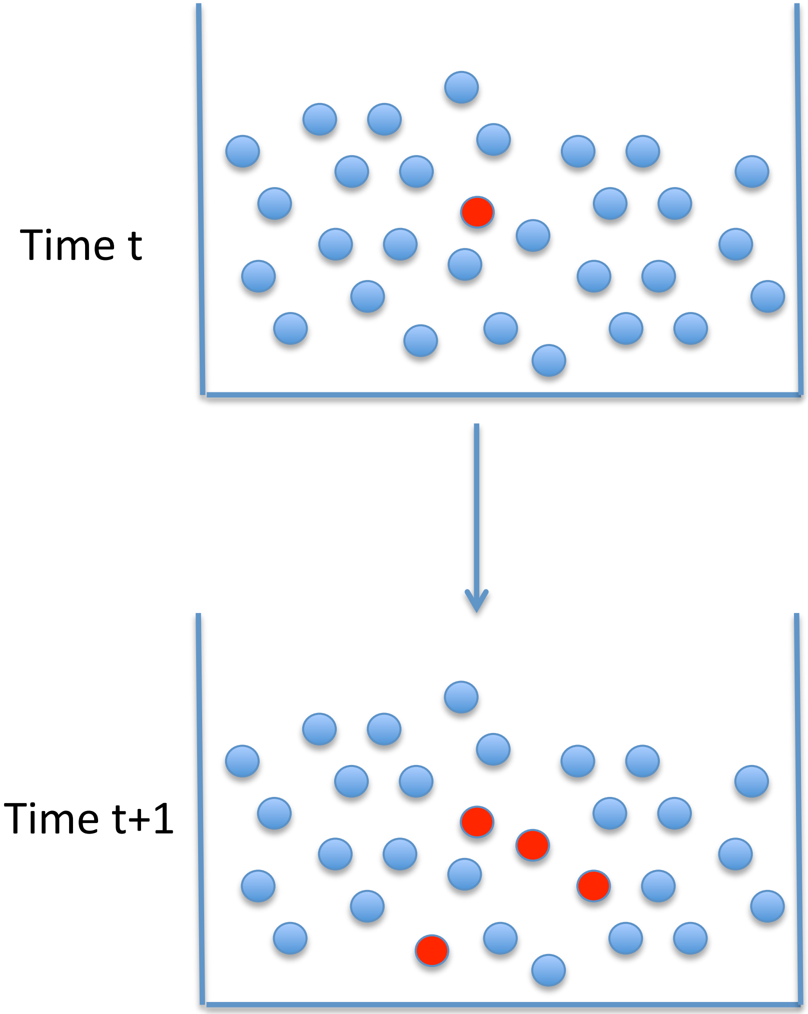
How is the SFS shaped by difference processes?



The geneaology of a sample



The Wright-Fisher model

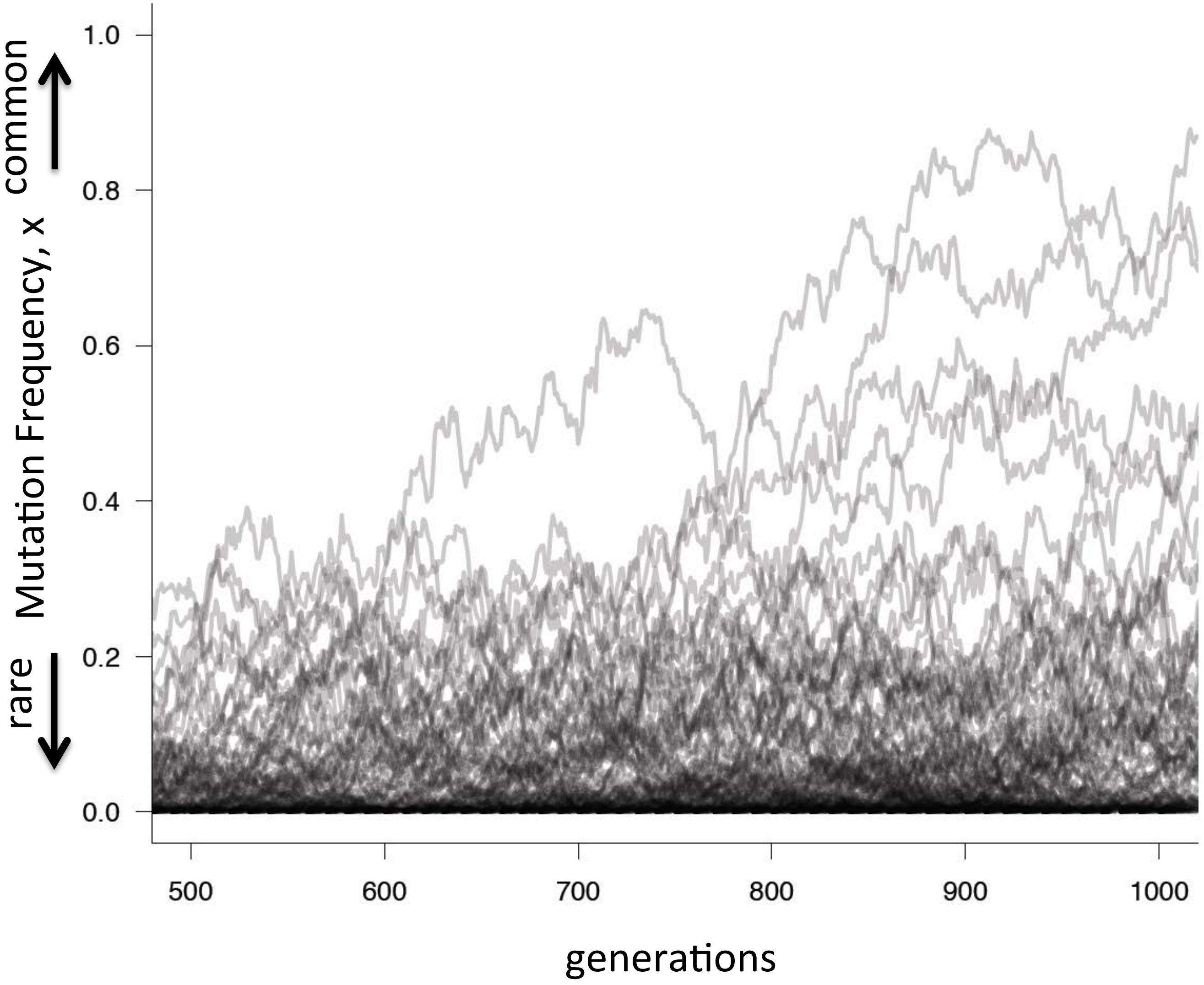


X(t) = number of individuals carrying mutation at time t

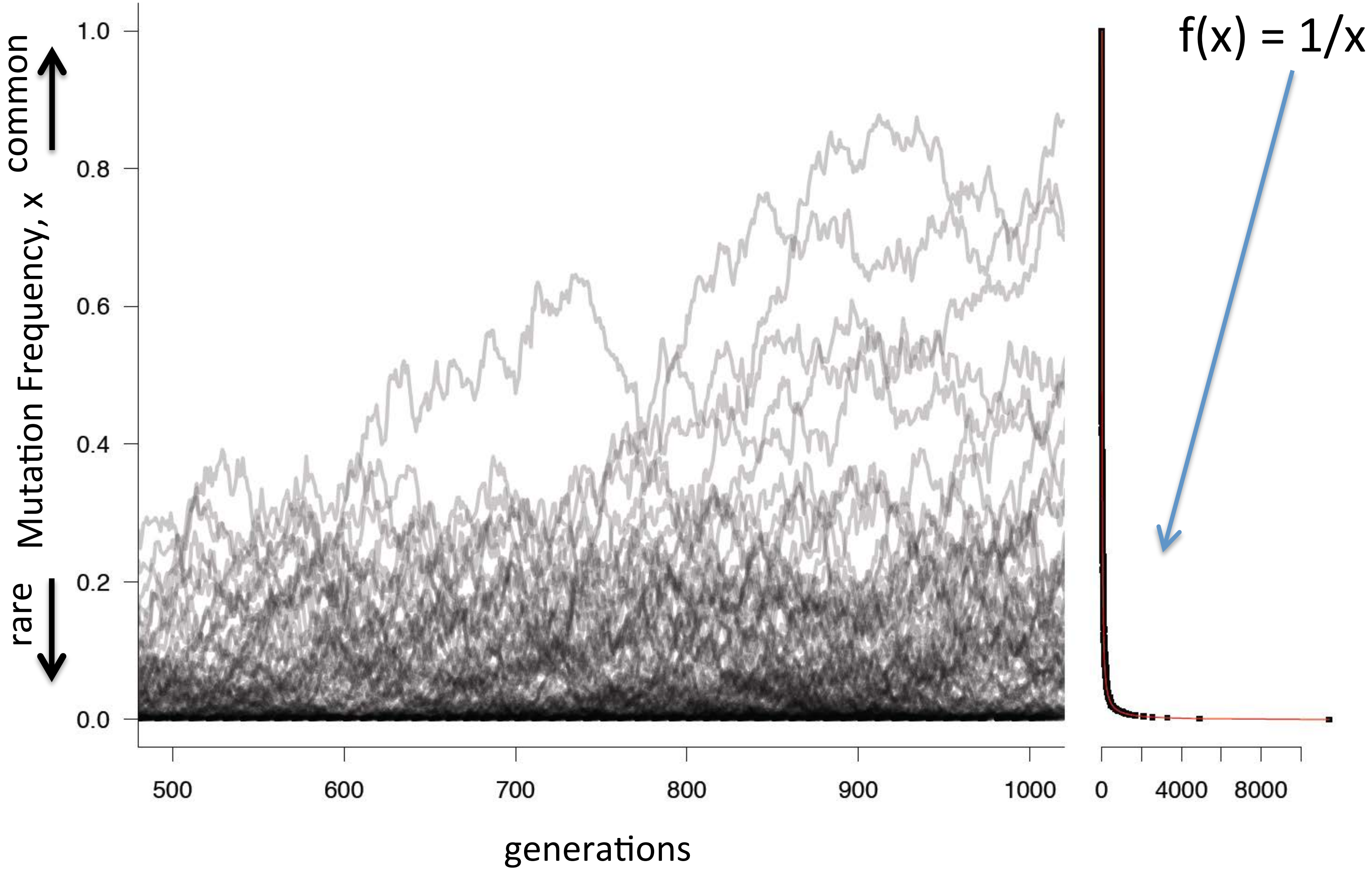
X(t) is a **Markov chain** with transition probabilities:

$$P_{ij} = \mathrm{Bin}(2N, i/2N)$$

If mutations are arriving at Poisson times



If mutations are arriving at Poisson times

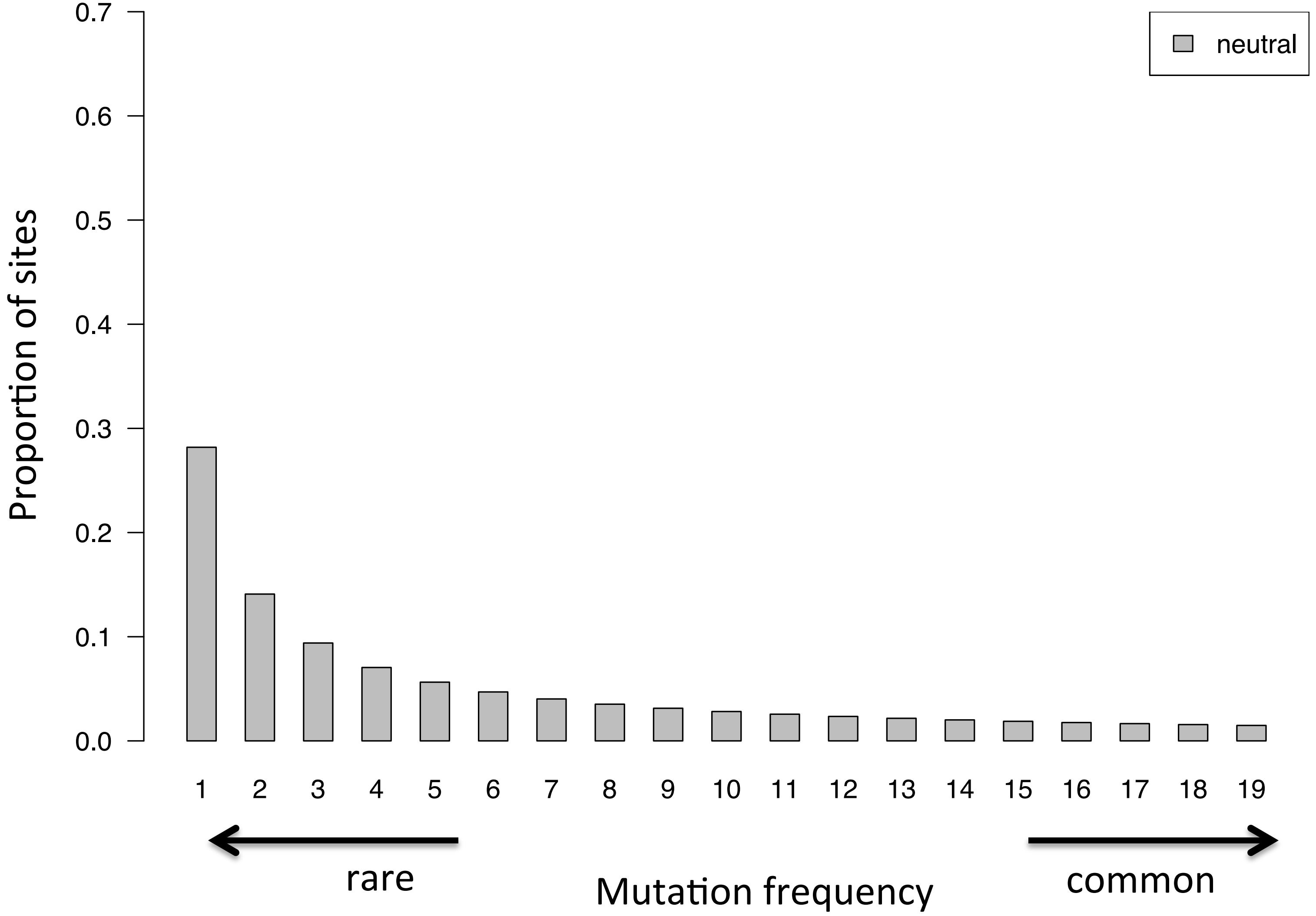


The SFS in a sample of size *n*

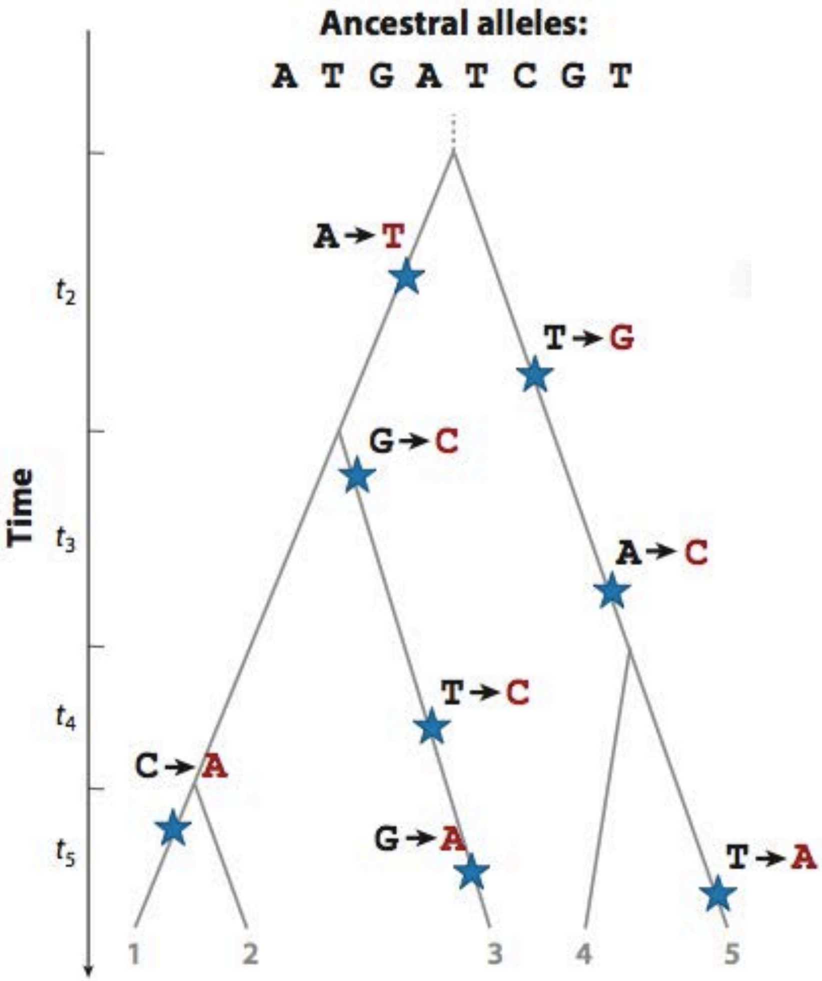
The expected number of variant sites at frequency *i*

$$
\begin{aligned}
\theta F(i,\gamma) &= \theta \int_0^1 f(q)\Pr(i \mid q)\,dq \\
&= \theta \int_0^1 \frac{1}{q}\binom{n}{i} q^i (1-q)^{n-i}\,dq \\
&= \frac{\theta}{i}
\end{aligned}
$$

The site frequency spectrum

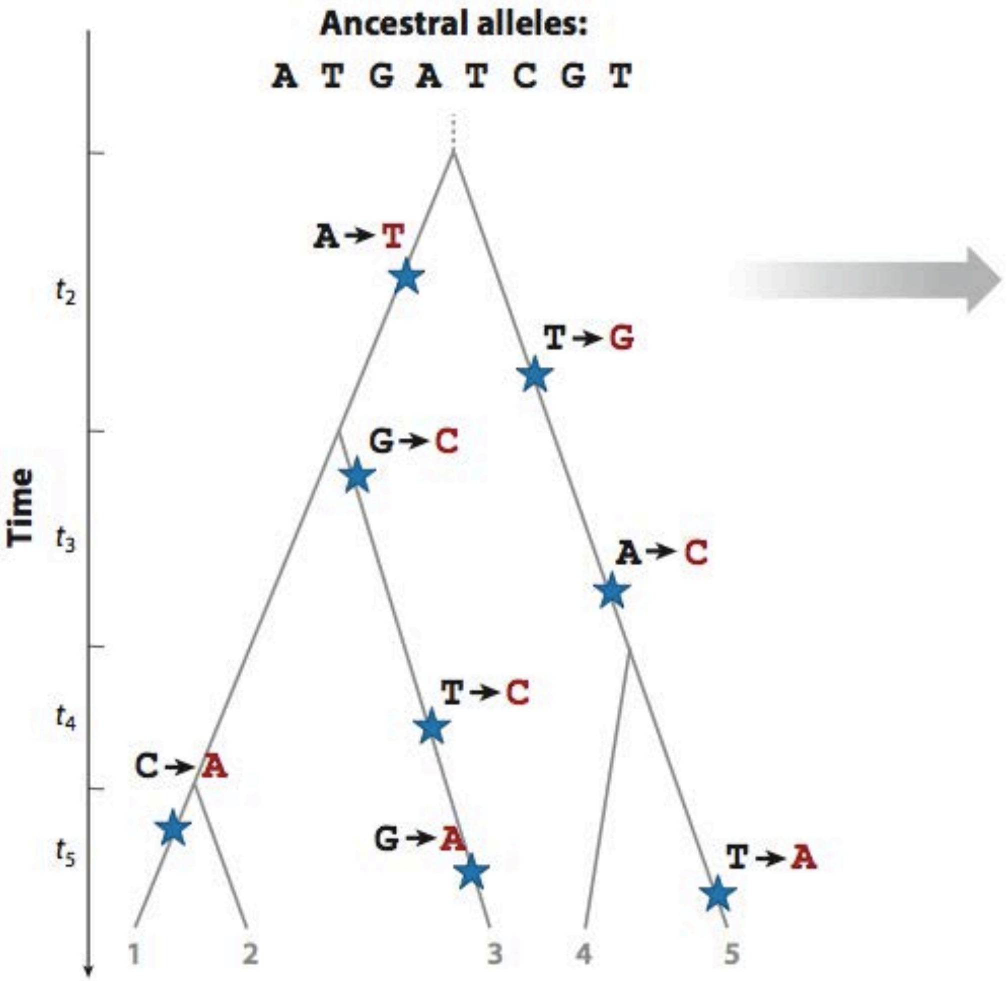


The site frequency spectrum from the coalescent



Beichman A, Huerta-Sanchez E, Lohmueller K. Annual Review of Ecology, Evolution, and Systematics (2019)

The site frequency spectrum from the coalescent



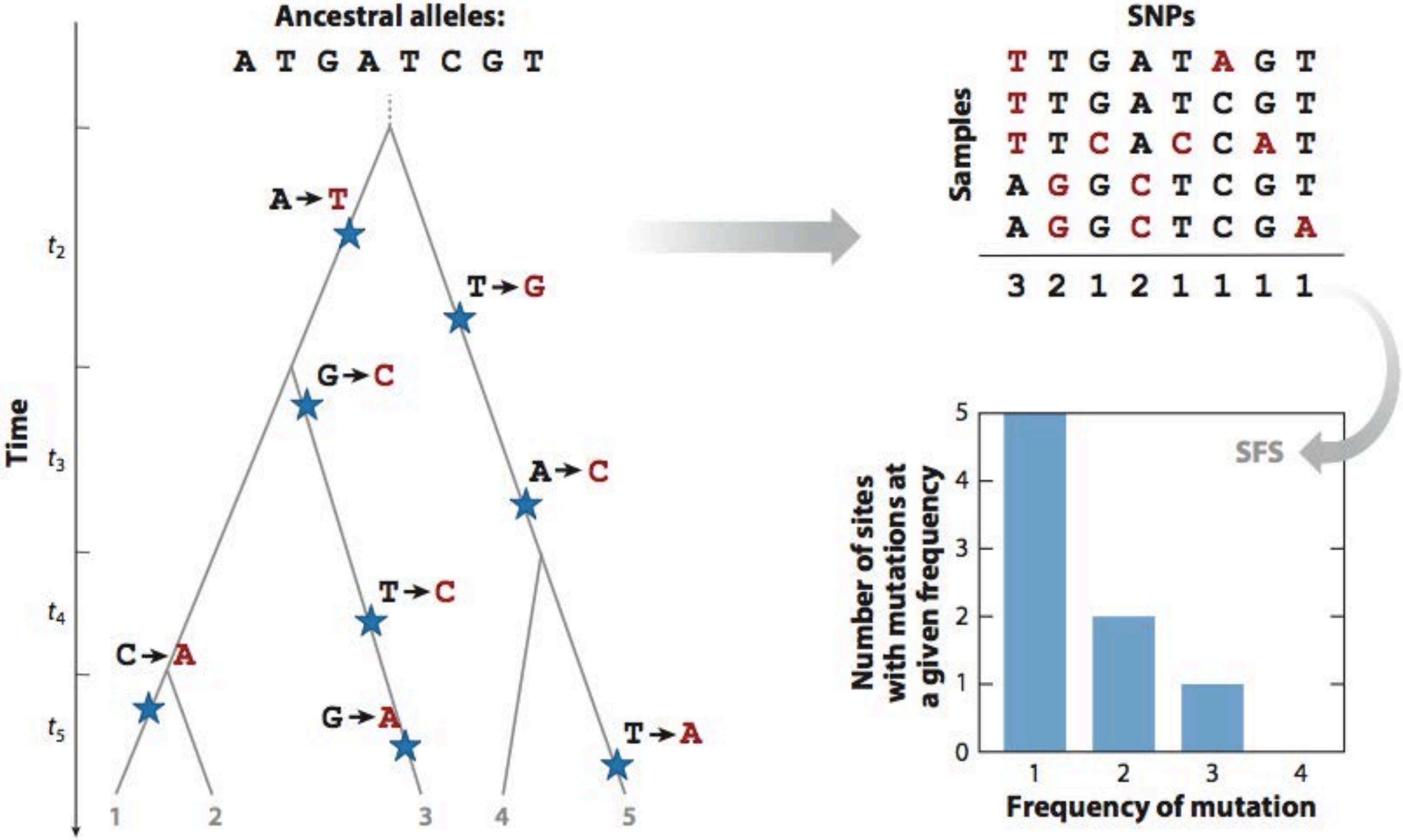
SNPs

Samples

T	T	G	A	T	A	G	T
T	T	G	A	T	C	G	T
T	T	C	A	C	C	A	T
A	G	G	C	T	C	G	T
A	G	G	C	T	C	G	A
3	2	1	2	1	1	1	1

Beichman A, Huerta-Sanchez E, Lohmueller K. Annual Review of Ecology, Evolution, and Systematics (2019)

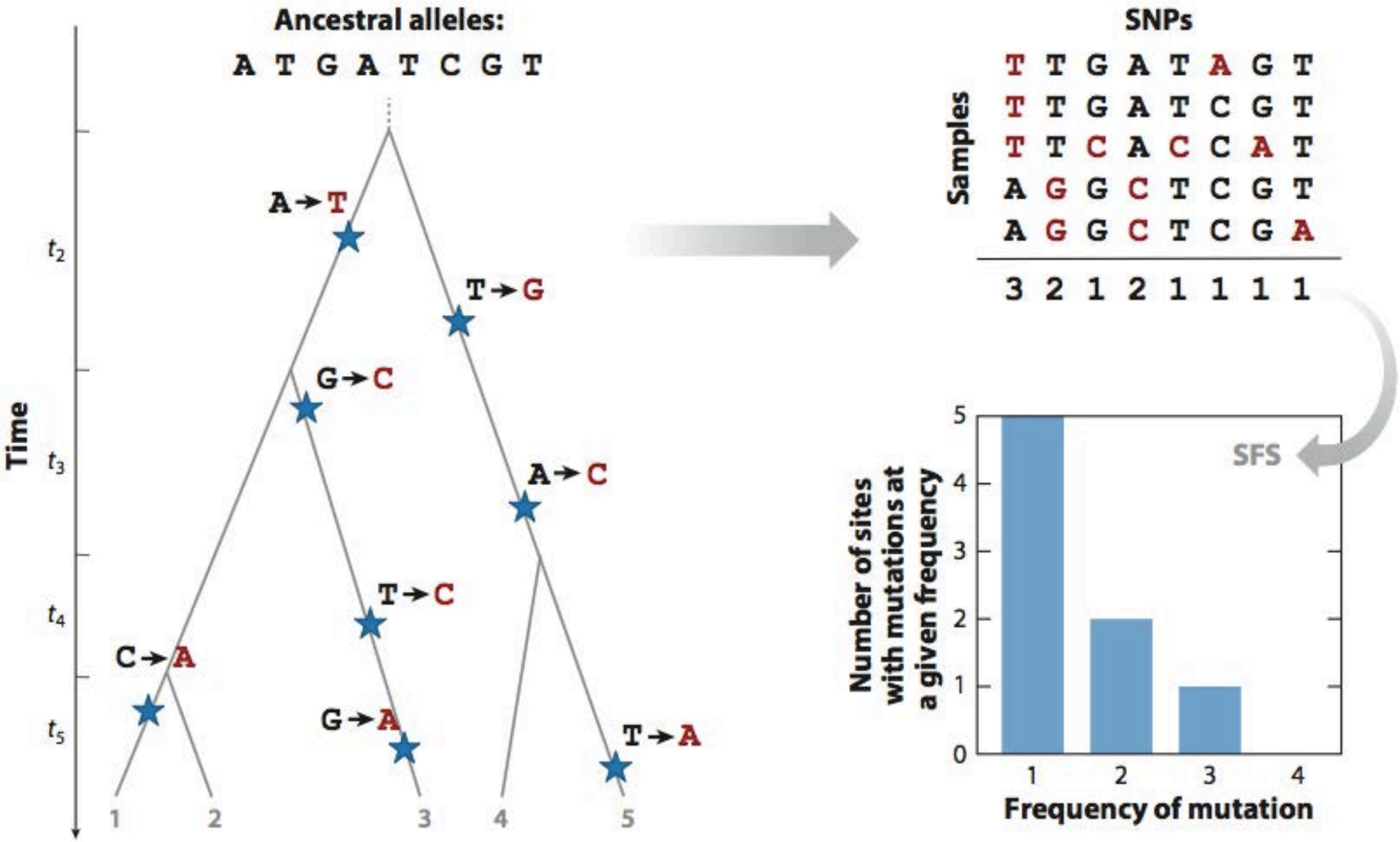
The site frequency spectrum from the coalescent



T	T	G	A	T	A	G	T
T	T	G	A	T	C	G	T
T	T	C	A	C	C	A	T
A	G	G	C	T	C	G	T
A	G	G	C	T	C	G	A
3	2	1	2	1	1	1	1

Beichman A, Huerta-Sanchez E, Lohmueller K. Annual Review of Ecology, Evolution, and Systematics (2019)

The site frequency spectrum from the coalescent

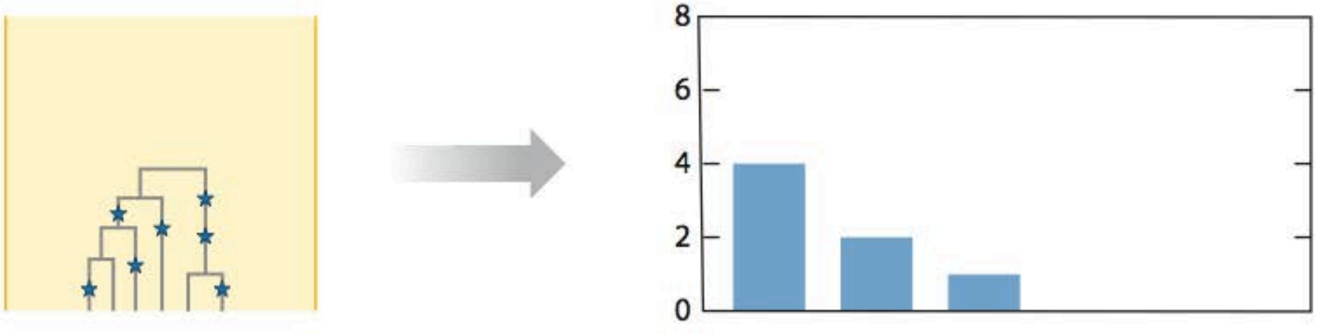


T	T	G	A	T	A	G	T
T	T	G	A	T	C	G	T
T	T	C	A	C	C	A	T
A	G	G	C	T	C	G	T
A	G	G	C	T	C	G	A
3	2	1	2	1	1	1	1

Beichman A, Huerta-Sanchez E, Lohmueller K. Annual Review of Ecology, Evolution, and Systematics (2019)

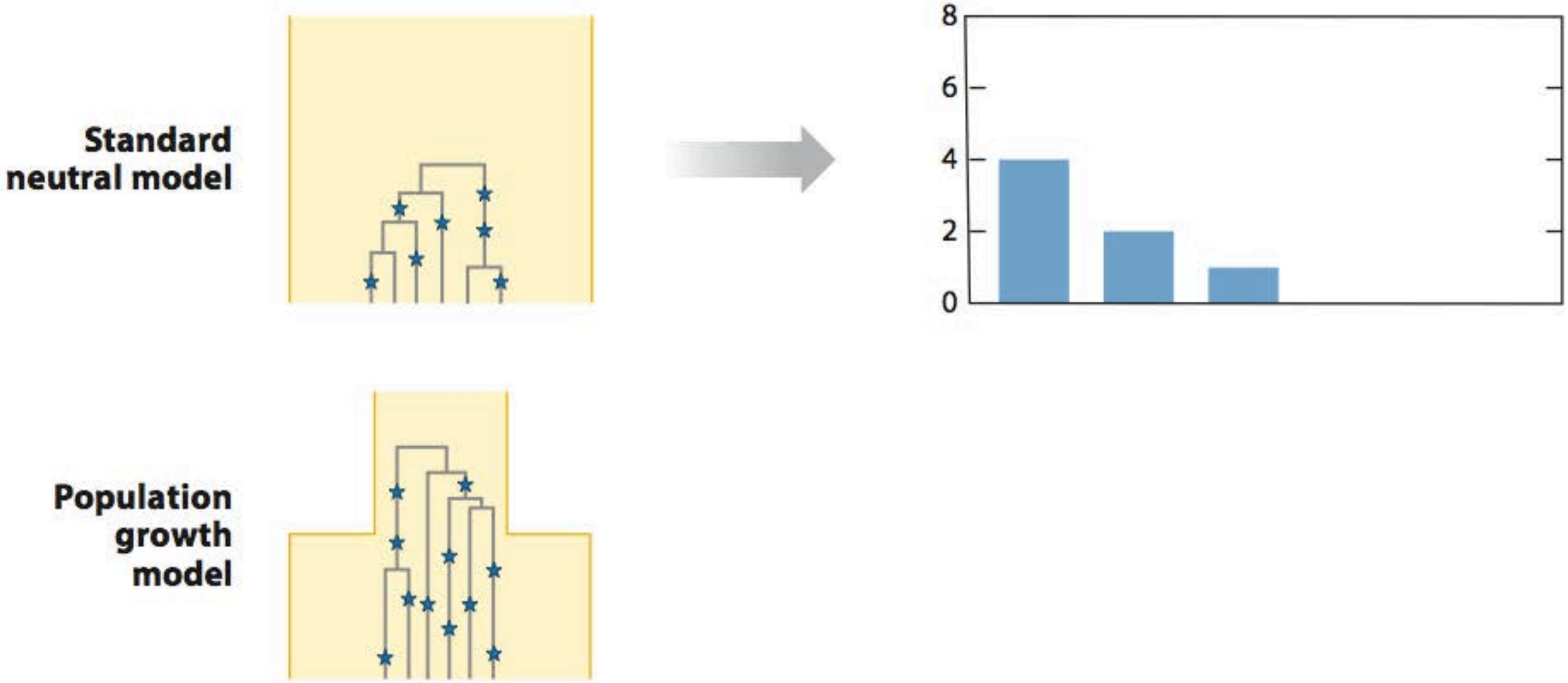
The effect of demography

Standard neutral model



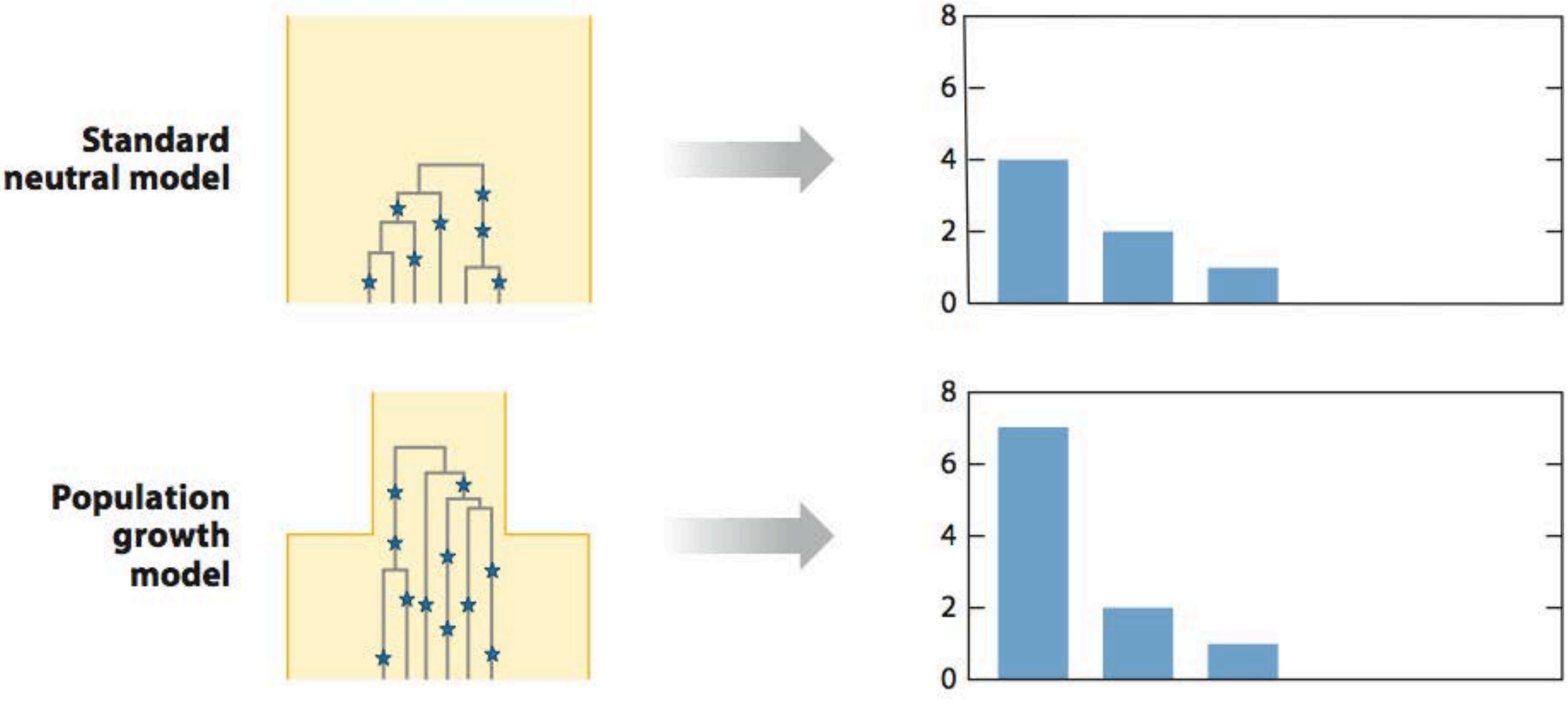
Beichman A, Huerta-Sanchez E, Lohmueller K. Annual Review of Ecology, Evolution, and Systematics (2019)

The effect of demography



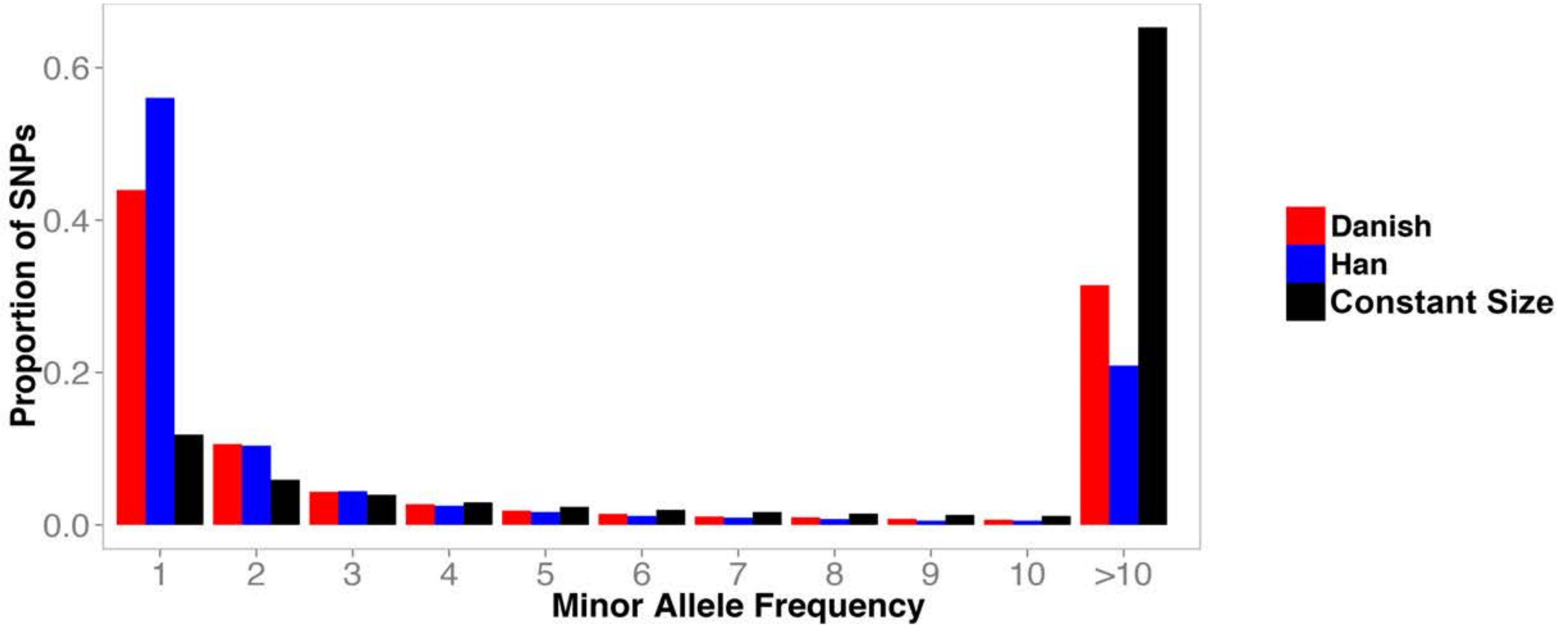
Beichman A, Huerta-Sanchez E, Lohmueller K. Annual Review of Ecology, Evolution, and Systematics (2019)

The effect of demography



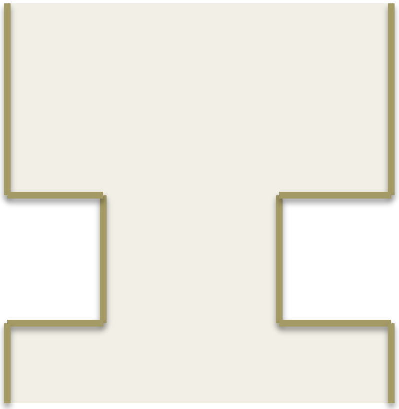
Beichman A, Huerta-Sanchez E, Lohmueller K. Annual Review of Ecology, Evolution, and Systematics (2019)

Data from 1500 Danes and 1500 Chinese individuals

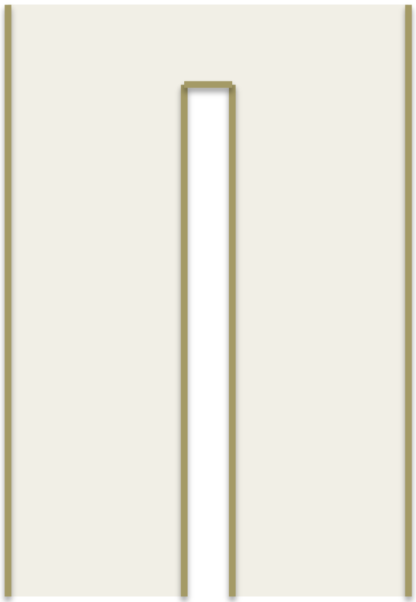


Synonymous sites

The effect of a bottleneck and population structure



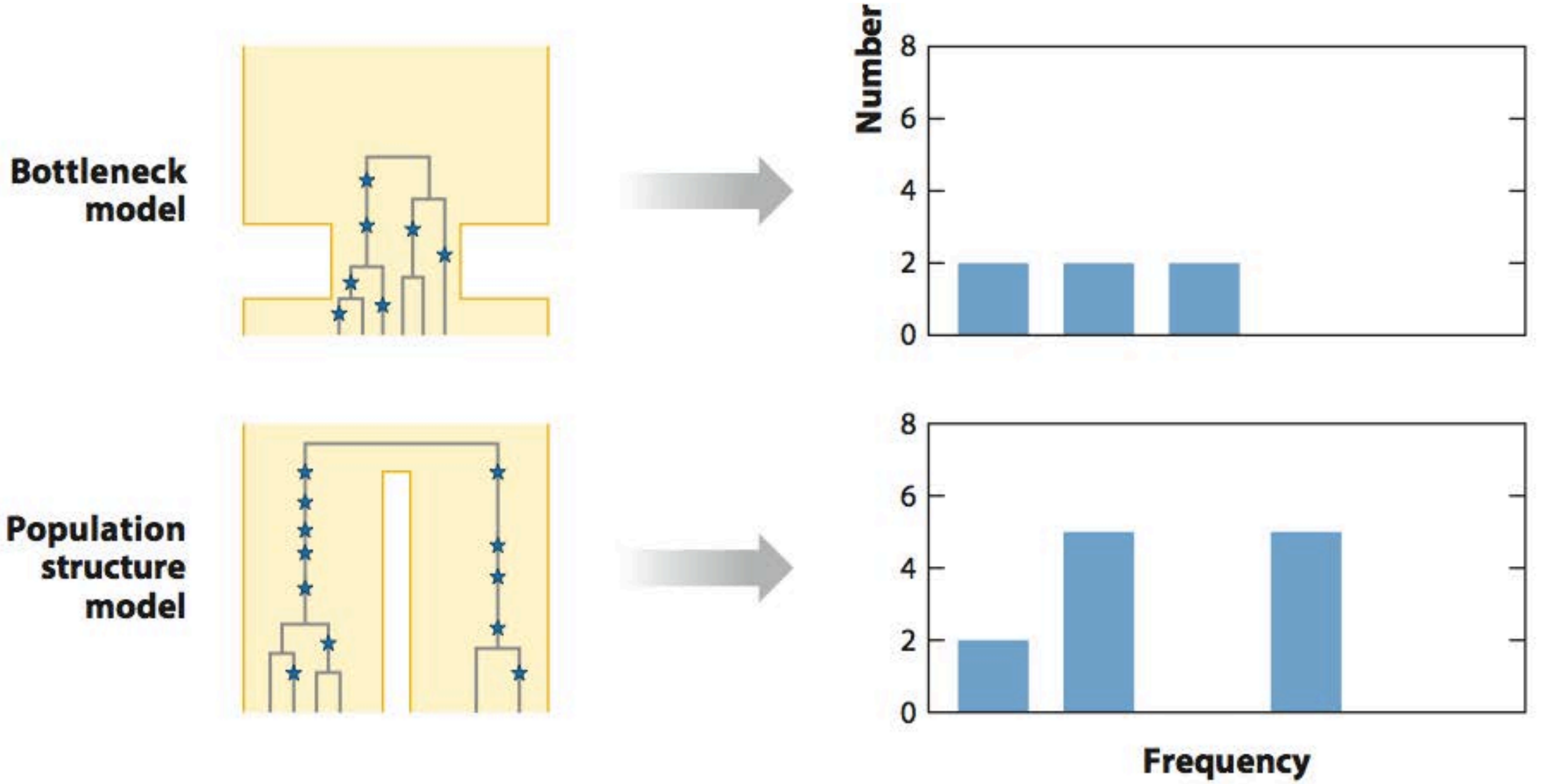
bottleneck



Population structure

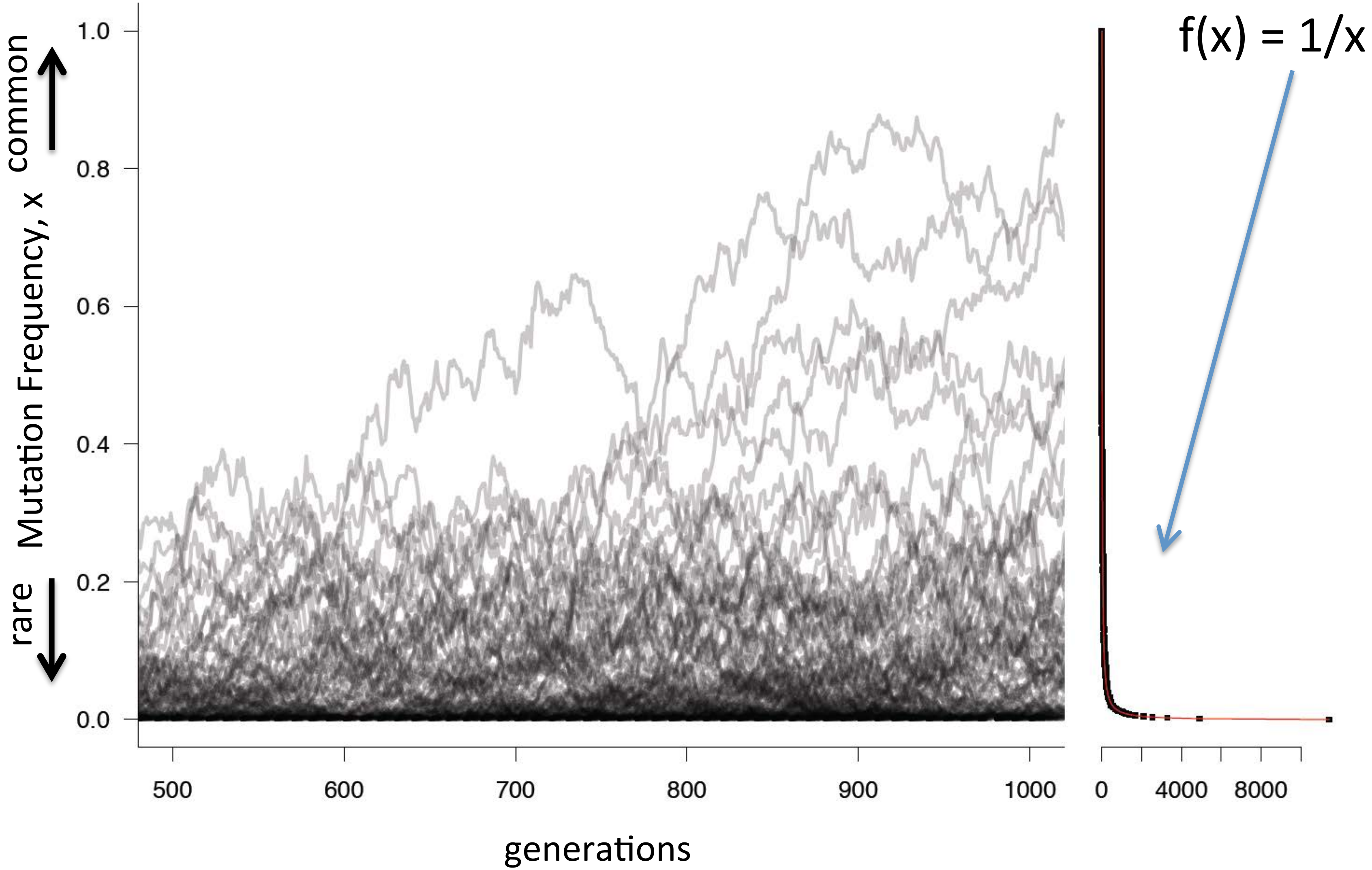
Beichman A, Huerta-Sanchez E, Lohmueller K. Annual Review of Ecology, Evolution, and Systematics (2019)

The effect of demography



Beichman A, Huerta-Sanchez E, Lohmueller K. Annual Review of Ecology, Evolution, and Systematics (2019)

If mutations are arriving at Poisson times



The SFS in a sample of size *n* with selection

The expected number of variant sites at frequency *i*

$$\begin{aligned}\theta F(i,\gamma) &= \int_0^1 \frac{f(q,\gamma)}{2} \cdot \Pr(i \mid q)\, dq \\ &= \int_0^1 \frac{1-e^{-2\gamma(1-q)}}{1-e^{-2\gamma}} \frac{1}{q(1-q)} \binom{n}{i} q^i (1-q)^{n-i} dq\end{aligned}$$

-

The SFS in a sample of size *n*

The expected number of variant sites at frequency *i*

$$\begin{aligned}\theta F(i,\gamma) &= \int_0^1 \frac{f(q,\gamma)}{2} \cdot \Pr(i \mid q)\, dq \\ &= \int_0^1 \frac{1-e^{-2\gamma(1-q)}}{1-e^{-2\gamma}} \frac{1}{q(1-q)} \binom{n}{i} q^i (1-q)^{n-i} dq\end{aligned}$$

The probability of seeing x_i sites at frequency *i*

$$p(X_i = x_i \mid \theta, \gamma) = e^{-\theta F(i,\gamma)} \frac{(\theta F(i,\gamma))^{x_i}}{x_i!}.$$

The SFS in a sample of size *n*

The expected number of variant sites at frequency *i*

$$\theta F(i,\gamma) = \int_0^1 \frac{f(q,\gamma)}{2} \cdot \Pr(i \mid q)\, dq$$

$$= \int_0^1 \frac{1-e^{-2\gamma(1-q)}}{1-e^{-2\gamma}} \frac{1}{q(1-q)} \binom{n}{i} q^i (1-q)^{n-i} dq$$

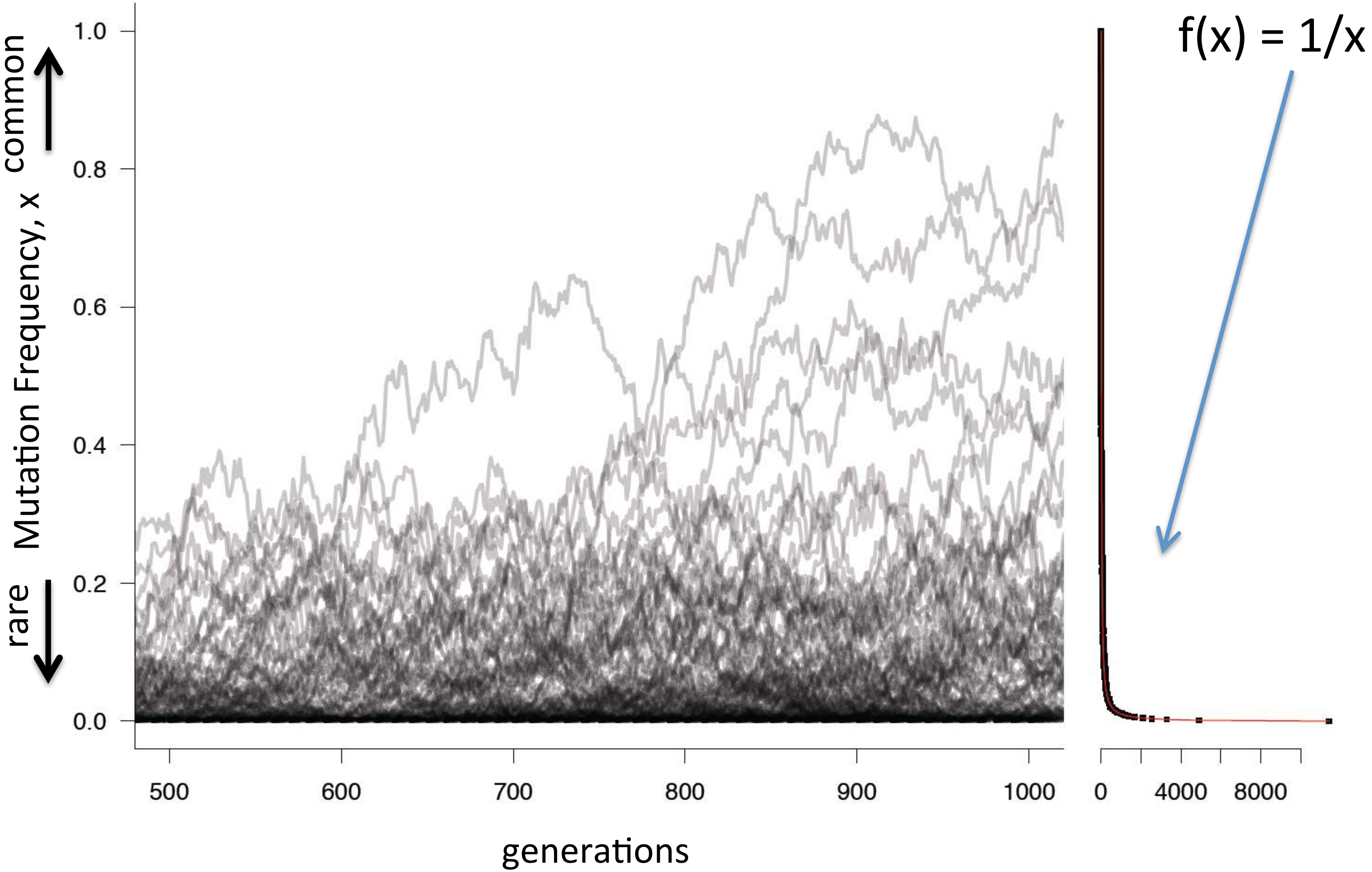
The probability of seeing x_i sites at frequency *i*

$$p(X_i = x_i \mid \theta, \gamma) = e^{-\theta F(i,\gamma)} \frac{(\theta F(i,\gamma))^{x_i}}{x_i!}.$$

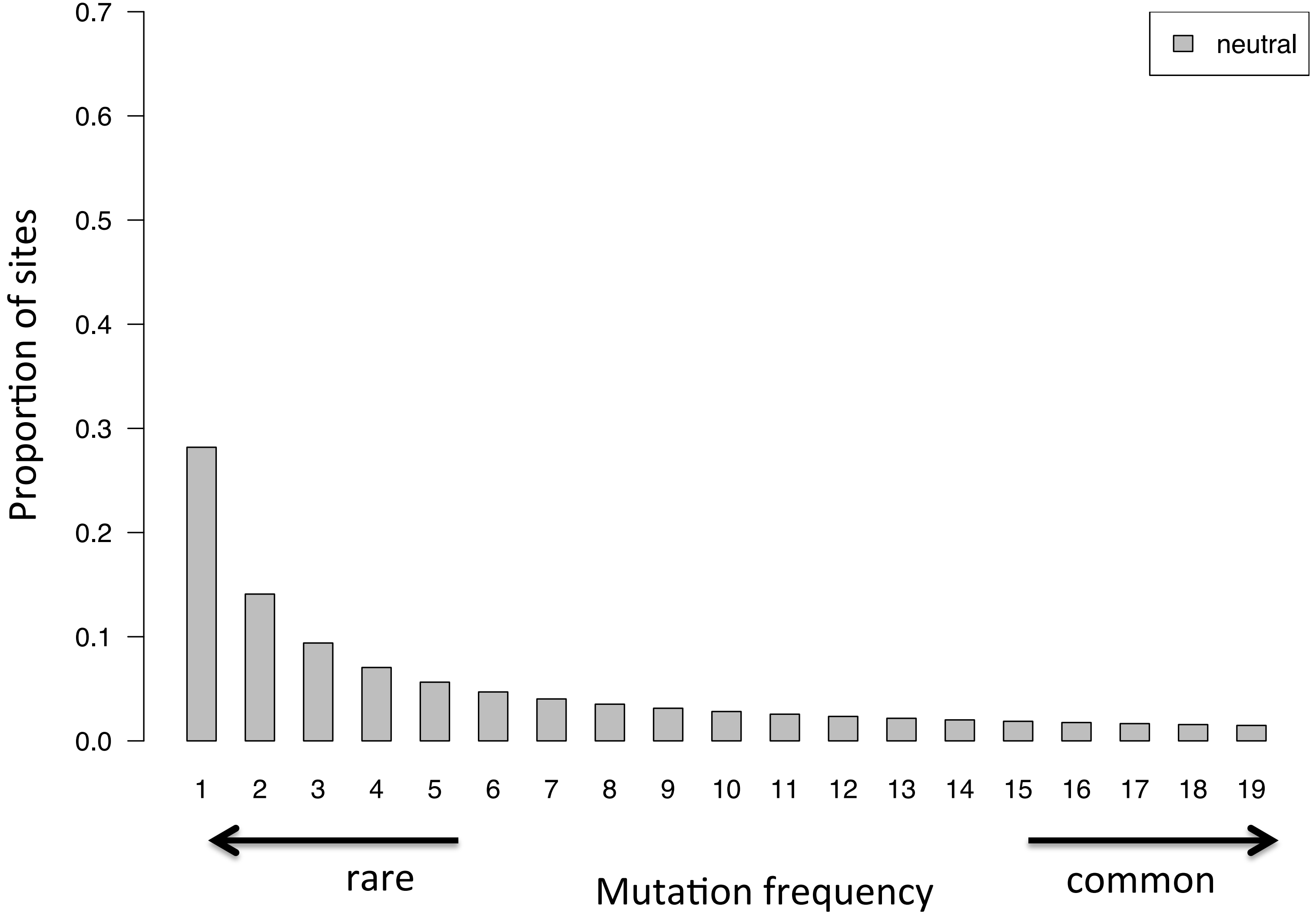
The likelihood function

$$L_u(\theta, \gamma \mid x) = \prod_{i=1}^{n-1} e^{-\theta F(i,y)} \frac{(\theta F(i,\gamma))^{x_i}}{x_i!}$$

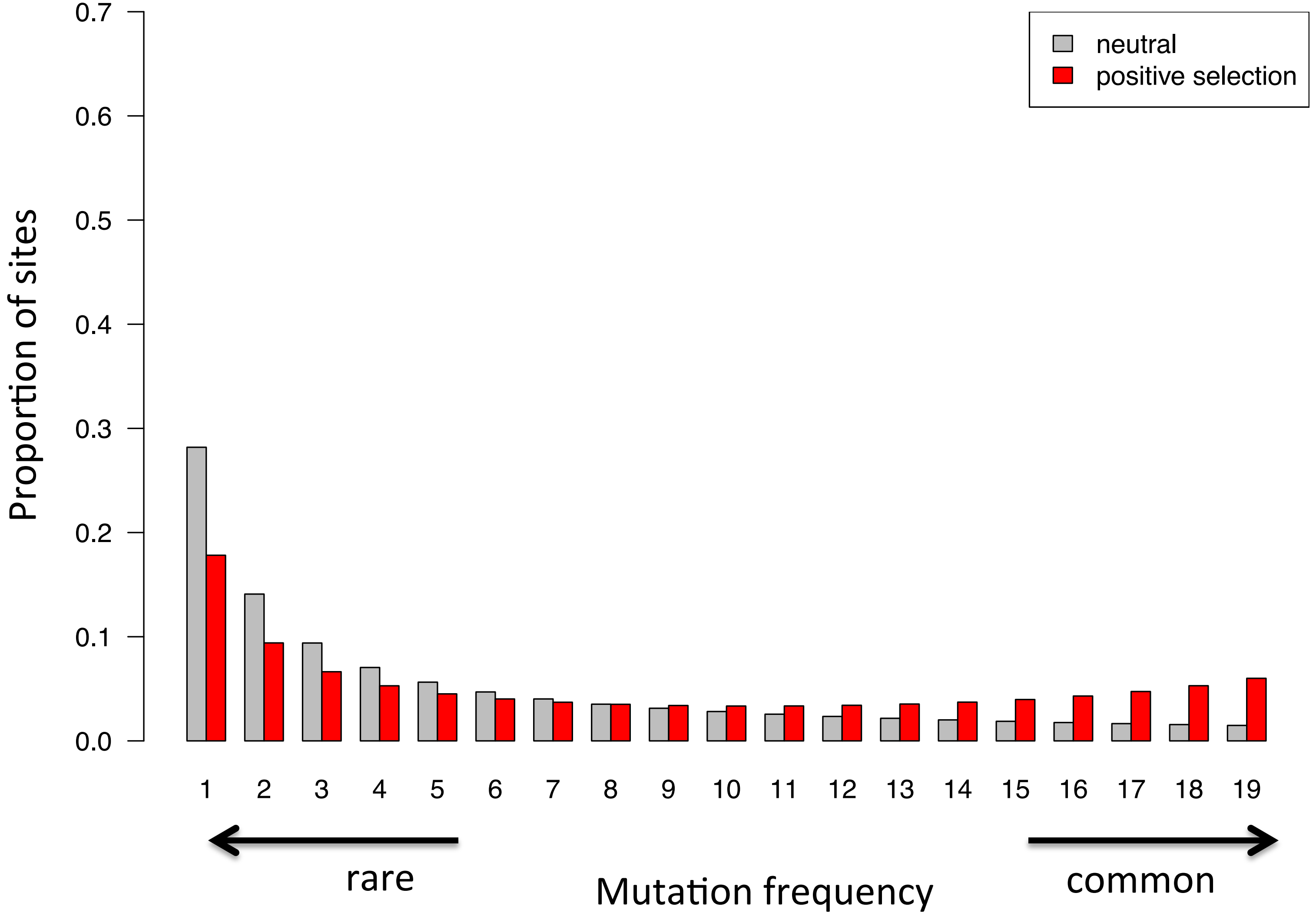
If mutations are arriving at Poisson times



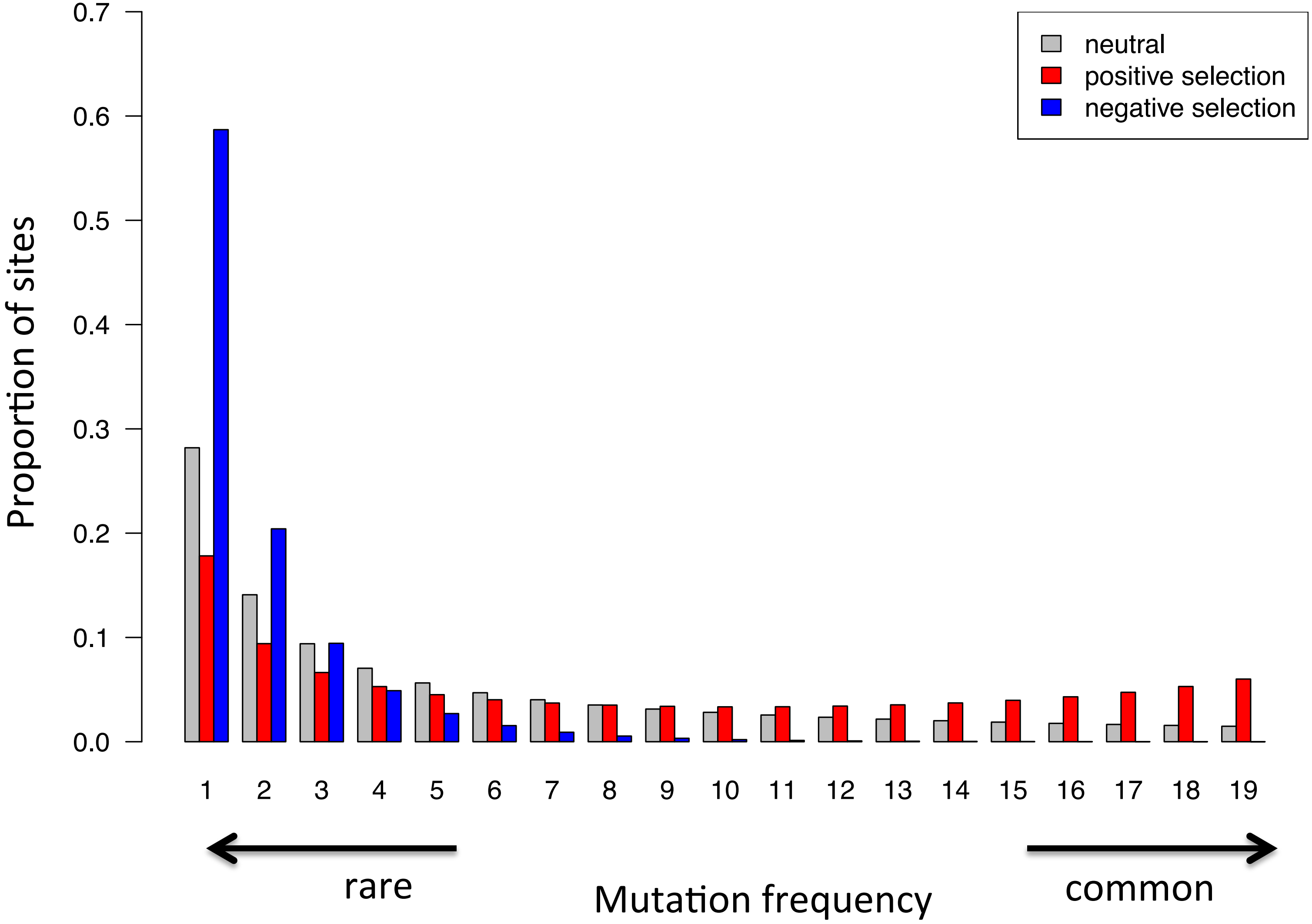
The site frequency spectrum



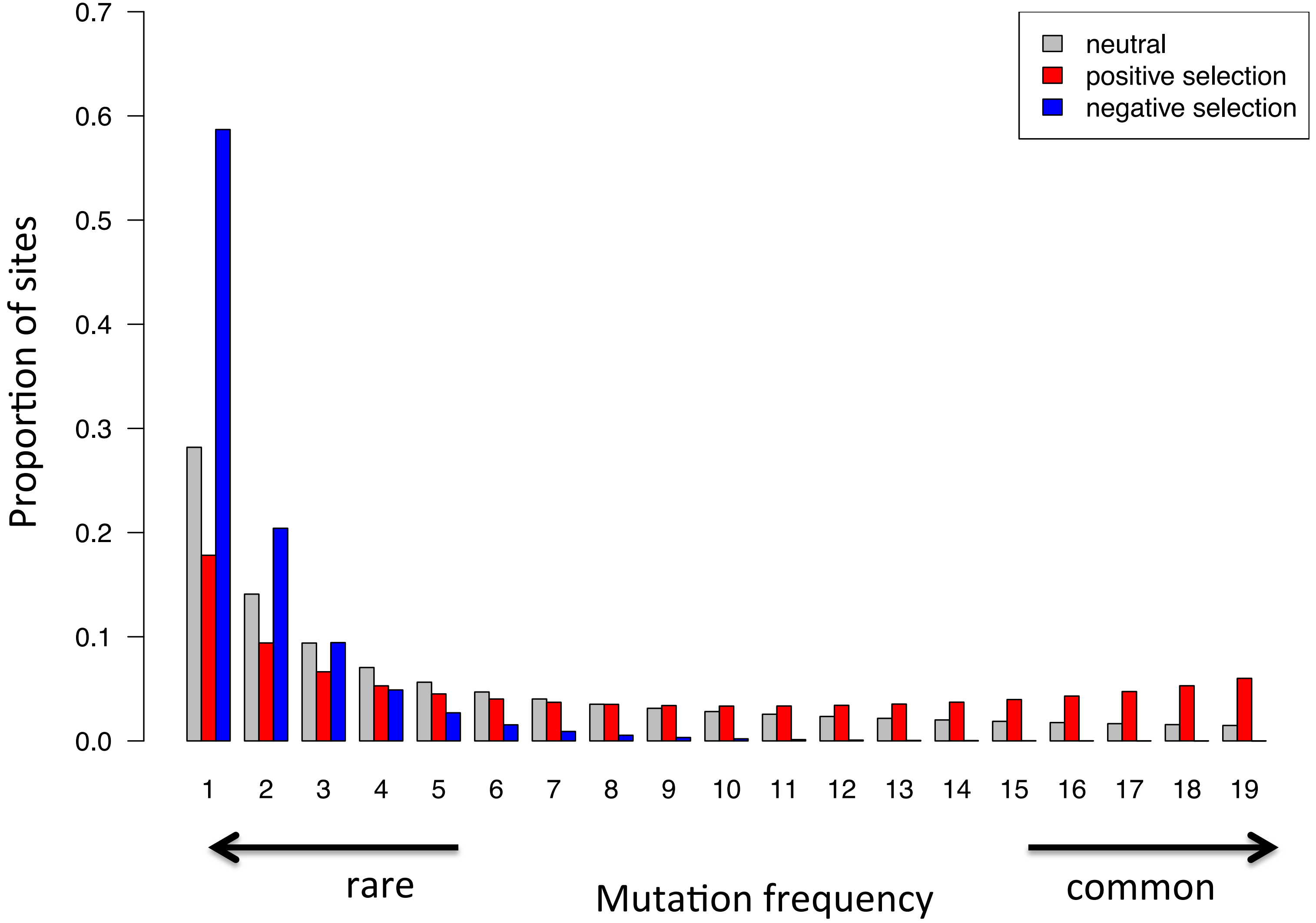
The site frequency spectrum



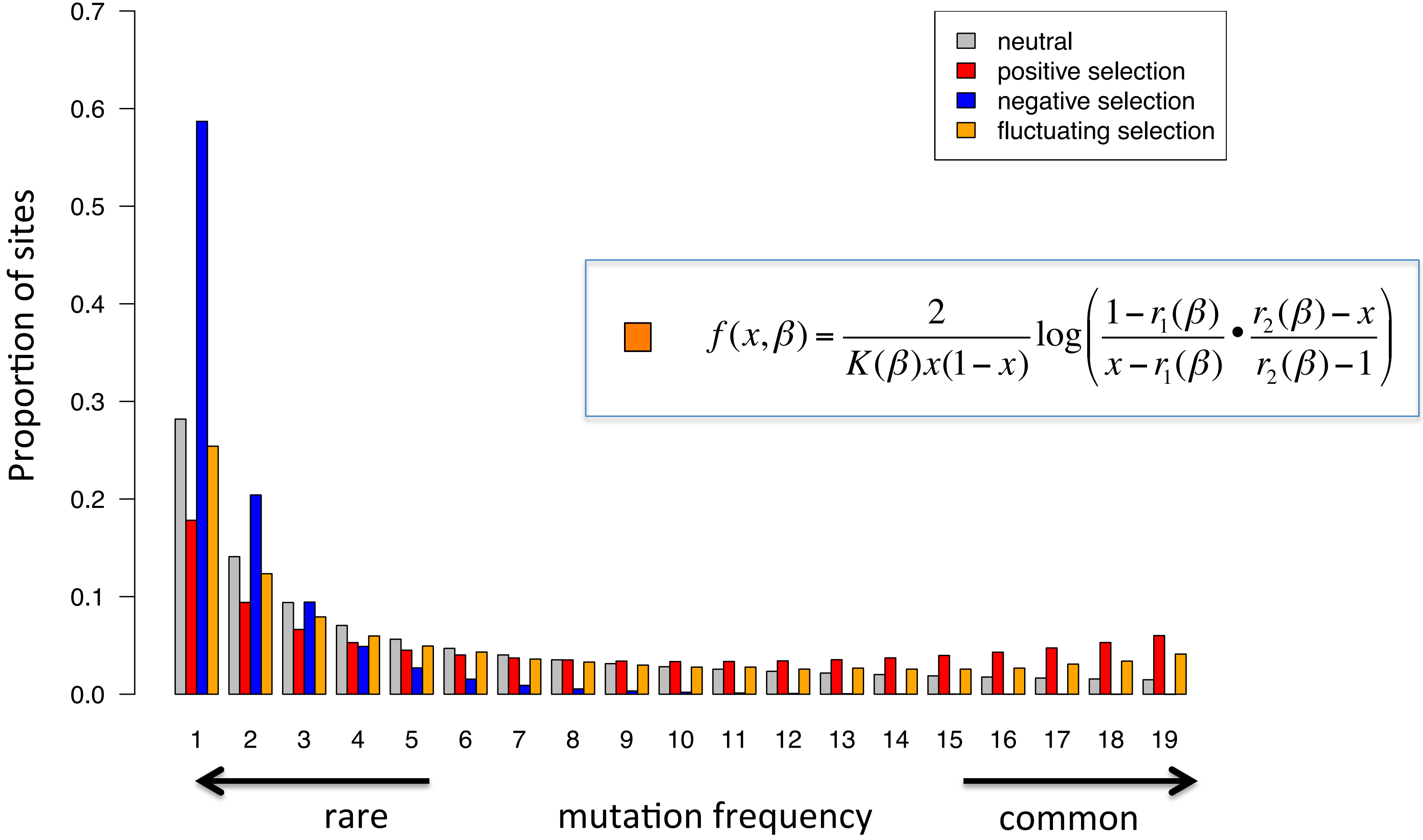
The site frequency spectrum



The site frequency spectrum



The site frequency spectrum



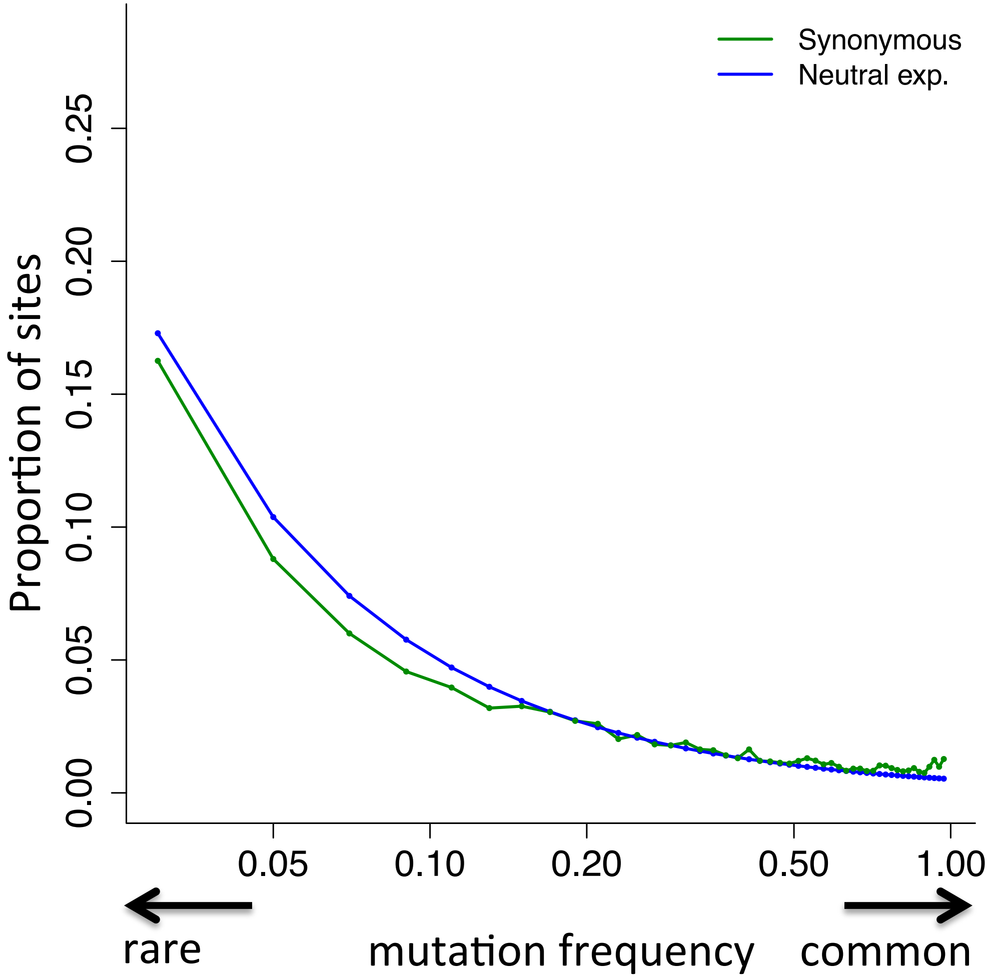
🟧 $$f(x,\beta) = \frac{2}{K(\beta)x(1-x)} \log\left(\frac{1-r_1(\beta)}{x-r_1(\beta)} \bullet \frac{r_2(\beta)-x}{r_2(\beta)-1} \right)$$

Huerta-Sanchez et al. (2008)

Population sample

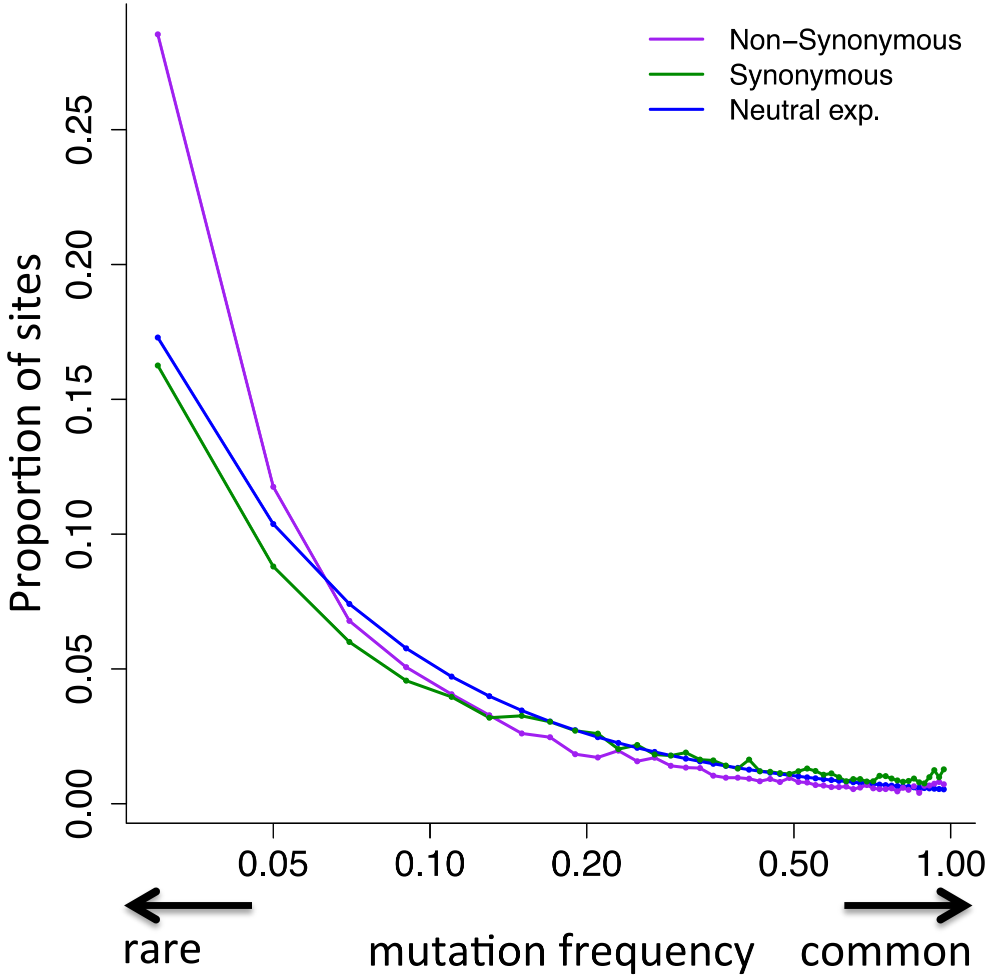
200 individuals of Danish nationality

An excess of rare mutations



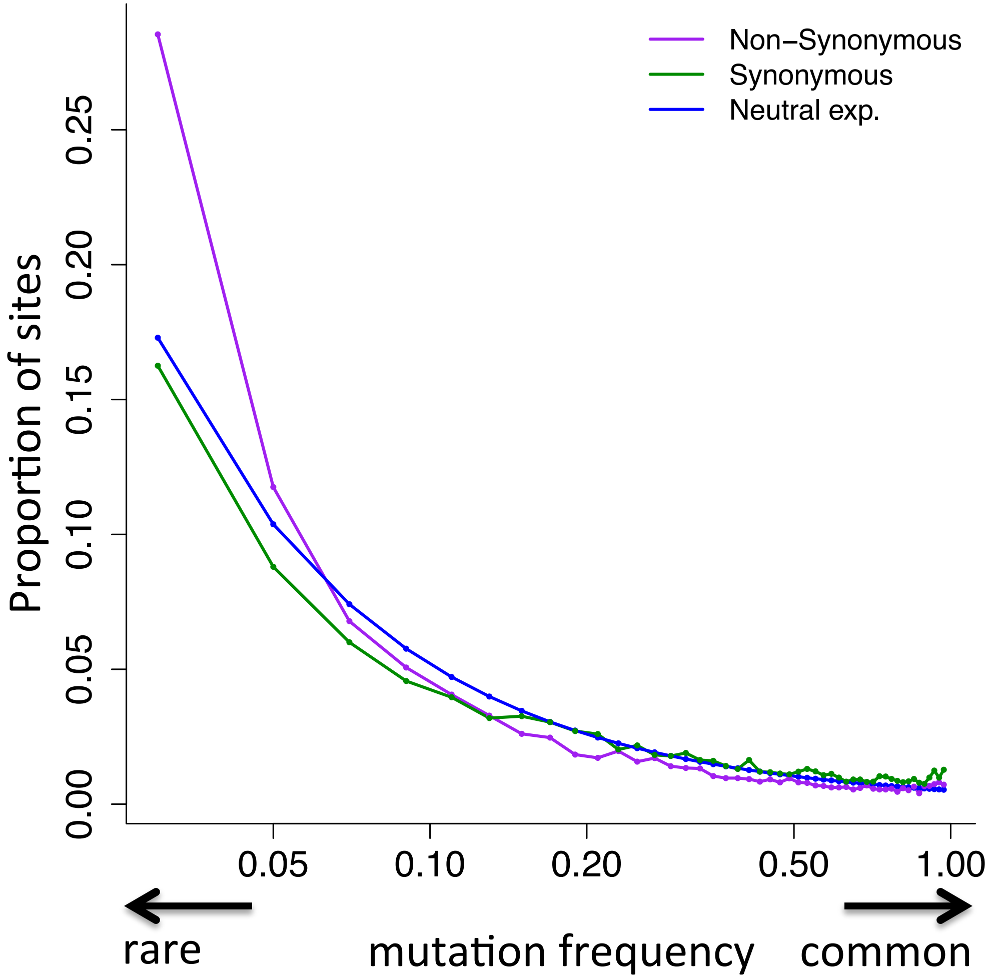
Li* Y. et al. (2010) Nature Genetics

An excess of rare mutations

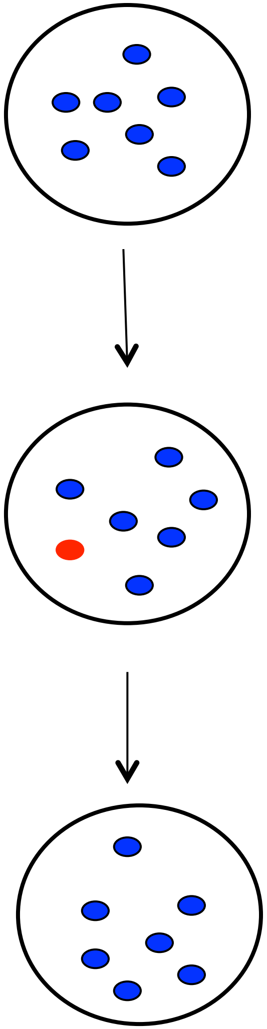


Li* Y. et al. (2010) Nature Genetics

An excess of rare mutations

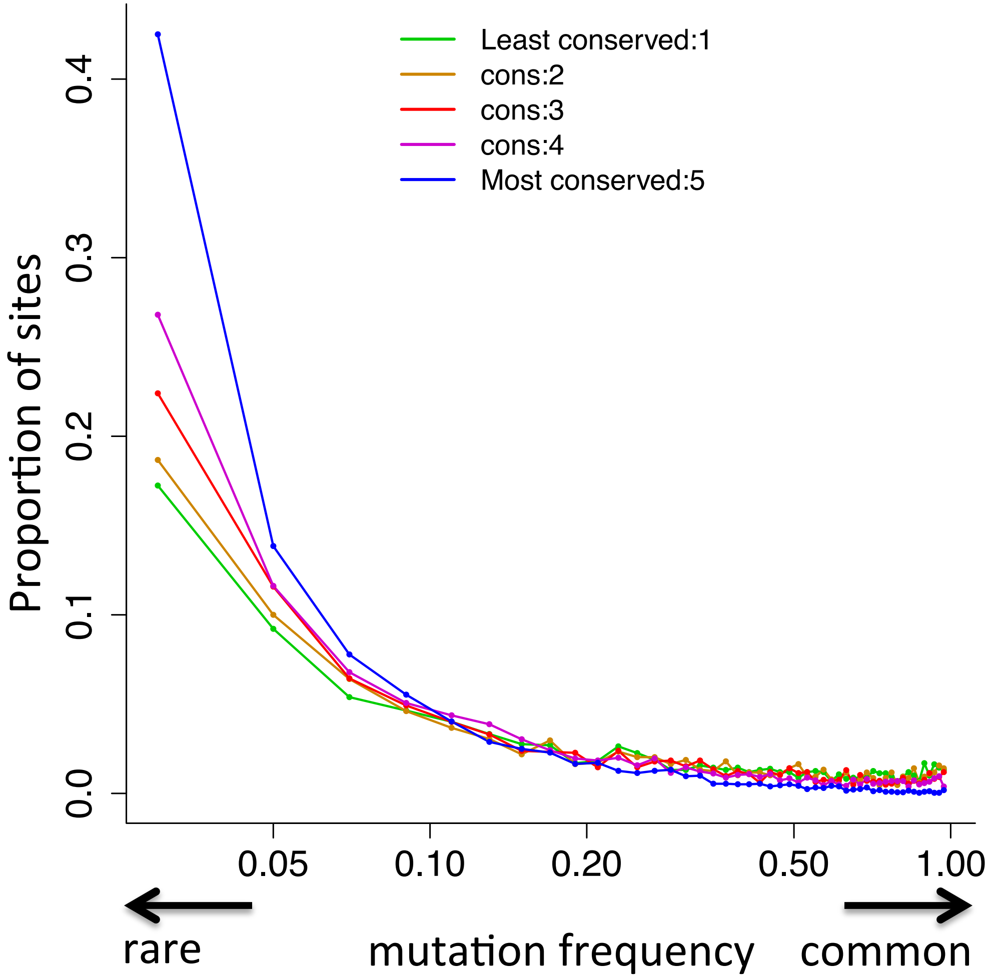


Negative Selection



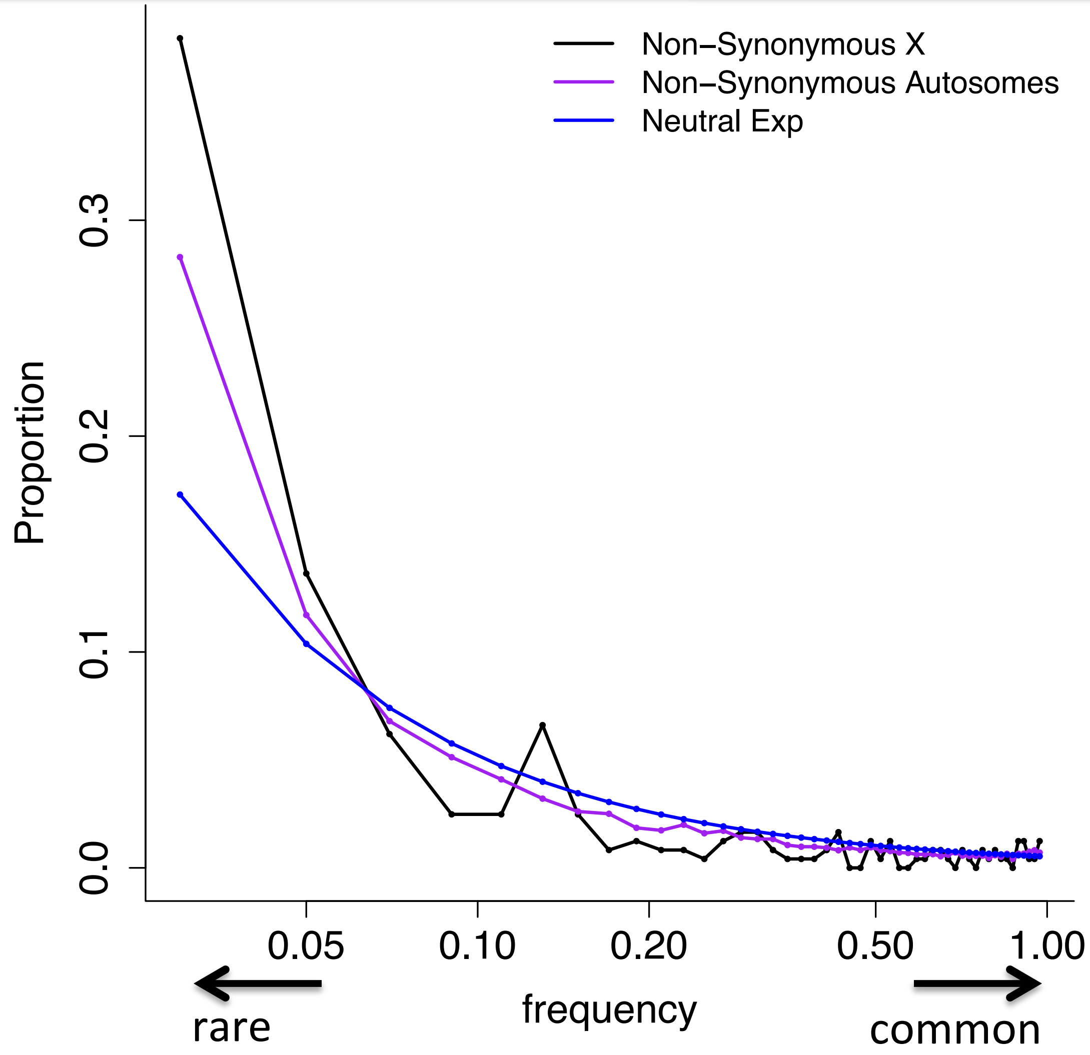
Li* Y. et al. (2010) Nature Genetics

An excess of rare deleterious mutations



Li* Y. et al. (2010) Nature Genetics

X chromosome



Li* Y. et al. (2010) Nature Genetics

Distribution of selective effects of non-synonymous mutations

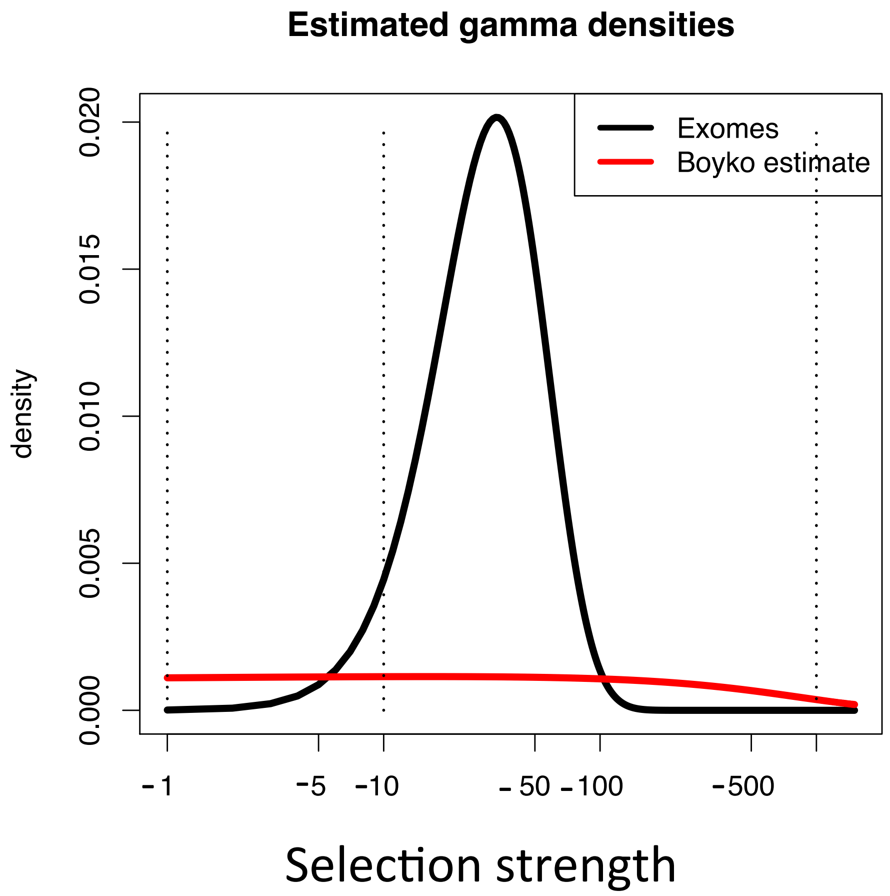
$$f(x) = k\frac{1}{x} + (1-k)\int f_{sel}\left(x,-\gamma\right) Gamma\left(\gamma;\alpha,\beta\right) d\gamma$$

A proportion *k* of mutations are neutral

A proportion *(1-k)* are deleterious

Li* Y. et al. (2010) Nature Genetics

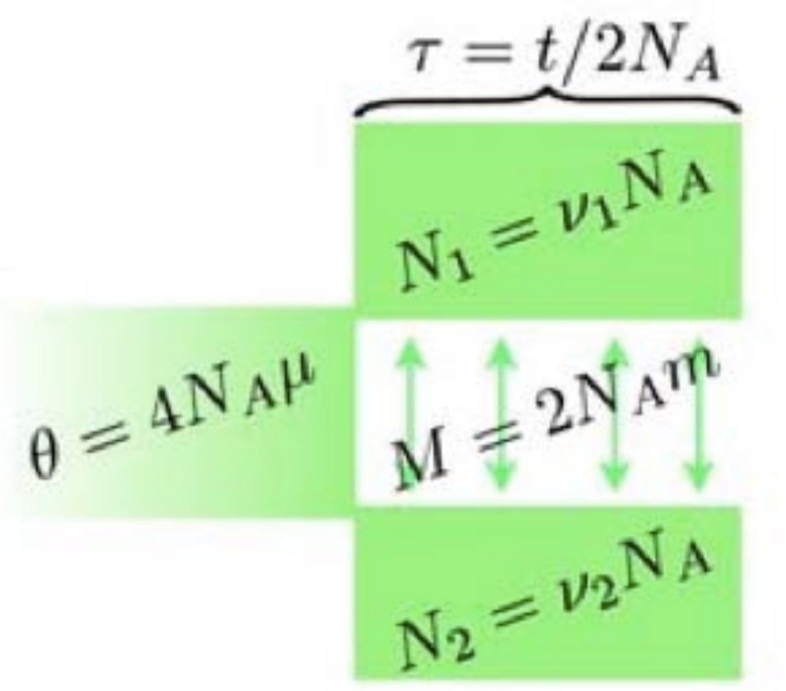
A larger proportion of weakly deleterious mutations



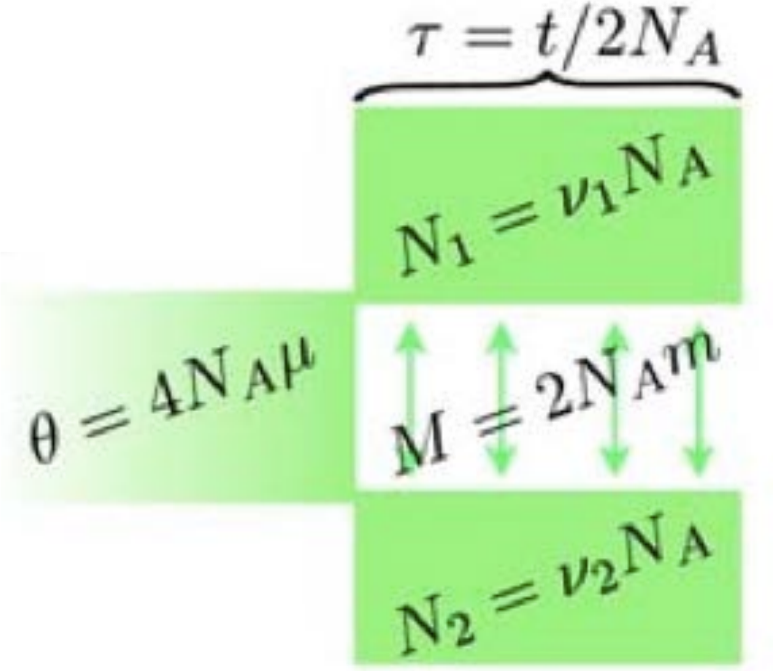
Boyko et al. (2008)
Plos Genetics

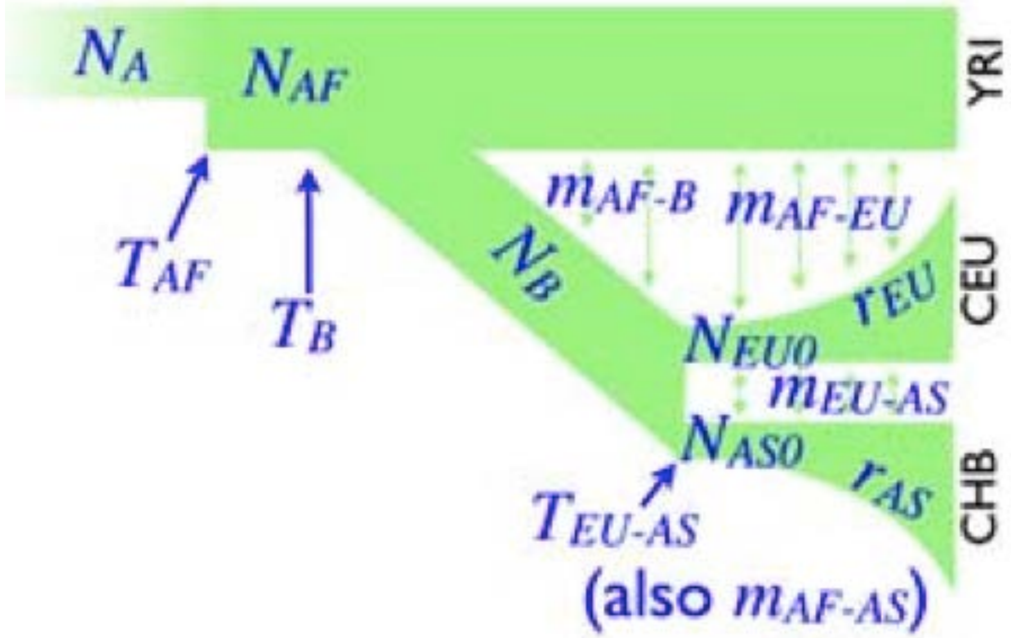
Li* Y. et al. (2010) Nature Genetics

Demographic inference, Gutenkunst et al. 2009

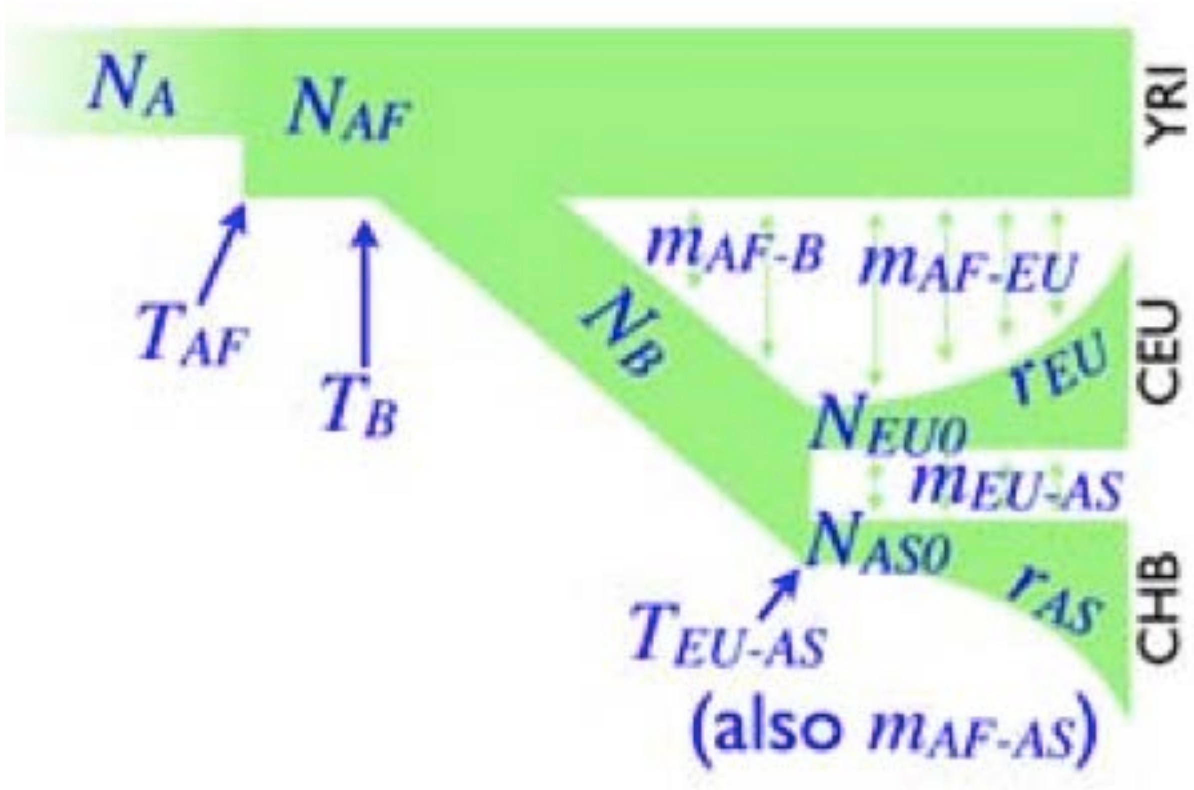


Demographic inference



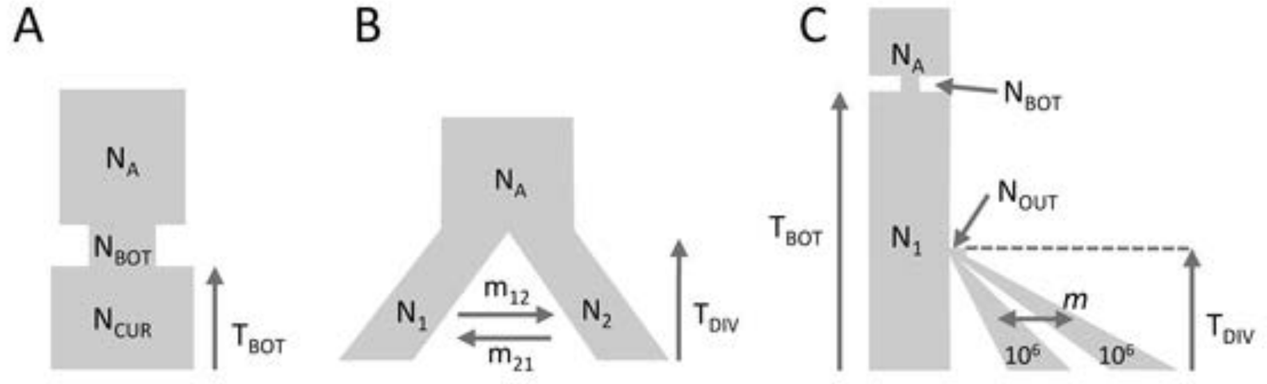


dadi

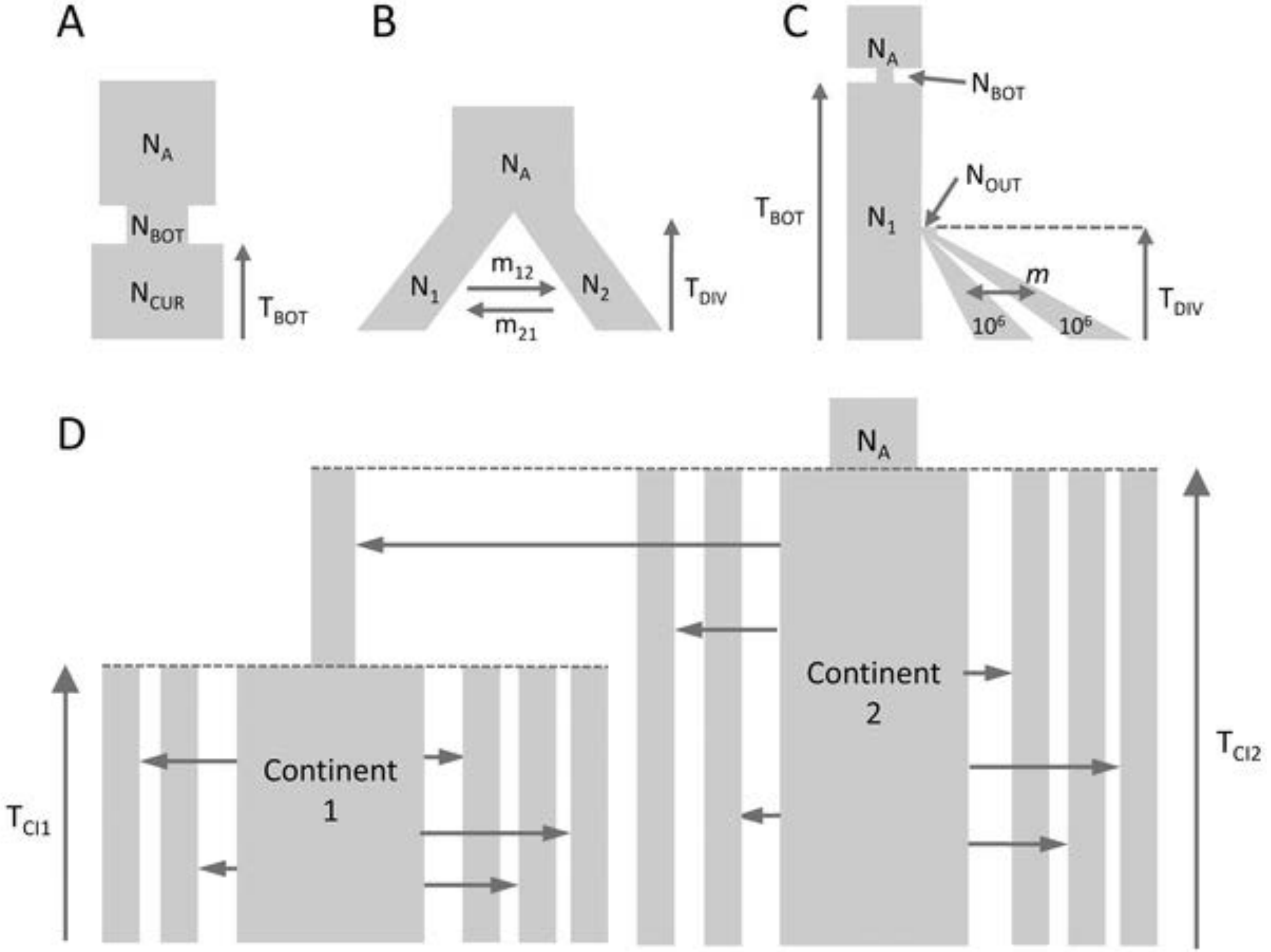


$$\frac{d}{dt} f\left(q,t;\Theta\right) = \frac{1}{2}\frac{d^2}{dq^2}\left\{V\left(q;\Theta\right) f\left(q,t;\Theta\right)\right\} - \frac{d}{dq}\left\{M\left(q;\Theta\right) f\left(q,t;\Theta\right)\right\}$$

dadi

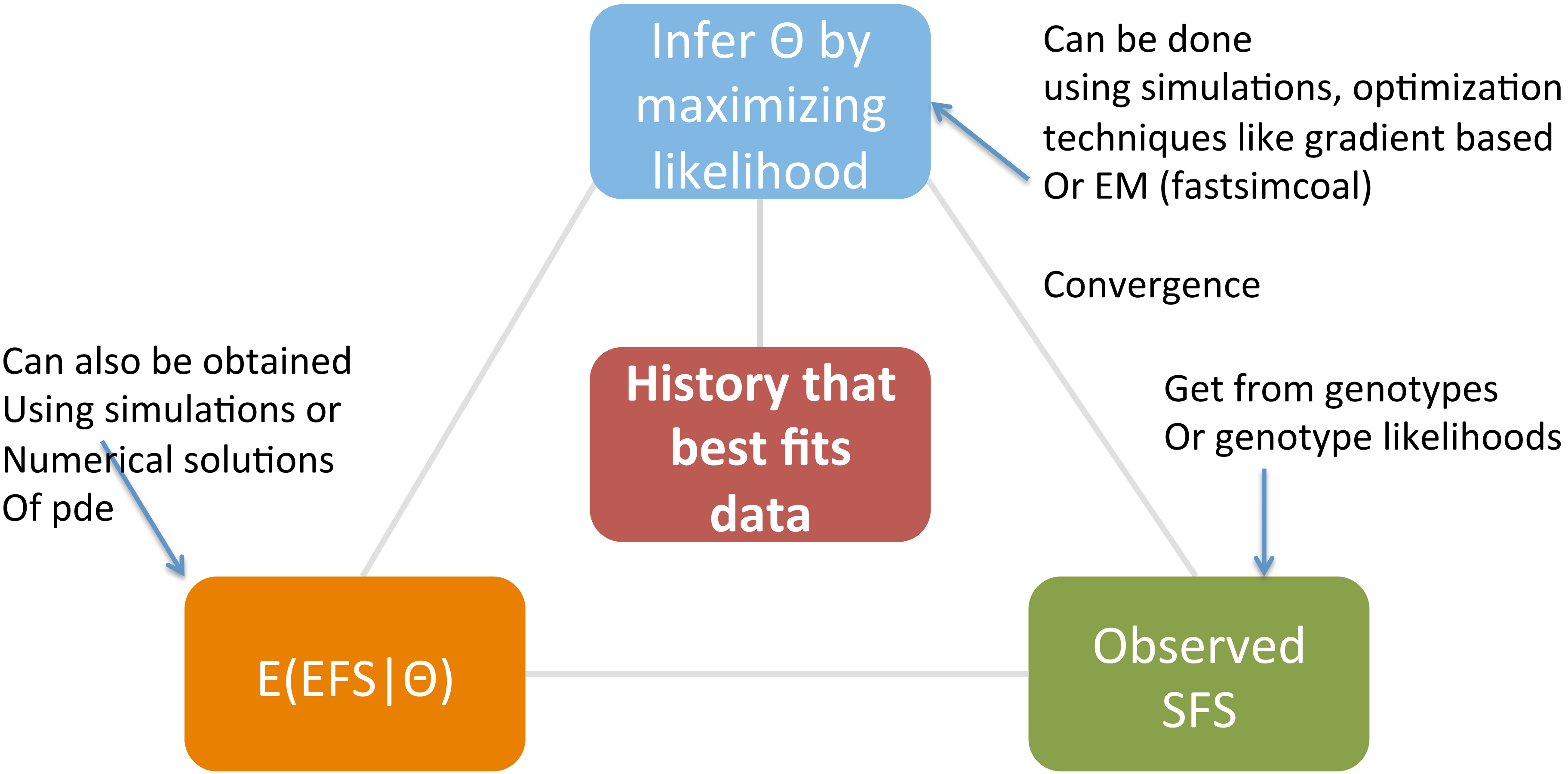

A
N_A
N_BOT
N_CUR
T_BOT
B
N_A
N_1
N_2
m_12
m_21
T_DIV
C
N_A
N_BOT
T_BOT
N_1
N_OUT
m
10^6
10^6
T_DIV

dadi



Fastsimcoalsim
Excoffier et al. 2013

Demographic inference with dadi of fastsimcoal



Infer Θ by maximizing likelihood

Can be done
using simulations, optimization
techniques like gradient based
Or EM (fastsimcoal)

Convergence

History that best fits data

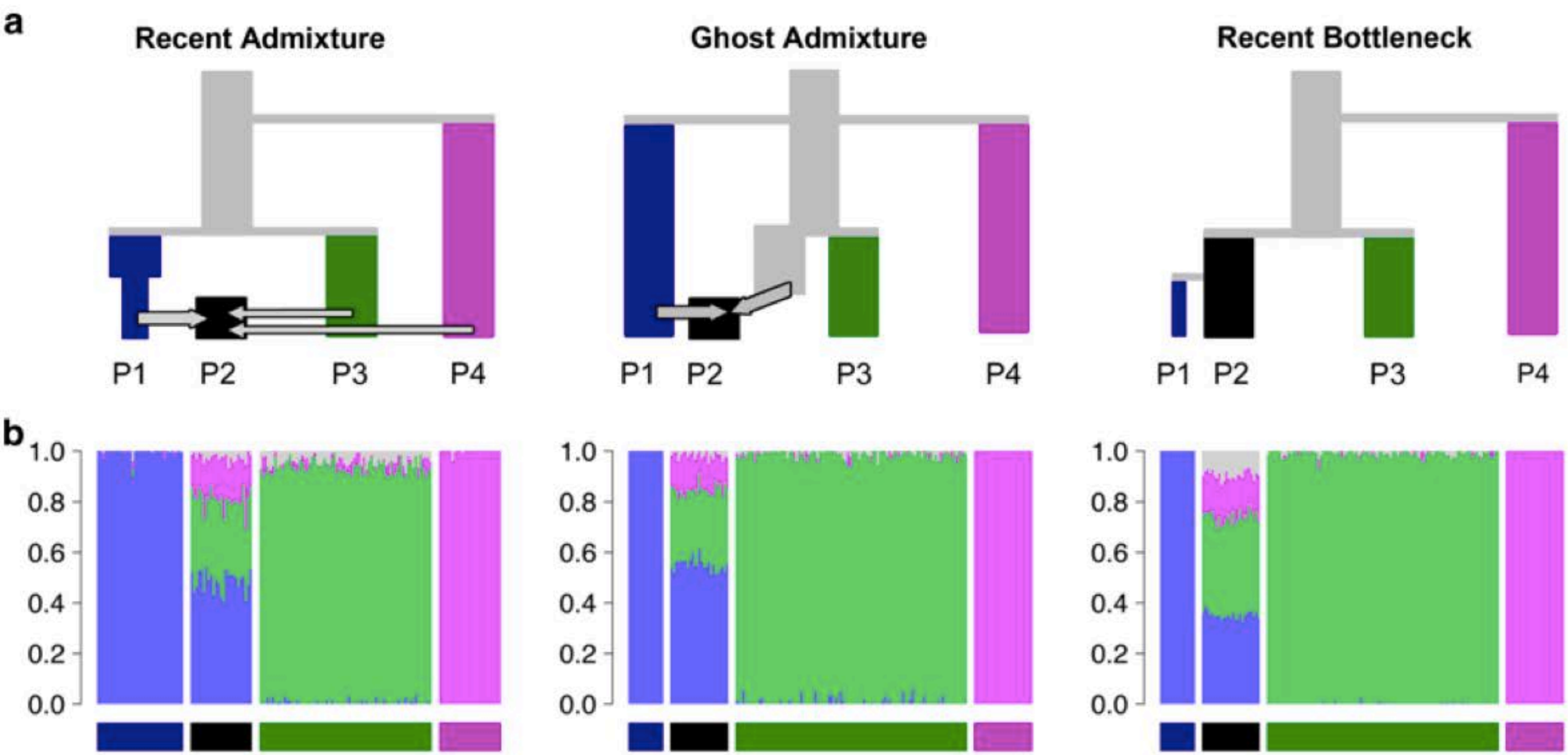
Can also be obtained
Using simulations or
Numerical solutions
Of pde

Get from genotypes
Or genotype likelihoods

E(EFS|Θ)

Observed SFS

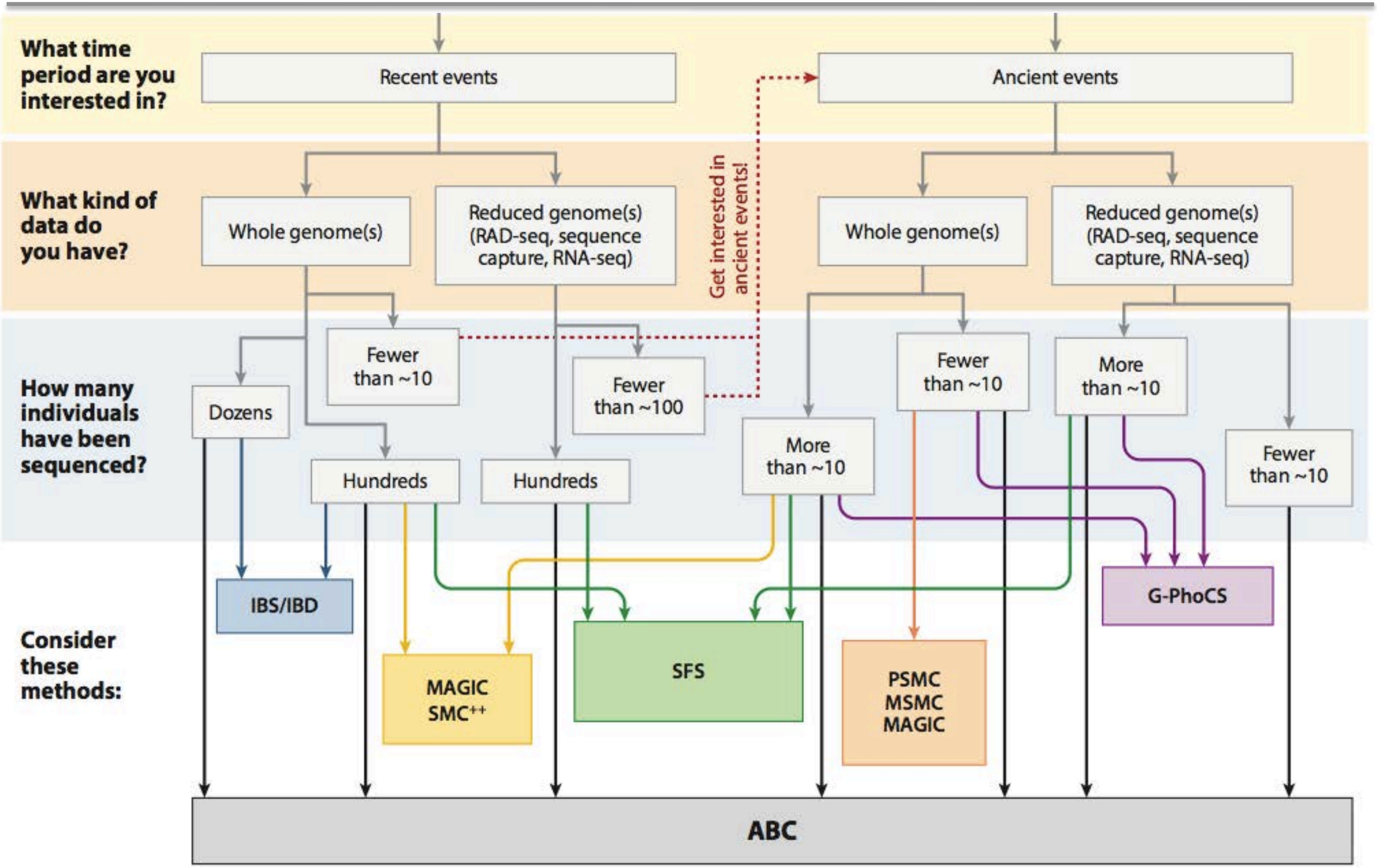
Benefits of model-based approach over Structure



A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots

Daniel J. Lawson, Lucy van Dorp & Daniel Falush

Decision Tree from Beichman et al. (2018) Annual Review of Ecology, Evolution, and Systematics


What time period are you interested in?
Recent events
Ancient events
What kind of data do you have?
Whole genome(s)
Reduced genome(s) (RAD-seq, sequence capture, RNA-seq)
Whole genome(s)
Reduced genome(s) (RAD-seq, sequence capture, RNA-seq)
Get interested in ancient events!
How many individuals have been sequenced?
Dozens
Fewer than ~10
Hundreds
Fewer than ~100
Hundreds
More than ~10
Fewer than ~10
More than ~10
Fewer than ~10
Consider these methods:
IBS/IBD
MAGIC
SMC++
SFS
PSMC
MSMC
MAGIC
G-PhoCS
ABC

We need to consider demography when detecting selection

- When the population size grows, genetic drift slows down and selection becomes more efficient
- A population bottleneck speeds up genetic drift and makes natural selection get less efficient
- Population size changes shift the site frequency spectrum away from the shape SFS(*n*)~1/*n* that is expected under neutrality
- Natural selection also shifts the SFS away from the equilibrium 1/*n* shape
- The SFS shape can be used to test for selection, but only after correcting for nonequilibrium demography