

# Evolution Moléculaire (Masters 1 Bioinformatique / Biotechnologie)

## Evolution moléculaire : neutre ou adaptative ?

**Maxime Bonhomme**

Laboratoire de Recherche en Sciences Végétales

16 février 2017

# Evolution Moléculaire : neutre ou adaptative ?

## 1 Théorie neutraliste

## 2 Sélection au niveau moléculaire

- sélection positive
- sélection purifiante
- sélection balancée

## 3 Tests de neutralité

- tests sur les fréquences alléliques
- polymorphisme et divergence
- méthodes phylogénétiques
- conclusion

## 4 References

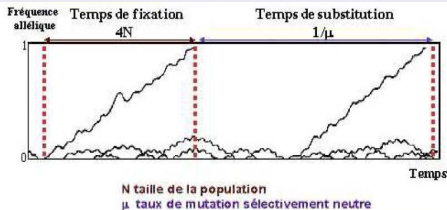
# Théorie neutraliste de l'évolution moléculaire

## Principaux résultats

- développée par Motoo Kimura (1968, 1969, 1983)
- selon la théorie, la majorité des polymorphismes moléculaires résulte de **l'évolution par dérive génétique** d'allèles mutants sélectivement neutres (ex : ADN non codant majoritaire, 3ème position des codons - mutation synonyme -) :
- à un gène donné, dans une population, les mutations se produisent à un taux par génération de  $2N\mu$  ( $2N$  = nombre de "copies" du gène dans une population diploïde,  $\mu$  = taux de mutation du gène)
- sous dérive chaque mutation a une probabilité de fixation = sa fréquence =  $\frac{1}{2N}$
- sous dérive chaque mutation a une probabilité d'élimination =  $1 - \frac{1}{2N}$
- le taux de substitution d'allèles neutres =  $2N\mu * \frac{1}{2N} = \mu$
- temps moyen de fixation d'une mutation dépend de la taille de la population :  **$4N$  générations**
- dans les populations de petites taille, temps de fixation plus court
- à l'équilibre mutation-dérive, la quantité de polymorphisme dans la population est déterminé par le produit  $N\mu$ , généralement mesuré par  $\theta = 4N\mu$

## Théorie neutraliste

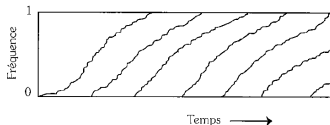
## Dynamique de remplacement des allèles neutres



De nouveaux allèles neutres apparaissent sans cesse par mutation. La plupart sont perdus, mais certains finissent par se fixer et remplacent les anciens. Cela arrive en moyenne toutes les  $1/\mu$  générations.  $\mu$  étant le taux de mutation neutre au locus. Le temps moyen entre l'apparition d'un nouvel allèle neutre destiné à remplacer l'ancien et le moment où il se fixe est de  $4N$  générations,  $N$  étant la taille de la population.



comparaison d'une population de petite taille (a) et de grande taille (b). Le taux de mutation neutre est le même, donc le taux de substitution aussi mais comme  $4N$  est différent, on trouve à tout moment plus de polymorphisme neutre dans la population de grande taille que dans celle de petite taille.



# Théorie neutraliste

## En pratique

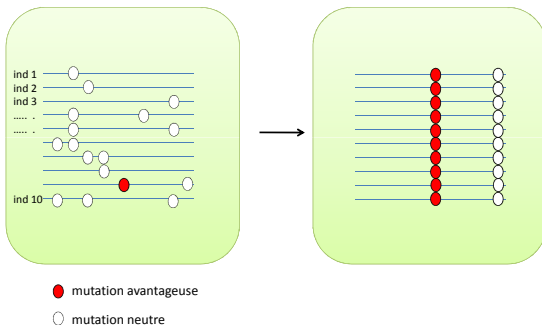
- **mutation neutre** : le plus souvent éliminée de la population, mais peut aussi se substituer à l'allèle sauvage, à cause des effets aléatoires de la **dérive génétique** dans les petites populations
- **mutation légèrement défavorable** : se comporte de manière similaire à une mutation neutre
- **mutation défavorable** : diminue en fréquence (**sélection négative**)
- **mutation favorable** : augmente en fréquence (**sélection positive**)
- les mutations favorables ou défavorables sont de toute façon sous l'emprise de la dérive génétique

## Quels sont les effets de la sélection naturelle ?

- affecte la **distribution des fréquences alléliques**, parfois rapidement
- affecte le **nombre d'allèles** maintenus (augmentation, diminution)
- affecte l'**hétérozygotie**
- affecte le **temps de résidence** des allèles dans les populations (divergence des populations)
- affecte la proportion de **changements synonymes et non-synonymes** le long des séquences

## Sélection positive

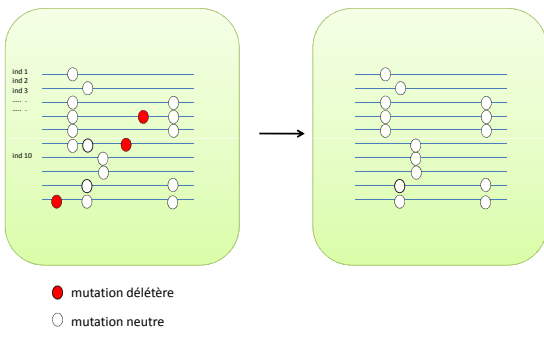
## Sélection positive (darwinienne, directionnelle)



- gènes ayant un rôle dans l'adaptation (ex : résistance aux insecticides chez le moustique, adaptation à la sécheresse)

## Sélection purifiante

## Sélection purifiante (stabilisante, background selection)

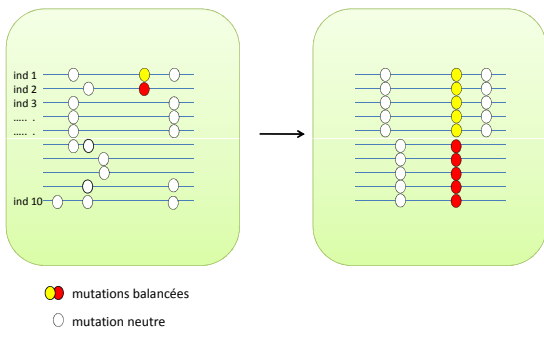


- gènes "domestiques" ("housekeeping genes") : les changements sont contre-sélectionnés



## Sélection balancée

## Sélection balancée (diversifiante)



- avantage à l'hétérozygote (overdominance)
- sélection fréquence dépendante (sélection de l'allèle rare, dynamique de fréquences cyclique)
- ex : anémie falciforme chez l'homme, gènes de l'immunité (maintien d'un fort polymorphisme)

## Impact des différentes formes de sélection sur la diversité génétique

Evolutionary factor	Intraspecific variability <sup>a</sup>	Interspecific variability	Ratio of interspecific to intraspecific variability	
			Frequency spectrum	
Increased mutation rate	Increases	Increases	No effect	No effect
Negative directional selection	Reduced	Reduced	Reduced if selection is not too strong	Increases the proportion of low frequency variants
Positive directional selection	May increase or decrease	Increased	Increased	Increases the proportion of high frequency variants
Balancing selection	Increases	May increase or decrease	Reduced	Increases the proportion of intermediate frequency variants
Selective sweep (linked neutral sites)	Decreased	No effect on mean rate of substitution, but the variance increases	Increased	Mostly increases the proportion of low frequency variants

<sup>a</sup>Note that selection also affects other features of the data not mentioned here, such as levels of LD, haplotype structure, and levels of population subdivision.

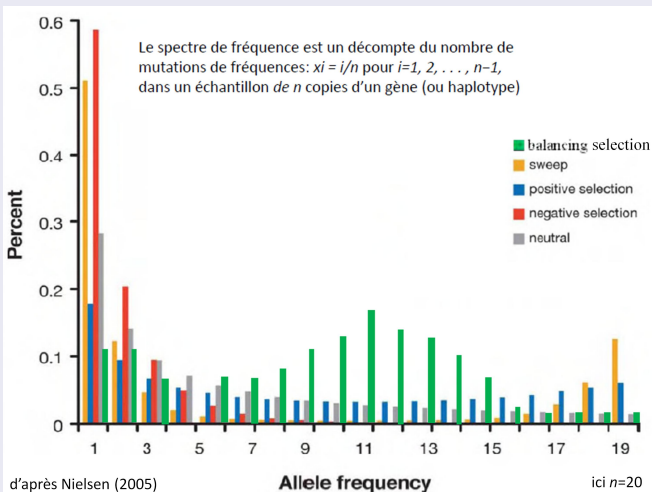
## Comment détecter la sélection naturelle ?

- **approche directe** : suivre expérimentalement une population au cours du temps
  - avantage : on connaît et on manipule l'agent de la sélection
  - nécessite des données qui sont assez rarement disponibles. Contraintes sur l'échelle de temps et taille d'échantillons.
  - quels organismes ?
- **approche alternative "indirecte"** : différentes signatures moléculaires de la sélection peuvent être utilisées pour tester le modèle neutre :
  - spectre de fréquences alléliques et diversité nucléotidique
  - polymorphisme / divergence
  - méthodes phylogénétiques ( $\frac{dN}{dS}$ )

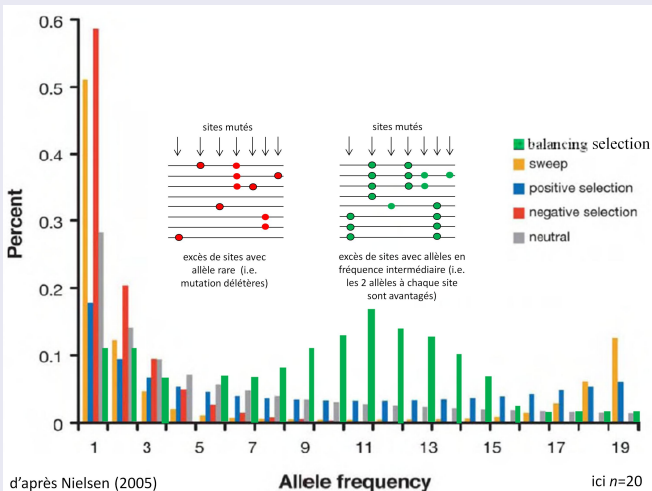
## Comment détecter la sélection naturelle ?

- importance de la théorie neutraliste :
  - un modèle "nul" (hypothèse "H0" qui décrit un monde sans sélection naturelle)
  - la diversité génétique n'est affectée que par la dérive, la mutation, la recombinaison et la migration
- **prédire avec le modèle nul ce qu'on devrait attendre** (hétérozygotie, nombre d'allèles, distribution des fréquences alléliques) et **tester l'ajustement de nos données au modèle**

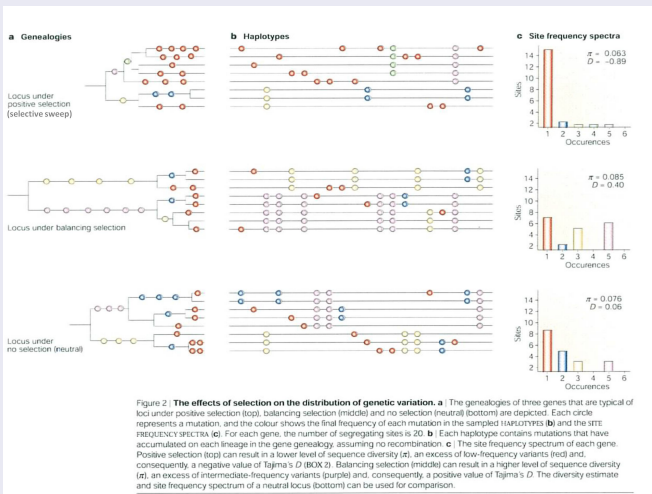
## Spectre de fréquences alléliques et sites nucléotidiques



## Spectre de fréquences alléliques et sites nucléotidiques



## Spectre de fréquences alléliques et sites nucléotidiques, et généalogie



# Test de Tajima sur la diversité nucléotidique

- compare deux estimateurs du paramètre  $\theta = 4N\mu$  :
  - sur la base du nombre de sites qui ségrégent ( $S$ )
  - sur la base de la diversité nucléotidique (hétérozygotie) ( $\pi$ )
- chaque mutation crée un nouveau site ségrégeant ( $S$ ) mais contribue très peu à la diversité nucléotidique ( $\pi$ )
- ces 2 estimateurs diffèrent donc par l'importance relative accordée aux variants rares et intermédiaires

$$D = \frac{\hat{\theta}_\pi - \hat{\theta}_S}{SE(\hat{\theta}_\pi - \hat{\theta}_S)} \quad (1)$$

- sous  $H_0 = \text{neutralité}$  :  $\mathbb{E}(D) = 0$  et  $\text{Var}(D) = 1$  (utilisation des lois normales et beta, ou simulations, pour effectuer le test)
- $D < 0 = \text{sélection purifiante}$ , présence de mutations légèrement délétères dans la population : excès d'allèles rares, forte contribution de  $S$  (possible aussi en cas d'expansion de la population, et de balayage sélectif)
- $D > 0 = \text{sélection balancée}$ , présence de mutations en fréquences intermédiaires : moins d'allèles rares mais beaucoup d'hétérozygotie, forte contribution de  $\pi$  (possible aussi en cas de goulot d'étranglement de la population)



# Test de Tajima sur la diversité nucléotidique

exemple du gène *CCR5* chez l'homme

## A strong signature of balancing selection in the 5' cis-regulatory region of *CCR5*

Michael J. Bamshad<sup>1\*</sup>, Srinivas Mummidi<sup>1§</sup>, Enrique Gonzalez<sup>1§</sup>, Seema S. Ahuja<sup>1§</sup>, Diane M. Dunn<sup>1</sup>, W. Scott Watkins<sup>1</sup>, Stephen Wooding<sup>1</sup>, Anne C. Stone<sup>1</sup>, Lynn B. Jorde<sup>1</sup>, Robert B. Weiss<sup>1</sup>, and Sunil K. Ahuja<sup>1§</sup>

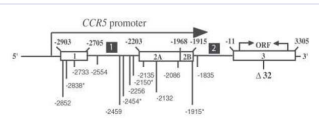


Table 1. Summary of sequence variation in the cis-regulatory region of *CCR5*

Population	n <sup>a</sup>	S	$\theta_W$	$\pi \pm SD, \%$	Tajima's D <sup>b</sup>	F <sub>s</sub> <sup>c</sup>
NIH panel	176	9	1.57	0.29 ± 0.17	—	— <sup>d</sup>
Old World panel	224	13	2.18	0.21 ± 0.13	0.667 (0.37)	0.02 (0.38)
Africans	62	12	2.56	0.22 ± 0.13	0.292 (0.38)	-0.81 (0.57)
Non-Africans	162	8	1.42	0.21 ± 0.12	2.08 (0.03)	2.57 (0.10)
Asians	54	6	1.32	0.20 ± 0.12	2.52 (0.01)	3.45 (0.06)
Europeans	48	7	1.58	0.22 ± 0.13	2.20 (0.02)	1.61 (0.17)
Indians	60	7	1.54	0.20 ± 0.12	1.85 (0.04)	2.34 (0.12)

<sup>a</sup>Number of chromosomes.

<sup>b</sup>P value is given in parentheses.

<sup>c</sup>Ethnic identity unlinked to samples, therefore haplotypes could not be estimated reliably.

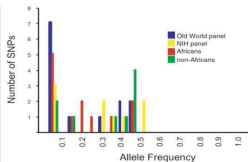
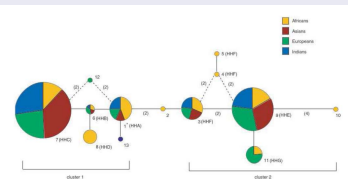
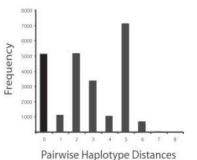


Fig. 2. Allele frequency spectrum for 13 SNPs found in the Old World and NIH panels, and for Africans and non-Africans. The frequency of the derived allele of each SNP is shown.

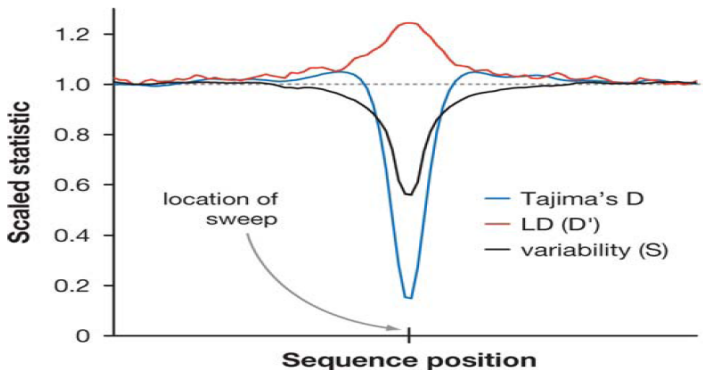


• des allèles trop divergents pour un modèle neutre: sélection balancée?

- avantage à long terme aux hétérozygotes à *CCR5*?
- un locus impliqué dans la résistance à d'autres maladies?

# Test de Tajima sur la diversité nucléotidique

détection d'un balayage selectif



**Figure 1**

The effect of a selective sweep on genetic variation. The figure is based on averaging over 100 simulations of a strong selective sweep. It illustrates how the number of variable sites (variability) is reduced, LD is increased, and the frequency spectrum, as measured by Tajima's D, is skewed, in the region around the selective sweep. All statistics are calculated in a sliding window along the sequence right after the advantageous allele has reached frequency 1 in the population. All statistics are also scaled so that the expected value under neutrality equals one.

## Polymorphisme et divergence des mutations synonymes et non synonymes

## test de McDonald-Kreitman

- polymorphisme neutre et divergence neutre sont deux facettes d'un même processus
- hypothèse du test : si mutations *syn* et *nonsyn* sont **neutres**, la proportion de polymorphismes *syn* et *nonsyn* dans une espèce devrait être égale à la proportion de différences *syn* et *nonsyn* entre espèces
- exemple d'un site nucléotidique chez 5 individus par espèce :
  - AAAAA chez espèce 1, GGGGG chez espèce 2 = une différence fixée
  - AGAGA chez espèce 1, AAAAA chez espèce 2 = un site polymorphe
- classification des sites nucléotidiques dans un tableau de contingence

Type de changement	Divergence	Polymorphisme
remplacement (non syn)	$N_1$	$N_2$
silencieux (syn)	$N_3$	$N_4$

- test de  $\chi^2$  d'indépendance ou test exact de Fisher
- McDonald and Kreitman (1991) : mise en évidence de la sélection sur le gène de l'*Adh* chez 3 espèces de *Drosophila*

# Comparaison des taux de substitution

## dégénérescence du code génétique

le code génétique										
		Deuxième lettre						ijk		
		U		C		A		G		
Première lettre (côté 5')	U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
		UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys	C
		UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop	A
		UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp	G
	C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
		CUC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
		CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
		CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
	A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
		AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
		AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	A
		AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg	G
	G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
		GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
		GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
		GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G
		codon d'initiation			codon de terminaison					

Troisième lettre (côté 3')

# Comparaison des taux de substitution

séquence 1: UCC CAU ACA  
 séquence 2: UCC CAC AAA

séquence 1: UCC CAU ACA  
 séquence 2: UCC CAC AAA  
 acide aminé: Ser His Thr  
 Lys

sites synonymes: 3

substitutions synonymes: 1

sites non synonymes: 6

substitutions non synonymes: 1

$$d_N = \frac{\text{nombre de substitutions non synonymes}}{\text{nombre de sites non synonymes}}$$

$$d_S = \frac{\text{nombre de substitutions synonymes}}{\text{nombre de sites synonymes}}$$

$$d_N / d_S = \omega \quad [\text{parfois appelé Ka/Ks}]$$

$0 < \omega < 1$ : sélection purifiante (séquence conservée)

$\omega = 1$  : neutralité

$\omega > 1$  : sélection positive (diversifiante)



$$d_N = 1/6$$

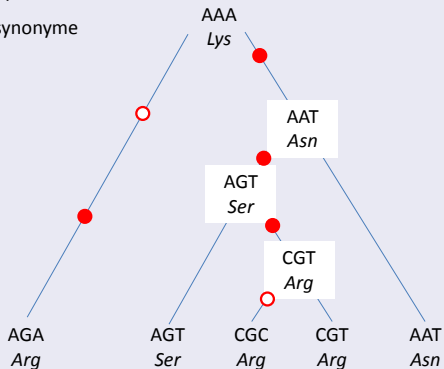
$$d_S = 1/3$$

$$d_N / d_S = \omega = 0.5$$

# Comparaison des taux de substitution

○ substitution synonyme

● substitution non-synonyme



on estime  $d_N$  et  $d_S$  sur chaque branche:

- il faut inférer la séquence ancestrale (ici l'acide aminé Lys)
- utilisation d'un espèce divergente (orthologue distant)

# Comparaison des taux de substitution

## un grand nombre de méthodes

- **méthodes heuristiques (approximations) :**

- inférence de la séquence ancestrale par parcimonie, estimation de  $d_N$  et  $d_S$  sur chaque branche, et approximation normale de  $d_N - d_S$  (Messier and Stewart, 1997)
- test de Fisher des  $d_N$  et  $d_S$  de toutes les branches (Zhang et al. 1997)
- pour les deux méthodes problème des erreurs sur la reconstruction de la séquence ancestrale
- pour éviter cela Zhang et al. (1998) calculent  $d_N$  et  $d_S$  pour chaque comparaison de séquences 2 à 2, et estiment les longueurs de branches indépendamment pour les taux de substitution *syn* et *non - syn*. Ensuite, ils comparent les longueurs de branches *syn* et *non - syn* ( $b_N$  et  $b_S$ ).

- **méthodes par vraisemblance (Likelihood methods) :**

- analyses plus rigoureuses car tiennent compte de l'incertitude sur la séquence ancestrale, en utilisant les **modèles de substitution des codons** et en **analysant toutes les séquences conjointement dans un arbre phylogénétique**
- avantage : on peut modéliser la **variation de  $\omega$  le long d'une branche** de l'arbre et le **long de la séquence**, jusqu'à estimer un  $\omega$  par site nucléotidique!
- donc on peut estimer la sélection dans le temps (branches internes) mais aussi sur certaines partie d'un gène (sites conservés, sites adaptatifs)
- exemple : sur 7645 gènes chez homme-chimpanzé-souris, 1547 ont un  $\omega > 1$  le long de la branche qui mène aux humains, et 1534 le long de la branche qui mène aux chimpanzé : pas les mêmes gènes

# Conclusion

## points principaux

- la sélection affecte les fréquences **alléliques** au cours du temps pour certains gènes; elle est responsable de l'**évolution adaptative** à l'échelle des gènes, et ceci se traduit à l'échelle macroscopique (cellulaire, physiologie, morphologie, etc...)
- la sélection peut prendre plusieurs formes selon la **distribution des valeurs sélectives** des allèles à un locus
- dans les populations naturelles, on ne connaît pas l'agent de la sélection : il est difficile de détecter la sélection **directement**
- les tests de mise en évidence de la sélection utilisent la **théorie neutraliste de l'évolution moléculaire** comme hypothèse "statistiquement" neutre (test de neutralité)



# Conclusion

la recherche de signatures de sélection permet de...

- identifier des gènes fonctionnellement importants
- mieux comprendre les contraintes fonctionnelles des différentes parties du génome, des gènes
- nous mettre sur la voie de gènes impliqués dans des maladies (défense, virulence, résistance, sensibilité)
- reconstituer les mécanismes évolutifs qui ont permis l'émergence de certaines adaptations

## Liste des tests usuels

Table 1 | **Commonly used tests of neutrality**

Test	Compares
<b><i>Tests based on allelic distribution and/or level of variability</i></b>	
Tajima's $D$	The number of nucleotide polymorphisms with the mean pairwise difference between sequences
Fu and Li's $D$ , $D^*$	The number of derived nucleotide variants observed only once in a sample with the total number of derived nucleotide variants
Fu and Li's $F$ , $F^*$	The number of derived nucleotide variants observed only once in a sample with the mean pairwise difference between sequences
Fay and Wu's $H$	The number of derived nucleotide variants at low and high frequencies with the number of variants at intermediate frequencies
<b><i>Tests based on comparisons of divergence and/or variability between different classes of mutation</i></b>	
$d_n/d_s$ , $K_a/K_s$	The ratios of non-synonymous and synonymous nucleotide substitutions in protein coding regions
HKA	The degree of polymorphism within and between species at two or more loci
MK	The ratios of synonymous and non-synonymous nucleotide substitutions in and between species

HKA, Hudson-Kreitman-Aguade; MK, McDonald-Kreitman.

## References



*pour aller plus loin...*

Computational Molecular Evolution, Ziheng Yang, Oxford Series in Ecology and Evolution



*Nielsen R. Molecular signatures of natural selection. Annual Review of Genetics. 2005. 39 :197-218*



*Bamshad M and Wooding SP. Signatures of natural selection in the human genome. Nature Review Genetics. 2003. 4 :99-211*



*Evolution Biologique. Ridley, De Boeck Universié*



*Principles of Population Genetics, 4th Edition. Hartl DL, Clark AG*