How to detect selection in DNA sequences?

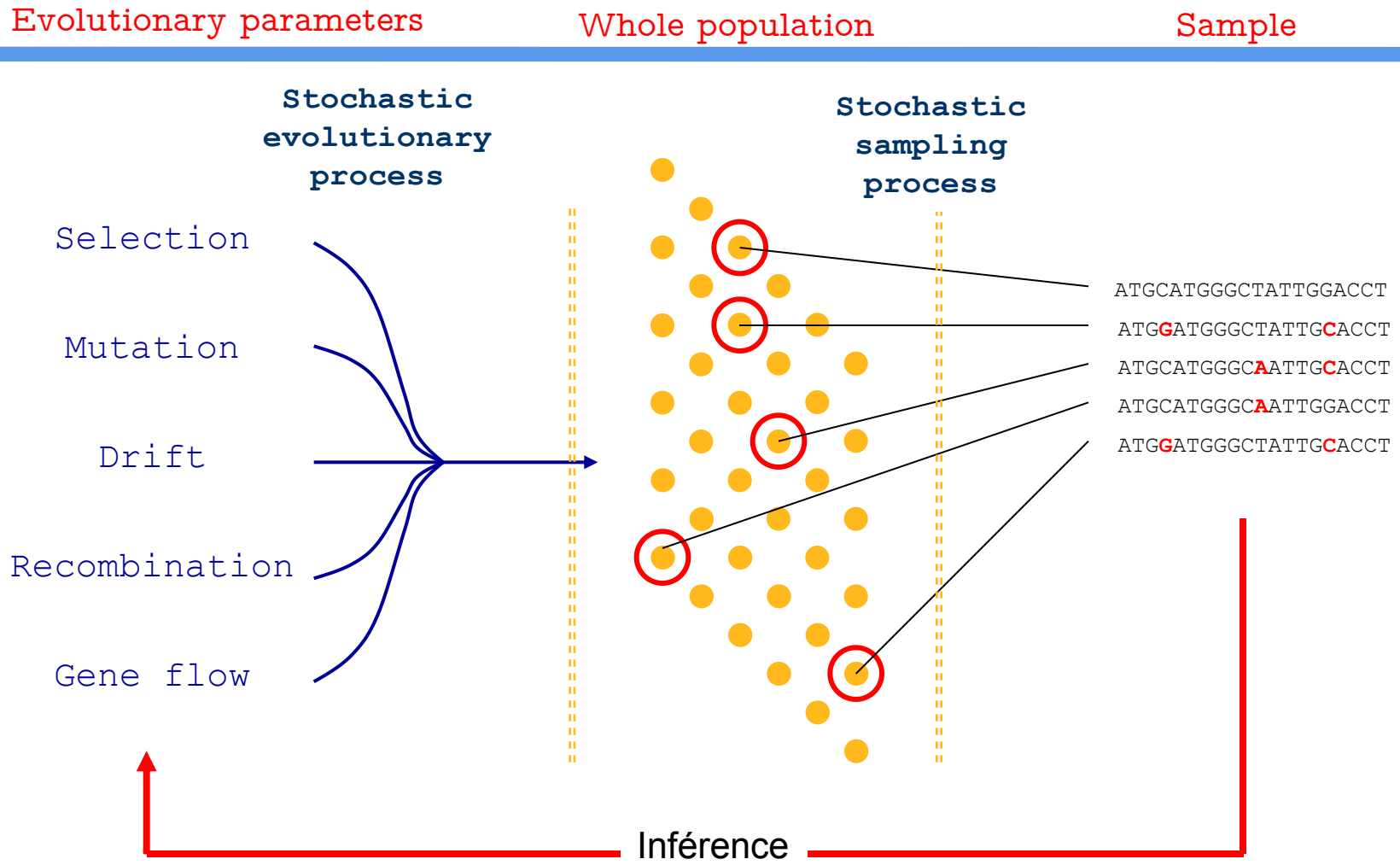
C. Lemaire

Norwich, April 5-6th 2017

MUTATION, RECOMBINATION AND SELECTION : POPULATION GENOMICS INFERENCES

- Polymorphic sites identification
- Polymorphism rate estimation
- What forces shape polymorphism?

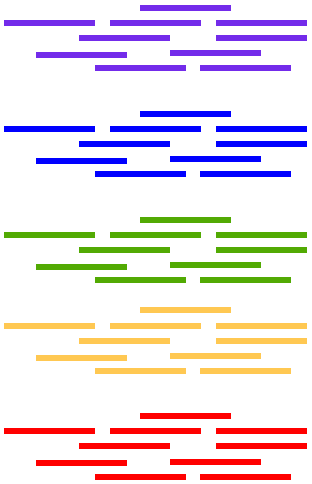
INFERENCES IN POPULATION GENOMICS



Polymorphism identification

DNA or RNA Extraction

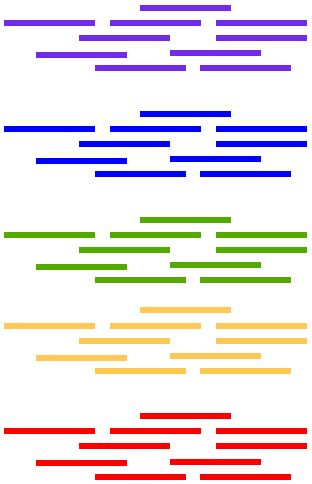
Sequencing
454
Illumina, …



Polymorphism identification

DNA or RNA Extraction

Sequencing
454
Illumina, …



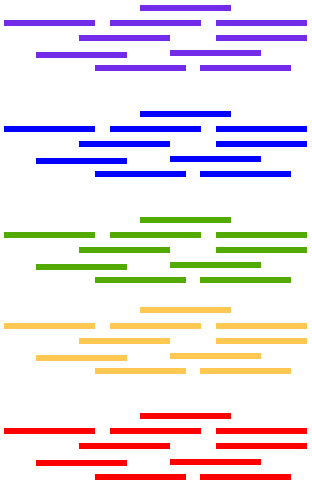
***de novo* Assembling**
ou
Alignment on reference

SNP calling
Indel calling

Polymorphism identification

DNA or RNA Extraction

Sequencing
454
Illumina, …



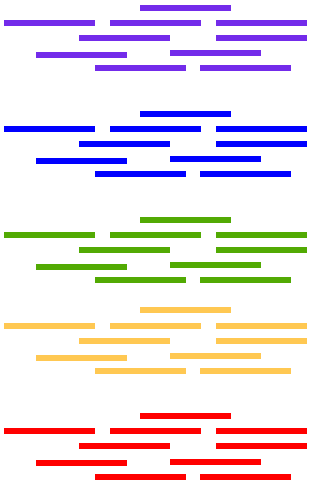
***de novo* Assembling**
ou
Alignment on reference

SNP calling
Indel calling

Polymorphism identification

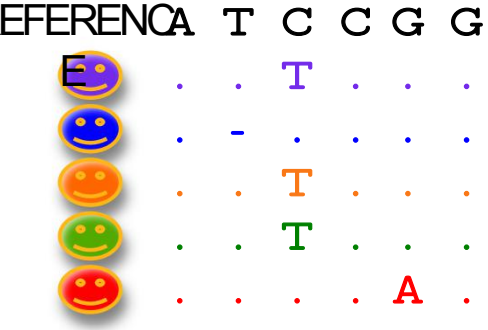
DNA or RNA Extraction

Sequencing
454
Illumina, …



***de novo* Assembling**
ou
Alignment on reference

SNP calling
Indel calling



Polymorphism identification

***de novo* Assembling**
or
Alignment on reference

SNP calling
Indel calling

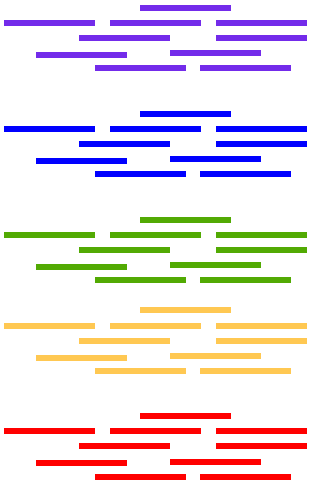
- Read quality
- Mate pair alignment
- Coverage thresholds min & max
- Max divergence rate (e.g. BWA 8% + 2 indels)

Keep information on overall number of sites (i.e. **non filtered**) for polymorphism rates estimates

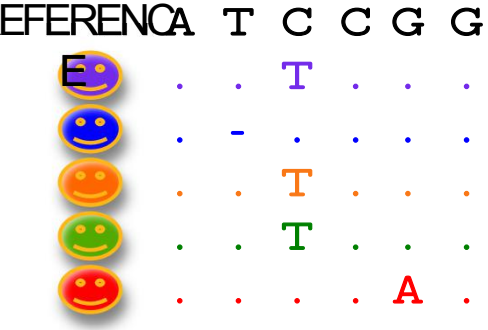
Polymorphism identification

DNA or RNA Extraction

Sequencing
454
Illumina, …



***de novo* Assembling**
ou
Alignment on reference



SNP calling
Indel calling

Rare but *C o o l* :
Structural Variants

Polymorphism identification

Divergence islands detection

- *de novo* assembling > alignment on référence (e.g. limit <8% divergence for BWA)

CONCLUSIONS

- Polymorphic sites are mainly « Single Nucleotide Polymorphism » or SNPs .
- WARNING: non covered regions and filters for SNP calling for polymorphism rates estimates.
- Coverage analysis: « Structural Variants » : duplications, deletions, insertion, divergence islands, that could have strong phenotypic effects.

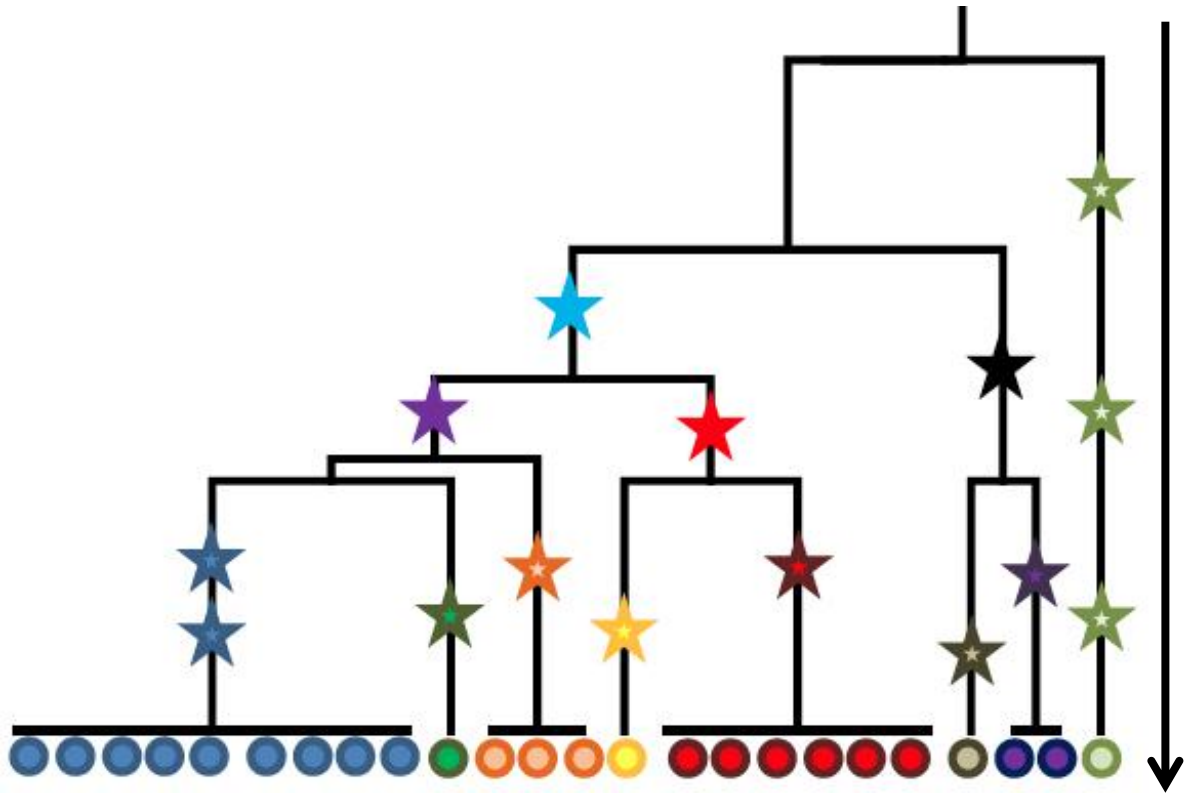
Polymorphism rate estimation

POLYMORPHISM RATE

CC GCA GAG TTA CTA ATC GAA $N = 6$ individuals
CG GCA GAG TTA CTA ATC GAA $L = 21$ sites on alignment
CC GCA AAG TTA CCA ATT GAA $S = 5$ polymorphic sites
CC GCA GAG TTA CCA ATC GAA
CC GCA AAG TTA CTA ATC GAG
CC GCA AAG TTA CTA ATC GAA

How to quantify polymorphism?

Polymorphism is a combination of mutation and time



Genealogical relationships between sequences

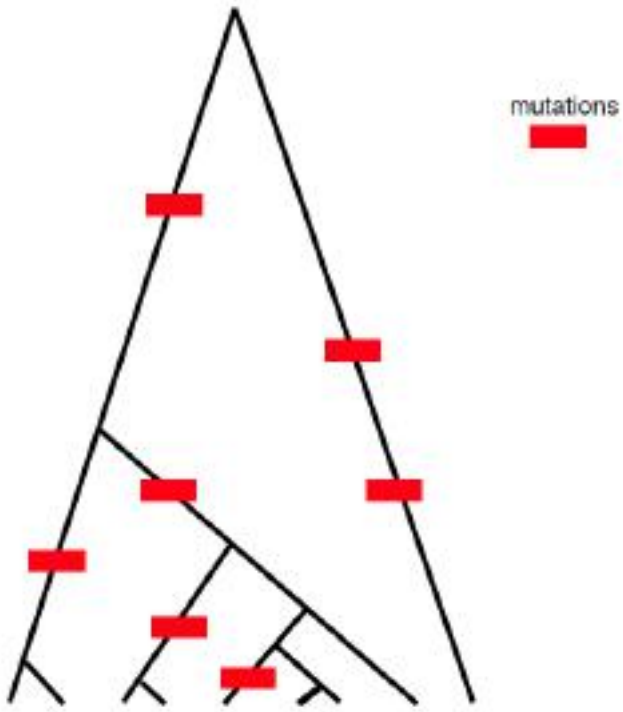
History (Time)

DNA sequences → Polymorphism

$P=f(T,\mu)$

Genealogical relationships between sequences

MRCA: Most Recent Common Ancestor

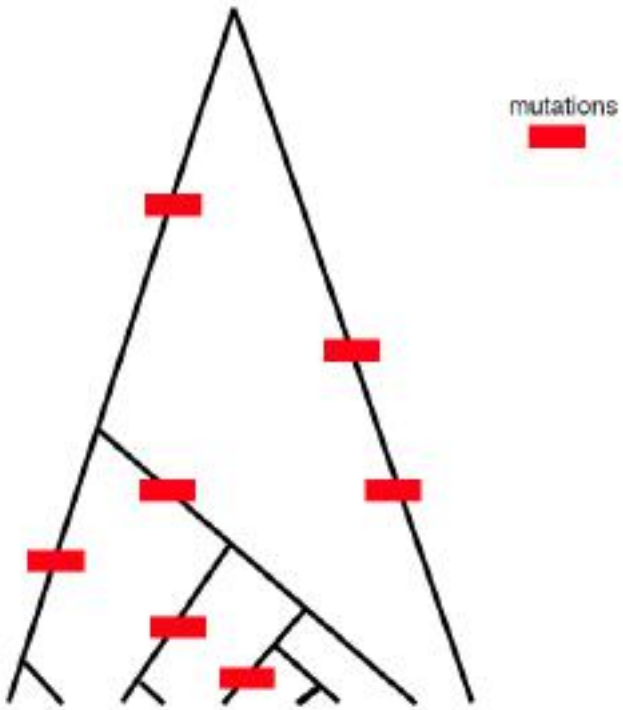


(Achaz, Introduction à la coalescence, 2005)

The coalescent, *i.e.* the genealogy of alleles at a locus, depends on evolutionary forces that shape polymorphisms.
Null hypothesis in evolution.

More details on the coalescence theory on:
http://www.sfu.ca/biology/courses/bisc869/869_lectures/MHP_Coalescent.pdf

MRCA: Most Recent Common Ancestor



(Achaz, Introduction à la coalescence, 2005)

Number of mutations
Poisson approximation of Binomial

$$P(k/t) = e^{-\mu t}\frac{(\mu t)^k}{k!}$$

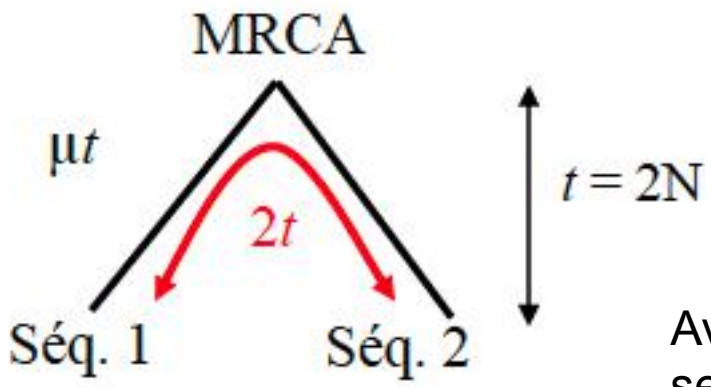
$$E(k/t) = \mu t$$

Poisson distribution

K: number of mutations
μ: mutation rate/gen/locus

Genealogical relationships between sequences

What is the expected number of differences between two sequences?



$$\underbrace{P(k/t) = e^{-\mu t}\frac{(\mu t)^k}{k!}}_{\text{Poisson distribution}} \qquad E(k/t) = \mu t$$

K: number of mutations
μ: mutation rate/gen/locus
T: nb generations

Average coalescence time between two sequences

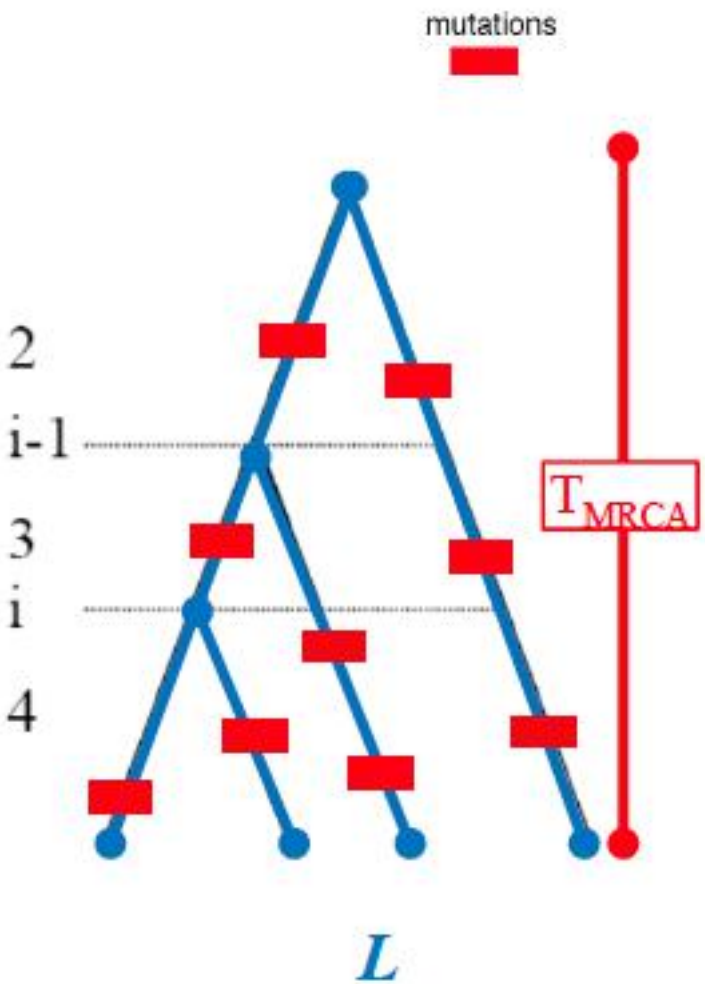
$$E(t_2) = 2N \qquad T_i = \frac{t_i}{2N} \qquad E(T_2) = 1$$

$$E(k_2) = 2 \times \mu \times E(t_2)$$

$$E(k_2) = 2 \times \mu \times 2N = 4N\mu = \theta$$

q Is the populational mutation rate.
Easier to estimate than *N* and *μ*

The Watterson's q_S



Length of a genealogy

$$L = a_n \times 4N \quad \text{with} \quad a_n = \sum_{i=1}^{n-1} \frac{1}{i}$$

Expected number of polymorphic sites

$$E(S) = \mu \times L$$

$$E(S) = \mu \times a_n \times 4N$$

$$E(S) = a_n \theta$$

$$\theta_S = \frac{S}{a_n}$$

The Tajima's P

P is the expected average number of differences between any pairs of sequences in the population

$$\Pi = \frac{1}{C_n^2} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij}$$

$$E(d_{ij}) = E(k_2) = 2 \times \mu \times 2N = 4N\mu = \theta$$

$$\theta_\Pi = E(\Pi)$$

Polymorphism rate estimation

POLYMORPHISM RATE

CC GCA GAG TTA CTA ATC GAA
CG GCA GAG TTA CTA ATC GAA
CC GCA AAG TTA CCA ATT GAA
CC GCA GAG TTA CCA ATC GAA
CC GCA AAG TTA CTA ATC GAG
CC GCA AAG TTA CTA ATC GAA

$N = 6$ individuals
$L = 21$ sites on alignment
$S = 5$ polymorphic sites

$$P = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{d_{ij}}{L.N.(N-1)/2}$$

$$Q = \frac{S}{L.\sum_{i=1}^{N-1} 1/i}$$

Polymorphism rate estimation

POLYMORPHISM RATE

CC GCA GAG TTA CTA ATC GAA
CG GCA GAG TTA CTA ATC GAA
CC GCA AAG TTA CCA ATT GAA
CC GCA GAG TTA CCA ATC GAA
CC GCA AAG TTA CTA ATC GAG
CC GCA AAG TTA CTA ATC GAA

$N = 6$ individuals
$L = 21$ sites on alignment
$S = 5$ polymorphic sites

$$P = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{p_{ij}}{LN.(N-1)/2}$$

$P = 0.102$

$$Q = \frac{S}{L \sum_{i=1}^{N-1} 1/i}$$

$Q = 0.104$

The Fu & Li's q_{he}

Internal (solid) and external (dashed) in a genealogy.

Estimator based on the mutations found on external branches.

Present in one copy: singleton

$$\eta_e = L_n \mu = 4N\mu$$

$$E(\eta_e) = \theta$$

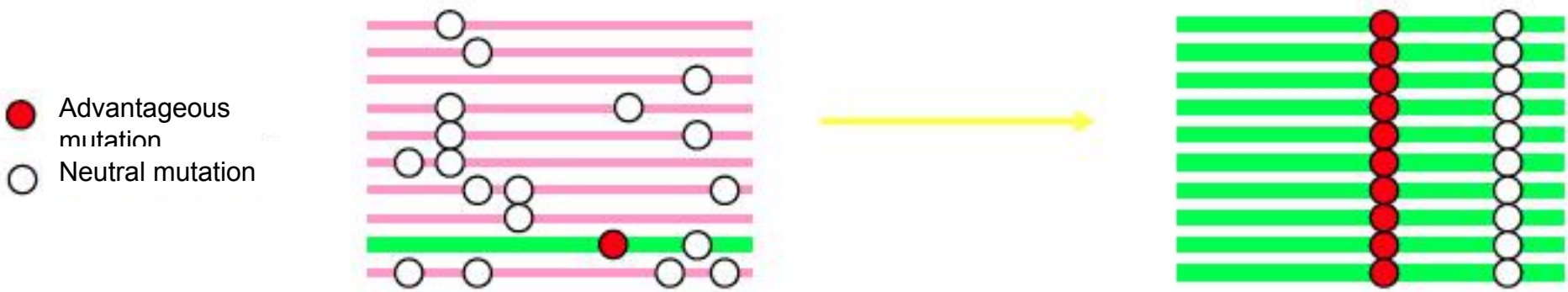
Theoretical expectations of the polymorphism rate

- Neutral mutations
- Constat size population

$P = Q = 4 N_e \mu$ diploïd population

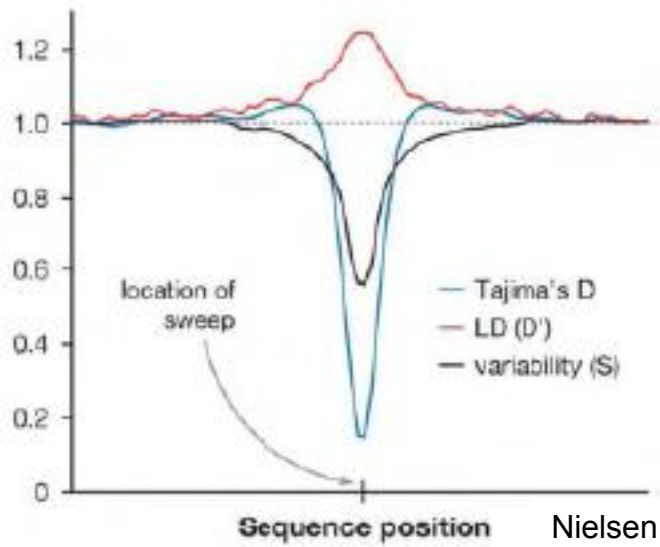
$P = Q = 2 N_e \mu$ haploïd population

Positive selection



An advantageous mutation will increase in frequency and eliminate all polymorphism around

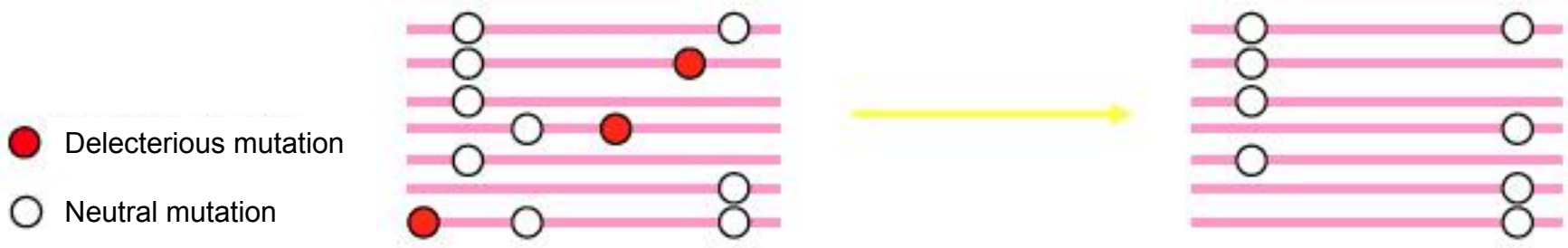
Selective sweeps



Nielsen (2005)

Purifying selection

Selection against deleterious mutations

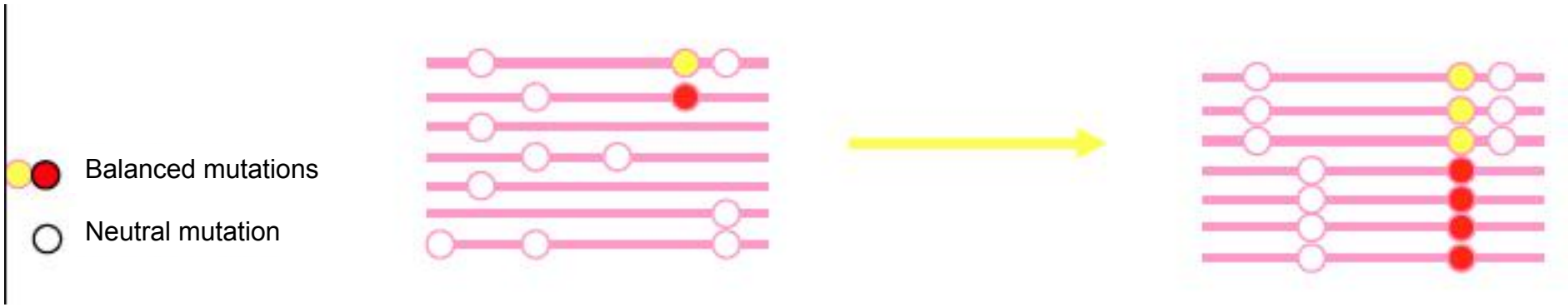


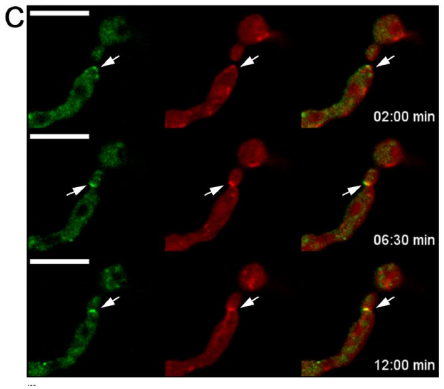
Related to gene functional constraints: Housekeeping genes

Selection against change!!

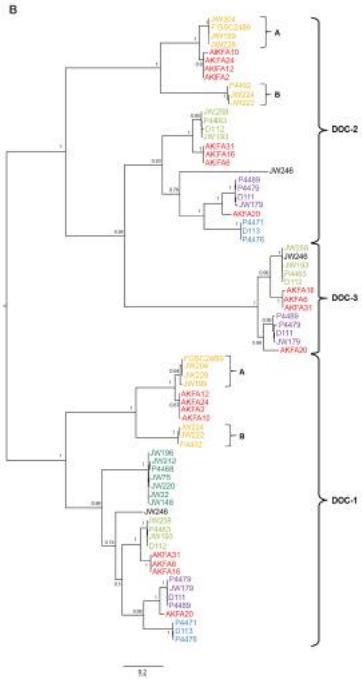
Balancing selection

Increases the variability into the population





Recognition genes doc-1 doc-2 and doc-3 in *Neurospora sp.*



Neutrality tests are based on comparisons between estimators of q

Watterson's q_S	$\theta_S = \frac{S}{a_n}$	$a_n = \sum_{i=1}^{n-1} \frac{1}{i}$
Tajima's q_P	$\theta_\Pi = E(\Pi)$	
Fu & Li's q_{he}	$\theta_{\eta_e} = E(\eta_e)$	

$$T = \frac{\widehat{\theta_1} - \widehat{\theta_2}}{\sqrt{var(\widehat{\theta_1} - \widehat{\theta_2})}}$$

Tajima's D (1983)

Tajima's q_P

$$\theta_\Pi = E(\Pi)$$

Watterson's q_S

$$\theta_S = \frac{S}{a_n} \qquad a_n = \sum_{i=1}^{n-1} \frac{1}{i}$$

$$D = \frac{\Pi - \frac{S}{a_n}}{\sqrt{var\left(\Pi - \frac{S}{a_n}\right)}}$$

D<0 Excess of rare variants -> Population expansion / selective sweep / purifying selection

D>0 Deficit of rare variants -> Population decline / Population structure/ balancing selection

Adaptive evolution of non-coding DNA in *Drosophila*

Peter Andolfatto[1]

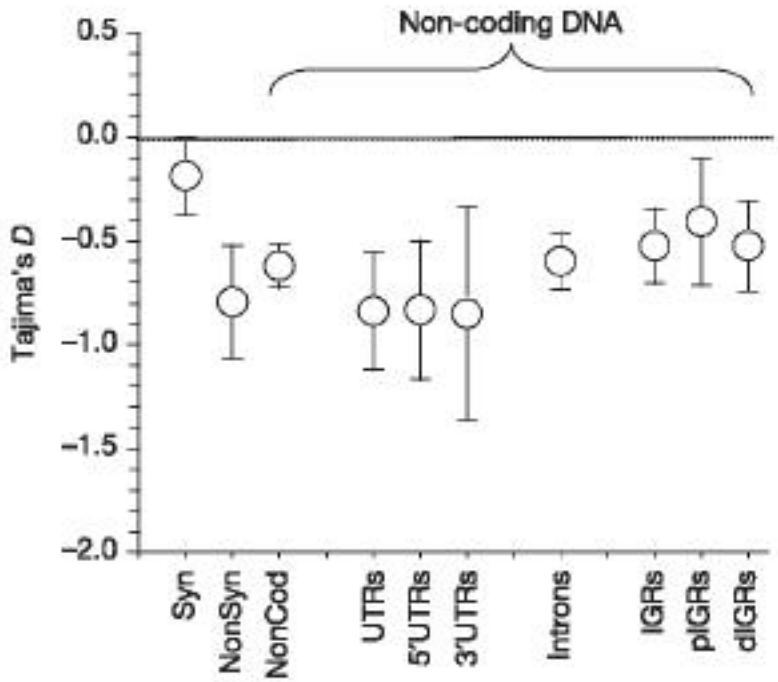


Figure 1 | Mean Tajima's *D* values for coding and non-coding DNA. Means across loci are given with bars indicating two standard errors. The expectation of *D* under the neutral model is shown as a dotted line. Syn, synonymous sites; NonSyn, non-synonymous sites; NonCod, pooled non-coding DNA.

Fu & Li's F (1993)

Tajima's q_P

$$\theta_{\Pi} = E(\Pi)$$

Fu & Li's q_{he}

$$\theta_{\eta_e} = E(\eta_e)$$

$$F = \frac{\Pi - \eta_e}{\sqrt{var(\Pi - \eta_e)}}$$

P Should be less affected by selection than q_{he}

Detection of recent selective sweeps

Fay & Wu's H (2000)

Tajima's q_P

$\theta_\Pi = E(\Pi)$

ξ_i: Number of mutations present *i* times

$\theta_\xi = iE(\xi_i)$

For *i* = 1 $\theta_{\eta_e} = E(\eta_e)$

$$\theta_H = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i^2 \xi_i$$

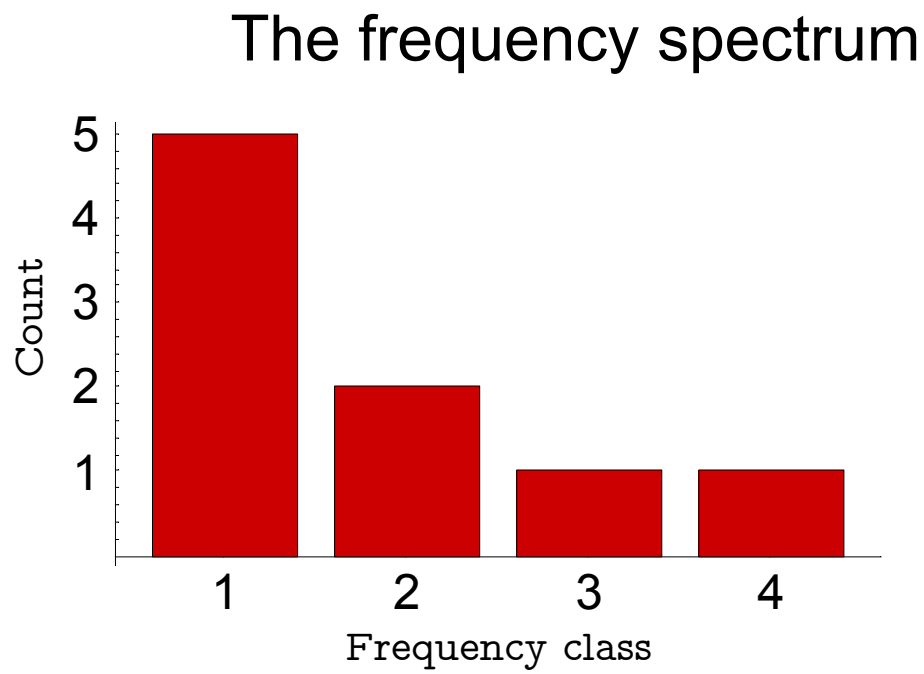
$$H = \frac{\Pi - \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i^2 \xi_i}{\sqrt{var\left(\Pi - \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i^2 \xi_i\right)}}$$

Detection of recent selective sweeps

The frequency spectrum: an example

Sequence \ Site	163	975	1972	2188	3529	4424	4961	5286	7019
1	A	G	G	C	T	T	A	A	A
2	A	T	G	C	T	C	G	A	A
3	G	T	G	T	T	C	A	C	G
4	A	G	G	C	T	C	A	A	G
5	A	G	A	C	C	C	G	A	A
Frequency class:	1	2	1	1	1	4	2	1	3

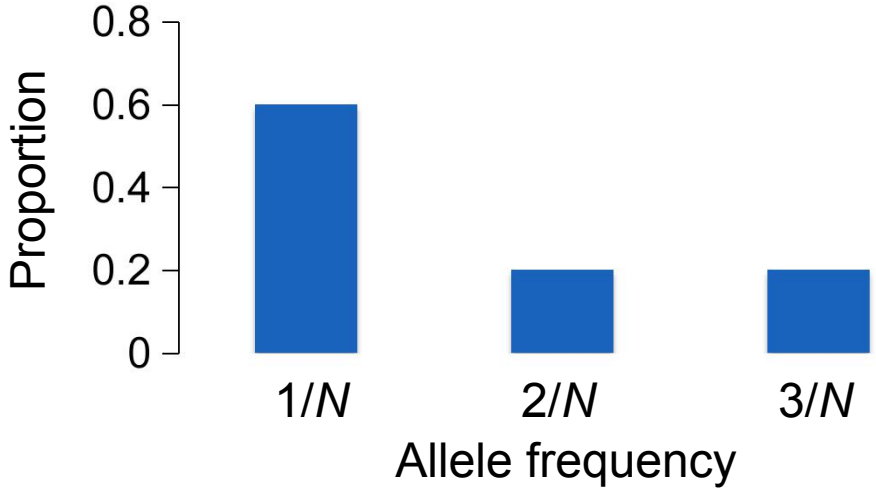
Ancestral Derived



SITE FREQUENCY SPECTRUM : SFS

CC GCA GAG TTA CTA ATC GAA
CG GCA GAG TTA CTA ATC GAA
CC GCA AAG TTA CCA ATT GAA
CC GCA GAG TTA CCA ATC GAA
CC GCA AAG TTA CTA ATC GAG
CC GCA AAG TTA CTA ATC GAA

$N = 6$ individuals
$L = 21$ sites on alignment
$S = 5$ polymorphic sites



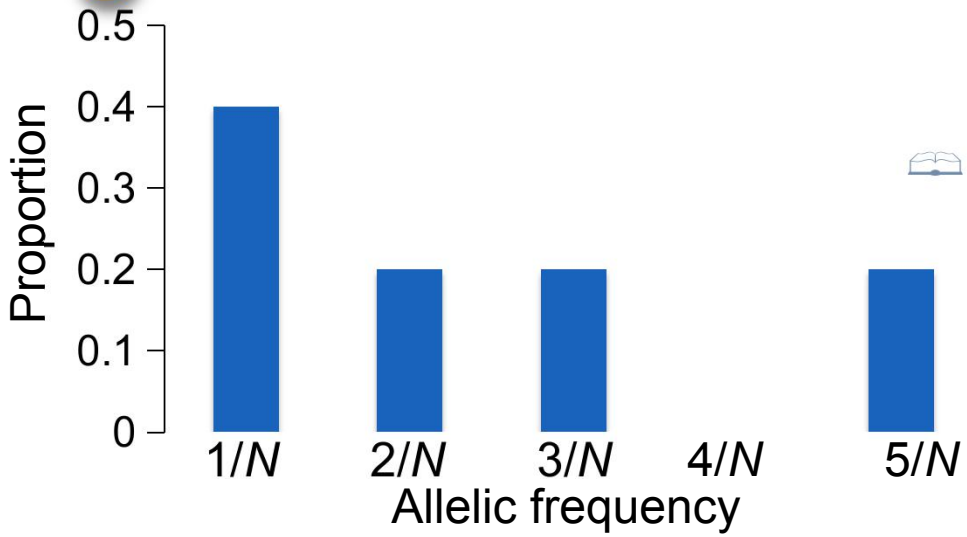
SFS can be « *folded* » when we only take into account frequences from $1/N$ to $N/2$

Polymorphism estimation

SITE FREQUENCY SPECTRUM : SFS

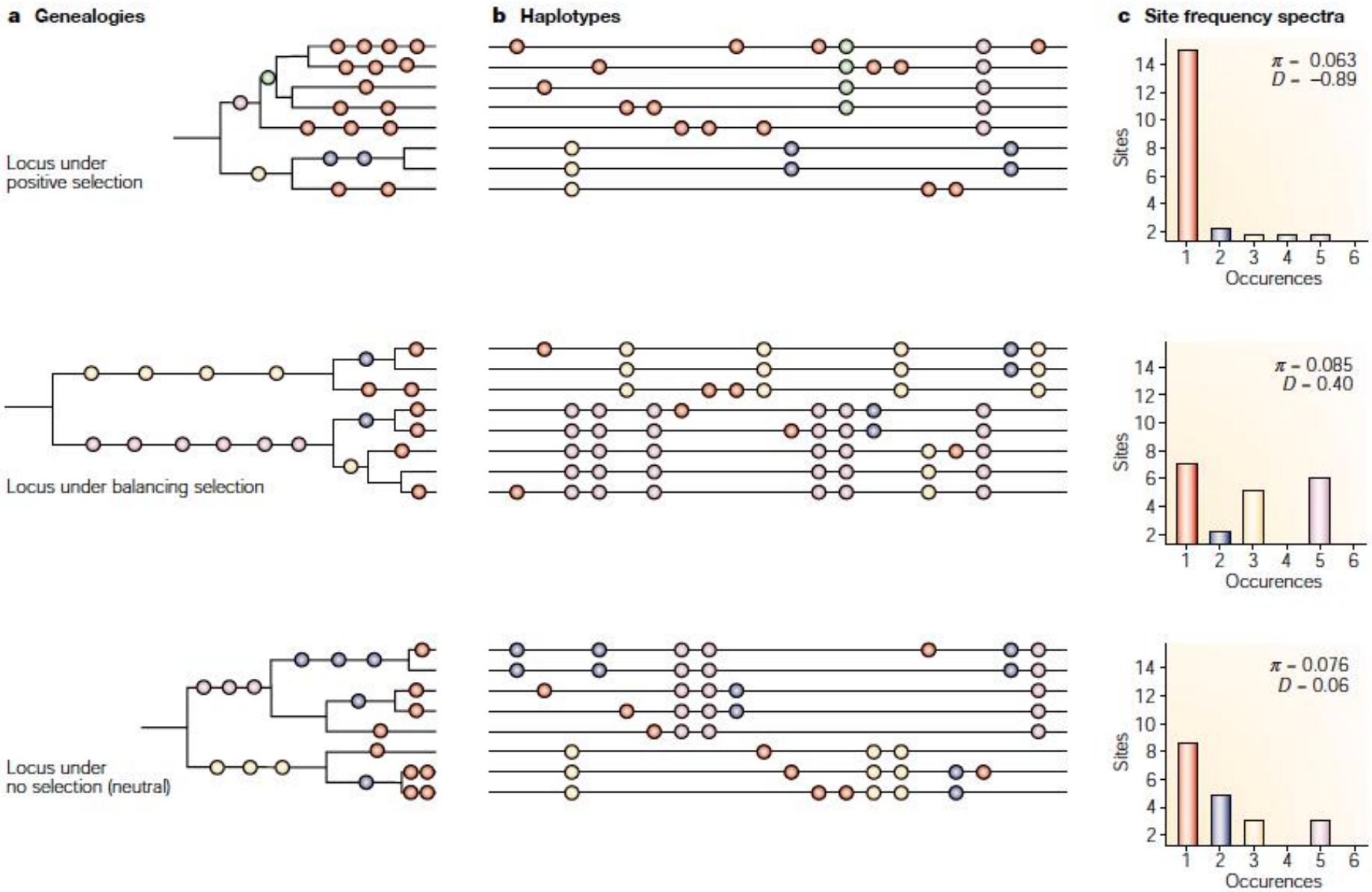
CC GCA GAG TTA CTA ATC GAA
CG GCA GAG TTA CTA ATC GAA
CC GCA AAG TTA CCA ATT GAA
CC GCA GAG TTA CCA ATC GAA
CC GCA AAG TTA CTA ATC GAG
CC GCA AAG TTA CTA ATC GAA

$N = 6$ individuals
$L = 21$ sites on alignment
$S = 5$ polymorphic sites

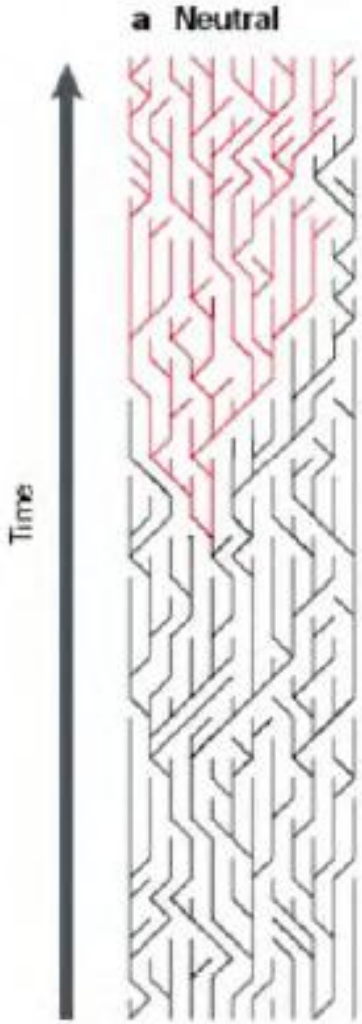


le SFS can be « *folded* » or « *unfolded* » if we know the ancestral state

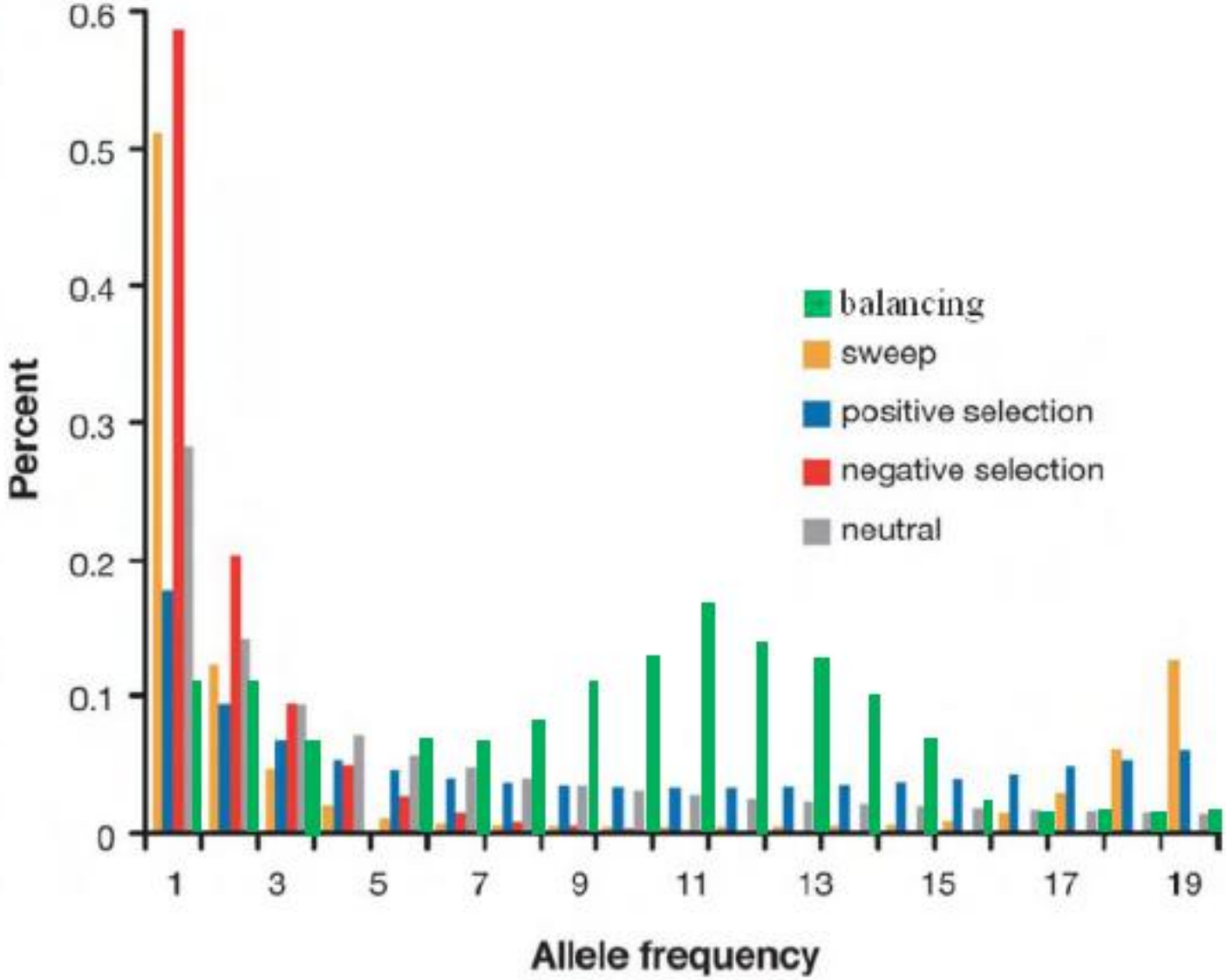
Evolutionary forces like selection shape the SFS



Neutrality

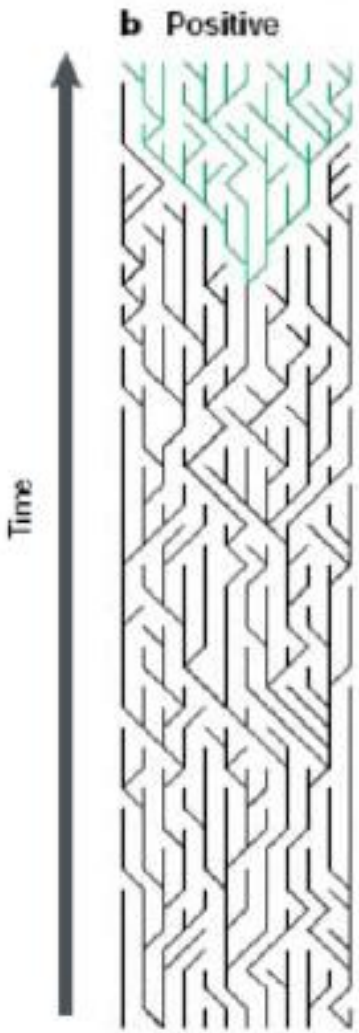


(Bamshad & Wooding)

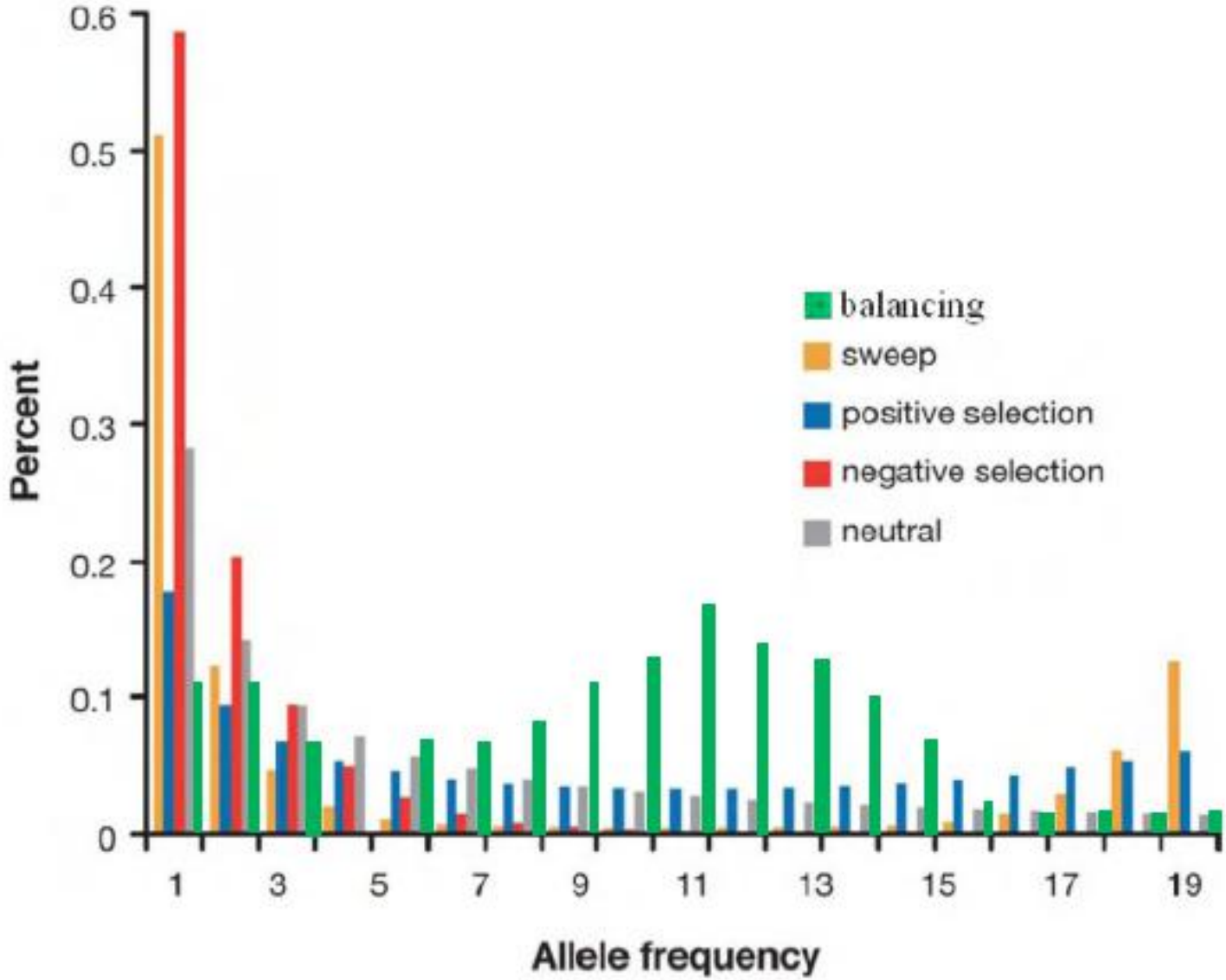


(d'après Nielsen, 2005)

Positive selection



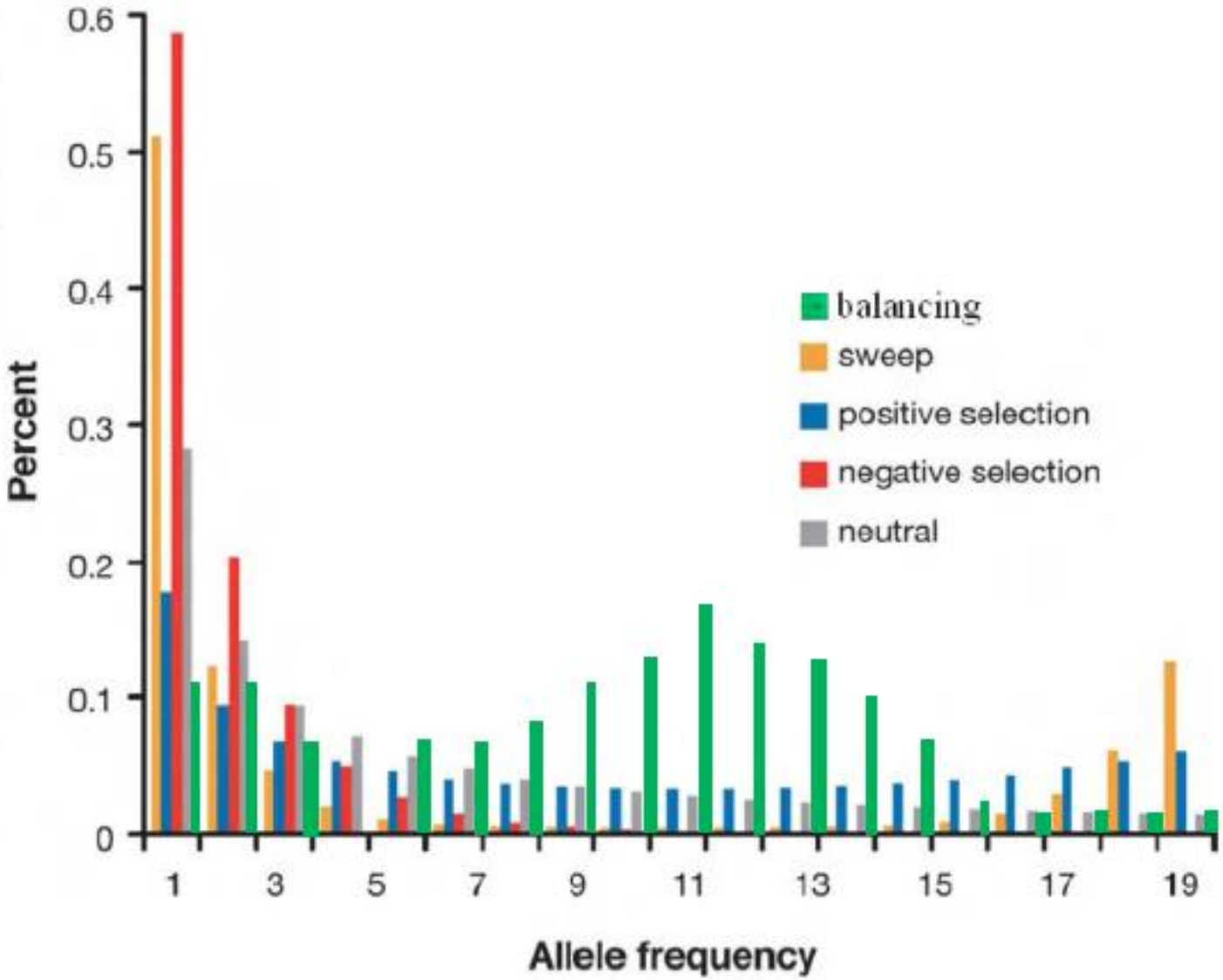
(Bamshad & Wooding)



(d'après Nielsen, 2005)

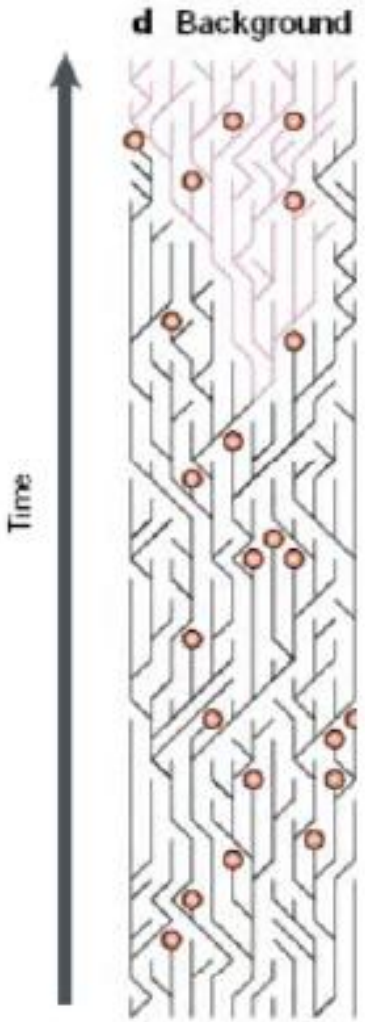
Balancing selection

(Bamshad & Wooding)

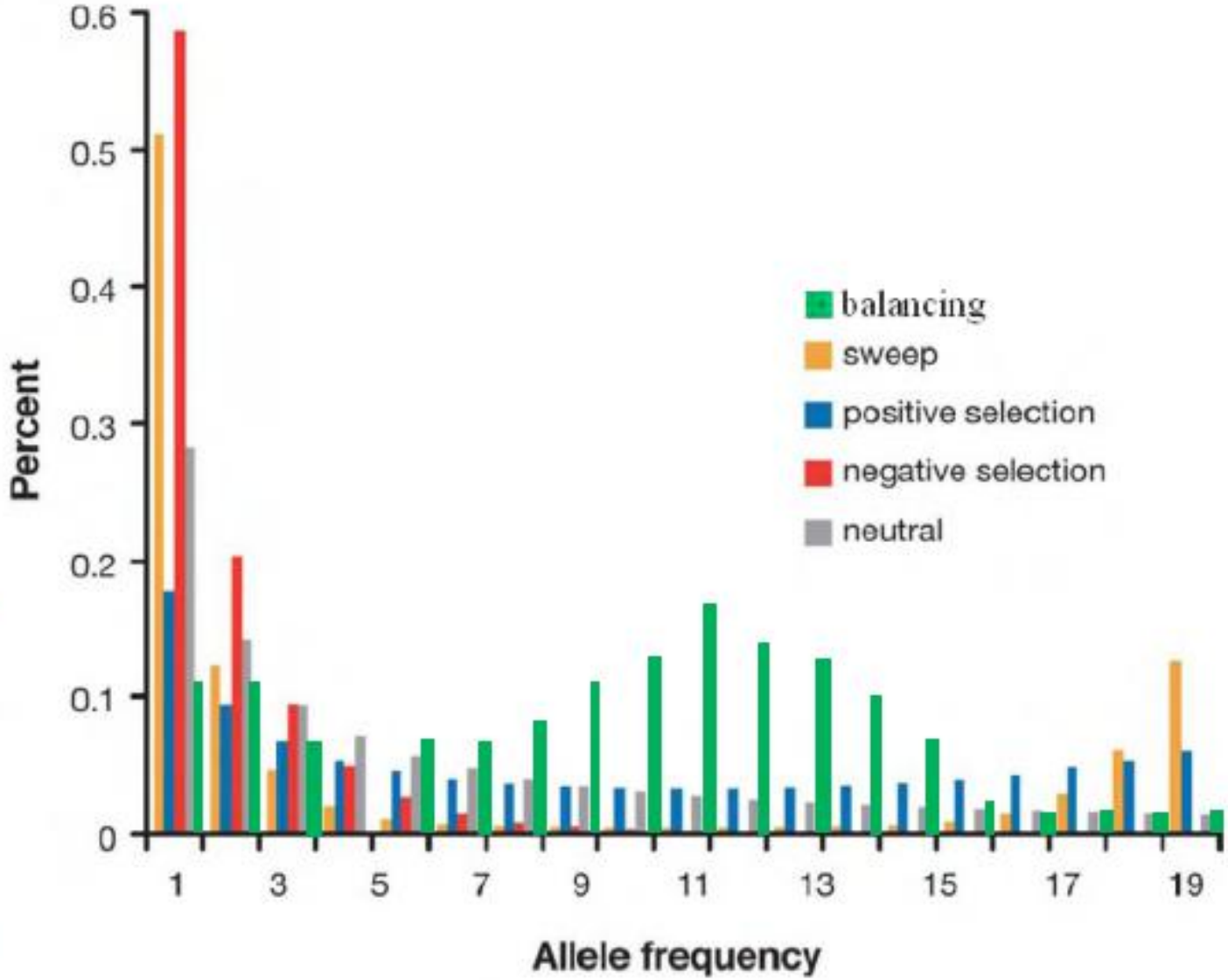


(d'après Nielsen, 2005)

Purifying/background selection



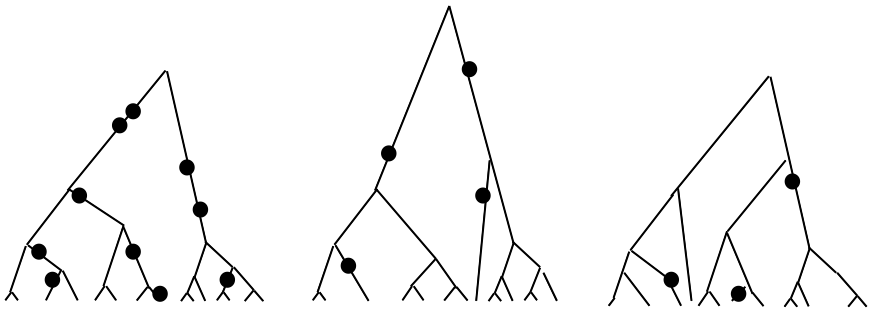
(Bamshad & Wooding)



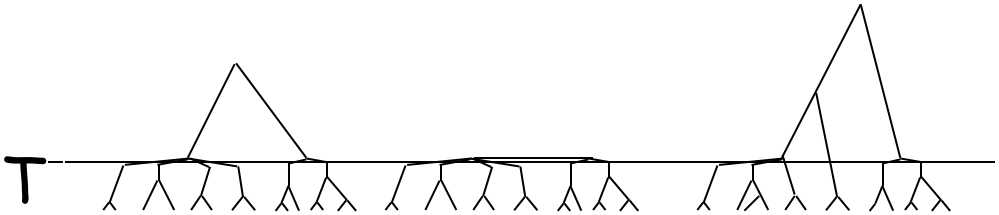
(d'après Nielsen, 2005)

Demography and selection can be confounded

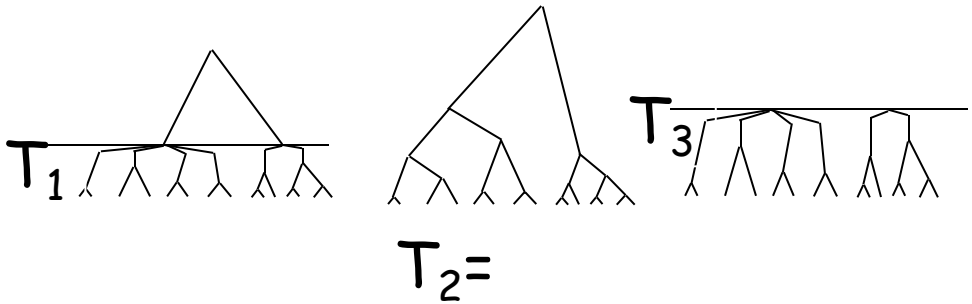
M_1: neutral, constant size
p parameters (q_1, …, q_p)



M_2: bottleneck
p+2 parameters (**T**, **S**, q_1, ..., q_p)



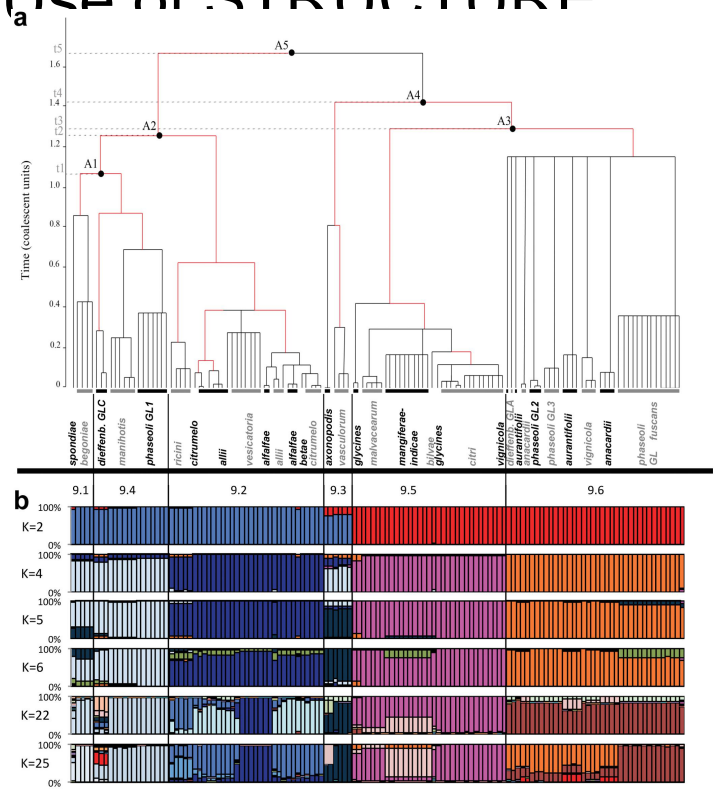
M_3: selective sweep
3*p* parameters
(**T_1**, **S_1**, q_1, ... , **T_p**, **S_p**, q_p)



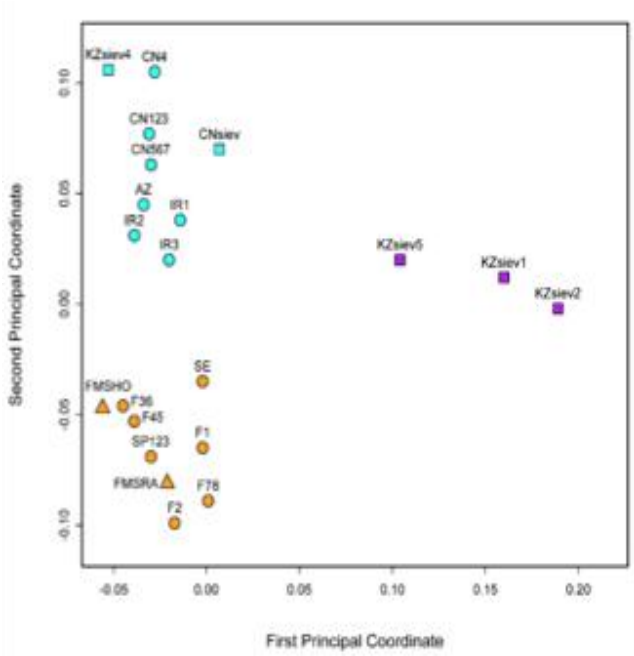
All the tests (Tajima's D, Fu & Li's F and Fay & Wu H) assume ONE population (perturbation of the SFS)

Important to figure out how many populations are in my dataset

Use of STRUCTURE or PCA



Xanthomonas axonopodis

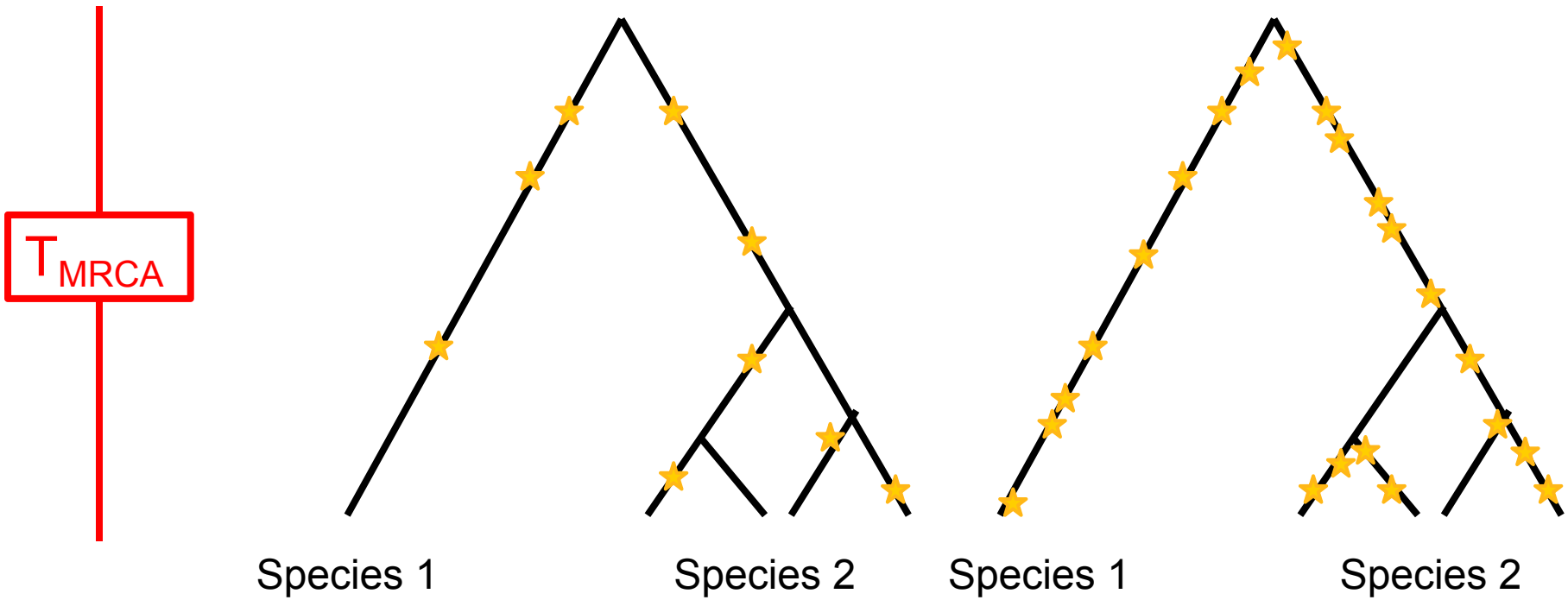


Venturia inaequalis

There are other tests based on polymorphism and divergence

To be continued….

The Hudson Kreitman & Agadé (HKA) test (Hudson, Kreitman & Agadé, 1987)



Under the neutral hypothesis

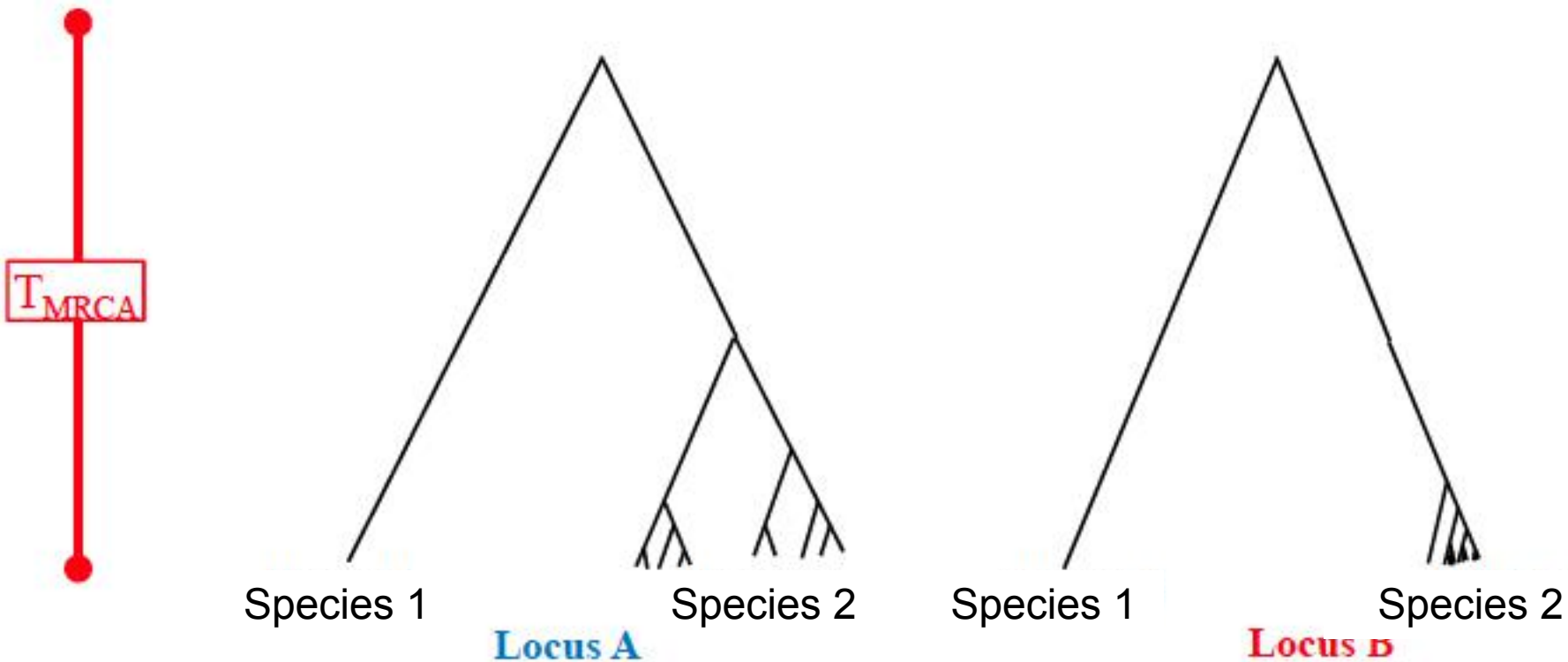
$D_A = 2T\mu_A$

$S_A = 4N\mu_A a_n = \theta_A a_n$

$D_B = 2T\mu_B$

$S_B = 4N\mu_B a_n = \theta_B a_n$

$$\frac{\theta_A}{D_A} = \frac{\theta_B}{D_B} = \frac{4N\mu_i a_n}{2T\mu_i} = \frac{4Na_n}{2T} \quad \text{constant } \forall \text{ locus}$$



$D_A = 2T\mu_A$

$S_A = 4N\mu_A a_n = \theta_A a_n$

$D_B = 2T\mu_B \sim D_A$

$S_B = 4N\mu_B a_n = \theta_B a_n < S_A$

$$\frac{\theta_A}{D_A} > \frac{\theta_B}{D_B}$$

Reduced polymorphism at locus B probably by selection because it can not be due to:

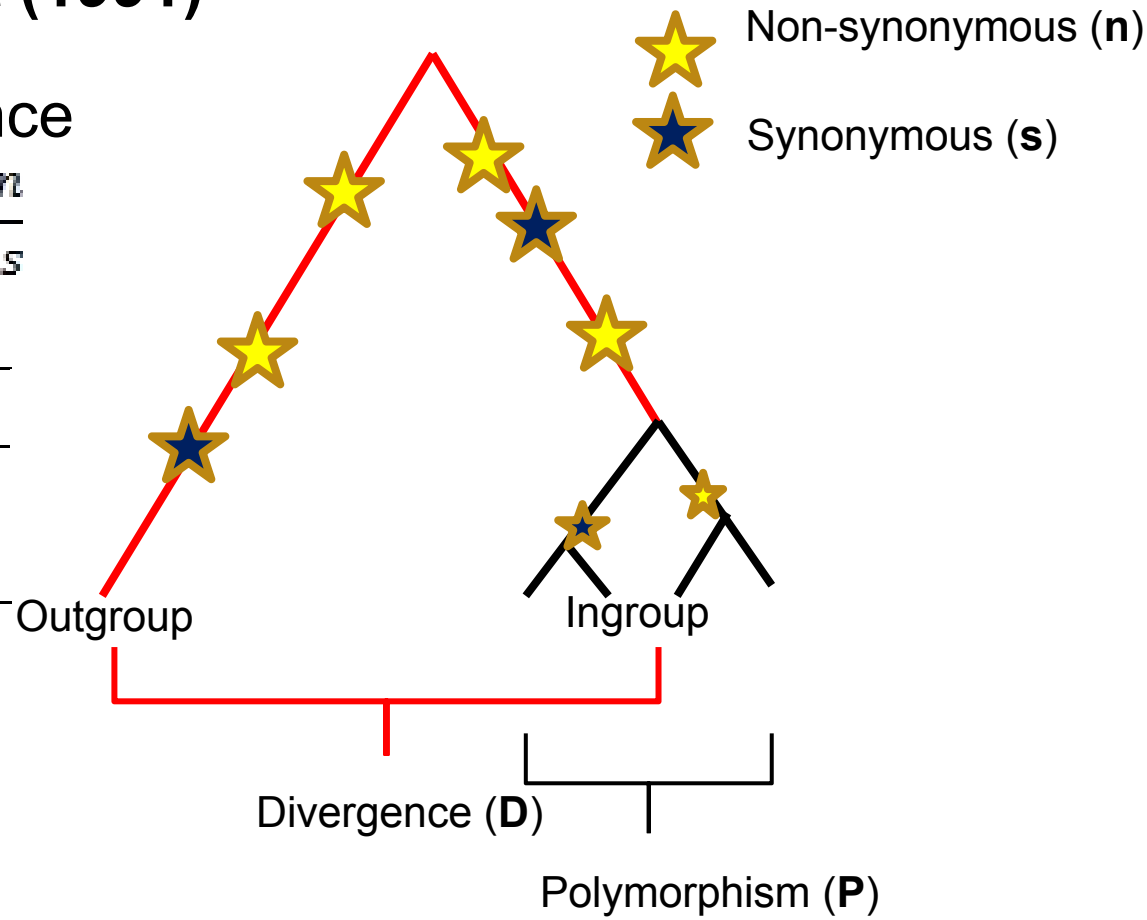
- **reduced population size (too much polymorphism at locus A) or**

•McDonald Kreitman test (1991)

•Test for adaptive divergence

•It estimates α $\alpha = 1 - \frac{DsPn}{DnPs}$

	Divergence	Polymorphism
NS	4	1
S	2	1



Mathematics		Biology
α = 0	Dn/Ds = Pn/Ps	Neutral
α < 0	Dn/Ds < Pn/Ps	**Polymorphism (Pn)** excess
α > 0	Dn/Ds > Pn/Ps	**Divergence (Dn)** excess

Mathematics		Biology
$\alpha = 0$	$Dn/Ds = Pn/Ps$	Neutral
$\alpha < 0$	$Dn/Ds < Pn/Ps$	**Polymorphism (Pn)** excess
$\alpha > 0$	$Dn/Ds > Pn/Ps$	**Divergence (Dn)** excess

- BUT there is estimation biases
  - Distribution of fitness effects

    slightly deleterious mutations which segregate within ingroup (increase Pn)

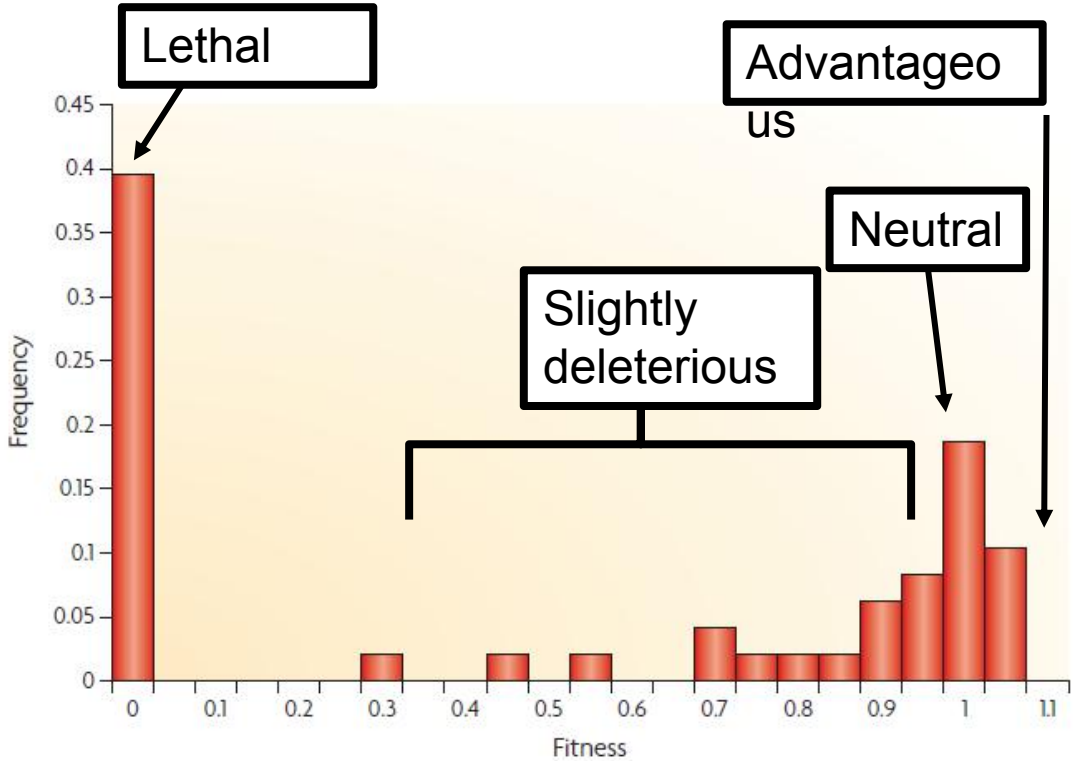
- **3 kind of mutation effects :**

- **Neutral:**
  - **Absence of effects on the fitness**

- **Advantageous:**
  - **Increase the fitness**

- **Deleterious (lethal + Slightly deleterious):**
  - **Complex effects (Fitness variation between 0 and 1)**

- **Underestimation of α**



Experimental distribution of mutation fitness effects from Stomatosis virus (from Eyre Walker et *al.*,2007)

Molecular evolution gives us a lot of information

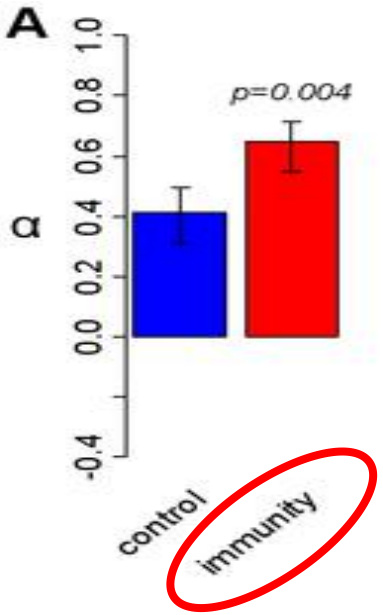
Example :

- Genes involved in interactions with other organism show higher adaptive rates than others

- Ex: Adaptive evolution for immunity genes in *Drosophila* (Obbard et *al*., 2009)

  - Adaptation for efficient immunity system against pathogen

- What about pathogen evolution ?



Adaptive rate comparison (From Obbard et *al*., 2009)