Génétique des Populations – Génétique Humaine

Polymorphisme – Coalescence – Tests de Neutralité

S. Mousset

Biométrie et Biologie Évolutive, UMR5558, Lyon I

Bioinformatique et Génomique Évolutive

29 avril 2010

Table des matières

Devenir des mutations dans les populations

Devenir des mutations :

- Origine des différences entre individus de la divergence entre les espèces.
- Comment évoluent-elles ?
- Quels sont les effets de la sélection ?
- Quel est le rôle de la taille des populations ?

Plan détaillé

Le modèle neutre de Wright-Fisher

Une population *théorique idéale*

- De taille constante
- À générations non chevauchantes
- Sans effets de la sélection
- Nombre de descendants $\sim P(1)$

L'effectif efficace d'une population

The effective population size N_e is the size of a theoretical ideal Wright-Fisher population that would most closely reflect the evolutionary behavior of a nonideal natural population of N individuals.

L'effectif efficace d'une population

Généralement $N_e << N$, $(\frac{N_e}{N} < 0.001)$:

- Variance du succès reproducteur > 1
- Variation stochastique de l'environnement
- Structure des populations
- Fluctuation de la taille des populations (démographie)
- Effets de la sélection
- ...

Plan détaillé

Le destin d'un nouvel allèle neutre

Probabilité de perte immédiate

La fréquence d'un nouvel allèle est

$$p_0 = \frac{1}{2N}$$

La probabilité qu'il soit immédiatement perdue est

$$\left(1 - \frac{1}{2N}\right)^{2N} = e^{2N \ln\left(1-\frac{1}{2N}\right)} \approx e^{-1}$$

Rq : cette probabilité est indépendante de N.

Le destin d'un nouvel allèle sélectionné

Probablité de perte immédiate

La probabilité d'échantillonnage d'un nouvel allèle ayant un coefficient de sélection s est

$$p_0 = \frac{1+s}{2N}$$

La probabilité qu'il soit immédiatement perdue est

$$\left(1 - \frac{1+s}{2N}\right)^{2N} = e^{2N \ln\left(1-\frac{1+s}{2N}\right)} \approx e^{-(1+s)}$$

Rq : Cette probabilité est indépendante de N.

Probabilité de fixation d'un nouvel allèle

Cas d'un allèle neutre

La probabilité de fixation p_f est la fréquence initiale de l'allèle $\frac{1}{2N}$.

Le temps de séjour à l'état polymorphe *sachant que l'allèle s'est fixé* est $4N_e$.

Si le taux de mutation neutre est μ, alors la vitesse de fixation des nouvelles mutations est :

$$2N\mu\frac{1}{2N} = \mu$$

Rq : ce taux d'évolution est indépendant de N.

Probabilité de fixation d'un nouvel allèle

Cas d'un allèle sélectionné

La probabilité de fixation p_f est

$$p_f = \frac{1 - e^{-\frac{2N_e s}{N(1-s)}}}{1 - e^{-\frac{4N_e s}{1-s}}} \approx \frac{\frac{2N_e s}{N}}{1 - e^{-4N_e s}}$$

Le temps de séjour à l'état polymorphe *sachant que l'allèle s'est fixé* est $\frac{4}{|s|} \ln 2N_e$

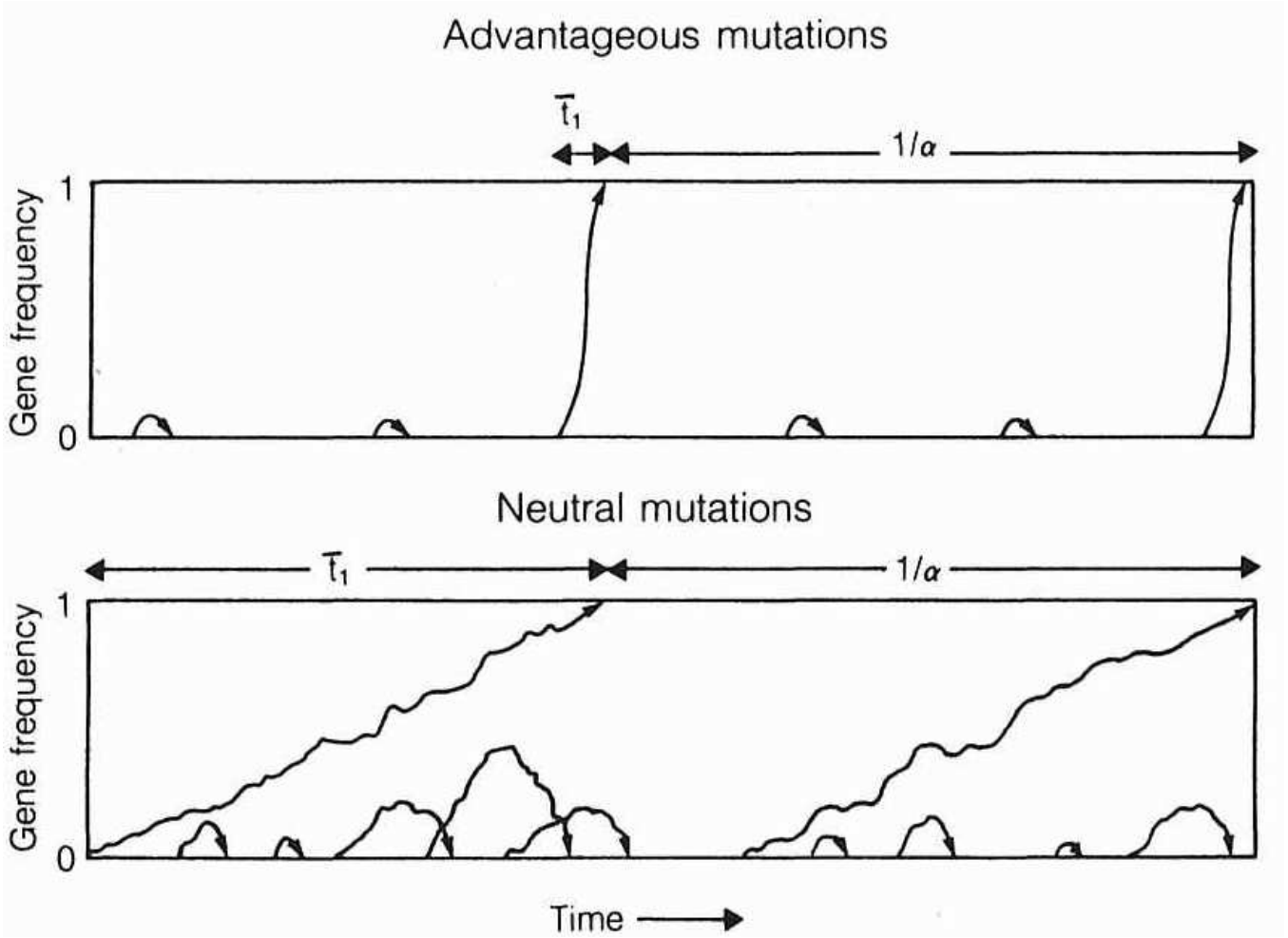
Le taux de fixation des mutations avantageuses tend vers

$$4N_e s \mu_b$$

Rq : ce taux d'évolution est *dépendant* de N_e.

Plan détaillé

Vitesse de fixation des mutations



- Le taux de substitution est égal au taux de mutation par locus (Kimura, 1968).
- Les mutations avantageuses et délétères se fixent rapidement $\left(\bar{t}_1 = \frac{4 \ln 2N_e}{s}\right)$
- Les mutations neutres se fixent lentement $(\bar{t}_1 = 4N_e)$
- Les mutations délétères sont maintenues à une fréquence faible (équilibre mutation-dérive).

$\Rightarrow$ L'essentiel du polymorphisme observable est neutre.

Plan détaillé

- La divergence résulte de la fixation de mutations initialement polymorphes.
- Une mutation fixée a subit le "filtre" de la sélection naturelle.
- Les mutations polymorphes sont soumise à la dérive génétique.

$\Rightarrow$ Contraster le patron de polymorphisme et de divergence nous renseigne sur les processus sélectifs.

$\Rightarrow$ Le polymorphisme nous renseigne sur les mécanismes démographiques.

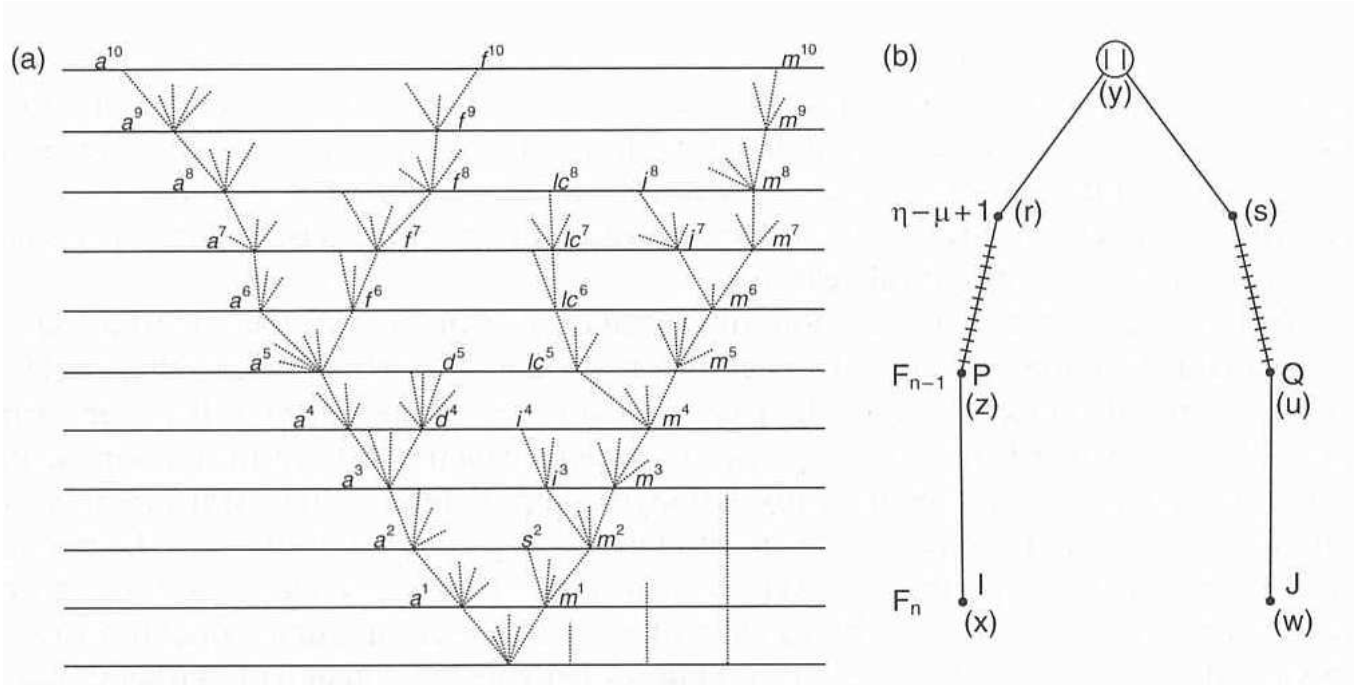
Table des matières

Qu'est-ce que le coalescent ?

- Un outil mathématique
- Un paradigme pour la la généalogie *attendue* d'un échantillon de lignée ($\neq$ phylogénie)
- Un outil pour l'évolution du polymorphisme dans les populations

Plan détaillé

Les origines du coalescent



Qu'est-ce que l'identité par ascendance ?

- Identité par ascendance (IBD) = Séquence ancestrale commune.
- Combien de temps jusqu'à un ancêtre commun dans une population de WF ?
- Combien de mutations entre deux séquences homologues ?

Plan détaillé

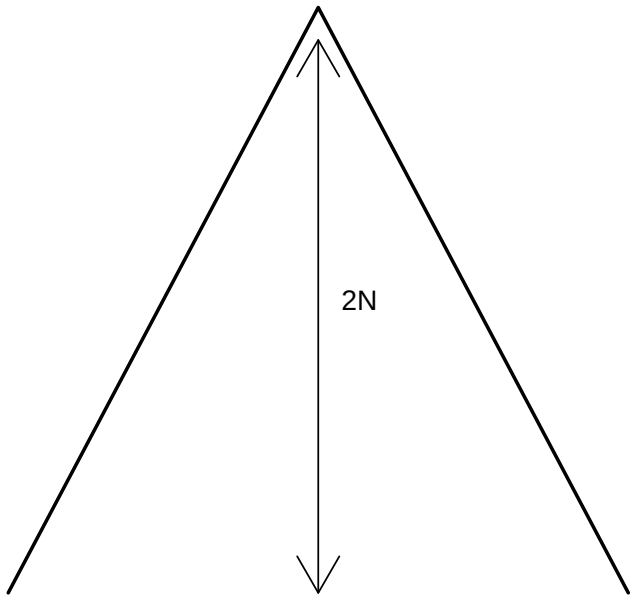
IBD dans une population de Wright-Fisher

- IBD à la première génération : $p_1 = \frac{1}{2N}$.
- IBD à la deuxième génération : $p_2 = \frac{2N-1}{2N}\frac{1}{2N}$.
- ...
- IBD à la $k$ième génération : $p_k = \left(\frac{2N-1}{2N}\right)^{(k-1)} \frac{1}{2N}$.

$\Rightarrow$ Le temps d'attente avant un événement d'IBD suit une loi géométrique de paramètre $\frac{1}{2N}$.

Temps moyen avant la coalescence de deux lignées

L'espérance du temps avant l'IBD est $2N$ générations.



$\pi = 4N\mu$

Si le taux de mutation par génération est μ, alors le nombre attendu de différences entre paires de séquences est :

$$\pi = 4N\mu$$

$\Rightarrow$ Un estimateur de $4N\mu$ basé sur un échantillon de n séquences orthologues est

$$\hat{\theta}_\pi = \hat{\pi} = \frac{1}{\binom{n}{2}} \sum_{i \leq j} k_{ij},$$

où k_{ij} est le nombre de différences entre les séquences i et j (Tajima, 1983).

Plan détaillé

Coalescence de n lignées, temps discret

À chaque génération, la probabilité que 2 lignées parmi n coalescent est

$$p_n = \binom{n}{2} \frac{1}{2N} = \frac{n(n-1)}{4N}$$

On note T_n la durée d'attente avant un événement de coalescence de 2 lignées, T_n suit une loi géométrique de paramètre $\frac{n(n-1)}{4N}$.

$$p(T_n \leq \tau) = 1 - \left(1 - \frac{n(n-1)}{4N}\right)^{\tau} = 1 - e^{\tau \ln\left(1 - \frac{n(n-1)}{4N}\right)}$$

Coalescence de n lignées, temps continu

$$p(T_n \leq \tau) = 1 - e^{\tau \ln\left(1 - \frac{n(n-1)}{4N}\right)}$$

Or $\ln(1+x) \sim x$ si $x \to 0$. Donc

$$p(T_n \leq \tau) \approx 1 - e^{-\frac{n(n-1)}{4N}\tau}$$

Si on compte le temps en unités de $4N$ générations ($4Nt = \tau$), on obtient :

$$p(T_n \leq t) \approx 1 - e^{-n(n-1)t}$$

Distributions des temps de coalescence

Le temps est compté en unités de $4N$ générations.

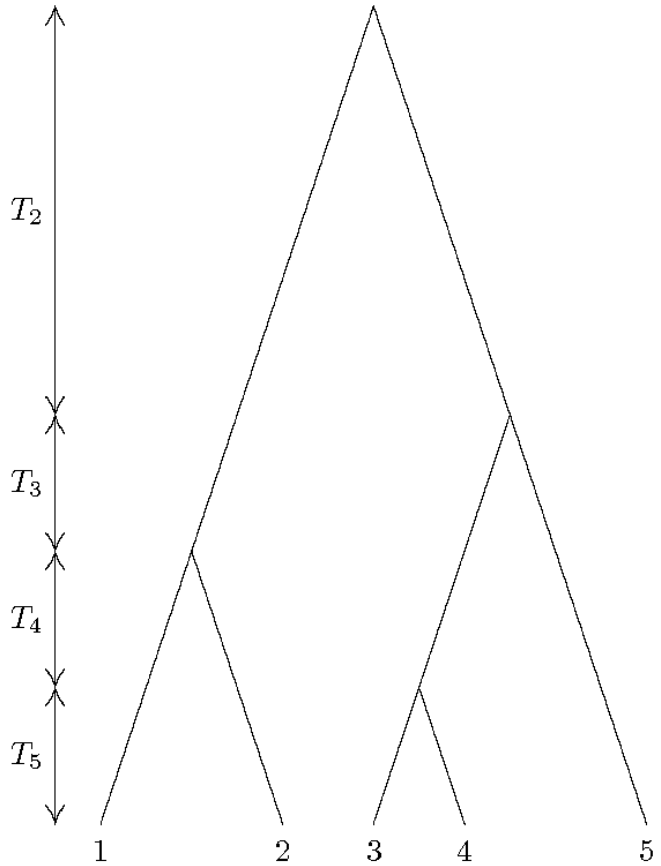
$$p(T_n \leq t) = 1 - e^{-n(n-1)t}$$

$\Rightarrow$ T_n suit une loi exponentielle de paramètre $\frac{1}{n(n-1)}$.

$$E(T_n) = \frac{1}{n(n-1)}, \qquad V(T_n) = \frac{1}{(n(n-1))^2}$$

Topologie de l'arbre de coalescence

La topologie du coalescent est aléatoire.



Plan détaillé

But : Mesurer le polymorphisme

- Le paramètre $\theta = 4N_e\mu$ joue un rôle prépondérant dans l'évolution des séquences.
- Comment estimer θ à partir du polymorphisme ?
- Quel est l'intervale de temps concerné par ces estimations ?

Hauteur moyenne de l'arbre de coalescence

La hauteur du coalescent est $H_n = \sum_{k=2}^{n} T_k$.

$$
\begin{aligned}
E(H_n) &= \sum_{k=2}^{n} \frac{1}{k(k-1)} = \sum_{k=2}^{n} \frac{1-k+k}{k(k-1)} \\
&= \sum_{k=2}^{n} \frac{1}{k-1} - \sum_{k=2}^{n} \frac{1}{k} = \sum_{k=1}^{n-1} \frac{1}{k} - \sum_{k=2}^{n} \frac{1}{k} \\
&= 1 - \frac{1}{n}
\end{aligned}
$$

$\Rightarrow$ Le temps de coalescence moyen d'une population entière est $4N$ générations.

Somme des longueurs des branches

La somme des longueurs des branches est $a_n = \sum_{k=2}^{n} kT_k$.

$$\begin{aligned} E(a_n) &= \sum_{k=2}^{n} \frac{k}{k(k-1)} = \sum_{k=2}^{n} \frac{1}{k-1} \\ &= \sum_{k=1}^{n-1} \frac{1}{k} \end{aligned}$$

$\Rightarrow$ L'espérance du nombre de mutations dans un échantillon de n séquences est $4N\mu E(a_n)$.

$\theta = 4N\mu$

Un estimateur de $4N\mu$ basé sur un échantillon de n séquences orthologues est

$$\hat{\theta}_w = \frac{S}{\sum_{k=1}^{n-1} \frac{1}{k}}$$

où S est le nombre de sites polymorphes (de mutations) dans l'échantillon de n séquences (Watterson, 1975).

Somme des longueurs des branches externes

- Un coalescent de n lignées possède n branches externes.
- Pour obtenir un coalescent de $n+1$ lignées, on prolonge $n-1$ branches externes par des branches de longueur T_{n+1}, la $n$ième branche externe est remplacée par 2 branches externes de longueur T_{n+1}.
- La branche externe enlevée est choisie au hasard parmi les n branches externes.

On note J_n la longueur des branches externes du coalescent.

$\eta = 4N\mu$

$$E(J_2) = 2E(T_2) = 1$$

$$E(J_{n+1}) = E(J_n) - \frac{E(J_n)}{n} + (n+1)E(T_{n+1}) = E(J_n) - \frac{E(J_n)}{n} + \frac{1}{n}$$

Par récurrence, on obtient $E(J_n) = 1$

$\Rightarrow$ Un estimateur de $4N\mu$ est

$$\hat{\theta}_\eta = \eta.$$

où η est le nombre de mutations externes (Fu et Li, 1993).

Table des matières

Qu'est-ce qu'un test de neutralité ?

Neutralité = Conformité à un modèle nul (ex : modèle de WF)

Un test de neutralité détecte des effets

- Sélectifs (Sélection positive, balancée...)
- Démographiques (Expansion de pop, goulots d'étranglement, structure...)

Plan détaillé

- Nous avons vu trois estimateurs de $\theta = 4N\mu$.
- Si la population évolue de façon neutre, on attend

  $\hat{\theta}_w \approx \hat{\theta}_\pi \approx \hat{\theta}_\eta \approx 4N\mu$

D de Tajima (Tajima 1989)

$$D = \frac{\hat{\theta}_\pi - \hat{\theta}_w}{\sqrt{V(\hat{\theta}_\pi - \hat{\theta}_w)}}$$

- $D < 0 \Rightarrow$ excès de mutations à faible fréquence.
- $D > 0 \Rightarrow$ excès de mutations à fréquence intermédiaire.

$\Rightarrow$ Le D de Tajima est utilisé pour détecter des effets démographiques (bottleneck, structuration de population) ou sélectifs (*selective sweep*, *balancing selection*).

D de Fu et Li (Fu et Li, 1993)

$$D = \frac{\hat{\theta}_w - \hat{\theta}_\eta}{\sqrt{V(\hat{\theta}_w - \hat{\theta}_\eta)}}$$

- $D < 0 \Rightarrow$ excès de mutations externes.
- $D > 0 \Rightarrow$ déficit de mutations externes.

$\Rightarrow$ Les applications du D de Fu et Li sont proches de celles du D de Tajima, mais la puissance de ces tests est variable.

Plan détaillé

Le test McDonald et Kreitman (1991)

- Contraste le polymorphisme et la divergence
- Compare les mutations silencieuses et non silencieuses à un locus donné (cf $\frac{K_A}{K_S}$)

Si le patron de substitution (polymorphisme) diffère du patron de fixation (divergence), alors on suspecte un effet de la sélection positive.

Le test McDonald et Kreitman

TABLE 2 Number of replacement and synonymous substitutions for fixed differences between species and polymorphisms within species

	Fixed	Polymorphic
Replacement	7	2
Synonymous	17	42

La p-value du test exact de Fisher est $p = 0.007327$.

$\Rightarrow$ il y a un excès de fixations de mutations non synonymes au locus *Adh*.

Le test HKA (Hudson, Kreitman et Aguadé, 1987)

- Données de polymorphisme et divergence chez une ou deux espèces
- L locus
- Applicable au polymorphisme/divergence silencieuse seulement

Le test HKA compare le rapport du polymorphisme $\theta = 4N\mu$ et de la divergence $D = 2\tau\mu$. Ce rapport doit être indépendant du taux de mutation et donc du locus.

Principe du test HKA

Les paramètres du modèle

- f ratio des effectifs des deux espèces (nécessite des données de polymorphisme dans deux espèces)
- T temps de spéciation
- $\theta_1, \ldots, \theta_L$ taux de mutation aux L locus

Les $L+2$ paramètres sont estimés à partir de 3L données.

Un test de type χ^2 est appliqué avec $2L-2$ degrés de liberté.

Estimations des paramètres du modèle HKA

$$
\begin{cases}
\displaystyle\sum_{i=1}^{L} S_i^A = \sum_{i=1}^{L} m_i^A \, l_i \, C(n_i^A) \, \hat{\theta}_i \\
\displaystyle\sum_{i=1}^{L} S_i^B = \hat{f} \sum_{i=1}^{L} m_i^B \, l_i \, C(n_i^B) \, \hat{\theta}_i \\
\displaystyle\sum_{i=1}^{L} D_i = \sum_{i=1}^{L} m_i^{AB} \, \hat{\theta}_i \left(\hat{T} + l_i \frac{1+\hat{f}}{2} \right) \\
S_i^A + S_i^B + D_i = \hat{\theta}_i \left[m_i^A \, l_i \, C(n_i^A) \right. \\
\left. + m_i^B \, l_i \, C(n_i^B) \, \hat{f} + m_i^{AB} \left(\hat{T} + l_i \, \frac{1+\hat{f}}{2} \right) \right], \quad i = 1, \ldots, L-1
\end{cases}
$$

Le test HKA

	5′ Flanking			*Adh* locus		
	Length	No. sites compared[a]	No. sites variable	Length[b]	No. sites compared[a]	No. sites variable
Within species ($n = 81$)[c]	4000	414	9	900	79	8
Between species[d]	4052	4052	210	900	324	18

La p-value du test χ^2 est $p = 0.016$.

$\Rightarrow$ Le ratio polymorphisme/divergence diffère entre les locus.

Plan détaillé

Les autres tests de neutralité

D'autres tests de neutralité ont été proposés pour mesurer l'écart aux prédictions neutres du coalescent.

- Tests haplotypiques
  - Nombre d'haplotyopes K, diversité haplotypique H
  - Fréquence de l'haplotype majeur
- Test basé sur le déséquilibre de liaison Z_{nS}
- ...

La puissance des tests pour détecter un écart à la neutralité est variable et dépend du type d'écart.

Puissance des tests

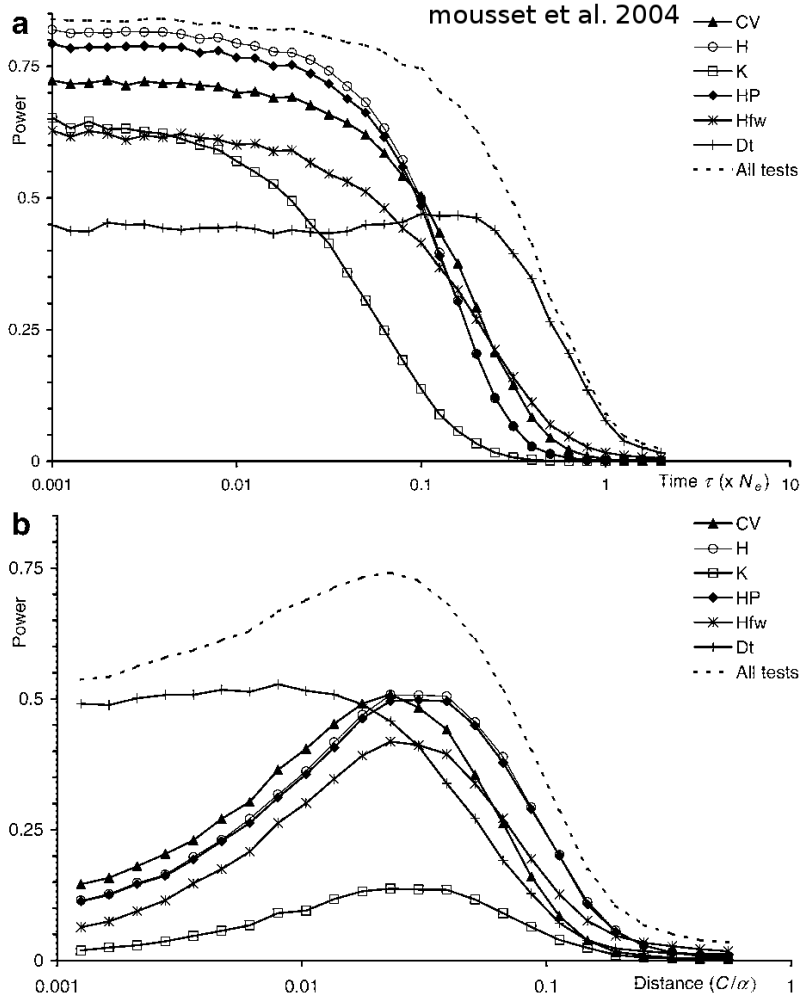


Table des matières

Conclusions

Le polymorphisme de séquence

- Le polymorphisme est majoritairement constitué de mutations neutres
- Le polymorphisme est transitoire, chaque allèle sera fixé ou perdu
- Le coalescent est un outil qui permet de prédire les patrons de polymorphisme
- On s'intéresse aux $4N_e$ dernières générations

Conclusions

Le coalescent

- Un outil mathématique permettant des dérivations analytiques.
- Un outil de simulations.
- Nécessité de préciser le modèle neutre.

Conclusions

Les inférences biologiques des tests de neutralité

- Plusieurs sources d'écart au modèle neutre.
- Écart génomique $\Leftrightarrow$ Effet démographique.
- Écart local $\Leftrightarrow$ Effet sélectif local.
- Puissances variables des tests de neutralité.