Evolution Moléculaire (Masters 1 Bioinformatique / Biotechnologie)
Evolution moléculaire : neutre ou adaptative ?

Maxime Bonhomme

Laboratoire de Recherche en Sciences Végétales

30 août 2020

Evolution Moléculaire : neutre ou adaptative ?

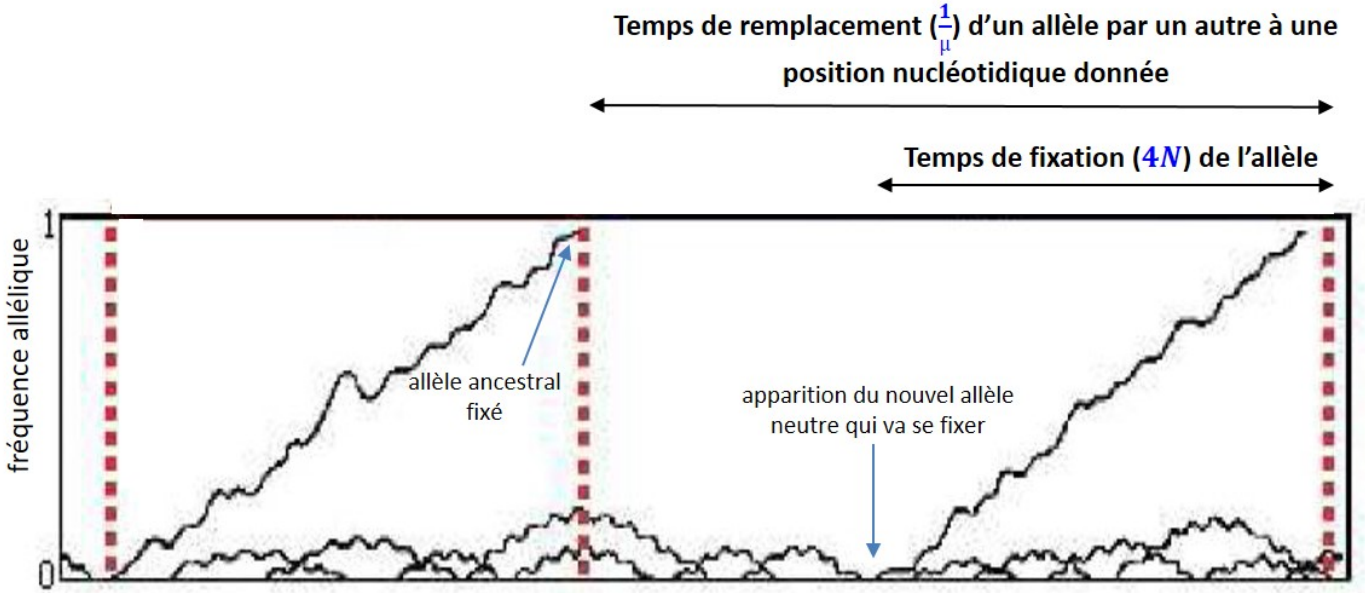
Théorie neutraliste de l'évolution moléculaire

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, μ = 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 fixation (remplacement) 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

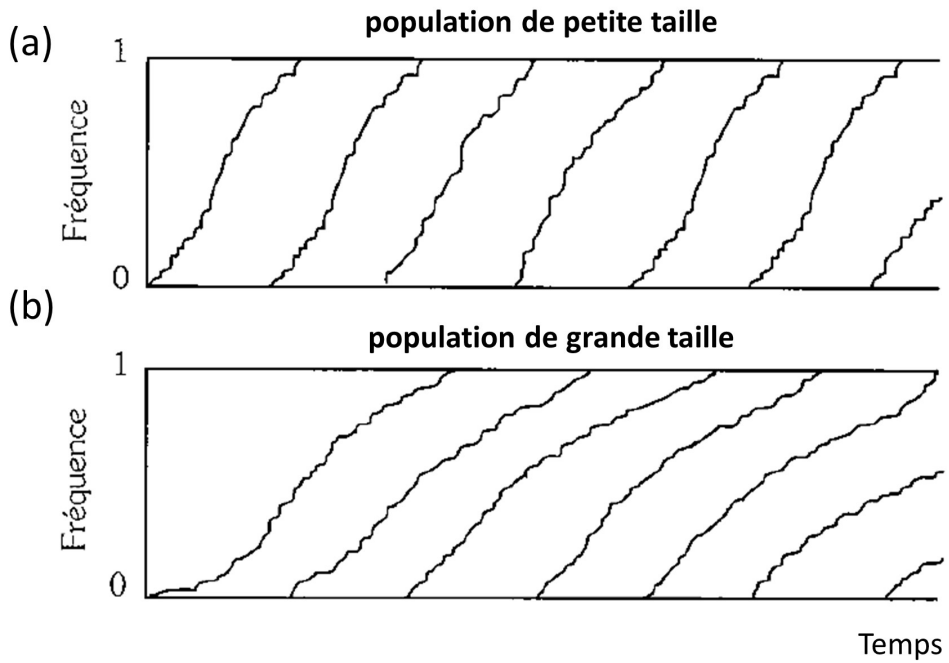
Dynamique de remplacement des allèles neutres



De nouveaux allèles neutres apparaissent sans cesse sur un locus par mutation. La plupart sont perdus, mais certains finissent par se fixer et remplacent les allèles ancestraux. Cela arrive en moyenne toute les $\frac{1}{\mu}$ générations, μ étant le taux de mutation neutre du locus. Le temps moyen entre l'apparition d'un nouvel allèle neutre destiné à remplacer l'allèle ancestral et le moment où il se fixe est de $4N$ générations.

Théorie neutraliste

La quantité de polymorphisme au locus dépend de la taille de la population



Le taux de mutation μ est le même, donc le taux de remplacement neutre aussi ($= \mu$), **mais comme $4N$ est différent, il y a tout moment plus de mutations neutres (polymorphisme $\theta = 4N\mu$) dans la population de grande taille que dans celle de petite taille ($\theta_b > \theta_a$).**

Théorie neutraliste

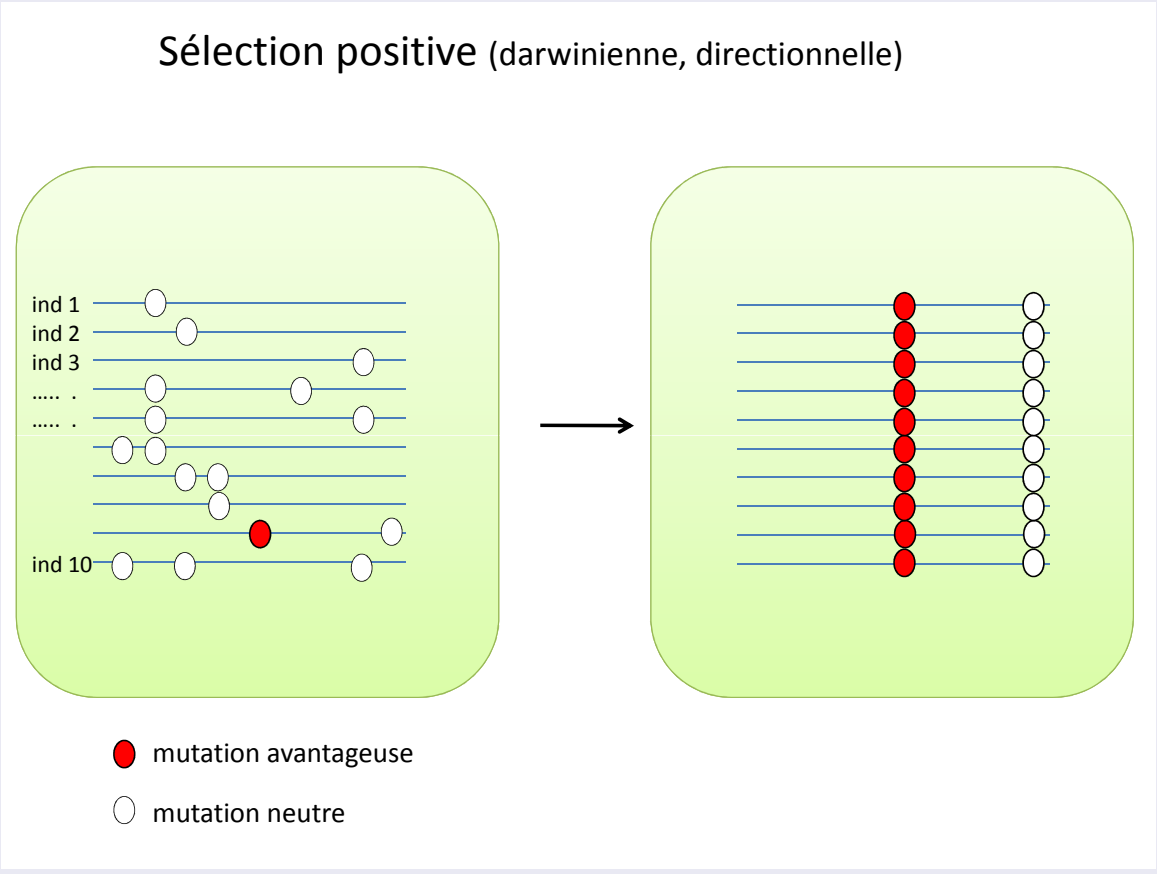
En pratique

- **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 ?

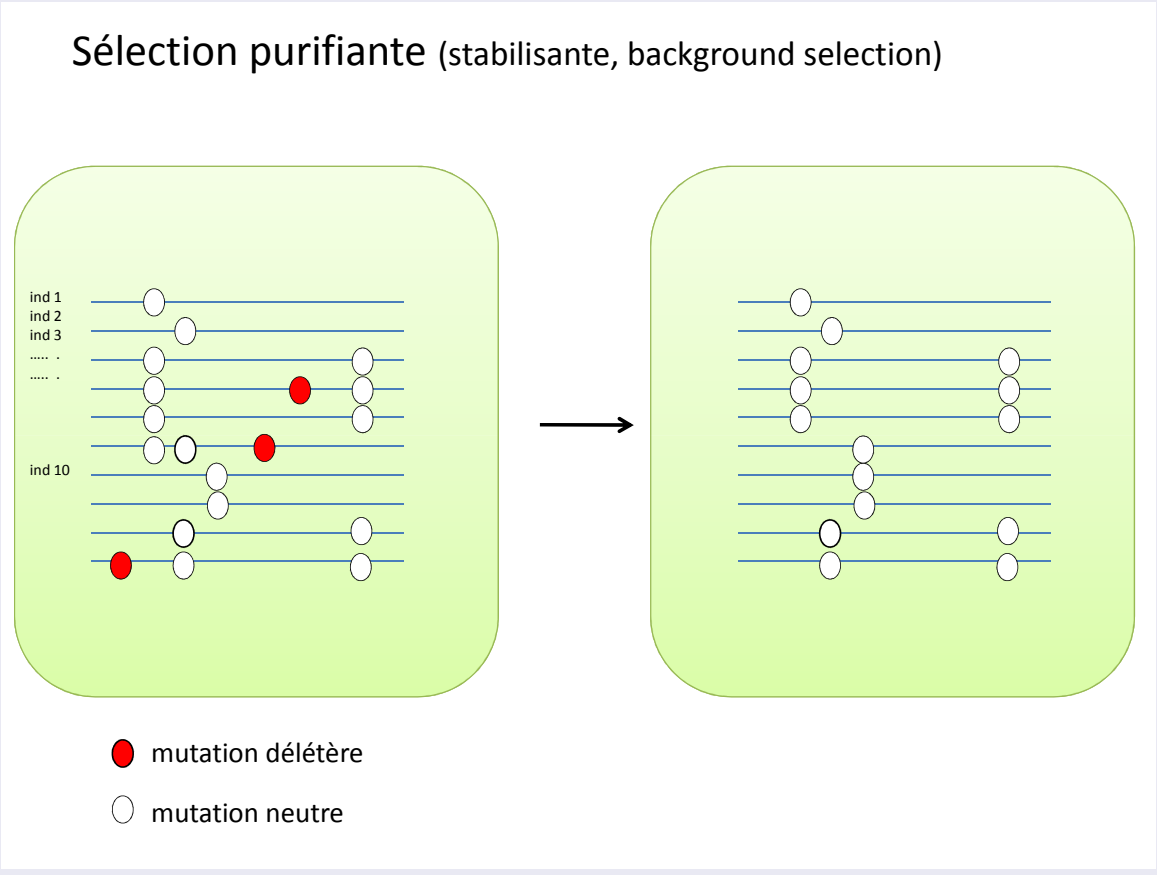
- affecte la **distribution des fréquences alléliques**, parfois rapidement
- affecte le **nombre d'allèles** maintenus (augmentation, diminution)
- affecte l'**hétérozygotie**
- affecte la proportion de **changements synonymes et non-synonymes** des séquences codantes

Sélection positive



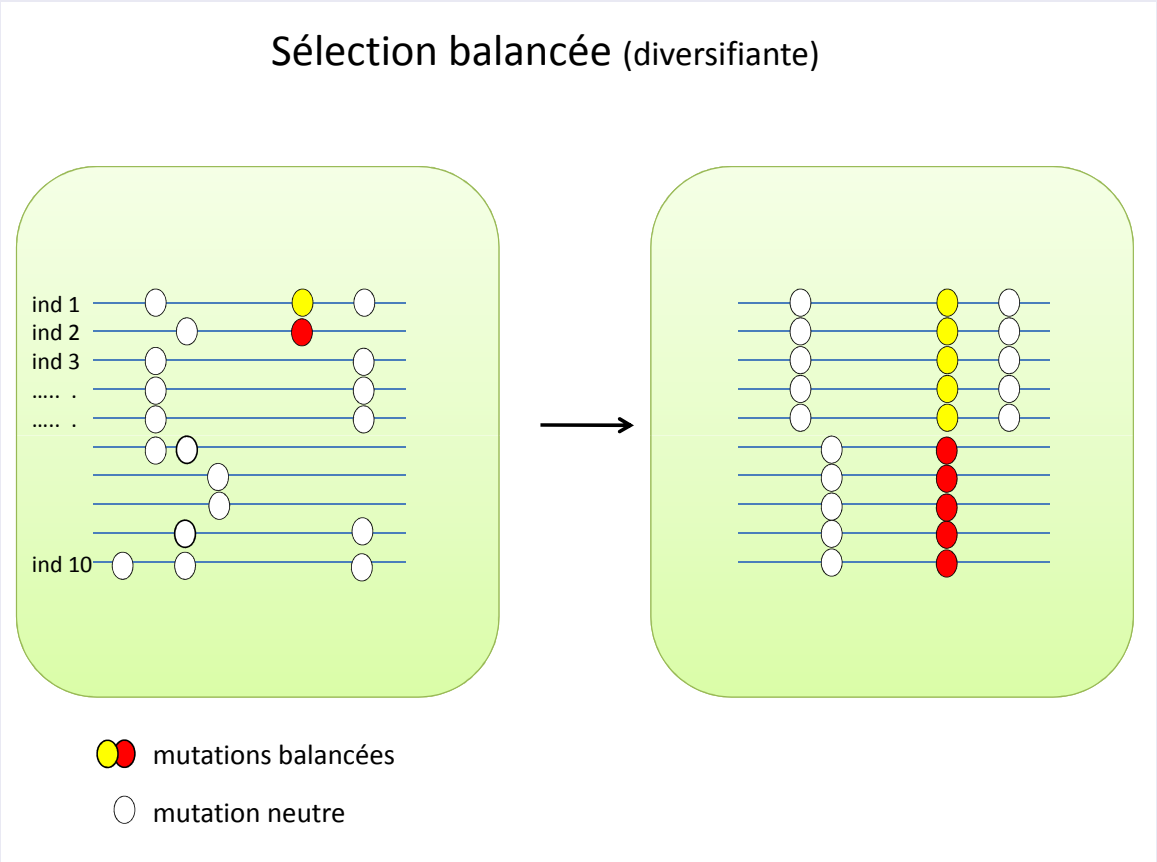
- gènes ayant un rôle dans l'adaptation (ex : résistance aux insecticides chez le moustique, adaptation à la sécheresse)

Sélection négative (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)

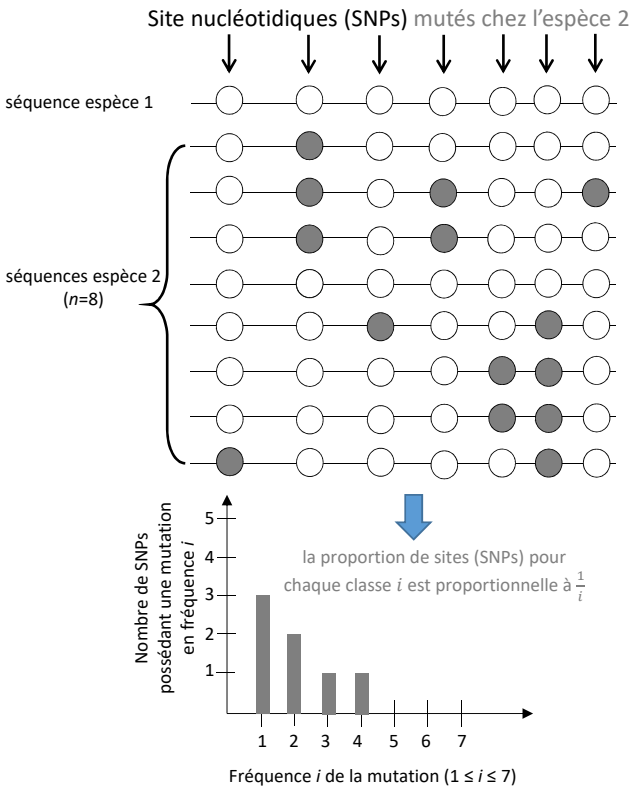
Comment détecter la sélection ?

- **approche directe** : suivre expérimentalement une population au cours du temps
  - avantage : on connaît et on manipule l'agent de la sélection.
  - "evolve and resequence"
  - contraintes sur l'échelle de temps et taille d'échantillons.
  - donc contraintes sur les organismes étudiés (ex : bactéries, drosophile, arabidopsis,..).
- **approche alternative "indirecte"** : différentes signatures moléculaires de la sélection sur les séquences d'ADN peuvent êtres utilisées :
  - spectre de fréquences alléliques et diversité nucléotidique
  - polymorphisme / divergence
  - méthodes phylogénétiques ($\frac{d_N}{d_S}$)
    - 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

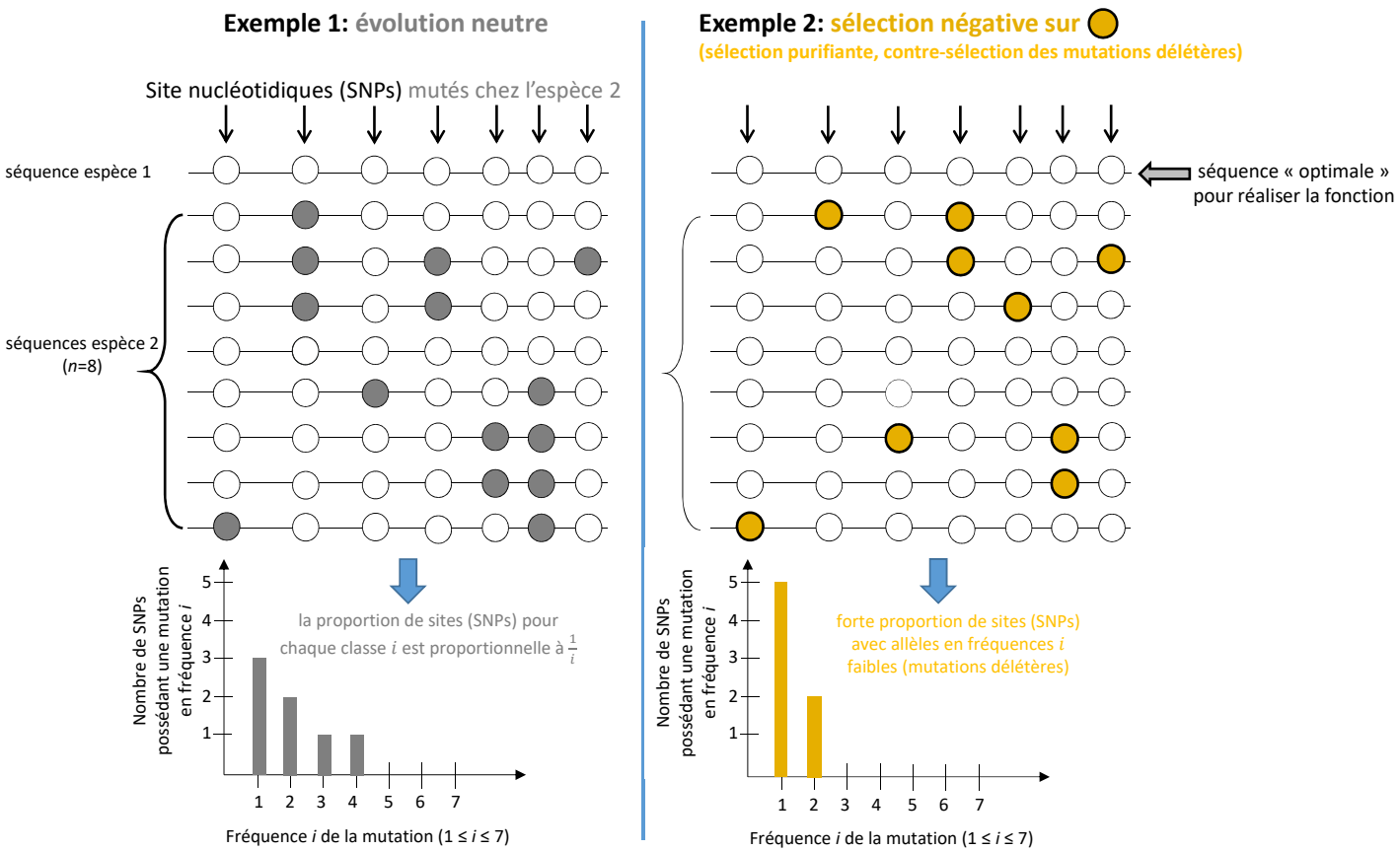
Neutralité

Exemple 1: évolution neutre



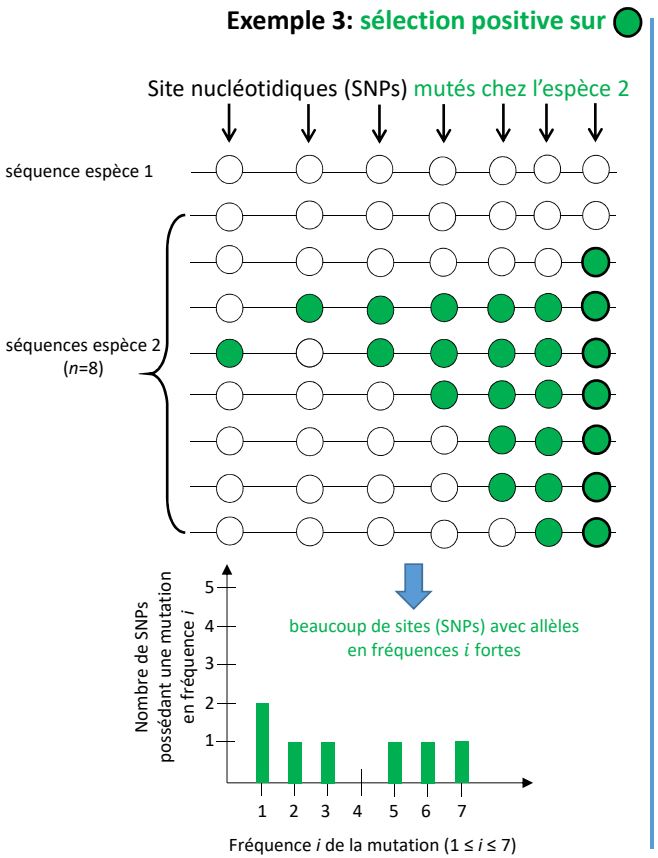
Spectre de fréquences alléliques et sites nucléotidiques

Sélection négative (purifiante)



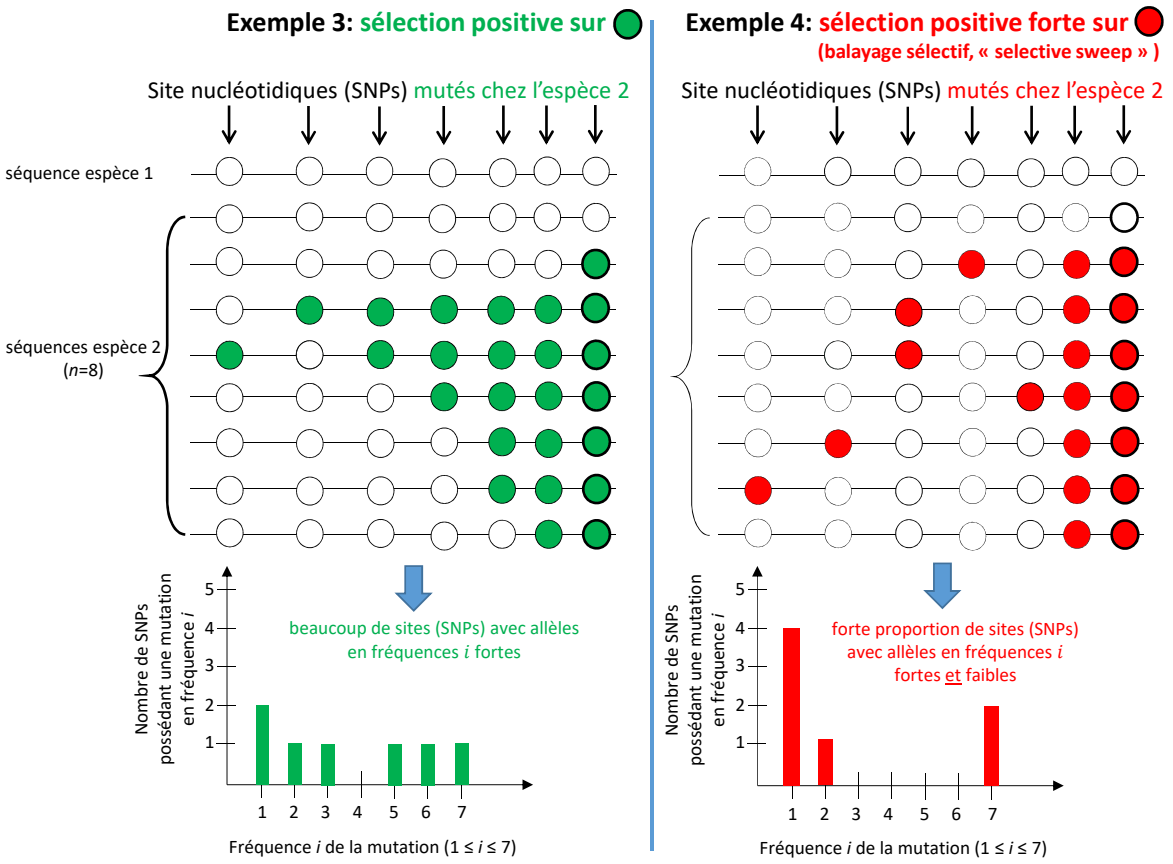
Spectre de fréquences alléliques et sites nucléotidiques

Sélection positive



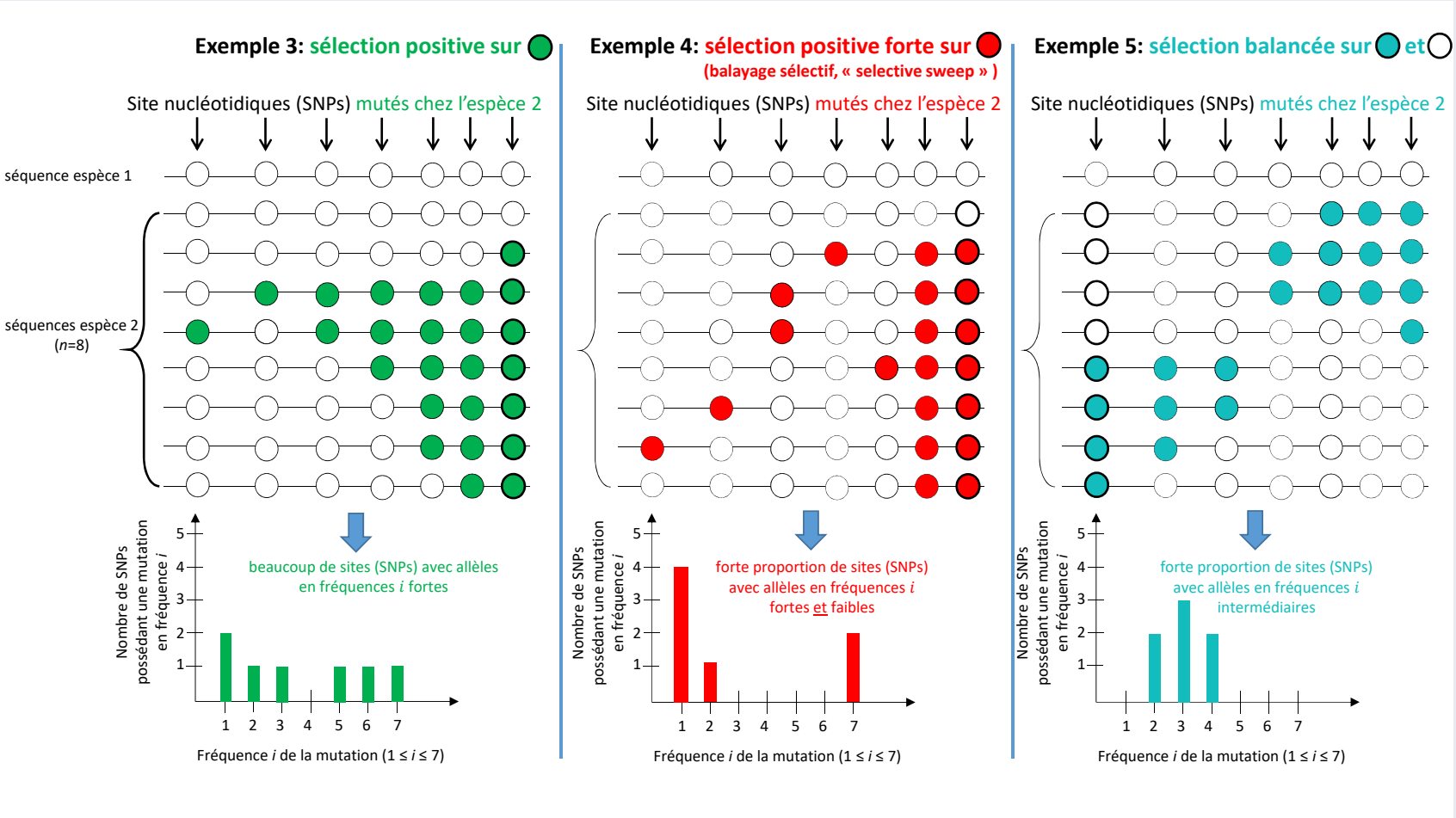
Spectre de fréquences alléliques et sites nucléotidiques

Sélection positive intense (balayage sélectif)



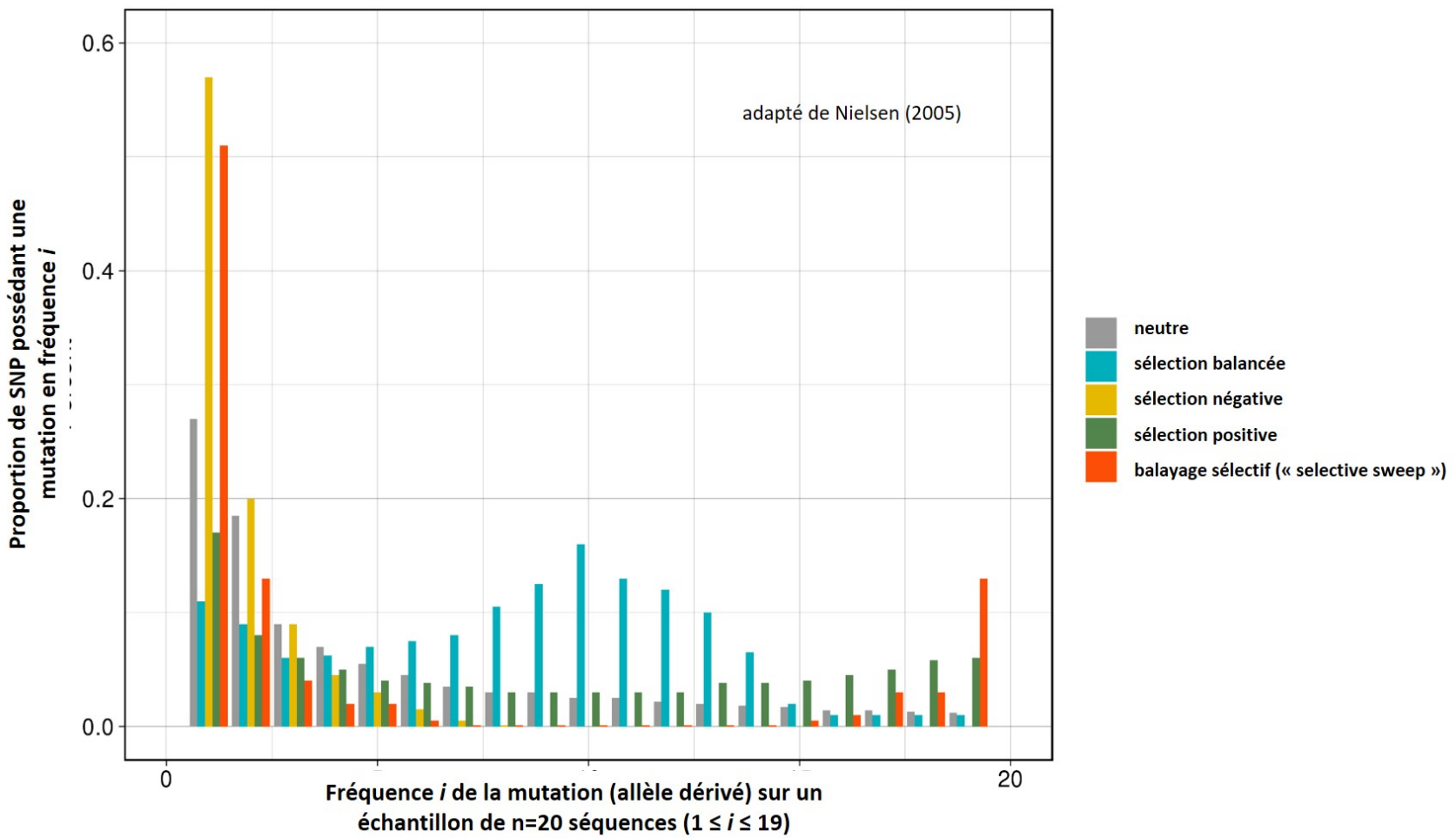
Spectre de fréquences alléliques et sites nucléotidiques

Sélection balancée



Spectre de fréquences alléliques et sites nucléotidiques

"Site Frequency Spectrum" (SFS) pour différentes formes de sélection



Spectre de fréquences alléliques et sites nucléotidiques, et généalogie

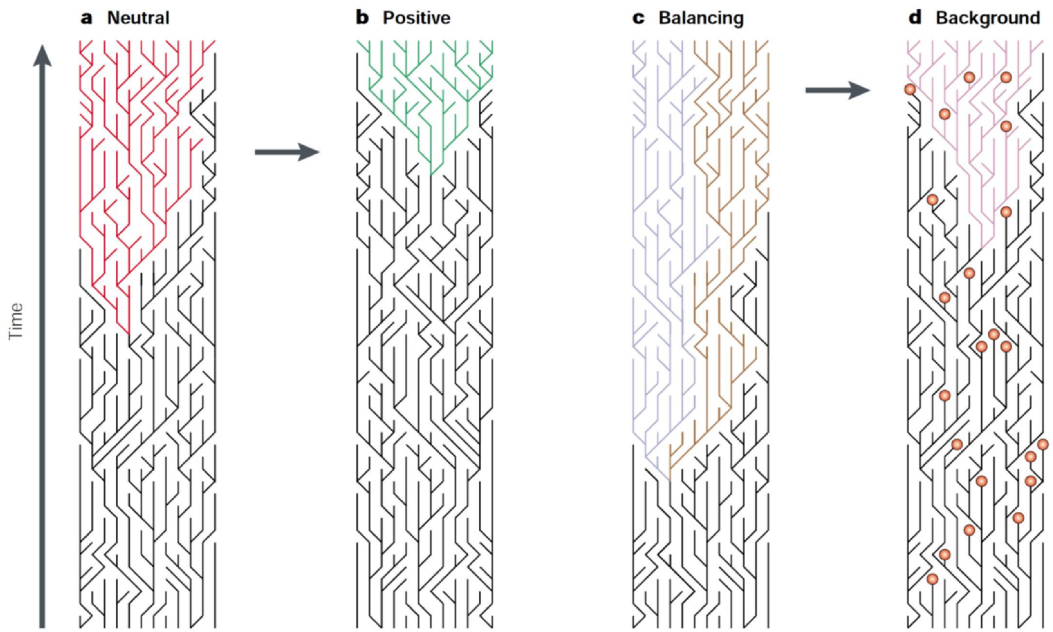
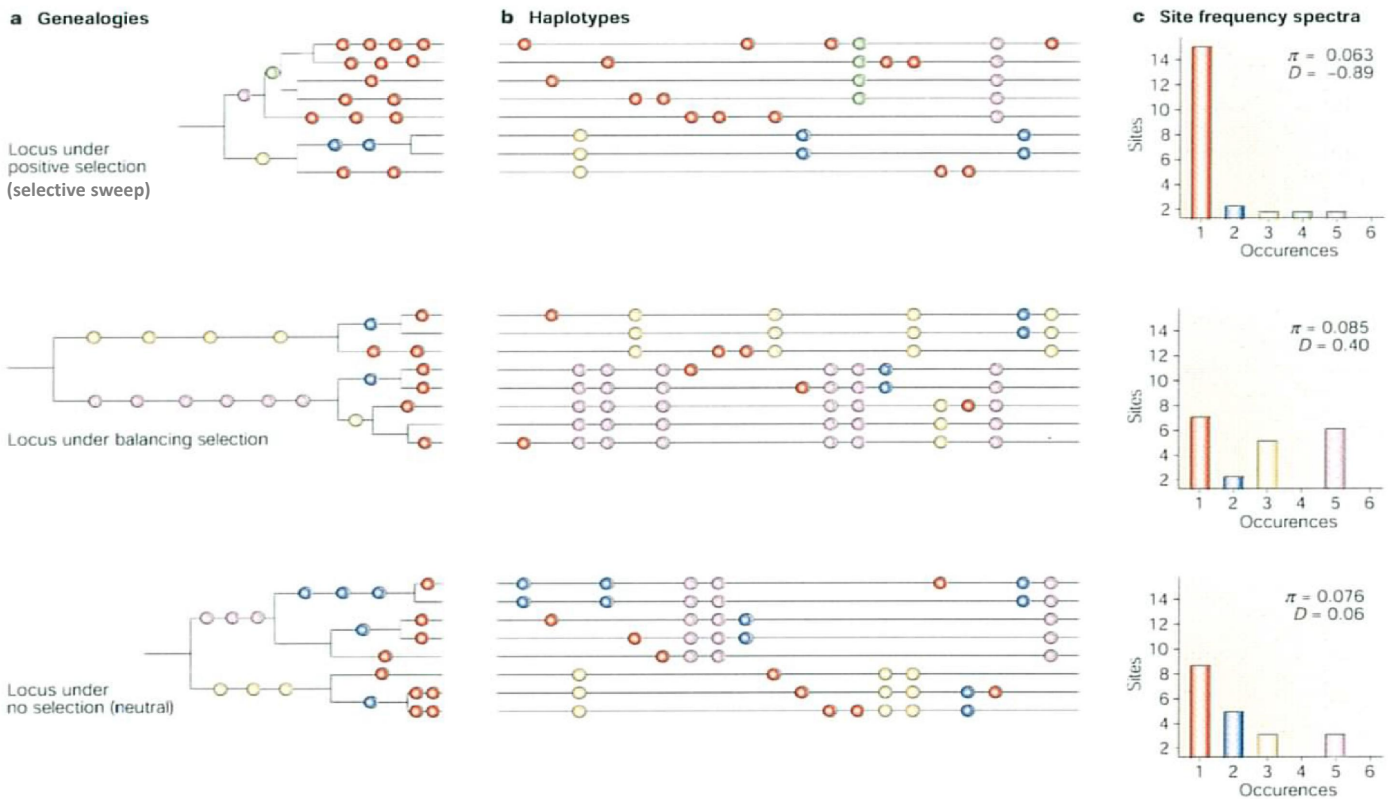


Figure 1 | **Effects of natural selection on gene genealogies and allele frequencies.** Each panel (**a**–**d**) represents the complete genealogy for a population of 12 haploid individuals. Each line traces the ancestry of a lineage, and coloured lines trace all descendants who have inherited an allele that is either neutral (**a**) or affected by natural selection (**b**–**d**) back to their common ancestor (that is, the coalescence of the genealogy). **a** | The genealogy of a neutral allele (red) as it drifts to FIXATION. **b** | The genealogy of an allele (green) that is driven to fixation more quickly after the onset of positive selection (arrow) compared with expectations under a neutral model. Note that the genealogy has a more recent coalescence. **c** | The genealogy of two alleles (blue and gold) under BALANCING SELECTION, which are driven neither to fixation nor to extinction. As a result, the genealogy of the two alleles has an older coalescence. **d** | The genealogy of an allele (purple) that drifts to fixation under the influence of BACKGROUND SELECTION. Each circle represents the elimination of a deleterious (that is, lethal) mutation by background selection. The coalescence of the lineage is more recent than expected under a neutral model because a linked deleterious mutation caused the extinction of one lineage (arrow) more quickly than would be predicted.

Bamshad & Wooding (2003)

Spectre de fréquences alléliques et sites nucléotidiques, et généalogie



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 nucléotidique (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§¶

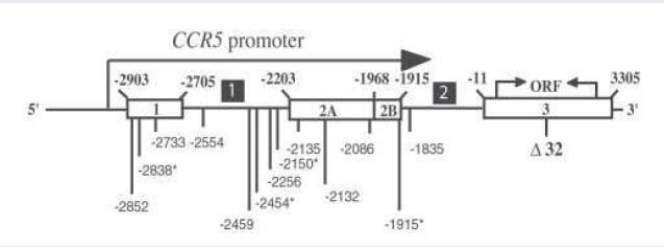
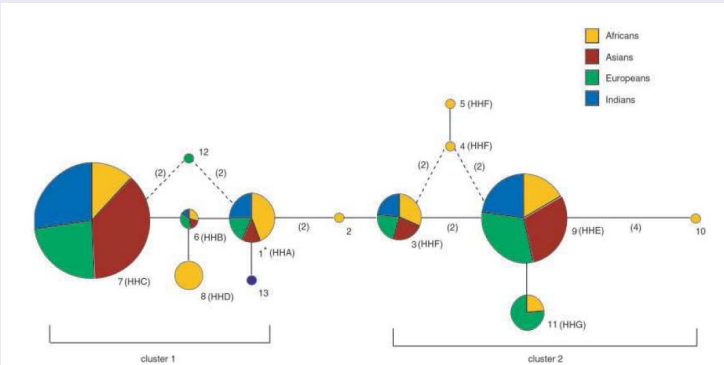


Table 1. Summary of sequence variation in the *cis*-regulatory region of *CCR5*

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.



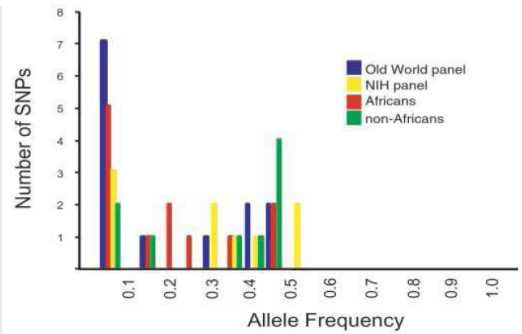
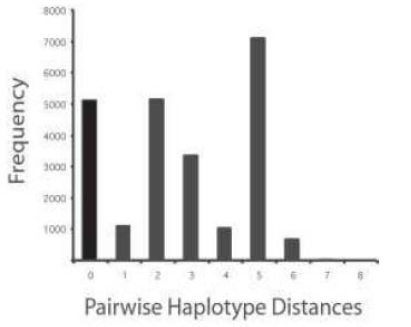


Fig. 2. Allele frequency spectrum for 13 SNPs found in the Old World and NIH panels, and for Africans and non-Africans. The frequency of the derived allele of each SNP is shown.



- 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

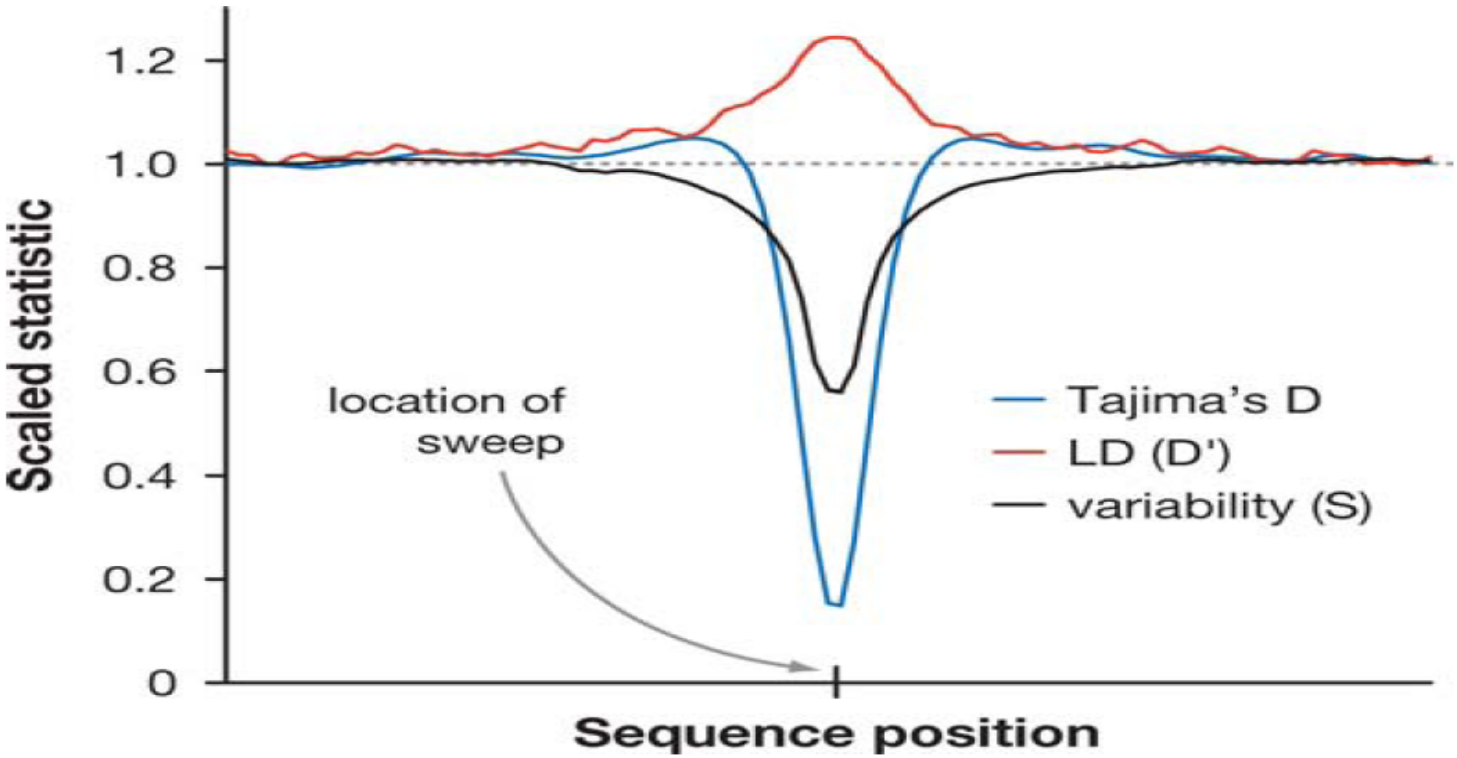


Figure 1

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

Comparaison des taux de substitution

dégénerescence du code génétique

le code génétique

Première lettre (côté 5')	Deuxième lettre U		Deuxième lettre C		Deuxième lettre A		Deuxième lettre G		Troisième lettre (côté 3')
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

Comparaison des taux de substitution pour deux séquences

séquence 1: TCC GCA CCA
séquence 2: TCA GCA ACA

séquence 1: TCC GCA CCA
séquence 2: TCA GCA ACA

séquence 1: *Ser* *Ala* *Pro*
séquence 2: *Ser* *Ala* *Thr*

- nombre moyen de sites synonymes par séquence: 3
- substitutions synonymes: 1
- nombre moyen de sites non synonymes par séquences: 6
- substitutions non synonymes: 1

d_N = 1/6

d_S = 1/3

$d_N / d_S = \omega = 0.5$

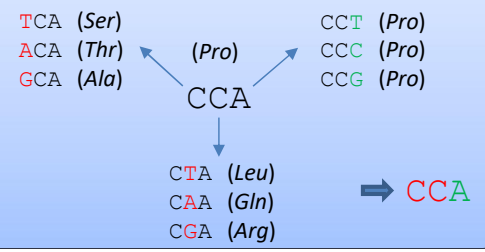
$$d_N = \frac{\text{nombre de substitutions non synonymes}}{\text{nombre de sites non synonymes}}$$

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

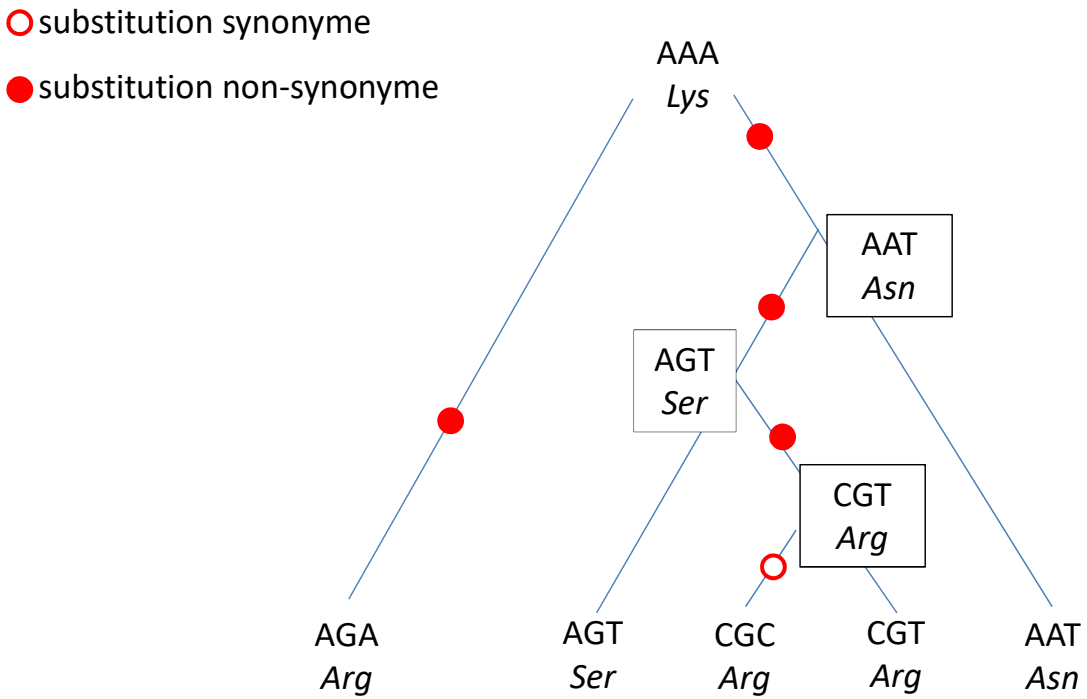
$d_N / d_S = \omega$ [aussi appelé Ka/Ks]

$0 < \omega < 1$: sélection purifiante (séquence conservée)
$\omega = 1$: neutralité
$\omega > 1$: sélection positive (diversifiante)

Quels sont les sites s et ns sur un codon ? : un cas simple



Comparaison des taux de substitution sur les branches de l'arbre



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 ω ¿ 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 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)

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
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