

# Utilisation des microsatellites pour inférer l'histoire démographique des populations

Les fréquences alléliques dans un échantillon de gènes prélevés dans une population résultent :

- des processus évolutifs du locus considéré (comment mute le locus, à quelle fréquence ?)
- de l'histoire/généalogie des gènes de l'échantillon considéré (dépend de l'histoire démographique de la population)

Cette dernière notion pas évidente a priori est apparue plus clairement avec le développement d'une théorie maintenant largement utilisée en génétique des populations : *la théorie de la coalescence*

Intérêt des microsatellites pour ces études :

- marqueurs co-dominants → connaissance de l'état allélique
- fort taux de mutation → nombreux allèles → phylogénie plus documentée

Progression historique :

- utilisation de statistiques résumant l'information ( $k$ ,  $H$ ,  $V$ )
- utilisation des données de base à chaque locus

# THEORIE DE LA COALESCENCE

Pionniers : KINGMAN J.F.C. (1982)  
GRIFFITHS R.C.  
TAVARÉ S.  
HUDSON R.R.

## GENETIQUE DES POPULATIONS THEORIQUE

Approche classique :

- Niveau d'étude : POPULATION (échantillon)
- Caractérisation par des fréquences : alléliques et génotypiques
- Etude des distributions de ces fréquences dans le temps et dans l'espace en incorporant des facteurs multiples (taille efficace, mutations, migrations, etc...)

Approche "coalescence" :

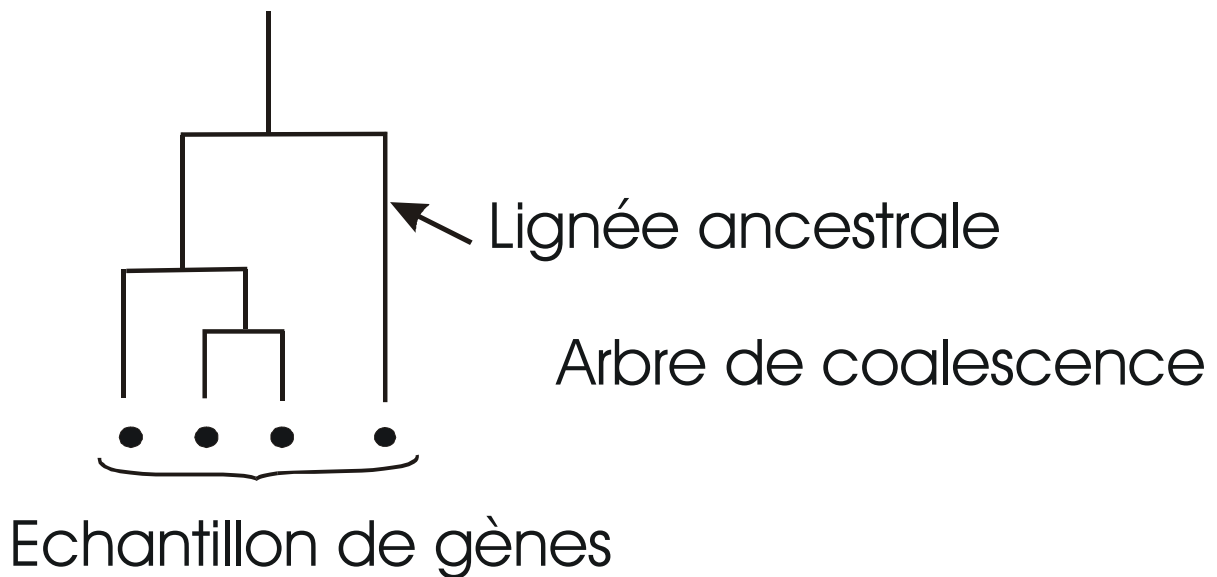
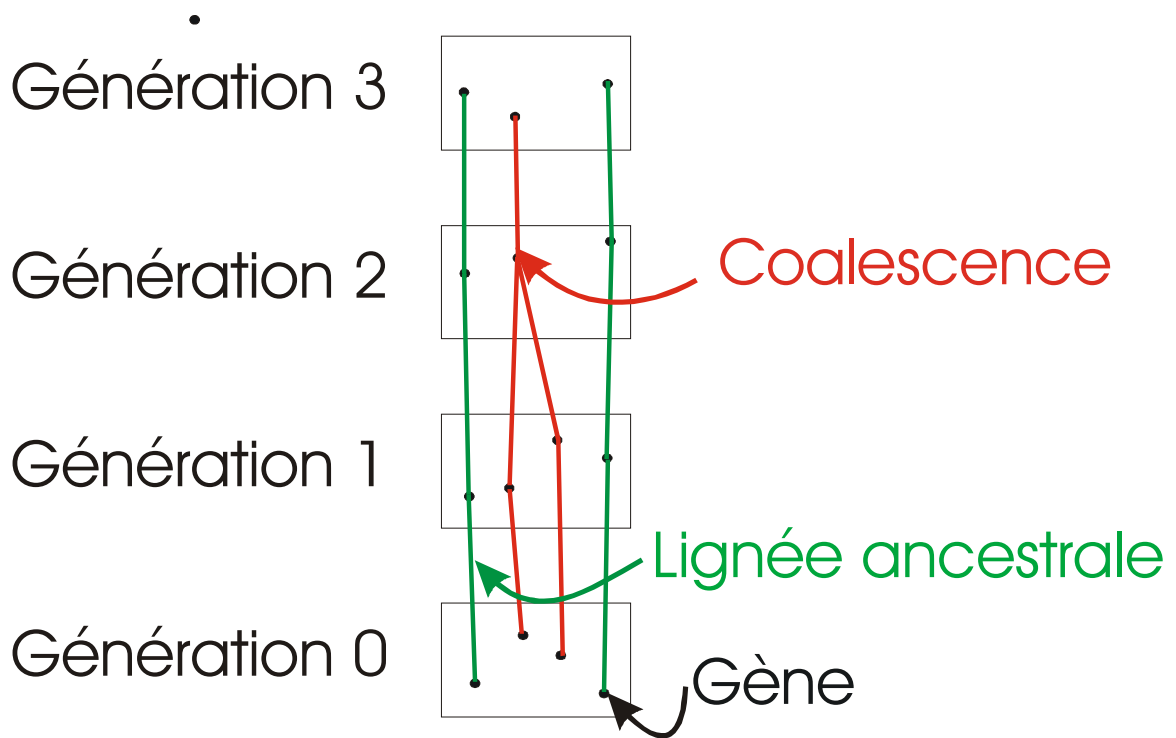
- Niveau d'étude : ECHANTILLON (population)
- Caractérisation du passé des gènes échantillonnés
- Prise en compte des relations phylogénétiques entre allèles

## Modèle de Wright-Fisher

- Générations séparées
- Les gènes présents à la génération  $g$  sont issus d'un tirage avec remise des gènes présents à la génération  $g-1$

## Définition de la coalescence

- On considère un *échantillon de gènes* prélevés à une génération donnée.
- On considère les *lignées ancestrales* de ces gènes
- Deux gènes observés à la génération  $g$  peuvent être la copie du même gène de la génération  $g-1$ . Dans ce cas, il y a *coalescence des lignées ancestrales* de ces deux gènes.
- Ce processus de coalescence des lignées ancestrales des gènes de l'échantillon se poursuit dans le passé jusqu'à l'*ancêtre commun le plus récent* de ces gènes. La généalogie de ces gènes se traduit donc par un arbre de coalescence.
- La théorie de la coalescence décrit ce processus de coalescence



## Longueur des branches de l'arbre de coalescence

HYPOTHESES :     - Organisme diploïde  
                      - Taille efficace constante ( $N_e$ )

- Cas de deux lignées ancestrales:

Proba (2 gènes soient la copie du même gène à la génération précédente) = Proba(2 lignées coalescent à la génération 1) =

$$1/2N_e$$

Proba (2 lignées coalescent  $k+1$  générations en arrière) = Proba (ne coalescent pas à 1) x Proba(ne coalescent pas à 2) x ...x Proba(ne coalescent pas à  $k$ ) x Proba (coalescent à  $k+1$ ) =

$$(1 - 1/2N_e)^k (1/2N_e) = (1/2N_e)e^{k \ln(1-1/2N_e)} \approx (1/2N_e)e^{-k/2N_e}$$

Le temps de coalescence de deux lignées ancestrales suit une loi de distribution exponentielle d'espérance  $2N_e$

## Longueur des branches de l'arbre de coalescence

- Cas de  $j$  lignées ancestrales:

Proba(2 gènes pris parmi  $j$  gènes soient la copie du même gène à la génération précédente) = Proba(2 lignées parmi  $J$  lignées coalescent à la génération 1) =

$$[j(j-1)/2]/2N_e = j(j-1)/4N_e$$

D'où Proba( $j$  lignées coalescent à  $k+1$ )  $\approx (j(j-1)/4N_e)e^{-kj(j-1)/4N_e}$

Le temps de la première coalescence dans  $j$  lignées ancestrales suit une loi de distribution exponentielle d'espérance  $4N_e/j(j-1)$

La longueur totale moyenne de l'arbre est égale à :

$$\begin{aligned} nT_n + (n-1)T_{n-1} + \dots + 2T_2 \\ = 4N_e \sum_{i=1}^{n-1} \frac{1}{i} \end{aligned}$$

## Age de l'ancêtre commun le plus récent

L'âge de l'ancêtre commun le plus récent de la généalogie est égal à la somme des temps entre coalescences successives

Dans un échantillon de  $n$  gènes, il y a  $n-1$  coalescences dont les durées sont égales à  $T_n, T_{n-1}, \dots, T_3$  et  $T_2$ .

Sachant que :  $E(T_j) = 4N_e/j(j-1)$

L'âge moyen du MRCA (*Most Recent Common Ancestor*) est donc égal à :

$$\sum_{j=2}^n \frac{4N_e}{j(j-1)} = 4N_e \left(1 - \frac{1}{n}\right)$$

## Mutations

HYPOTHESE : Taux de mutation  $\mu$  par gène par génération

Dans le modèle de Wright-Fisher, le nombre attendu de mutations par génération est égal à  $2N_e\mu$ .

Le nombre de mutations le long d'une branche de l'arbre de coalescence de longueur  $L$  suit une loi de Poisson de paramètre  $\mu L$ .

La probabilité d'observer  $m$  mutations est égale à :

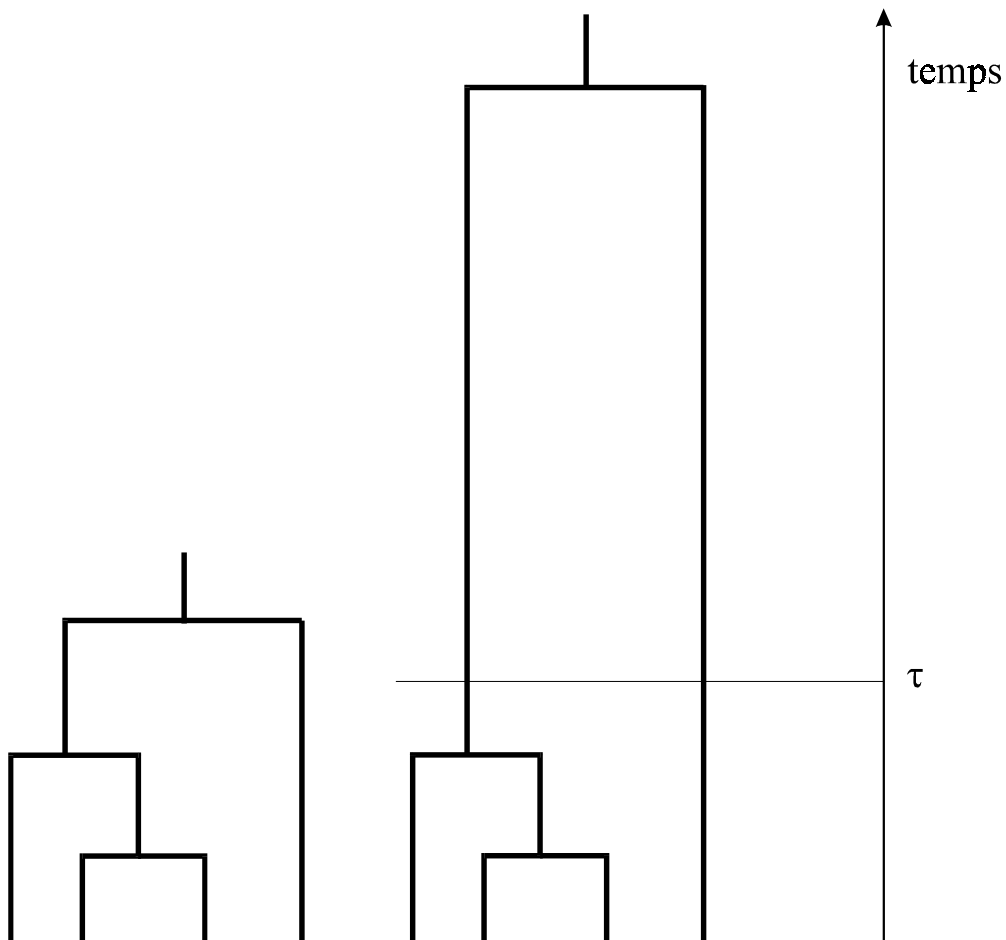
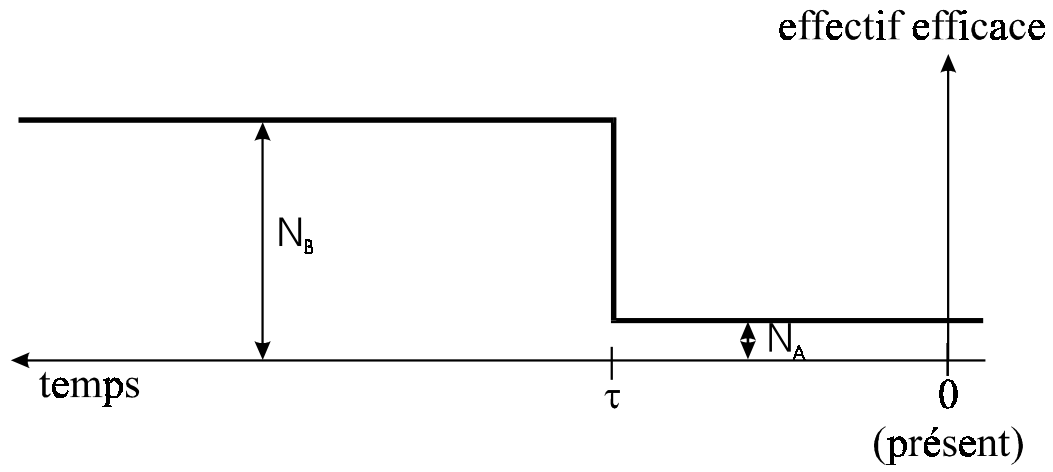
$$\frac{e^{-\mu L} (-\mu L)^m}{m!}$$

Le nombre total de mutations attendu sur l'arbre complet est obtenu en remplaçant  $L$  par la longueur totale de l'arbre. L'espérance du nombre total de mutations vaut donc :

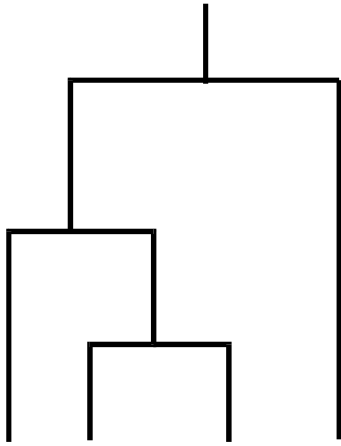
$$4N_e\mu \sum_{i=1}^{n-1} \frac{1}{i} = \theta \sum_{i=1}^{n-1} \frac{1}{i}$$

en utilisant le paramètre classique  $\theta = 4N_e\mu$

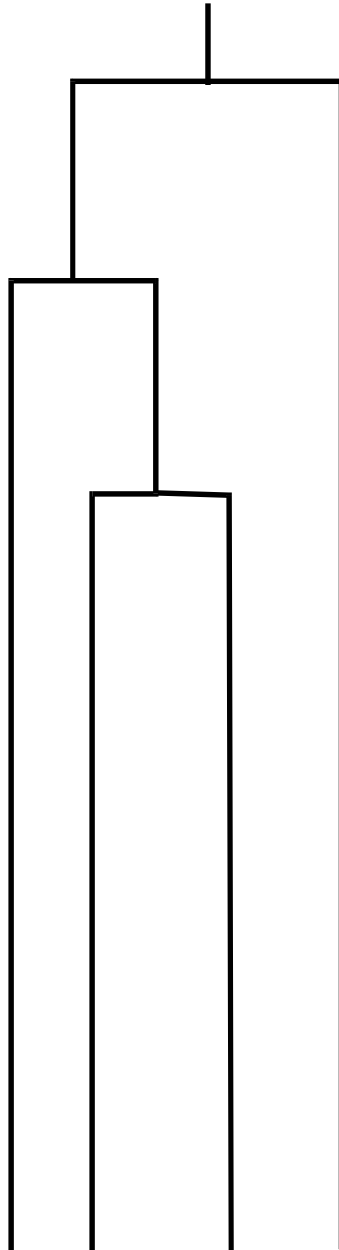
# Variation de l'effectif efficace



Effectif constant



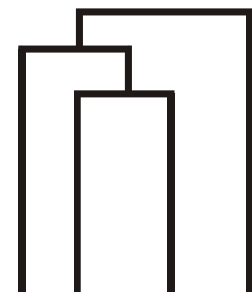
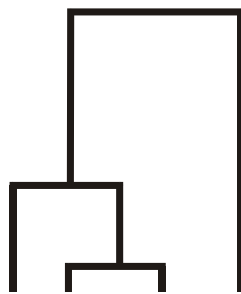
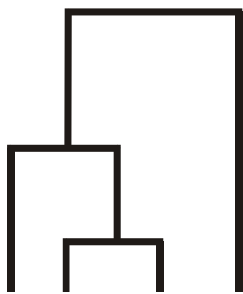
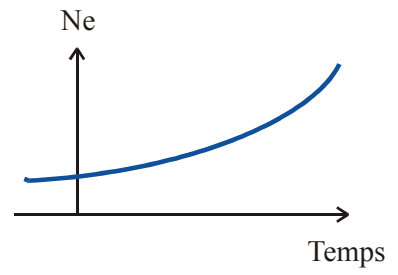
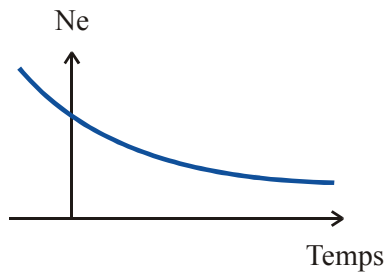
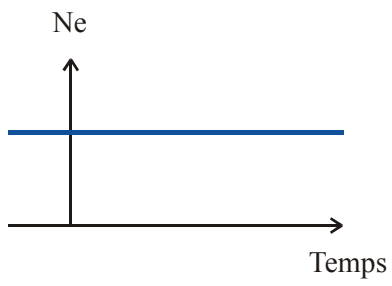
Effectif croissant



Augmentation de l'effectif :

- Allongement des branches terminales
- Mutations concentrées dans les branches terminales
- Allèles rares plus nombreux

## Variation progressive de l'effectif



Par rapport à un effectif stable dans le temps,  
la croissance ou la décroissance de l'effectif  
engendre une modification de la distribution  
des fréquences alléliques dans un échantillon

Pour estimer la tendance démographique récente d'une population,  
différentes approches sont possibles :

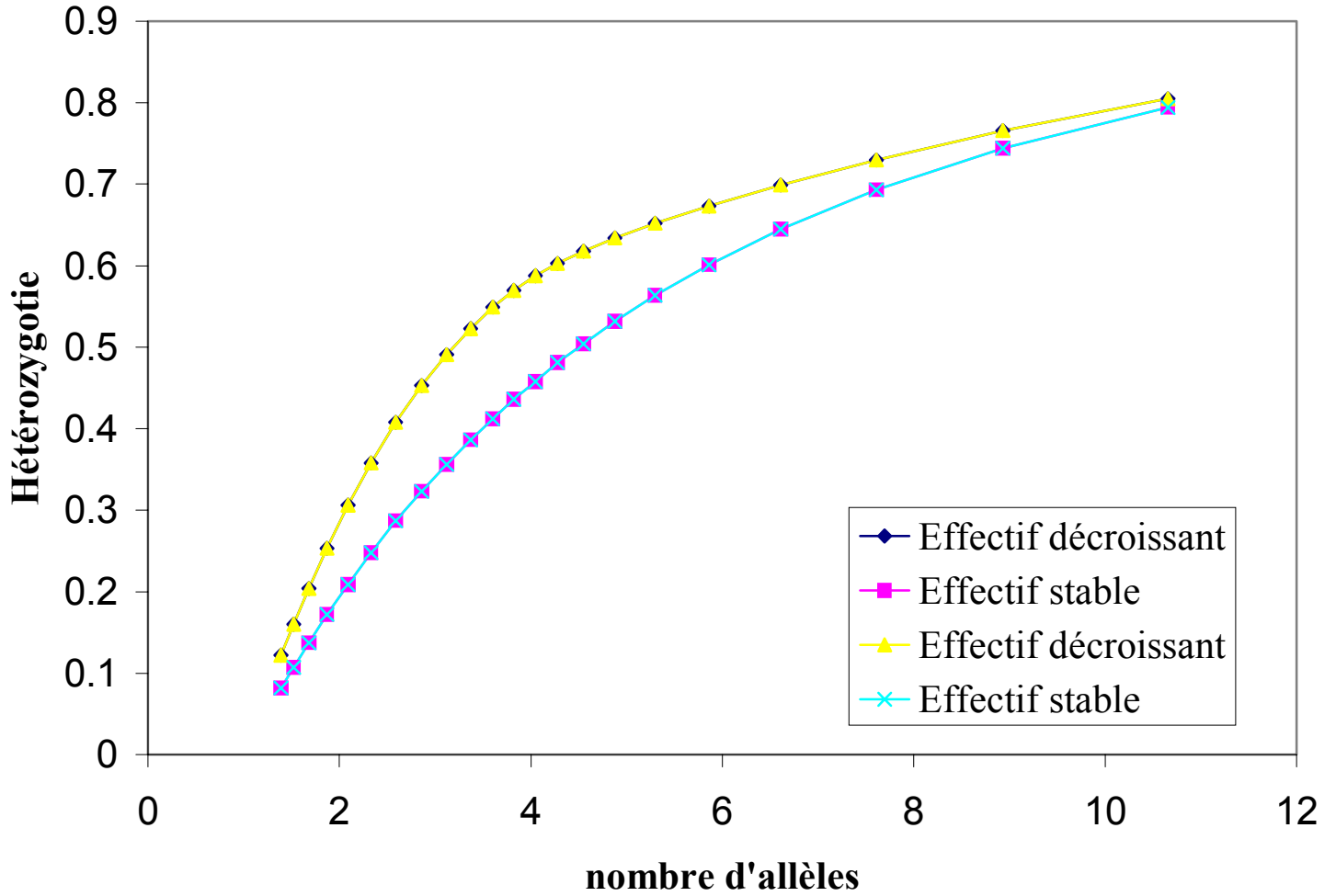
- Comparaison entre diversités géniques observée et attendue
- Comparaison de la probabilité d'obtenir l'échantillon observé sous  
diverses hypothèses d'histoire démographique
- Approche bayésienne

## Comparaison entre diversités géniques observée et attendue

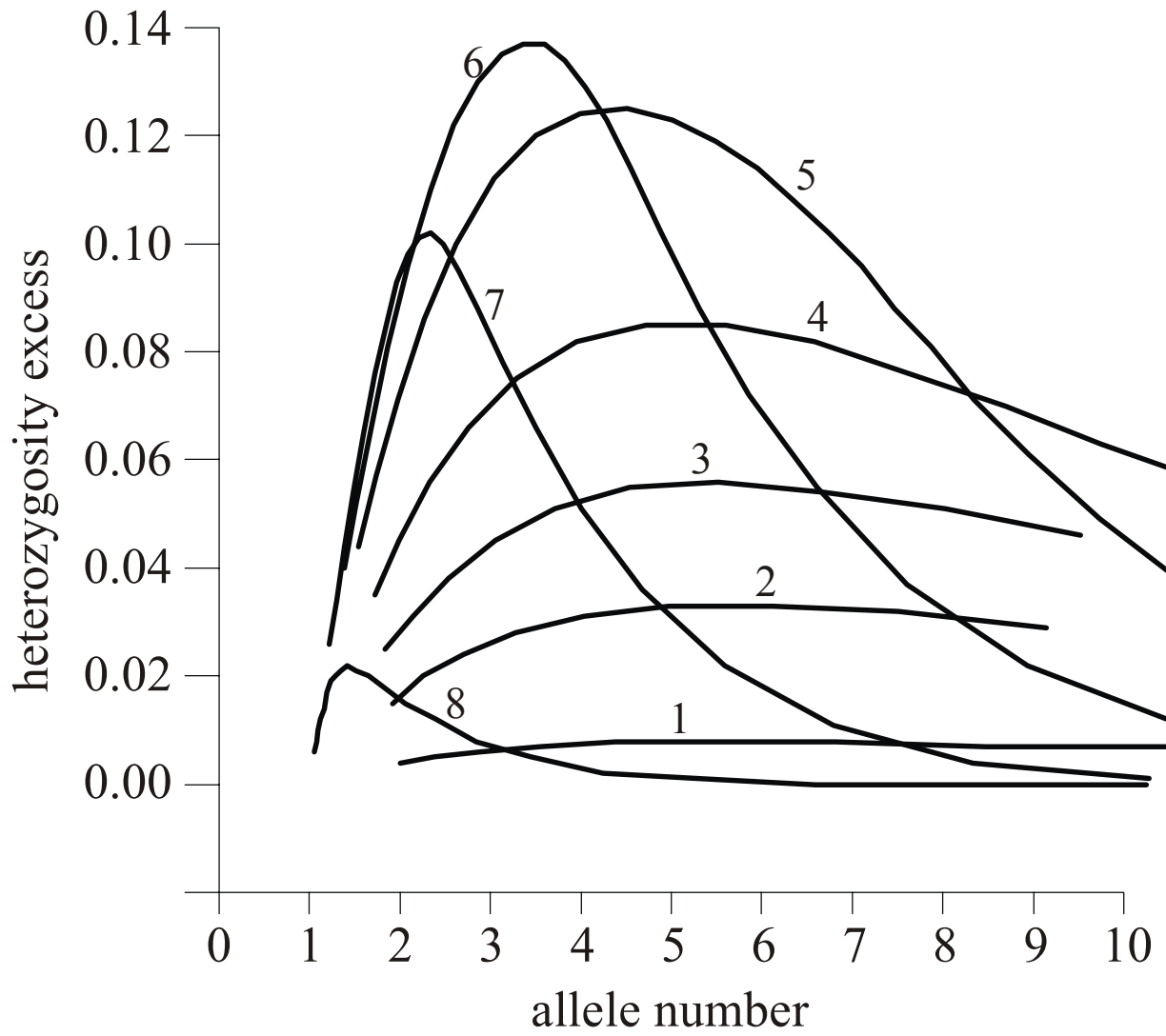
- La principale différence dans les distributions de fréquences alléliques entre populations d'effectif stable et d'effectif croissant/décroissant concerne la proportion d'allèles rares.
- Les allèles rares sont aussi importants que les allèles fréquents dans le paramètre "nombre d'allèles" mais leur rôle est mineur dans le paramètre "diversité génique" ( $H = 1 - \sum_i p_i^2$ ).
- Il en résulte qu'une variation d'effectif efficace modifie plus rapidement le nombre d'allèles que la diversité génique.
- Si la population a eu une histoire démographique stable, pour un modèle mutationnel donné, on peut calculer la diversité génique ( $H_e$ ) à partir du nombre d'allèles et de la taille de l'échantillon.
- Si la population a subi une réduction démographique, elle a perdu un nombre important d'allèles rares et la diversité génique observée sera plus élevée que la diversité génique attendue (sous l'hypothèse d'un effectif stable).

Vérification ...

# Infinite Allele Model

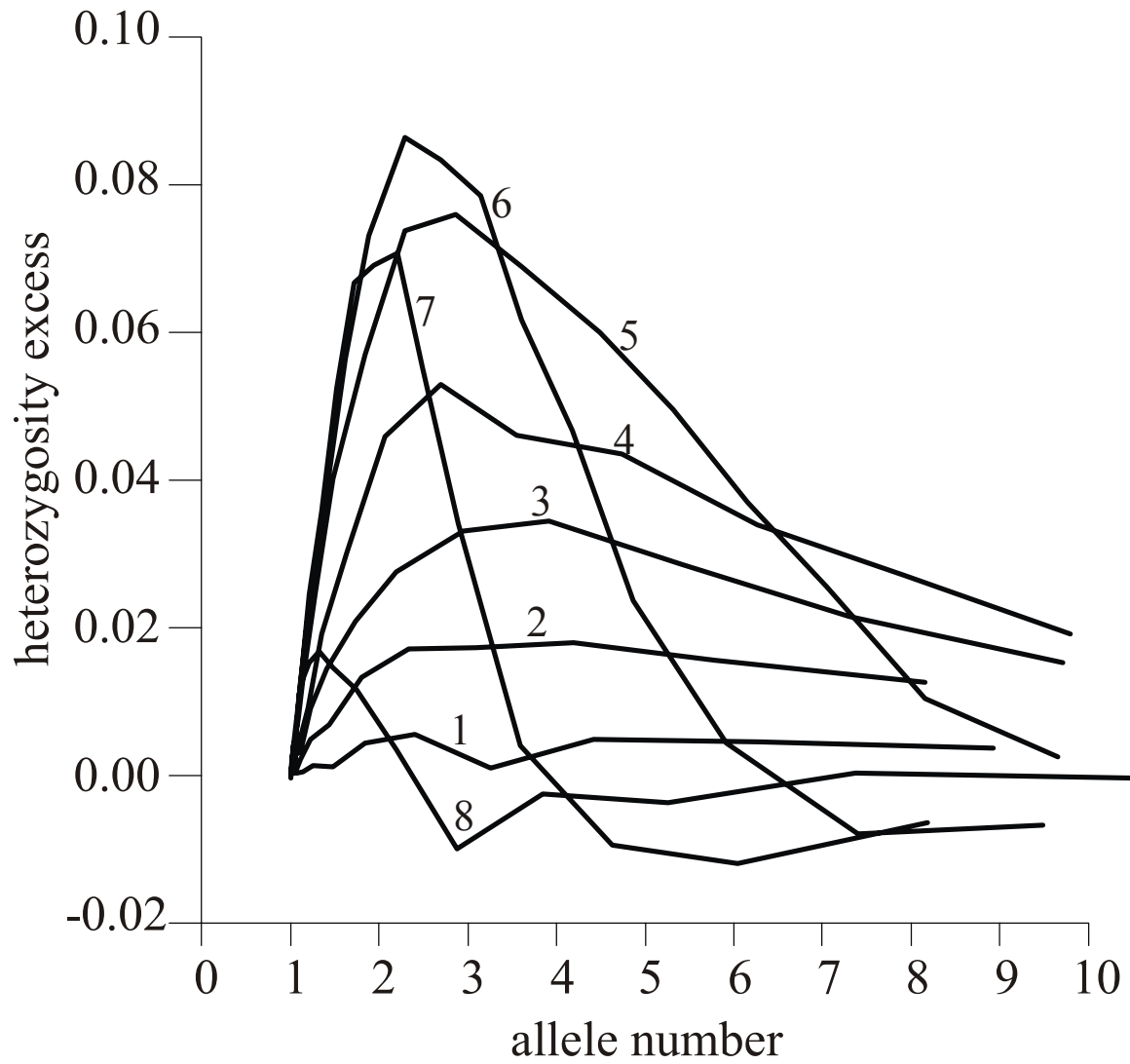


### A. Infinite Allele Model ( $\alpha=100$ )

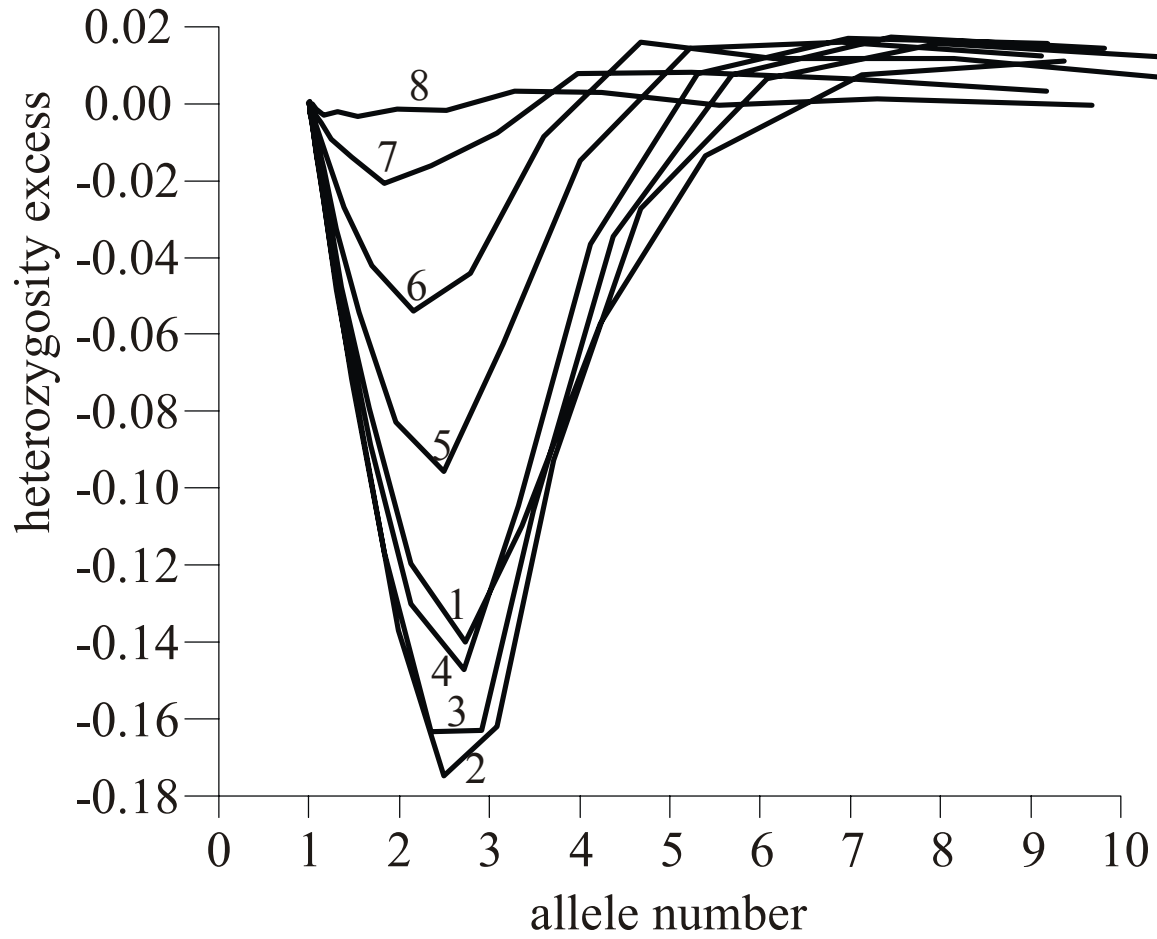




### D. Stepwise Mutation Model (mixed, $\alpha=100$ )



### C. Stepwise Mutation Model (strict, $\alpha=0.01$ )



## Comparaison entre diversités géniques observée et attendue

*Question 1 : distribution de la diversité génique attendue ?*

Elle est obtenue par simulation du processus de coalescence d'un échantillon de  $n$  gènes (taille de l'échantillon).

*Question 1bis : comment tenir compte du fait que la distribution de probabilité de la diversité génique est conditionnée par le nombre  $k$  d'allèles observés dans l'échantillon ?*

La méthode utilisée pour la simulation est spécifique.

*Question 2 : quel modèle mutationnel utiliser ?*

Deux modèles (IAM et SMM) considérés comme donnant des réponses extrêmes sont utilisés conjointement.

*Question 3 : comment procéder pour tester une des hypothèses d'histoire démographique ?*

Trois tests sont disponibles :

- Test du signe
- Test des différences standardisées (au moins 20 à 30 locus)
- Test de Wilcoxon

# SMM

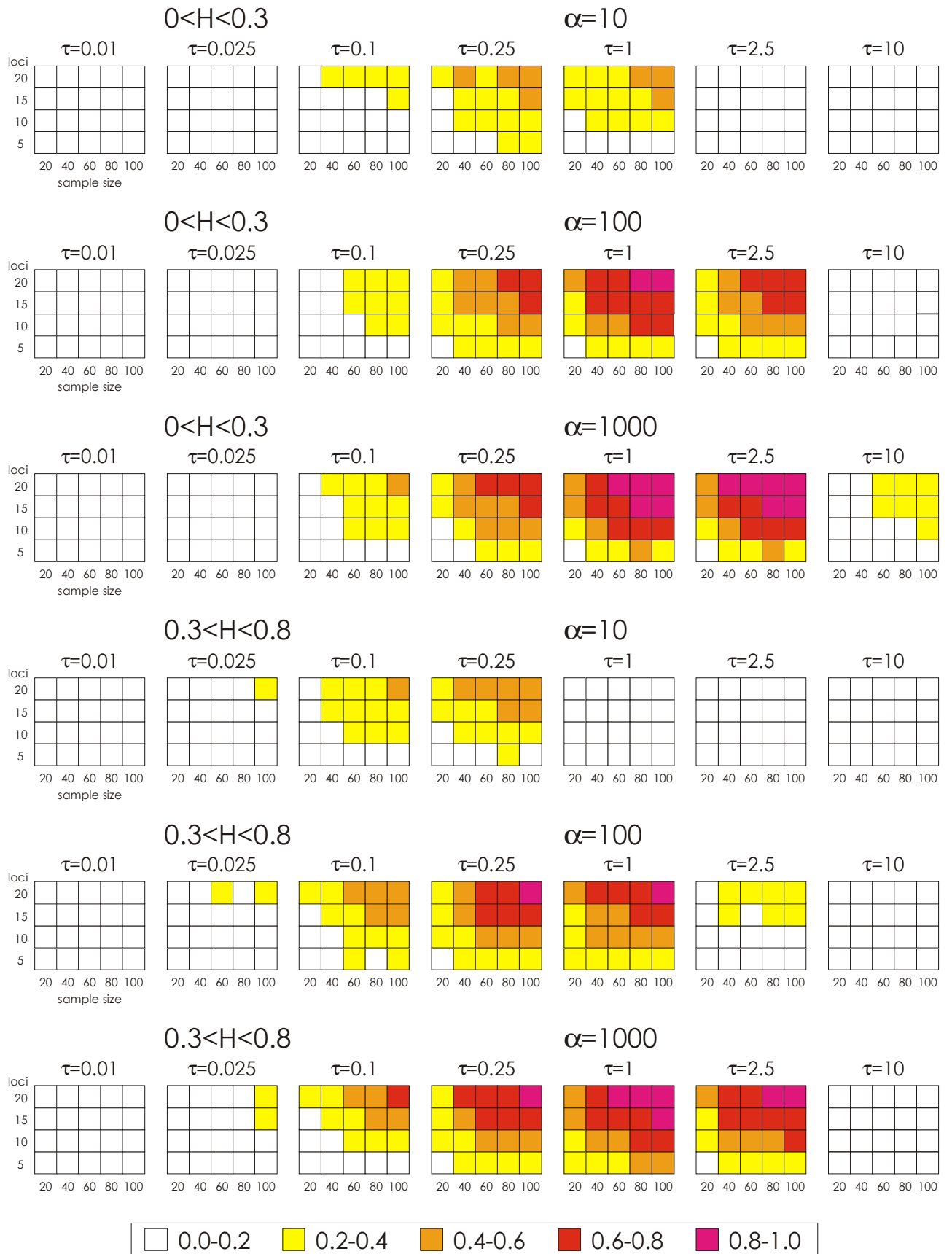


figure 3b

## Calcul de la probabilité d'un échantillon de gènes

Hypothèses :

- Modèle de Wright-Fisher (organismes diploïdes)
- Effectif efficace stable
- Locus à  $d$  états alléliques possibles
- $p_{ij}$  : probabilité qu'une mutation change un allèle  $i$  en allèle  $j$  ( $i, j \in [1 \dots d]$ )

Notations :

- $N_e$  : effectif efficace,  $\mu$  : taux de mutations,  $\theta = 4N_e\mu$
- $\mathbf{n} = (n_1, \dots, n_d)$  : configuration de l'échantillon ( $n = \sum_{i=1}^d n_i$ )
- $\mathbf{e}_i = (0, \dots, 0, 1, 0, \dots, 0)$  échantillon réduit à un seul allèle de type  $i$ .
- $q(\mathbf{n})$  probabilité de la configuration  $\mathbf{n}$

Formule de récurrence fondamentale

$$q(\mathbf{n}) = \frac{\theta}{\theta + n - 1} \sum_{i,j \in \{1, \dots, d\}, n_j > 0, i \neq j} \frac{n_i + 1}{n} p_{ij} q(\mathbf{n} + \mathbf{e}_i - \mathbf{e}_j) + \frac{n - 1}{\theta + n - 1} \sum_{j \in \{1, \dots, d\}, n_j > 0} \frac{n_j - 1}{n - 1} q(\mathbf{n} - \mathbf{e}_j)$$

$$\frac{\theta}{\theta + n - 1} = \frac{4N_e\mu}{4N_e\mu + n - 1} = \frac{\mu}{\mu + \frac{n-1}{4N_e}} = \frac{n\mu}{n\mu + \frac{n(n-1)}{(2)(2N_e)}}$$

$n\mu$  : probabilité d'une mutation dans les  $n$  lignées ancestrales de l'échantillon

$\frac{n(n-1)}{(2)(2N_e)}$  : probabilité d'une coalescence

$\frac{\theta}{\theta + n - 1}$  : probabilité pour que le dernier événement (mutation ou coalescence)

qui a affecté l'échantillon ait été une mutation

$\frac{n-1}{\theta + n - 1}$  : probabilité pour que cet événement ait été une coalescence

$\sum_{i,j \in \{1, \dots, d\}, n_j > 0, i \neq j} \frac{n_i + 1}{n} p_{ij} q(\mathbf{n} + \mathbf{e}_i - \mathbf{e}_j)$  considère toutes les configurations possibles qui

par mutation d'un allèle peuvent aboutir à la configuration actuelle et

$\sum_{j \in \{1, \dots, d\}, n_j > 0} \frac{n_j - 1}{n - 1} q(\mathbf{n} - \mathbf{e}_j)$  considère toutes les configurations possibles qui par

coalescence de deux lignées (portant le même allèle) peuvent aboutir à la configuration actuelle

A part pour des valeurs de  $n$  et de  $d$  faibles et des valeurs de  $P$  particulières, l'utilisation directe de cette formule est impraticable.

Griffiths et Tavaré ont proposé un algorithme de calcul fondé sur la simulation et qui s'apparente à de l'*importance sampling*.

La formule de récurrence peut être réécrite sous la forme :

$$q(\mathbf{n}) = f(\mathbf{n}) \left( \sum_{i,j \in \{1, \dots, d\}, n_j > 0, i \neq j} \lambda_{ij}(\mathbf{n}) q(\mathbf{n} + \mathbf{e}_i - \mathbf{e}_j) + \sum_{j \in \{1, \dots, d\}, n_j > 0} \mu_j(\mathbf{n}) q(\mathbf{n} - \mathbf{e}_j) \right)$$

En partant d'une configuration  $\mathbf{n}$ , il est possible de simuler un arbre de coalescence complet (jusqu'à l'ancêtre commun le plus récent).

Si on simule cet arbre en prenant

- $\lambda_{ij}$  comme probabilité d'une mutation d'un allèle d'état  $i$  vers un allèle d'état  $j$
- $\mu_j$  comme probabilité d'une coalescence de deux lignées ancestrales d'état  $j$

Griffiths et Tavaré ont montré que :

$$q(\mathbf{n}) = E_n \prod_{j=0}^{\tau} f(\mathbf{N}(j))$$

Pour calculer  $q(\mathbf{n})$ , il suffit de simuler un grand nombre d'arbre selon les règles ci-dessus et de prendre la moyenne des produits des coefficients  $f(\cdot)$ .

Sous réserve que les locus soient indépendants, il est possible de calculer la probabilité d'un échantillon multilocus (produit des probabilités pour chaque locus).

## Approche fondée sur le calcul de la probabilité d'un échantillon

Comparaison de modèles démographiques :

- effectif stable (référence)
- effectif ayant cru/décru de façon abrupte : deux paramètres  $\alpha$  et  $\tau$   
( $\alpha = N_{\text{actuel}}/N_{\text{passé}}$  et  $\tau =$  génération du changement d'effectif)
- autres types d'évolution démographique (linéaire, exponentielle, ...)

sur la base des probabilités de la configuration de l'échantillon sous l'un ou l'autre modèle.

Recherche des valeurs des paramètres qui maximisent la probabilité de l'échantillon (approche du maximum de vraisemblance).

Limites de l'approche :

- celles du modèle de Wright-Fisher
- calcul pour des valeurs ponctuelles des paramètres
- éventuelle lourdeur des calculs

## Approche bayésienne (principe)

Les **données**, quantités observables, résultent d'un processus. Ce processus est décrit par un **modèle** qui fait intervenir un ou plusieurs **paramètres** qui ne sont pas eux-mêmes directement observables.

En statistique bayésienne, les conclusions sur les paramètres sont effectuées en terme de distribution de probabilité. Ces distributions de probabilité sont conditionnées par les valeurs observées.

### Notations

$\theta$  : le ou les paramètres du modèle.

$y$  : les données

Le **modèle** fournit la *distribution de probabilité conjointe*  $p(\theta, y)$  qui peut s'écrire comme le produit d'une *distribution a priori*  $p(\theta)$  et d'une *distribution d'échantillonnage*  $p(y|\theta)$  :

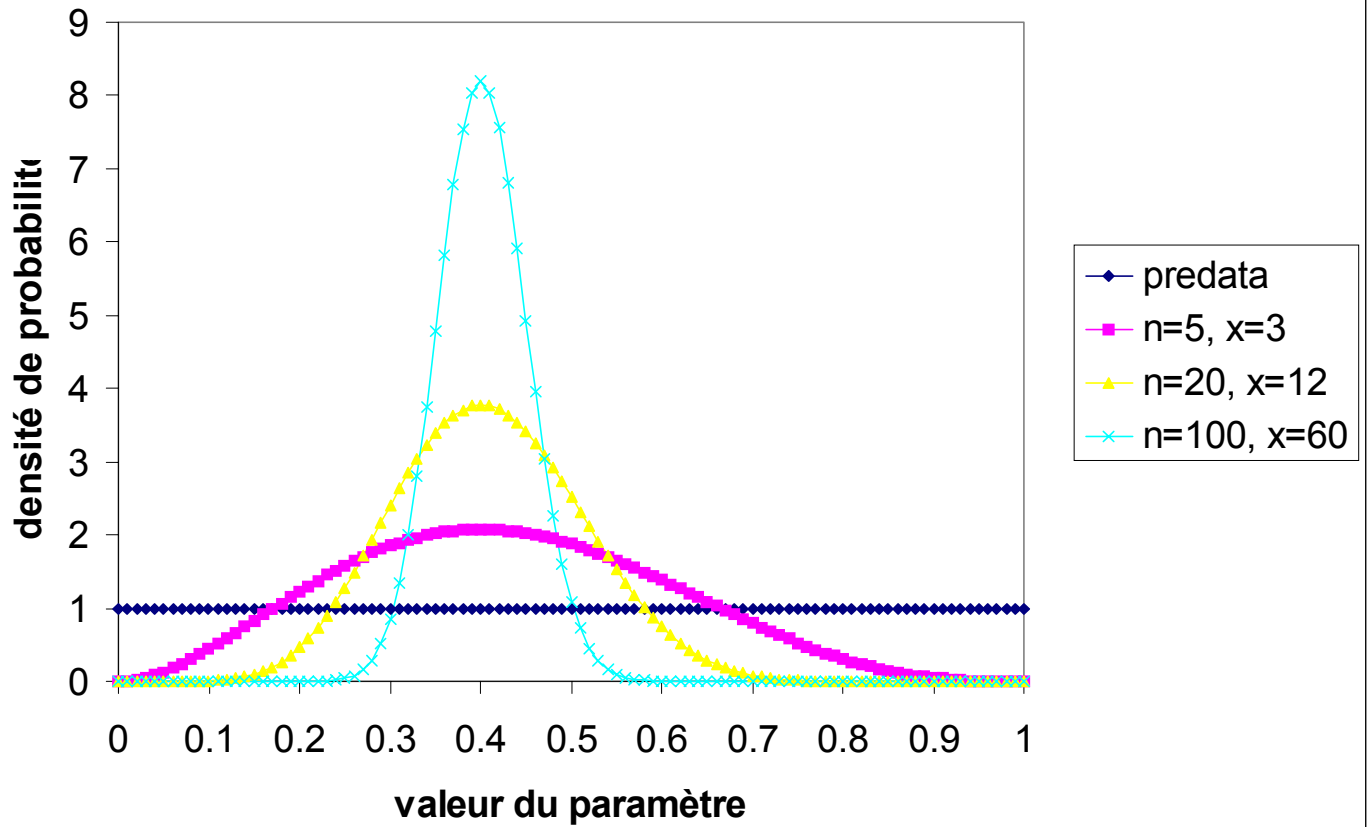
$$p(\theta, y) = p(\theta)p(y|\theta)$$

En conditionnant sur les données observées  $y$ , et en utilisant la propriété classique de la probabilité conditionnelle (règle de Bayes), on obtient la *densité a posteriori* :

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)} \propto p(\theta)p(y|\theta)$$

Pour un modèle donné, l'approche bayésienne consiste à partir d'une "idée" *a priori* sur les paramètres du modèle (formalisée par une densité de probabilité) et à la préciser en fonction des données observées.

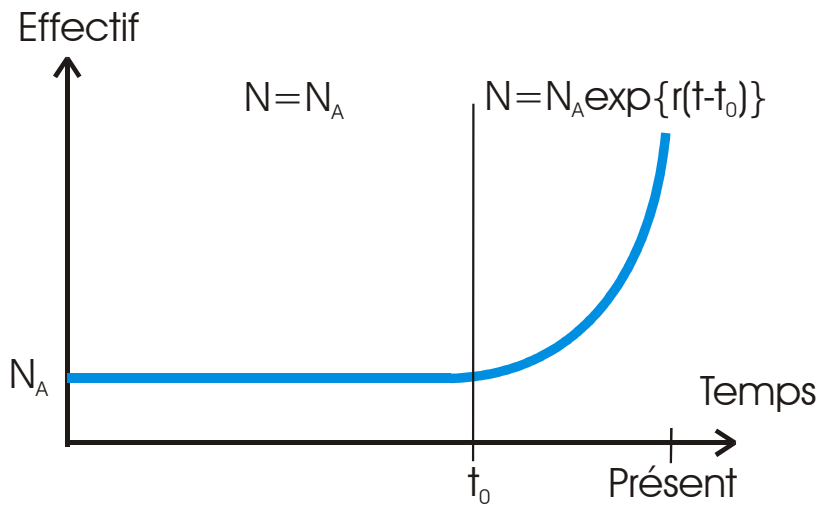
## Estimation du paramètre d'une loi binomiale



## Approche bayésienne (exemple)

*Pritchard J.K., Seielstad M.T., Perez-Lezaun A., Feldman M.W. 1999  
Population growth of human Y chromosomes: a study of Y chromosome  
microsatellites. Mol. Biol. Evol. 16(12):1791-1798*

### **Modèle démographique :**



Paramètres :  $N_A$ ,  $r$ ,  $t_0$

### **Modèle mutationnel :**

- $\mu$  constant, strict SMM
- $\mu$  constant, "SMM géométrique"
- $\mu$  variable, " SMM géométrique"
- $\mu$  constant, strict SMM, contraintes de taille

Paramètres :  $\mu$  (taux de mutation),  $\sigma$  (écart-type de la loi géométrique)

## Définition de densité *a priori* sur les paramètres

Exemple pour  $N_A$  : LogNormale( $m=8,5$ ,  $V=2$ ) ce qui correspond à une moyenne autour de 36000 et environ 95% des valeurs comprises entre 100 et 231000.

## Données disponibles :

445 chromosomes Y prélevés sur différents groupes régionaux

Chaque chromosome est typé à 8 locus microsatellites (2 tri et 6 tétra)

Information utilisée :

- taille de l'échantillon ( $n=445$ )
- nombre d'haplotypes ( $k=316$ )
- moyenne des variance du nombre de répétition par locus ( $V=1.149$ )
- moyenne de la diversité génique ( $H=0.6358$ )

## Estimation des densités a posteriori

Utilisation d'un algorithme d'*acceptation - rejet*

1. Simulation indépendante de chaque paramètre selon sa densité *a priori*
2. Simulation d'un échantillon de  $n$  chromosomes par simulation du processus de coalescence en utilisant les valeurs de paramètres obtenues en 1
3. Calcul de  $k^*$ ,  $V^*$  et  $H^*$
4. Si les écarts relatifs  $|k-k^*|/k$ ,  $|V-V^*|/V$  et  $|H-H^*|/H$  sont tous inférieurs à une valeur donnée  $\delta$ , enregistrer les valeurs des paramètres
5. Retourner en 1

Cette procédure fournit un échantillon tiré de la distribution a posteriori des paramètres, conditionné par le fait que  $(k,V,H)$  est à au plus  $\delta$  des valeurs observées. En faisant tendre  $\delta$  vers 0, cela correspondra à conditionner par les valeurs observées, mais l'algorithme devient alors de moins en moins efficace. Valeur utilisée par les auteurs  $\delta = 0.1$  (taux d'acceptation de l'ordre de  $10^{-3}$ ).

## Résultats :

Modèle	NA	r	$t_0$
Strict SMM	1500 [100-4900]	0.008 [0.002-0.021]	18000 [6000-43000]
Géométrique	1000 [50-3500]	0.008 [0.002-0.022]	18000 [7000-41000]
Contrainte taille	2000 [200-6500]	0.008 [0.002-0.022]	17000 [6000-43000]
Pre-Data	36000 [100-231000]	0.005 [0.000-0.019]	20000 [600-75000]

### Test du modèle à démographie constante

Modification de la densité a priori sur le paramètre  $t_0$ :

avec probabilité 0.5,  $t_0 = 0$  (effectif constant)

avec probabilité 0.5,  $t_0$  suit une loi exponentielle ( $m=20000$ )

La valeur  $t_0 = 0$  a été retenue dans moins de 1% des cas d'acceptation des valeurs de paramètres dans l'algorithme d'acceptation – rejet

## Approche bayésienne (suite)

Algorithme d'*acceptation - rejet* : valable pour un nombre faible de "données".

Exemple précédent : information résumée dans  $k$ ,  $V$  et  $H$

Si on veut utiliser davantage d'information, il faut avoir recours à d'autres méthodes de simulation, e.g. méthodes de *simulation par chaînes de Markov*.

### Principe des MCMC (Monte Carlo Markov Chain)

Simulation d'une marche aléatoire dans l'espace des paramètres ( $\theta$ ) qui converge vers une distribution stationnaire de la densité a posteriori  $p(\theta|y)$ .

### Algorithme de Metropolis

- 1) Partir d'une valeur initiale "plausible"  $\theta^0$  du (des) paramètre(s)
- 2) Pour  $t = 1, 2, \dots$ 
  - a) Simuler une valeur  $\theta^*$  à partir d'une distribution *symétrique* ( $J_t(\theta^a|\theta^b) = J_t(\theta^b|\theta^a)$ )
  - b) Calculer le rapport  $r = \frac{p(\theta^*|y)}{p(\theta^{t-1}|y)}$
  - c) Si  $r > 1$ , prendre  $\theta^t = \theta^*$ ,  
Si  $r < 1$ , prendre  $\theta^t = \theta^*$  avec la probabilité  $r$ , sinon prendre  $\theta^t = \theta^{t-1}$

### Algorithme de Metropolis-Hastings

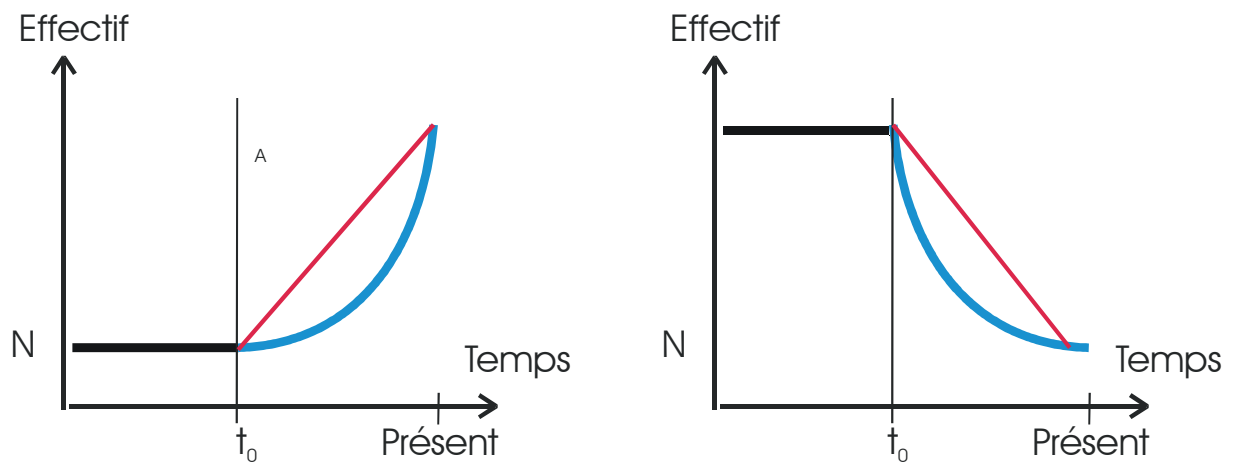
Généralisation de l'algorithme précédent au cas où la *jumping distribution* n'est pas symétrique. Il suffit de modifier la formule du rapport  $r$ .

$$r = \frac{p(\theta^*|y) / J_t(\theta^*|\theta^{t-1})}{p(\theta^{t-1}|y) / J_t(\theta^{t-1}|\theta^*)}$$

## Approche bayésienne (exemple)

*Beaumont M.A. 1999 Detecting population expansion and decline using microsatellites. Genetics 153:2013-2029*

### Modèle démographique



### Modèle mutationnel : strict SMM

Paramètres du modèle:

- $r = N_{(\text{présent})} / N_{(t_0)}$
- $t_0$
- généalogie des gènes échantillonnés
- temps entre événements (coalescence et mutations)
- $\theta (=4N_{(\text{présent})}\mu)$ .

Données : Pour chaque locus, configuration de l'échantillon

## Analyse utilisant l'algorithme de Metropolis-Hastings

Etape 1: construction d'un arbre de coalescence compatible avec les données

Etapes suivantes:

- a) modification d'un ou plusieurs paramètres
- b) calcul du rapport  $r$
- c) acceptation-rejet des nouvelles valeurs des paramètres selon la règle

Les modifications des paramètres portent

- dans 95% des cas sur l'arbre de coalescence (généalogie des gènes)  
exemple: addition de deux mutations de signes opposés (SMM), échanges de portions de lignées ancestrales adjacentes dans le temps, ...
- dans 5% des cas sur l'un ou plusieurs des autres paramètres  
exemple: tirage d'une nouvelle date pour un événement donné, tirage d'une nouvelle valeur de  $\theta$  ou de  $t_0$ , ...

Application de cette méthode aux données sur le *Northern hairy nose wombat*.

Cette analyse suggère que la population relique actuelle a subi un déclin rapide, et que ce déclin aurait commencé il y a plus longtemps que ce qu'on pense généralement.

Intérêt de la méthode: prise en compte possible des locus monomorphes

Limites de l'approche

- celles du modèle de Wright-Fisher et de la théorie de la coalescence
- celles des MCMC
- complexité des calculs et de la programmation