Evolution moléculaire : neutre ou adaptative ?

M1 : MABS

UE : Evolution Moléculaire

Maxime Bonhomme

UMR CNRS-UPS 5546, Laboratoire de Recherche en Sciences Végétales, Toulouse

11 septembre 2011

Evolution Moléculaire : neutre ou adaptative ?

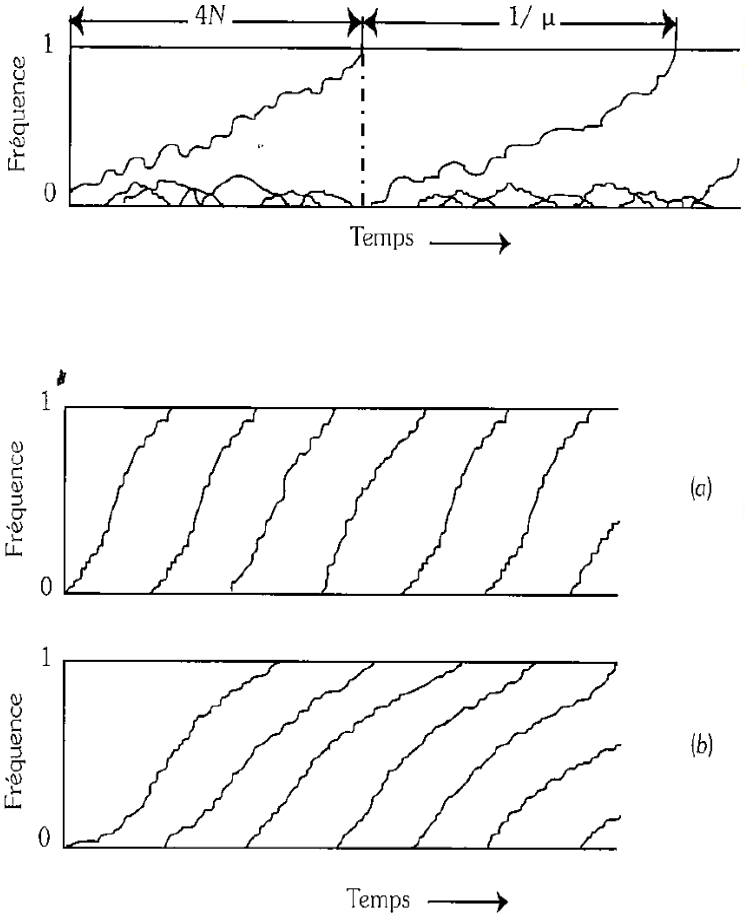
Théorie neutraliste de l'évolution moléculaire

Principaux résultats

- développée par Motoo Kimura
- 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 -) :
- les mutations se produisent à un taux par génération de $2N\mu$ (N = taille efficace, μ = taux de mutation)
- 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 dépends 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



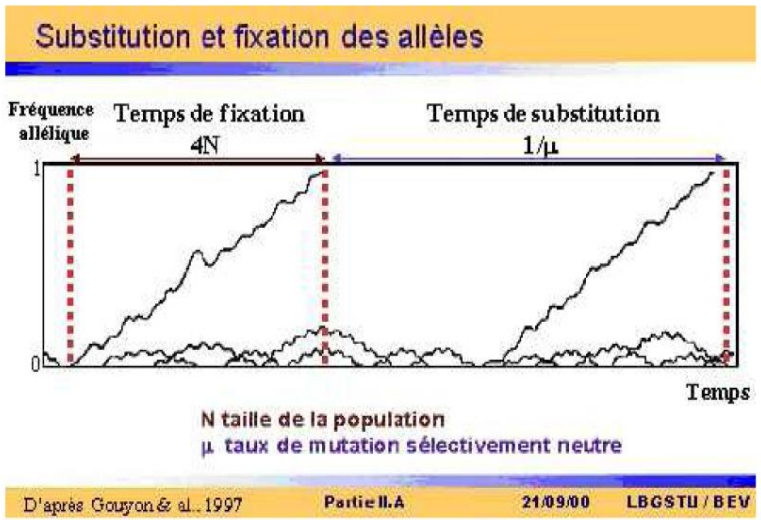


Figure 6 – Processus de la substitution des allèles neutres (d'après Kimura).
En haut : schématisation du processus. De nouveaux allèles neutres apparaissent sans cesse par mutation. La plupart sont perdus, mais certains finissent par se fixer et remplacent les anciens. Cela arrive en moyenne toutes les 1/μ générations, μ 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.
En bas : 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

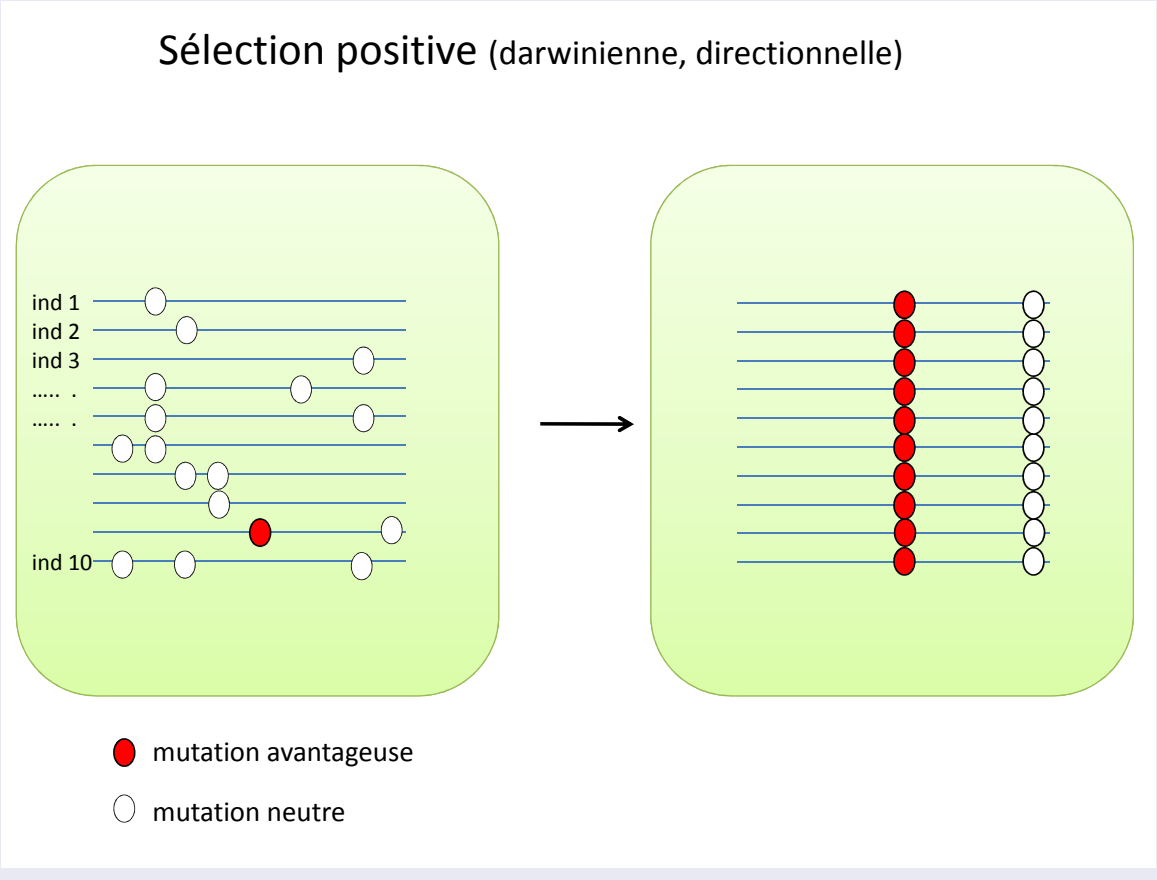
En pratique

- **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 ?

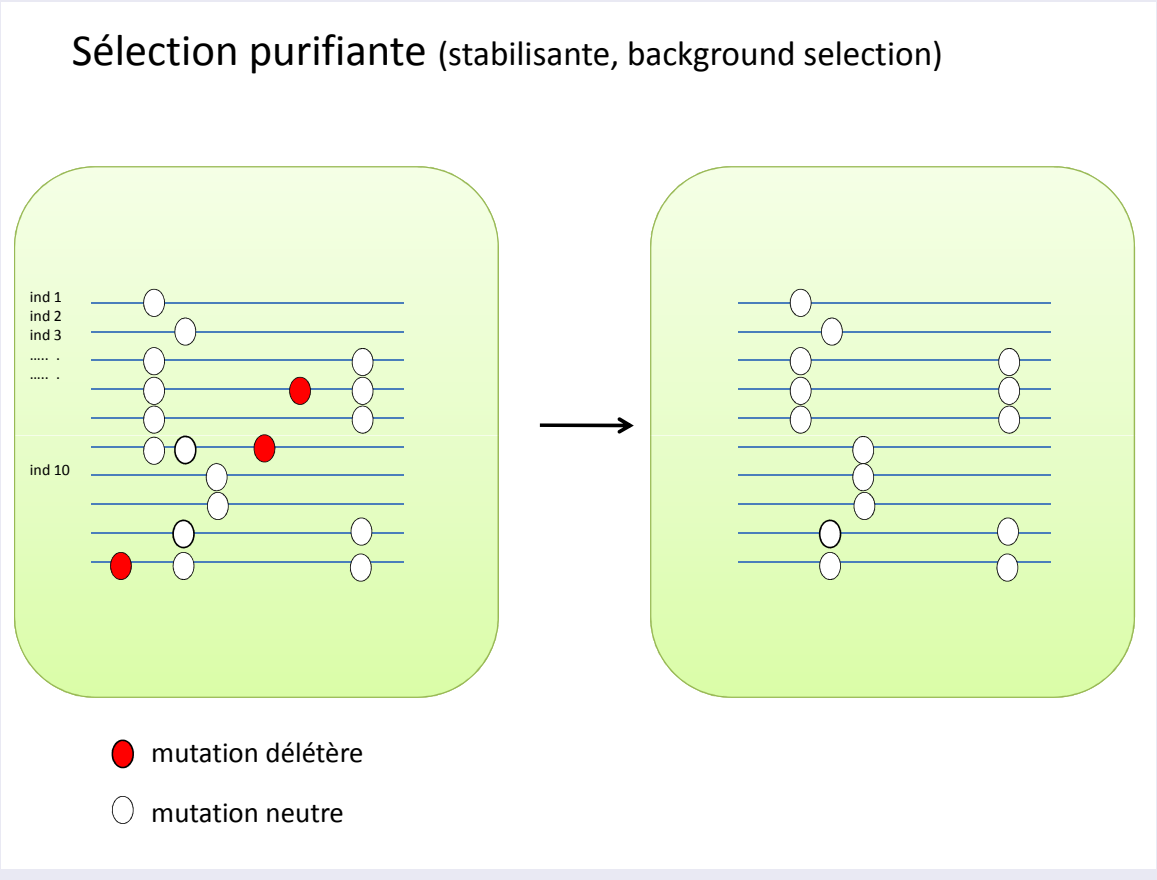
- affecte la **distribution des fréquences alléliques**, parfois rapidement
- affecte le **nombre d'allèles** maintenus (augmentation, diminution)
- affecte l'**hétérozygotie**
- affecte le **temps de résidence** des allèles dans les populations (divergence des populations)
- affecte la proportion de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

Sélection positive



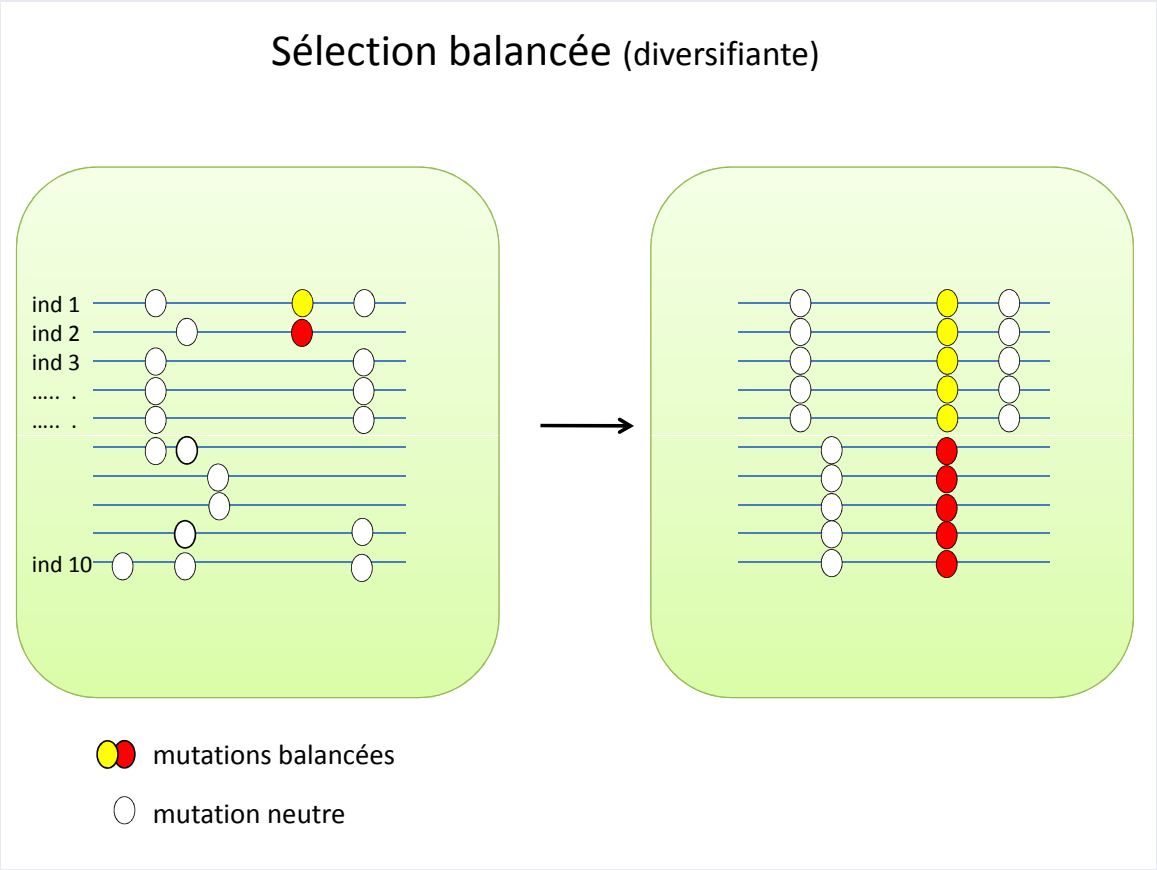
- gènes ayant un rôle dans l'adaptation (ex : résistance aux insecticides chez le moustique, adaptation à la sécheresse)

Sélection purifiante



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

Sélection balancée



- 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[a]	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

[a]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
  - nécessite des données exceptionnellement rares. Contraintes sur l'échelle de temps et taille d'échantillons
  - comment distinguer sélection de changements de l'environnement et des autres forces évolutives ?
  - effet de sélection faibles (non détectables à l'échelle de quelques générations) peuvent être importants à long terme ?
  - quels organismes ?
- **approche alternative "'indirecte"** : différentes signatures moléculaires de la sélection peuvent êtres utilisées pour tester le modèle neutre :
  - spectre de fréquences alléliques et diversité nucléotidique
  - polymorphisme / divergence
  - méthodes phylogénétiques ($\frac{d_N}{d_S}$)

Comment détecter la sélection naturelle ?

- importance de la théorie neutraliste :
  - un modèle "nul" 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**
- comparer la vraisemblance d'un modèle sans sélection naturelle avec la vraisemblance d'un modèle qui l'intègre (likelihood-ratio test) : l'amélioration permet-elle de significativement mieux décrire nos données ?

Spectre de fréquences alléliques

Le spectre de fréquence est un décompte du nombre de mutations de fréquences: *xi = i/n* pour *i=1, 2, . . . , n−1,* dans un échantillon *de n* copies d'un gène (ou haplotype)

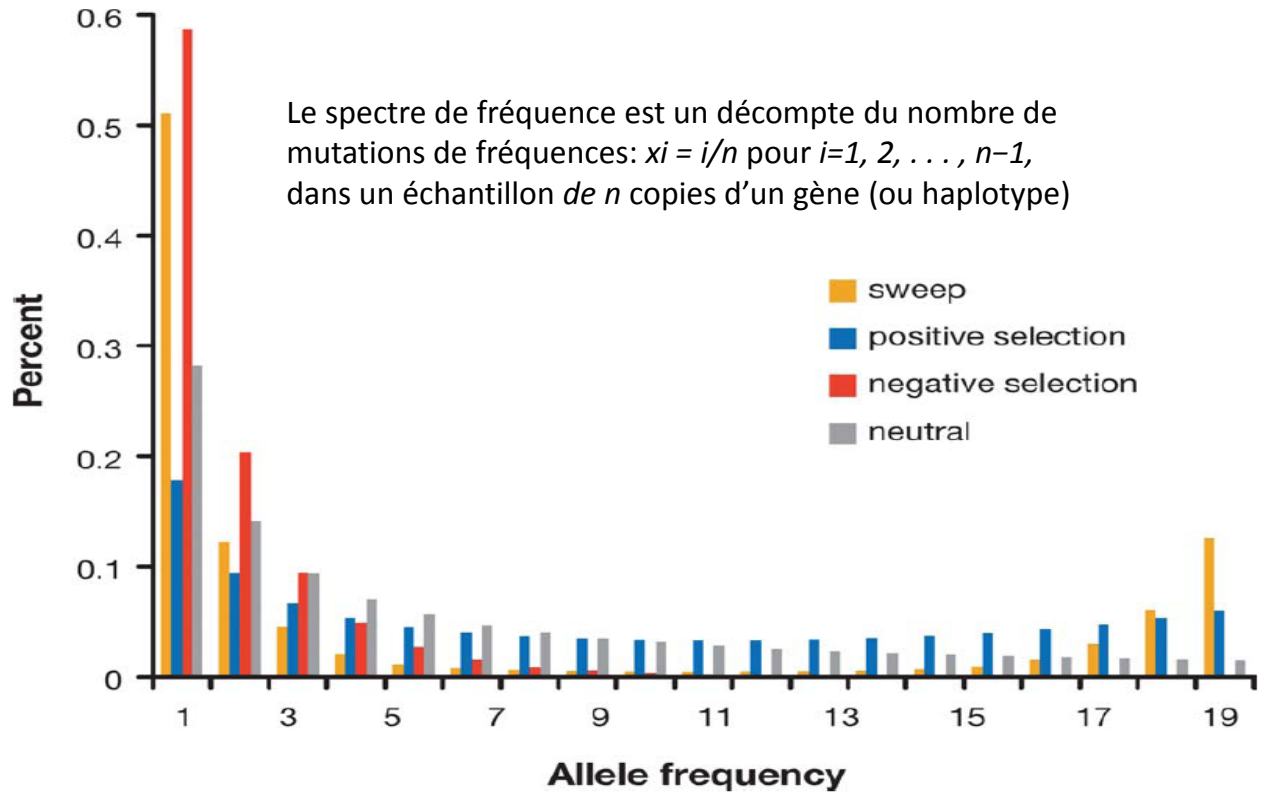
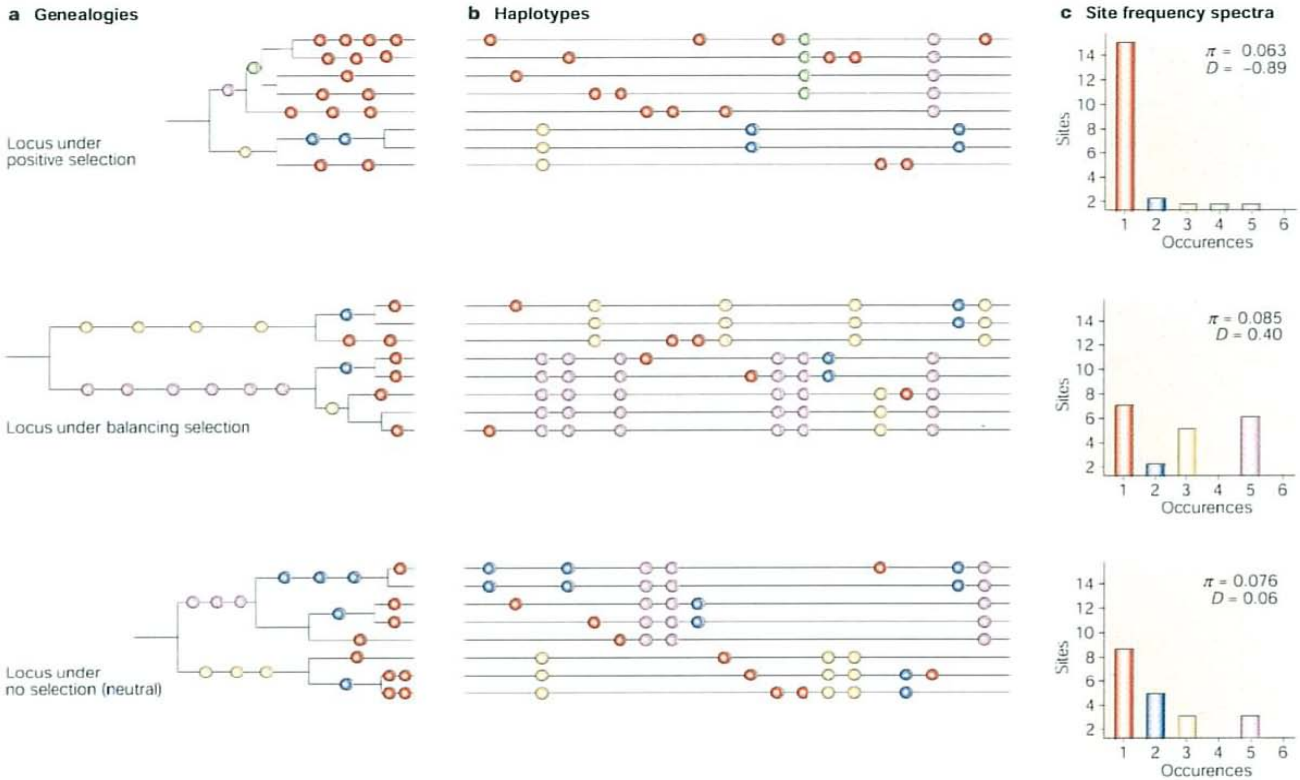


Figure 2

The frequency spectrum under a selective sweep, negative selection, neutrality, and positive selection. The frequency spectra under negative and positive selection are calculated using the PRF model by Sawyer & Hartl (88) for mutations with $2Ns = -5$ and 5, respectively, where N is the population size and s is the selection coefficient. For the selective sweep, the frequency spectrum is calculated in a window around the location of the adaptive mutation immediately after it has reached fixation in the population. In all cases, a demographic model of a population of constant size with no population subdivision is assumed.

Nielsen (2005)

Spectre de fréquences alléliques et sites nucléotidiques



Bamshad & Wooding (2003)

Figure 2 | **The effects of selection on the distribution of genetic variation. a** | The genealogies of three genes that are typical of loci under positive selection (top), balancing selection (middle) and no selection (neutral) (bottom) are depicted. Each circle represents a mutation, and the colour shows the final frequency of each mutation in the sampled HAPLOTYPES (**b**) and the SITE FREQUENCY SPECTRA (**c**). For each gene, the number of segregating sites is 20. **b** | Each haplotype contains mutations that have accumulated on each lineage in the gene genealogy, assuming no recombination. **c** | The site frequency spectrum of each gene. Positive selection (top) can result in a lower level of sequence diversity (π), an excess of low-frequency variants (red) and, consequently, a negative value of Tajima's *D* (BOX 2). Balancing selection (middle) can result in a higher level of sequence diversity (π), an excess of intermediate-frequency variants (purple) and, consequently, a positive value of Tajima's *D*. The diversity estimate and site frequency spectrum of a neutral locus (bottom) can be used for comparison.

Test de Tajima sur la diversité nucléotidique

- compare deux estimateurs du paramètre $\theta = 4N\mu$:
  - sur la base du nombre de sites qui ségrègent (S)
  - sur la base de la diversité nucléotidique (hétérozygotie) (π)
- chaque mutation crée un nouveau site ségrégeant (S) mais contribue très peu à la diversité nulcéotidique (π)
- 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 $H0$ = **neutralité** : $\mathbb{E}(D) = 0$ et $\mathrm{Var}(D) = 1$ (utilisation des lois normales et beta, ou simulations, pour effectuer le test)
- $D < 0$ = **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)
- $D > 0$ = **sélection balancée**, présence de mutations en fréquences intermédiaires : moins d'allèles rares mais beaucoup d'hétérozygotie, forte contribution de π (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*†‡, Srinivas Mummidi§¶, Enrique Gonzalez§¶, Seema S. Ahuja§¶, Diane M. Dunn†, W. Scott Watkins†, Stephen Wooding†, Anne C. Stone‖, Lynn B. Jorde†, Robert B. Weiss†, and Sunil K. Ahuja§¶

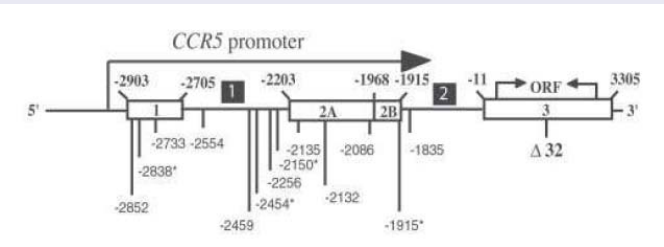
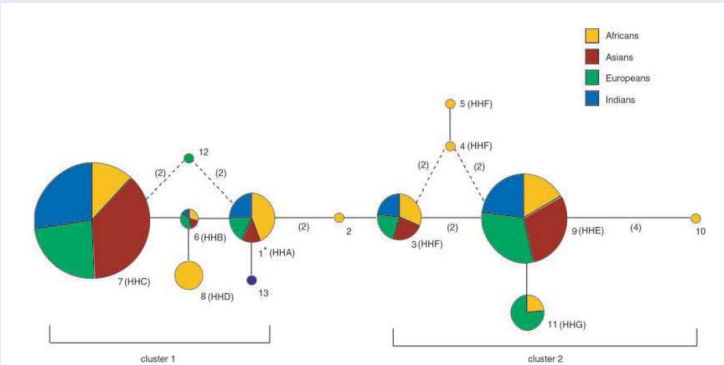


Table 1. Summary of sequence variation in the *cis*-regulatory region of *CCR5*

Population	*n**	*S*	θ_W	$\pi \pm$ SD, %	Tajima's *D*†	*Fs*†
NIH panel	176	9	1.57	0.29 ± 0.17	—	—‡
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)

*Number of chromosomes.
†*P* value is given in parentheses.
‡Ethnic identity unlinked to samples, therefore haplotypes could not be estimated reliably.



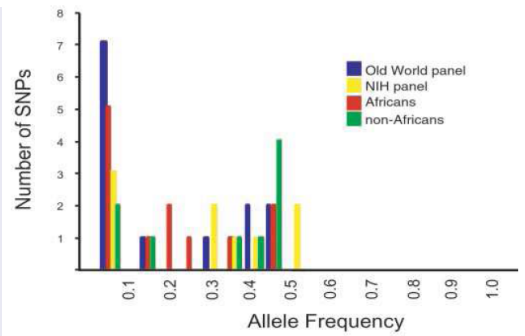
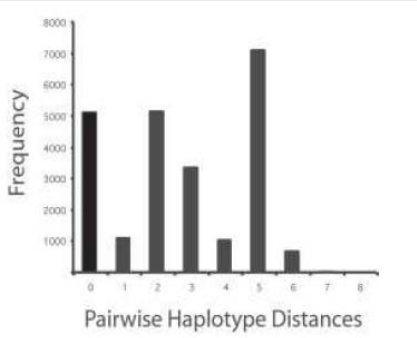


Fig. 2. Allele frequency spectrum for 13 SNPs found in the Old World and NIH panels, and for Africans and non-Africans. The frequency of the derived allele of each SNP is shown.



- 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

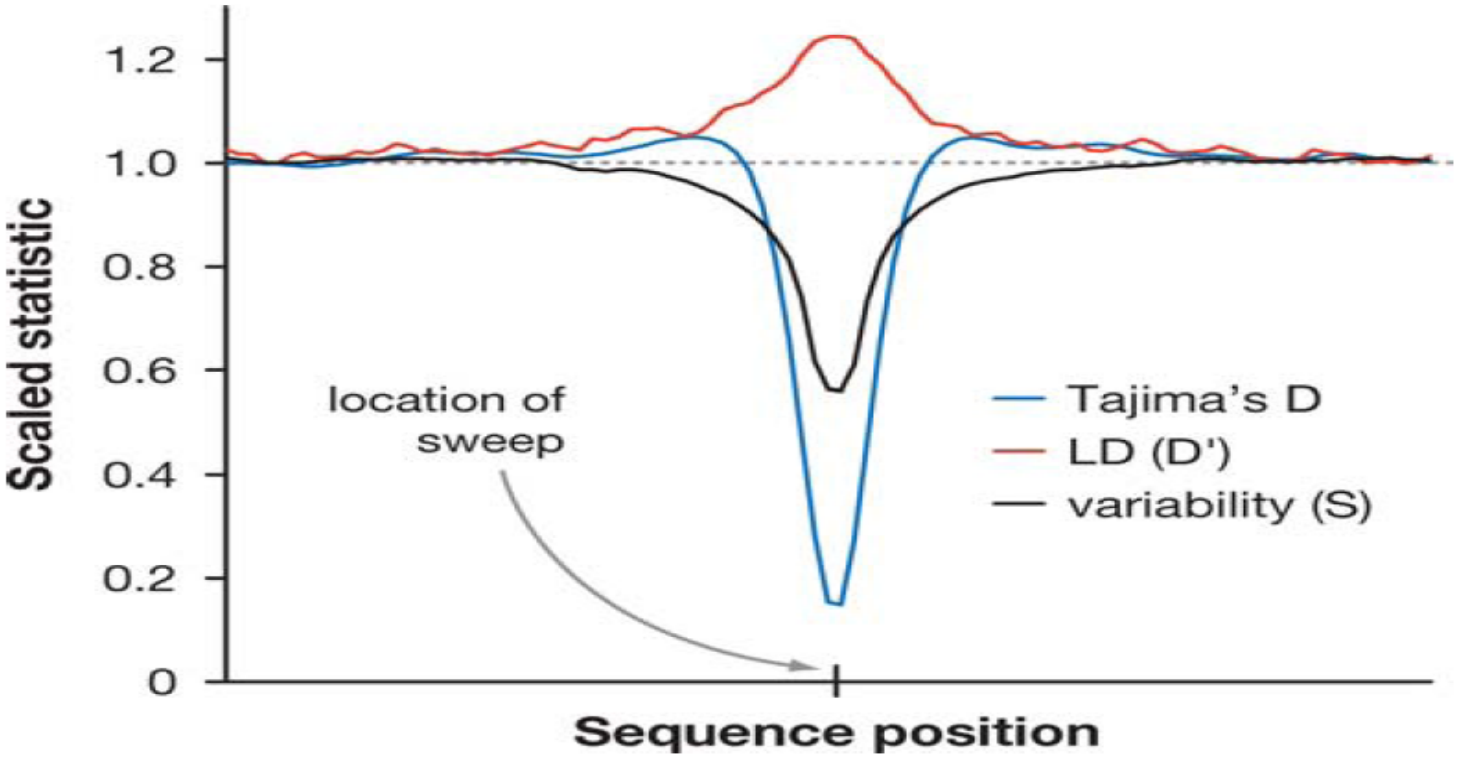


Figure 1

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 χ^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 Drosophile

Hétérogénéité du polymorphisme et de la divergence entre gènes

test HKA (Hudson-Kreitman-Aguade)

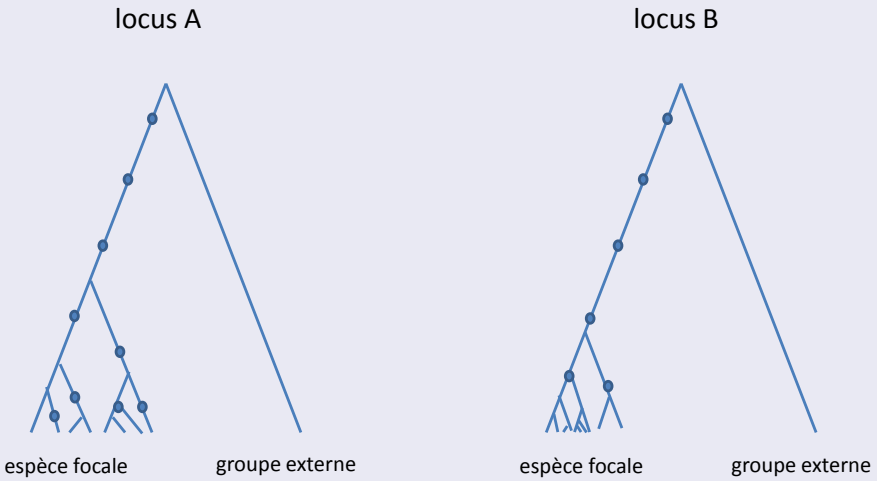
- polymorphisme neutre et divergence neutre sont deux facettes d'un même processus
- test de compatibilité du polymorphisme et de la divergence sur de multiples séquences de gènes non liés
- hypothèse du test :
  - à un locus avec fort taux de mutation : polymorphisme et divergence forts
  - à un locus avec faible taux de mutation : polymorphisme et divergence faibles
- pour L loci :
  - S_i^A et S_i^B = nombre de sites ségrégeant dans chaque espèce A et B, au locus i (mesure du polymorphisme)
  - D_i = nombre de sites divergents entre les espèces A et B, au locus i (mesure de la divergence)
- sous $H0$:

$$X^2 = \sum_{i=1}^{L} \frac{(S_i^A - \mathbb{E}(S_i^A))^2}{\mathrm{Var}(S_i^A)} + \sum_{i=1}^{L} \frac{(S_i^B - \mathbb{E}(S_i^B))^2}{\mathrm{Var}(S_i^B)} + \sum_{i=1}^{L} \frac{(D_i - \mathbb{E}(D_i))^2}{\mathrm{Var}(D_i)} \qquad (2)$$

- la statistique X^2 suit une loi de χ^2 à $3L - (L+2) = 2L - 2$ ddl **($3L$ observations moins L paramètres θ pour chaque gène pour une espèce, moins le ratio des 2 tailles de populations, moins le temps de divergence)**

Hétérogénéité du polymorphisme et de la divergence entre gènes

test HKA (Hudson-Kreitman-Aguade)



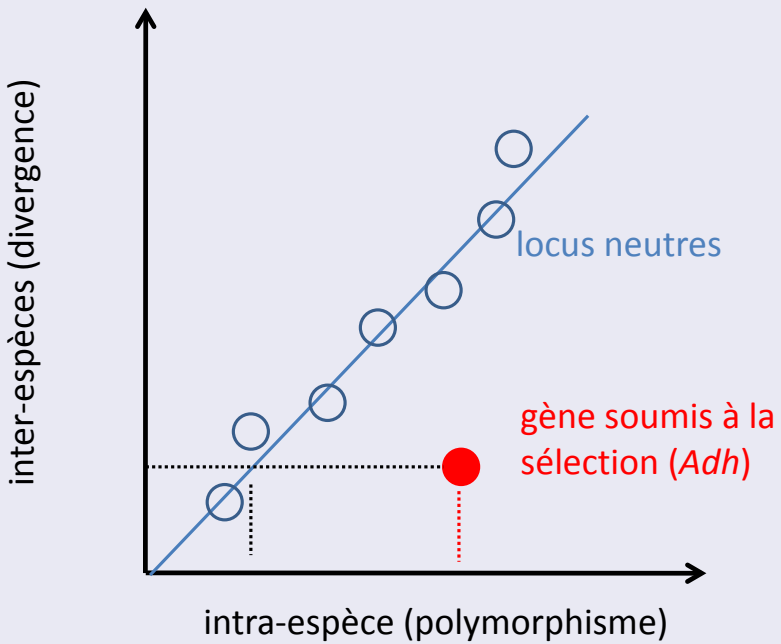
- les paramètres du modèle neutres: la taille N de la population, le taux de mutation μ
- le polymorphisme au locus se résume au paramètre $4N\mu$
- le polymorphisme réduit au locus B ne peut être expliqué par:
  - N petit (car le locus A a beaucoup de polymorphisme)
  - μ petit (car la distance au groupe externe serait alors réduite)

$\longrightarrow$ La sélection a influencé le polymorphisme au locus B

Hétérogénéité du polymorphisme et de la divergence entre gènes

gène de l'Adh

corrélation de la variation inter et intra-spécifique



	intra	inter	ratio
Adh	34	43	0.8
Locus neutre	30	77	0.4

« Trop » de polymorphisme au sein des espèces pour la divergence observée

Comparaison des taux de substitution

dégénerescence du code génétique

1re base	2e base U		2e base C		2e base A		2e base G		3e base
U	UUU	F Phe	UCU	S Ser	UAU	Y Tyr	UGU	C Cys	U
	UUC	F Phe	UCC	S Ser	UAC	Y Tyr	UGC	C Cys	C
	UUA	L Leu	UCA	S Ser	UAA	STOP Ocre	UGA	STOP Opale / U Sec / W Trp	A
	UUG	L Leu / START	UCG	S Ser	UAG	STOP Ambre / O Pyl	UGG	W Trp	G
C	CUU	L Leu	CCU	P Pro	CAU	H His	CGU	R Arg	U
	CUC	L Leu	CCC	P Pro	CAC	H His	CGC	R Arg	C
	CUA	L Leu	CCA	P Pro	CAA	Q Gln	CGA	R Arg	A
	CUG	L Leu	CCG	P Pro	CAG	Q Gln	CGG	R Arg	G
A	AUU	I Ile	ACU	T Thr	AAU	N Asn	AGU	S Ser	U
	AUC	I Ile	ACC	T Thr	AAC	N Asn	AGC	S Ser	C
	AUA	I Ile	ACA	T Thr	AAA	K Lys	AGA	R Arg	A
	AUG	M Met & START	ACG	T Thr	AAG	K Lys	AGG	R Arg	G
G	GUU	V Val	GCU	A Ala	GAU	D Asp	GGU	G Gly	U
	GUC	V Val	GCC	A Ala	GAC	D Asp	GGC	G Gly	C
	GUA	V Val	GCA	A Ala	GAA	E Glu	GGA	G Gly	A
	GUG	V Val / START	GCG	A Ala	GAG	E Glu	GGG	G Gly	G

- Acide aminé apolaire
- Acide aminé polaire
- Acide aminé acide
- Acide aminé basique
- Codon STOP

Comparaison des taux de substitution

Substitutions synonymes:

Séquence 1:	UUU	CAU	CGU
Séquence 2:	UUU	CAC	CGU
Acide aminé	*Phe*	*His*	*Arg*

Substitutions non-synonymes:

Séquence 1:	UUU	CAU	CGU
Séquence 2:	UUU	CAG	CGU
Acide aminé	*Phe*	*His*	*Arg*
		Gln	

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

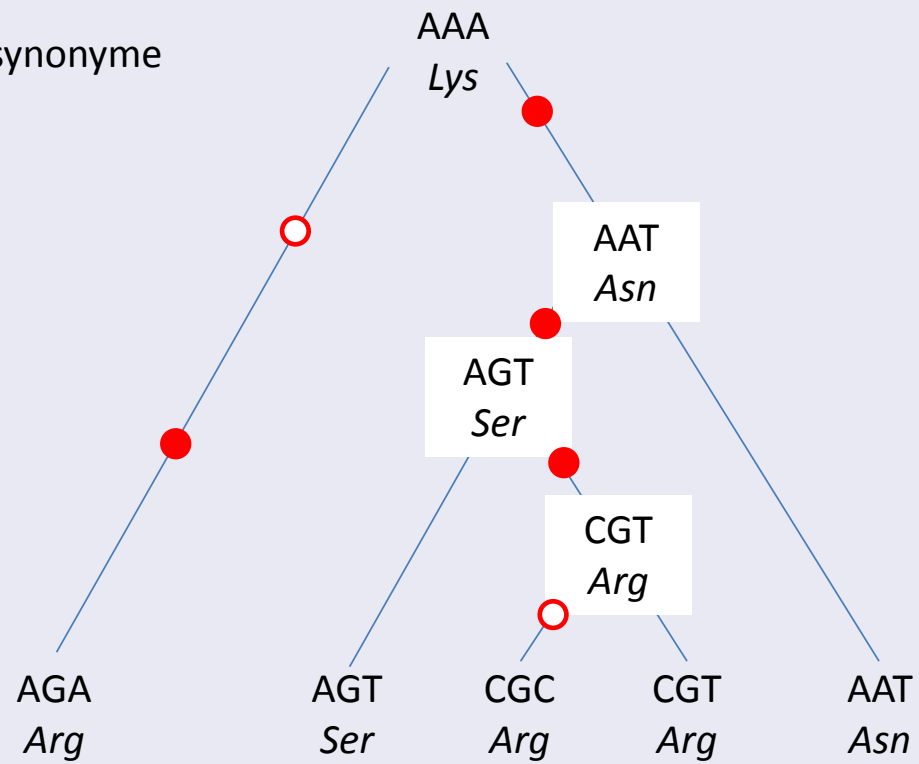
$$d_S = \frac{\text{nombre de substitutions synonymes}}{\text{nombre de sites synonymes}}$$

$d_N / d_S = \omega$

$0 < \omega < 1$: site conservé (sélection purifiante)
$\omega = 1$: neutralité
$\omega > 1$: sélection positive

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 comptent 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 ω le long d'une branche** de l'arbre et le **long de la séquence**, jusqu'à estimer un ω 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 s'applique sur les **phénotypes** dans une population. Elle affecte les fréquences **alléliques**, et c'est *via* ce changement qu'elle affecte la fréquence des phénotypes et est responsable de **l'évolution adaptative**
- la sélection peut prendre plusieurs formes selon **la distribution des valeurs sélectives** des allèles à un locus
- dans les populations naturelles, on ne connait 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)
- les tests de neutralité sont basés sur des informations de nature très différentes et peuvent fournir des résultats différents

Conclusion

la recherche de signatures de sélection permet de...

- mieux comprendre les contraintes fonctionnelles des différentes parties du génome, des gènes
- préciser l'étendue génomique locale des effets de la sélection
- identifier des gènes fonctionnellement importants
- 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
- contribuer à dresser l'"arbre de la Vie"

Liste des tests usuels

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

Test	Compares
Tests based on allelic distribution and/or level of variability	
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
Tests based on comparisons of divergence and/or variabillty between different classes of mutation	
d_N/d_S, K_a/K_s	The ratios of non-synonymous and synonmyous 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