Génétique des populations: Dérive, migration, extinction et F-statistiques

Module ENS 2009

Raphaël Leblois, leblois@mnhn.fr

Maitre de conférence Muséum National d'Histoire Naturelle,

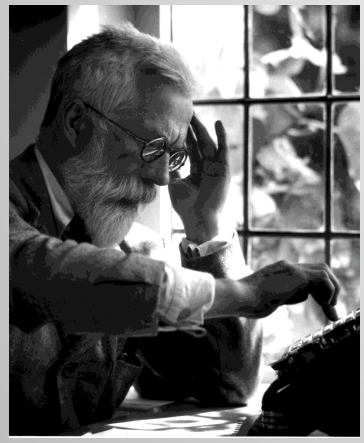
UMR7205 : Origine, Structure et Evolution de la biodiversité

Les mécanismes de l'évolution

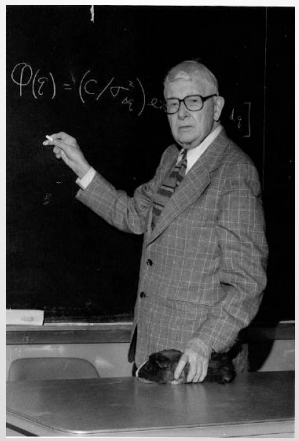
- Pour expliquer les mécanismes responsables des changements évolutifs, il faut développer une modélisation mathématique de l'évolution. Grâce aux modèles, on peut prédire la vitesse des changements dans une population, et comprendre comment ces changements dépendent des divers facteurs qui agissent dans les populations.

Les mécanismes de l'évolution

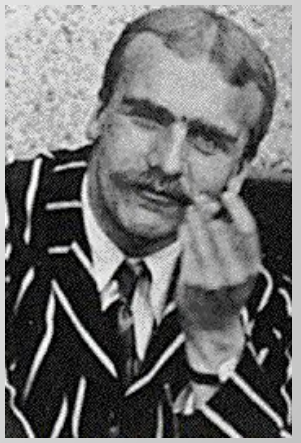
- Les mathématiques de l'évolution ont commencé à être développés dans les années 1920-1930



Ronald A Fisher (1890-1962)



Sewall Wright (1889-1988)



John BS Haldane (1892-1964)

La génétique des populations

- Décrire les génotypes, estimer leur fréquence et celle des allèles, déterminer leur distribution au sein des individus, des populations, et entre les populations

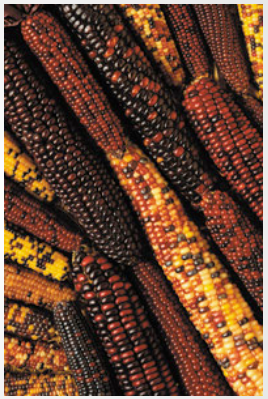
- Prédire et comprendre l'évolution des fréquences des gènes dans les populations sous l'effet de différents facteurs, ou « forces évolutives »

La génétique des populations

- Théorique : nécessaire pour tester des hypothèses, ou des modèles verbaux, et produire de nouvelles hypothèses
- Expérimentale : consiste à tester des modèles et leurs hypothèses dans des conditions contrôlées
- Empirique : description de la distribution du polymorphisme dans des populations naturelles, inférence de l'histoire (démographique et adaptative) des populations naturelles

Répartition de la variabilité génétique

- A l'intérieur même des individus
- Entre individus

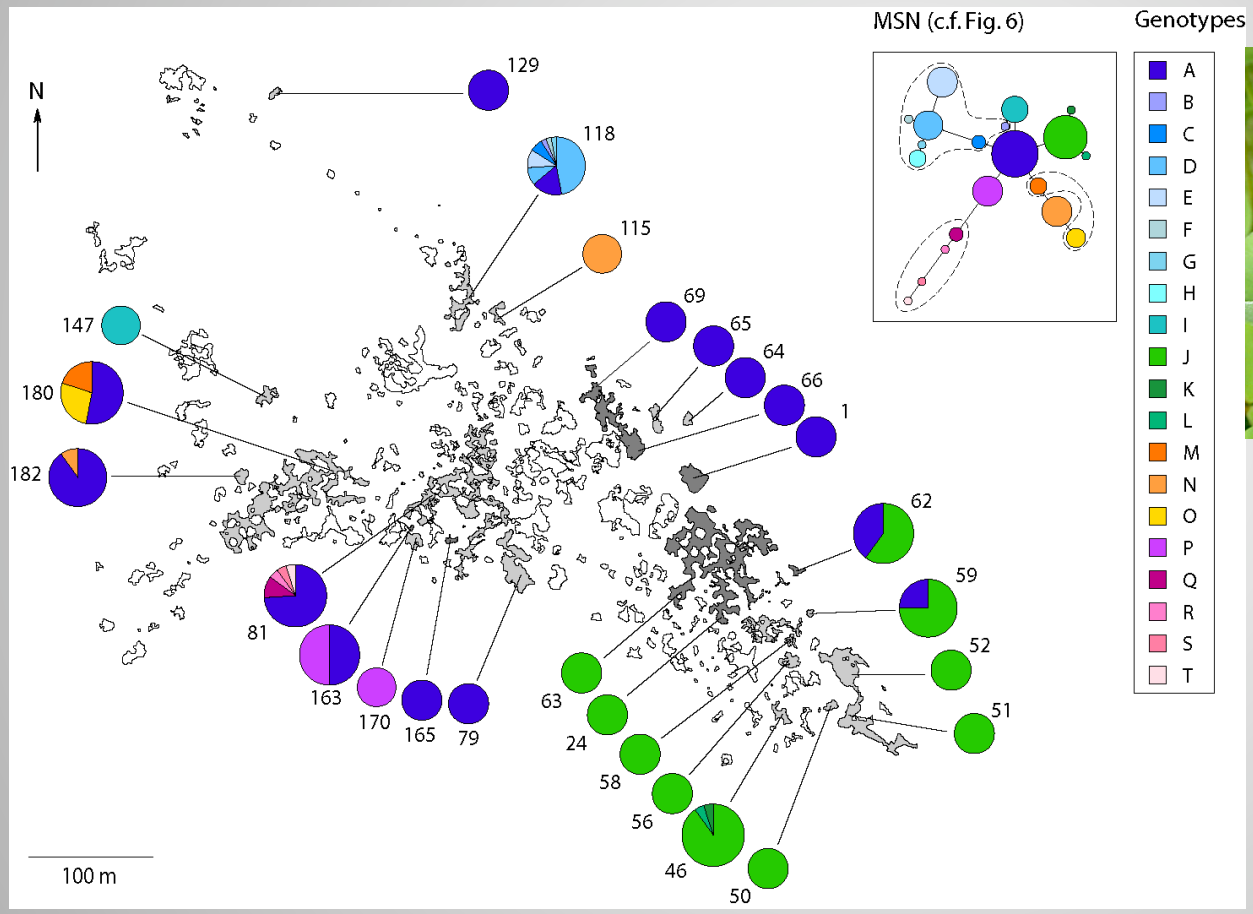


Zea mais

Cepea nemoralis

Répartition de la variabilité génétique

Variation entre sous-populations

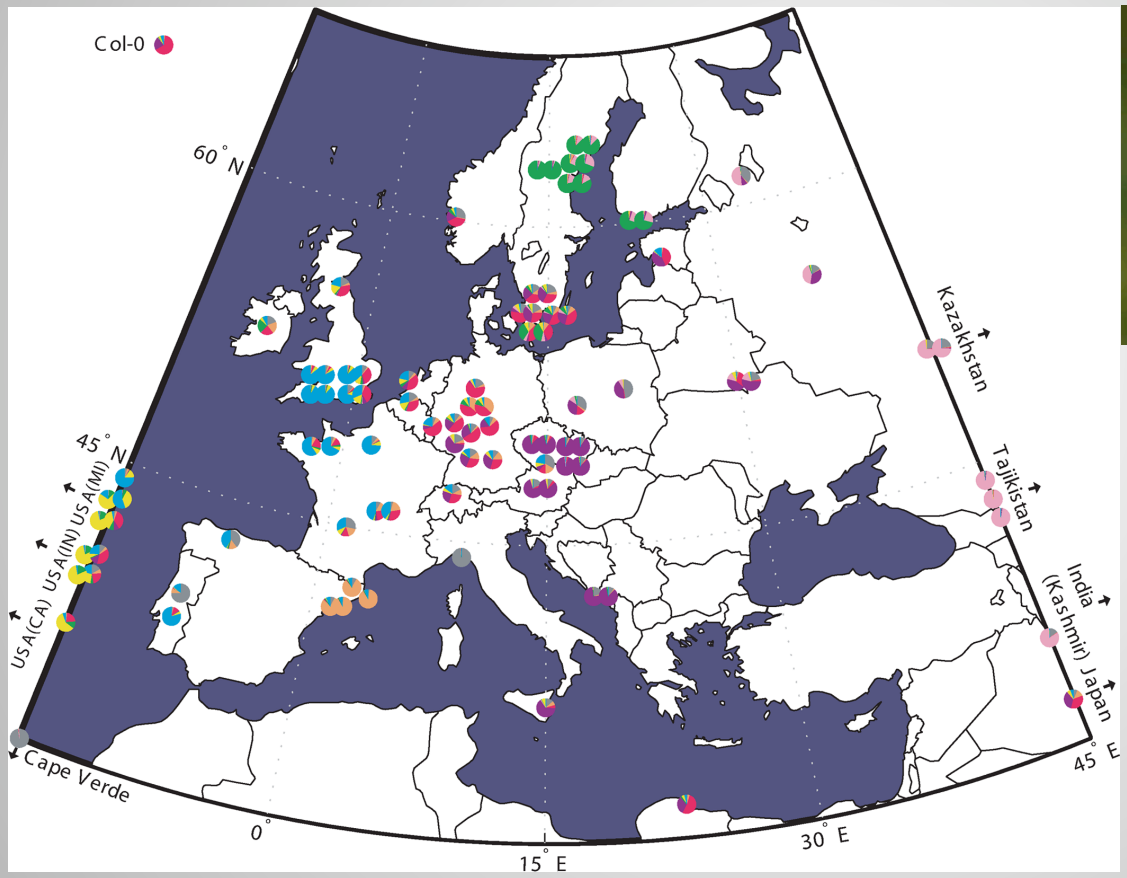


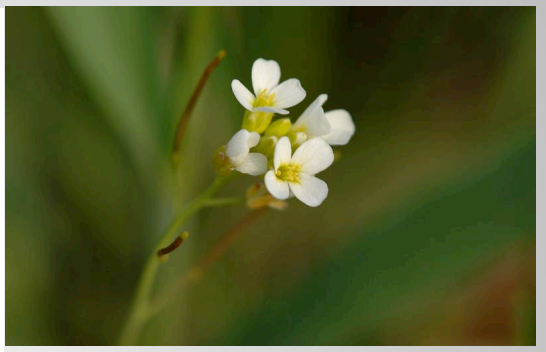
Marsilea strigosa

- Peu de polymorphisme au sein des mares, mais une forte différenciation entre mares

Répartition de la variabilité génétique

Variation entre populations





Arabidopsis thaliana

- 17 000 polymorphismes (SNPs ou indels) issus de l'analyse de 876 séquences obtenues parmi 96 individus (44 000 000 pb)

Quelques définitions

- **Gène :** copie d'une information génétique, portée par une séquence de nucléotides.
- **Locus :** emplacement d'un gène sur un chromosome
- **Allèle (ou état allélique) :** classe de gènes homologues (au même locus) tous équivalents. Deux gènes sont dans le même état allélique si l'information qu'ils portent est codée par la même séquence d'ADN ou s'ils sont la copie exacte d'un ancêtre commun.

Un individu diploïde possède donc deux gènes homologues à un locus (**son génotype**) et ces gènes peuvent avoir le même état allélique (individu/génotype **homozygote**) ou non (individu/génotype **hétérozygote**)

Qu’est-ce qu’un marqueur ?

- On utilise des marqueurs génétiques pour analyser de façon empirique la distribution du polymorphisme, à l’intérieur des individus, des populations, entre populations…

- Un marqueur génétique « idéal » doit :
  - ✓ Avoir un déterminisme simple et connu (ex: mendélien)
  - ✓ Être polymorphe
  - ✓ Être co-dominant
  - ✓ Être neutre vis-à-vis de la sélection naturelle (si l’on s’intéresse a la démographie uniquement)

Les marqueurs génétiques

- Allozymes / Isozymes = Marqueurs enzymatiques (électrophorèse sur gel d'amidon, d'acrylamide)

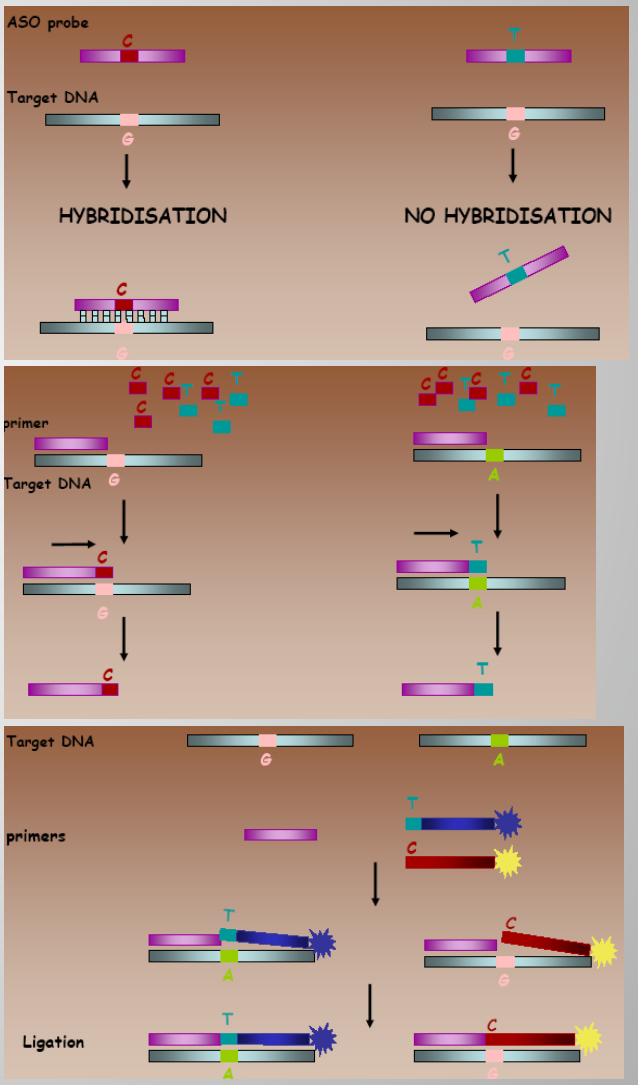
- Marqueurs microsatellites (répétitions en tandem de courts motifs de base (ex.: … AGAGAGAGAGAG…), dispersés dans les génomes

Les marqueurs génétiques

- RFLPs : *Restriction Fragment Length Polymorphisms*
- RAPDs : *Randomly Amplified Polymorphic DNAs*
- AFLPs : *Amplified Fragment Length Polymorphisms*
- SNPs : *Single Nucleotide Polymorphisms*

SNPs
(single nucleotide polymorphism)

- Hybridation Allèle Spécifique
- Extension d'amorce (mini-séquençage)
- Ligation Allèle Spécifique
- …



Les marqueurs génétiques

- Projet **HapMap** : bâtir une carte haplotypique du génome humain pour décrire la distribution du polymorphisme sur l'ensemble du génome

2002-2009 : plus de 25 millions de SNPs ont déjà été génotypés dans 11 populations humaines.

F-statistiques et génétique de populations

- Génétique des populations =
  - Comprendre et prédire l'effet des différentes forces évolutives (mutation, dérive , migration, sélection) sur l'évolution des fréquences des gènes dans le temps et l'espace
  - Pour cela, on utilise la modélisation mathématique de l'évolution au sein des populations

- Les F-statistiques sont des outils simple et puisant souvent utilisés dans les modèles de génétique des populations et par extension dans l'analyses des données de polymorphisme

Rappel :
Fréquences alléliques et génotypiques dans une population panmictique haploïde

Considérons un locus bi-allélique (A / a) dans population isolée de N individus **haploïdes** (et donc *N* gènes).

Les fréquences génotypiques sont égales aux fréquences alléliques, et définies à un moment *t* comme :

- $p[t]=D[t]=\text{Fréq(A)}=N_A/N$
- $q[t]=H[t]=\text{Fréq(a)}= N_a/N$

On peut vérifier que l'on a bien $D[t] + H[t] = p[t] + q[t] = 1$

Rappel :
Fréquences alléliques et génotypiques dans une population panmictique diploïde

Dans population de N individus **diploïde** (donc $2N$ gènes).

Les fréquences génotypiques sont définies à un moment t comme :

- $D[t]$=Fréq(AA)=N_{AA}/N
- $H[t]$=Fréq(Aa)= N_{Aa}/N
- $R[t]$=Fréq(aa)= N_{aa}/N

On peut vérifier que l'on a bien $D[t] + H[t] + R[t] = 1$

Et les fréquences alléliques :

- $p[t]=(2N_{AA}+N_{aa})\ /2N=D[t]+H[t]/2$
- $q[t]=(2N_{aa}+N_{aa})\ /2N=R[t]+H[t]/2$

et on a bien $p[t] + q[t]= D[t] + H[t] + R[t] = 1$

Rappel :
Evolution des fréquences alléliques et génotypiques dans une population diploïde

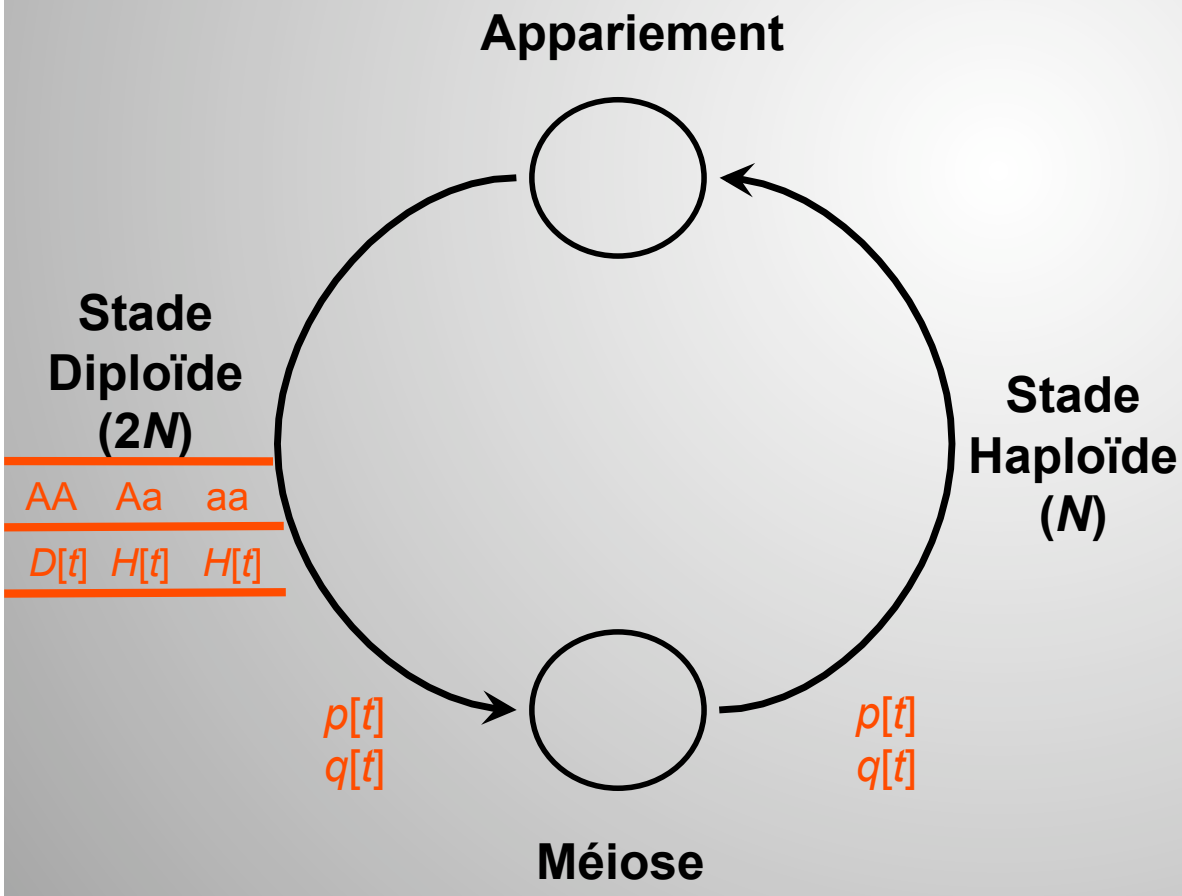
Quelles sont les fréquences génotypiques à la génération suivante?

On suppose :

- ✓ Croisement au hasard des gamètes (panmixie)
- ✓ Absence de mutation
- ✓ Absence de sélection
- ✓ Et une population de grande taille (infinie)

Rappel :
Evolution des fréquences alléliques et génotypiques dans une population diploïde

Hypothèses: panmixie, pas de mutation, pas de sélection, population de taille infinie



- En l'absence de sélection et de mutation, les fréquences alléliques parmi les gamètes sont égales aux fréquences alléliques parmi les adultes qui les ont produits

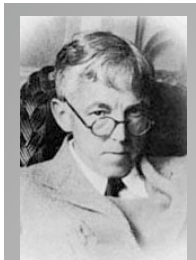
Rappel :
Evolution des fréquences alléliques et génotypiques dans une population diploïde

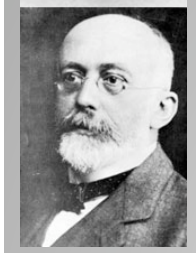
Hypothèses: panmixie, pas de mutation, pas de sélection, population de taille infinie

- Appariement au hasard des gamètes
- Proportions de Hardy-Weinberg

AA	$D[t+1] = p[t]^2$
Aa	$H[t+1] = 2\ p[t]\ q[t]$
aa	$R[t+1] = q[t]^2$

Gamètes mâles	Gamètes femelles : A, $p[t]$	Gamètes femelles : a, $q[t]$
A, $p[t]$	AA, $p[t]^2$	Aa, $p[t]q[t]$
a, $q[t]$	aA, $q[t]p[t]$	aa, $q[t]^2$





Rappel :
Evolution des fréquences alléliques et génotypiques dans une population diploïde
Loi de Hardy-Weinberg

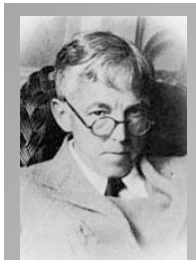
Conclusions :

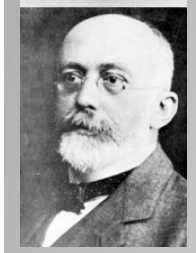
• les proportions de Hardy-Weinberg (fréquences génotypiques à l'équilibre en fonction des fréquences alléliques, p^2, $2pq$, q^2) sont atteintes en une génération pour une espèce diploïde

De plus

$$p[t+1] = D[t+1] + H[t+1]/2 = p[t]^2 + p[t]q[t] = p[t]\,(\,p[t] + q[t]\,) = p[t]$$
$$q[t+1] = R[t+1] + H[t+1]/2 = q[t]^2 + p[t]q[t] = q[t]\,(\,q[t] + p[t]\,) = q[t]$$

• les fréquences alléliques sont constantes au cours du temps



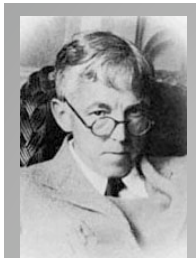


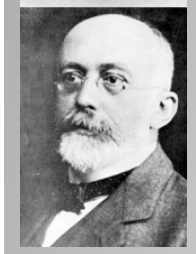
Rappel :
Evolution des fréquences alléliques et génotypiques dans une population diploïde
Loi de Hardy-Weinberg

Mais ceci n'est valable que sous les hypothèses de la loi de HW :

- population panmictique (croisement entre individus au hasard)
- générations non chevauchantes (tous les individus participant à la reproduction appartiennent à la même génération)
- population isolée (aucun gène introduit par des individus migrants)
- population de taille infinie
- pas de mutation
- pas de sélection

Ces hypothèses rarement vérifiées en populations naturelles





Rappel :
Loi de Hardy-Weinberg
Test de conformité à HW

- Pour un locus bi-allélique *A*, *a* :

	Attendu	*Observé*
AA	Np^2	N_1
Aa	$2Npq$	N_2
aa	Nq^2	N_3

On calcule une statistique *X*, qui mesure l'écart entre l'attendu et l'observé, qui suit un Chi2 à 1 d.d.l. (nombre de génotypes possibles – nombre d'allèles) :

$$X = \sum_{\text{génotypes}} \frac{\left(\text{effectif attendu} - \text{effectif observé}\right)^2}{\text{effectif attendu}}$$

Ecarts à Hardy Weinberg
Evolution des fréquences génotypiques : effet du régime de reproduction

En autofécondation totale :

AA	$D[t]$
Aa	$H[t]$
aa	$R[t]$

- Chacun des $H[t]$ individus hétérozygotes a 50% d'individus hétérozygotes dans sa descendance

Gamètes mâles \ Gamètes femelles	A	a
A	AA	Aa
a	aA	aa

Ecarts à Hardy Weinberg
Evolution des fréquences génotypiques : effet du régime de reproduction

Après une génération :

AA	100% AA
Aa	25% AA, 50% Aa, 25% aa
aa	100% aa

On a donc :

- $D[t+1] = D[t] + H[t] / 4$
- $H[t+1] = H[t] / 2$
- $R[t+1] = R[t] + H[t] / 4$

- Comment évoluent les fréquences alléliques ?

Ecarts à Hardy Weinberg
Evolution des fréquences génotypiques : effet du régime de reproduction

- Évolution des fréquences alléliques

$$p[t+1] = D[t+1] + H[t+1] / 2 = D[t] + H[t] / 4 + H[t] / 4$$
$$= D[t] + H[t] / 2$$
$$= p[t]$$

$$q[t+1] = D[t+1] + H[t+1] / 2 = D[t] + H[t] / 2 = q[t]$$

Les fréquences alléliques restent constantes

Comment évoluent les fréquences génotypiques ?

Ecarts à Hardy Weinberg
Evolution des fréquences génotypiques : effet du régime de reproduction

Puisque

- $D[t+1] = D[t] + H[t] / 4$
- $H[t+1] = H[t] / 2$
- $R[t+1] = R[t] + H[t] / 4$

On a donc $H[t] = H[t-1] / 2 = H[t-2] / 2^2 = \ldots = H[0] / 2^t$

- $H[t]$ tend vers 0 (plus d'hétérozygotes : tous les individus sont homozygotes)

- $p = D[0] + H[0] / 2 = D[t] + H[t] / 2$, d'où $D[t] = p - H[t] / 2$

- et donc $D[t]$ et $R[t]$ tendent vers p et q, respectivement

Ecarts à Hardy Weinberg
Evolution des fréquences génotypiques : effet du régime de reproduction

- Au bout d'un temps suffisamment long

AA	*p*
Aa	0
aa	*q*

Ecarts à Hardy Weinberg
Evolution des fréquences génotypiques : effet du régime de reproduction

En autofécondation partielle (chaque individus produit un taux *s* de descendants en autofécondation et (1 - *s*) en allofécondation), on s'attend à un déficit en hétérozygote par rapport aux proportions d'Hardy-Weinberg :

On définit F_{IS} tel que $H_{obs} = 2pq\ (\ 1\text{-}F_{IS}\ ) \rightarrow F_{IS} = 1 - H_{obs} / 2pq$

Or on a $p = D + H/2$ et $q = R + H/2 \rightarrow D = p^2 + pqF_{IS}$ et $R = q^2 + pqF_{IS}$

Les fréquences génotypiques de la population sont donc

AA	p^2+pqF_{IS}
Aa	$2pq(1\text{-}F_{IS})$
aa	q^2+pqF_{IS}

Ecarts à Hardy Weinberg
Evolution des fréquences génotypiques : effet du régime de reproduction

En autofécondation partielle (chaque individus produit des descendant en autofécondation avec un taux *s* et le reste (1 - *s*) en allofécondation) :

Quel que soit le FIS :

$$p = D + H/2 = p^2 + pqF_{IS} + pq(1-F_{IS}) = p^2 + pq = p$$
$$q = R + H/2 = q^2 + pqF_{IS} + pq(1-F_{IS}) = q^2 + pq = q$$

Les fréquences alléliques sont toujours constantes au cours du temps

Ecarts à Hardy Weinberg
Evolution des fréquences génotypiques : effet du régime de reproduction

- L'effet de la consanguinité du régime de reproduction est de modifier la composition génotypique de la population, mais pas de faire varier les fréquences alléliques, en l'absence de toute autre force évolutive

Ecarts à Hardy Weinberg
Evolution des fréquences génotypiques : effet du régime de reproduction

En autofécondation partielle : quel est la relation entre le F_{IS} et le taux d'autofécondation s?

Comment évoluent les fréquences génotypiques en fonction de s ?

- $D[t+1] = s\,D[t] + s\,H[t] / 4 + (1 - s)\,p[t]^2$
-
- $H[t+1] = s\,H[t] / 2 + (1 - s)\,2p[t]q[t]$
-
- $R[t+1] = s\,R[t] + s\,H[t] / 4 + (1 - s)\,q[t]^2$

Ecarts à Hardy Weinberg
Evolution des fréquences génotypiques : effet du régime de reproduction

En autofécondation partielle : quel est la relation entre le F_{IS} et le taux d'autofécondation s?

A l'équilibre

$H = s\,H / 2 + (1 - s)\,2pq$
$H = 2pq\,(1 - s) / (1 - s / 2)$

Or on a défini F_{IS} tel que : $H_{obs} = 2\,pq\,(1 - F_{IS})$

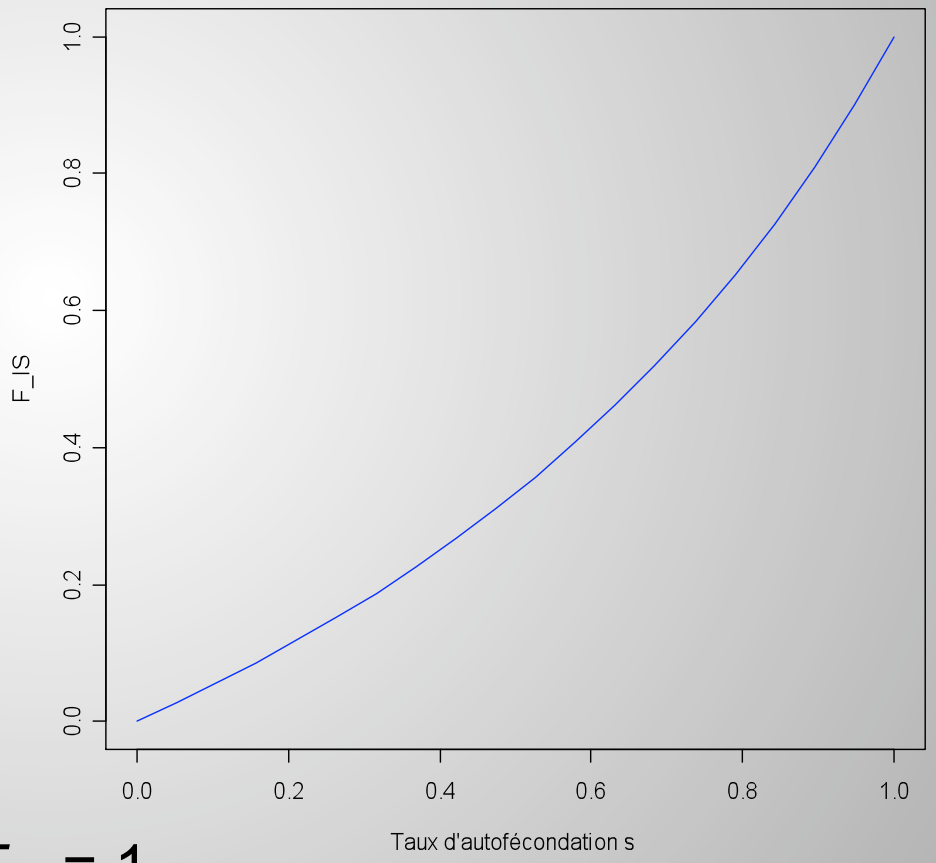
D'où :

$$F_{IS} = \frac{s}{2 - s}$$

Ecarts à Hardy Weinberg
Evolution des fréquences génotypiques : effet du régime de reproduction

En autofécondation partielle : quel est la relation entre le F_{IS} et le taux d'autofécondation *s*?

$$F_{IS} = \frac{s}{2 - s}$$



Si *s* = 0 (panmixie) F_{IS} = 0
Si *s* = 1 (autogamie complète) F_{IS} = 1

Ecarts à Hardy Weinberg
Evolution des fréquences génotypiques : effet du régime de reproduction

• Le F_{IS} mesure l'écart à la panmixie (déficit d'hétérozygote) dû à la consanguinité du régime de reproduction au sein d'une population

• On peut déduire de la valeur du F_{IS} le taux d'autofécondation moyen de la population :

$$s = \frac{1 + F_{IS}}{2F_{IS}}$$

Evolution des fréquences alléliques dans les population naturelles

Il existe 4 "forces évolutives" qui agissent en interactions et font évoluer les fréquences alléliques et génotypiques en populations naturelles :

- **la mutation**
- **la dérive génétique (variations stochastiques des fréquences alléliques dues aux effets d'échantillonnages)**
- **la migration (ou flux de gènes)**
- **la sélection naturelle**

Evolution des fréquences alléliques dans les population naturelles

Il existe 4 "forces évolutives" qui agissent en interactions et font évoluer les fréquences alléliques et génotypiques en populations naturelles :

- **la mutation**
- **la dérive génétique (variations stochastiques des fréquences alléliques dues aux effets d'échantillonnages)**
- **la migration (ou flux de gènes)**
- **la sélection naturelle**

La mutation

- La mutation est la source fondamentale de nouvelle variation génétique

- La mutation, c'est la modification spontanée d'un état allélique en un autre, par :
  - Substitutions nucléotidiques
  - Délétions
  - Insertion
  - Inversions chromosomiques
  - Translocations chromosomiques

La mutation

- Il existe des différences importantes du taux de mutation entre organismes (mais aussi entre gènes)

Espèce	Taille du génome (pb)	Taux de mutation par pb et par réplication	Taux de mutation par génome et par réplication
Escherichia coli	4.6×10^{6}	5.4×10^{-10}	0.0025
Bactériophage λ	4.9×10^{4}	7.7×10^{-8}	0.0038
Caenorhabditis elegans	8.0×10^{7}	2.3×10^{-10}	0.018
Souris	2.7×10^{9}	1.8×10^{-10}	0.49
Homme	3.2×10^{9}	5.0×10^{-11}	0.16

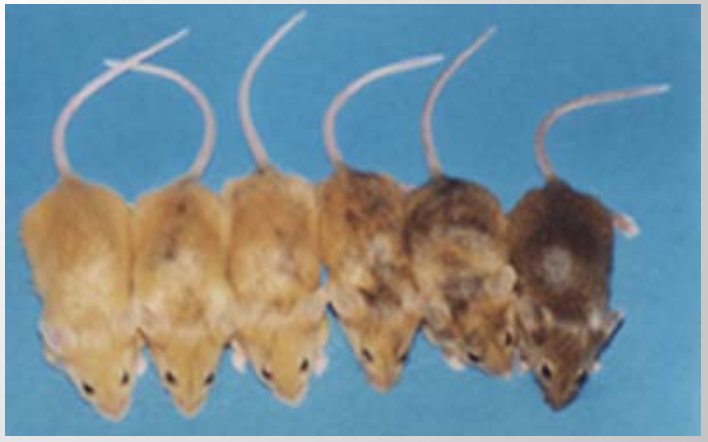
D'après Drake *et al*. 1998

La mutation

- Le taux de mutation peut aussi dépendre de l'allèle considéré. Par exemple, chez la souris, pour les gènes impliqués dans la couleur du pelage :

11.2×10^{-6} par locus et par gamète (type sauvage vers type mutant)

2.5×10^{-6} par locus et par gamète (type mutant vers type sauvage)



Les mutations qui touchent l'expression de la fonction sauvage (mutations 'avant' ou *forward*) sont souvent plus fréquentes que les mutations qui restaurent la fonction sauvage (mutations 'arrière' ou *backward*)

La mutation

- Taux de mutation de A vers a : μ
- Taux de mutation de a vers A : v
- Si p[t] est la fréquence de A au temps t, alors à la génération suivante, après la méiose, en l'absence de sélection :
- $p[t+1] = (1 - \mu)p[t] + v\, q[t]$
- $q[t+1] = (1 - v)q[t] + \mu\, p[t]$

- A l'équilibre ?

La mutation

- Taux de mutation de A vers a : μ
- Taux de mutation de a vers A : ν
- Si p[t] est la fréquence de A au temps *t*, alors à la génération suivante, après la méiose, en l'absence de sélection :
- $p[t+1] = (1 - \mu)p[t] + \nu\, q[t]$
- $q[t+1] = (1 - \nu)q[t] + \mu\, p[t]$
- A l'équilibre : $\hat{p} = \dfrac{\nu}{\mu + \nu}$

La mutation

$$\hat{p} = \frac{\nu}{\mu + \nu}$$

- p = 0 et p = 1 ne sont pas des valeurs d'équilibre. La fixation n'est pas stable quand il y a des mutations…
- Mais quel est le taux d'approche de l'équilibre ?

La mutation

$$
\begin{aligned}
p[t+1] - \hat{p} &= (1-\mu)p[t] + \nu(1-p[t]) - \hat{p} \\
&= (1-\mu-\nu)p[t] + \nu - \hat{p} \\
&= (1-\mu-\nu)p[t] - (1-\mu-\nu)\hat{p} \quad \text{puisque } \hat{p} = (1-\mu)\hat{p} + \nu(1-\hat{p}) \\
&= (1-\mu-\nu)(p[t] - \hat{p}) \\
&= (1-\mu-\nu)^2(p[t-1] - \hat{p}) \\
&= (1-\mu-\nu)^{t+1}(p[0] - \hat{p})
\end{aligned}
$$

- La fréquence d'équilibre est atteinte (seulement) au taux de (μ + v)

La mutation

- Quelle est la fréquence d'équilibre lorsque $\mu = v = 10^{-6}$?

$$\hat{p} = \frac{\nu}{\mu + \nu} \longrightarrow 0{,}5$$

La mutation

- Quelle est la fréquence d'équilibre lorsque $\mu = v = 10^{-6}$ ➡ 0.5
- Mais si on part de la fréquence initiale $p[0] = 0$ alors, après 10 000, 100 000 générations, on a:

  $p[10\ 000] = 0.0099$ $p[100\ 000] = 0.0906$

- Combien de générations faut-il pour atteindre 90% de la valeur d'équilibre ?

La mutation

- Quelle est la fréquence d'équilibre lorsque $\mu = v = 10^{-6}$ ➡ 0.5
- Mais si on part de la fréquence initiale $p[0] = 0$ alors, après 10 000, 100 000 générations, on a:

  $p[10\ 000] = 0.0099$ $p[100\ 000] = 0.0906$

- Combien de générations faut-il pour atteindre 90% de la valeur d'équilibre ?
  - 1.15×10^6 générations !!
- Il faut donc environ $1 / \mu$ générations pour que la population soit proche de l'équilibre (environs 15 à 30 000 000 d'années chez l'homme !)

La mutation

- C’est une force de **faible intensité** (les évènements arrivent sur une échelle de temps bien différente par rapport à la migration, la dérive, la sélection)
- C’est la seule **source de variation**, mais le devenir des mutations va dépendre des autres forces évolutives
- Les choses sont différentes si l’on considère que les caractères sont déterminés par beaucoup de gènes à des locus différents (génétique quantitative, la semaine prochaine…)

Evolution des fréquences alléliques en populations naturelles

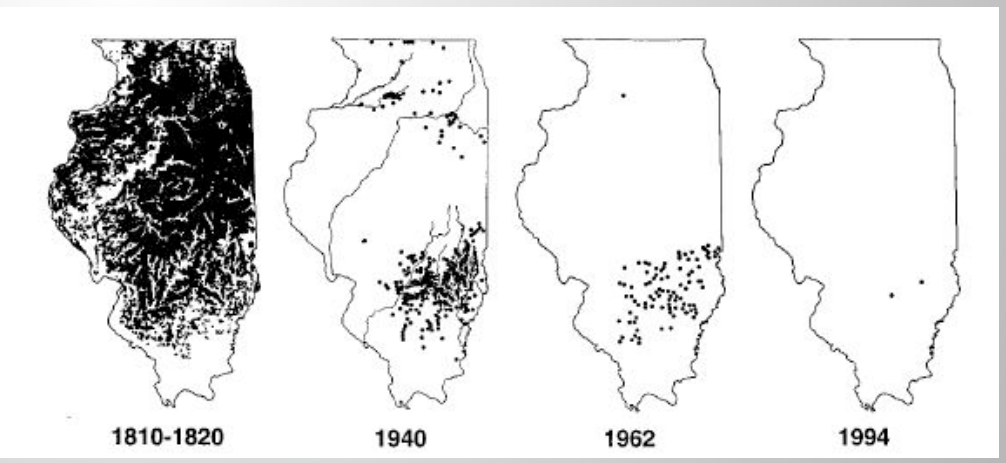
Il existe 4 "forces évolutives" qui agissent en interactions et font évoluer les fréquences alléliques en populations naturelles :

- **la mutation**
- **la dérive génétique (variations stochastiques des fréquences alléliques dues aux effets d'échantillonnages)**
- **la migration (ou flux de gènes)**
- **la sélection naturelle**

Évolution en populations finies

- On a supposé que les populations ont une taille infinie
- Or les populations naturelles sont rarement de taille infinie...

Tympanuchus cupido
Tétra des prairies



Illinois

Évolution en populations finies

- On a supposé que les populations ont une taille infinie : par exemple on a supposé que les fréquences alléliques chez les descendants étaient exactement égales à celles des parents en l'absence de mutation et de migration (modèle déterministe)

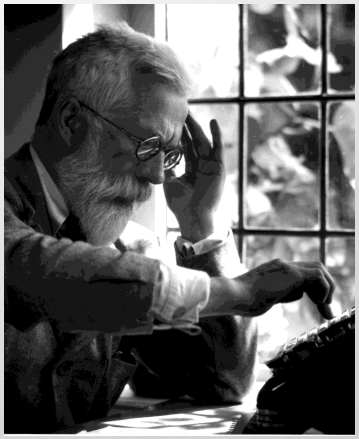
- Or dans les populations de taille finie, les fréquences peuvent varier d'une génération à l'autre simplement sous l'effet du hasard (modèle stochastique)

- On appelle cette variation de fréquence aléatoire due à l'effet d'échantillonnage dans une population de taille finie la dérive génétique

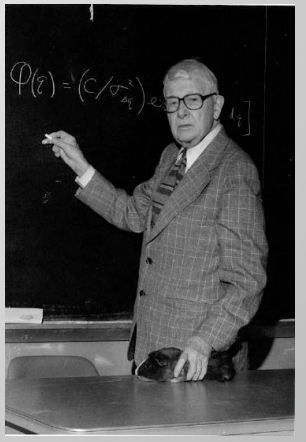
Évolution en populations finies

- De la même façon que dans une population de taille finie et constante chaque individu (asexué) ne donne pas exactement un descendant et un seul, chaque gène ne contribue pas par une copie et une seule à la composition génétique de la génération suivante, mais par un nombre variable de copies qui suit une loi de probabilité d'espérance égale à un.

Le modèle de Wright-Fisher

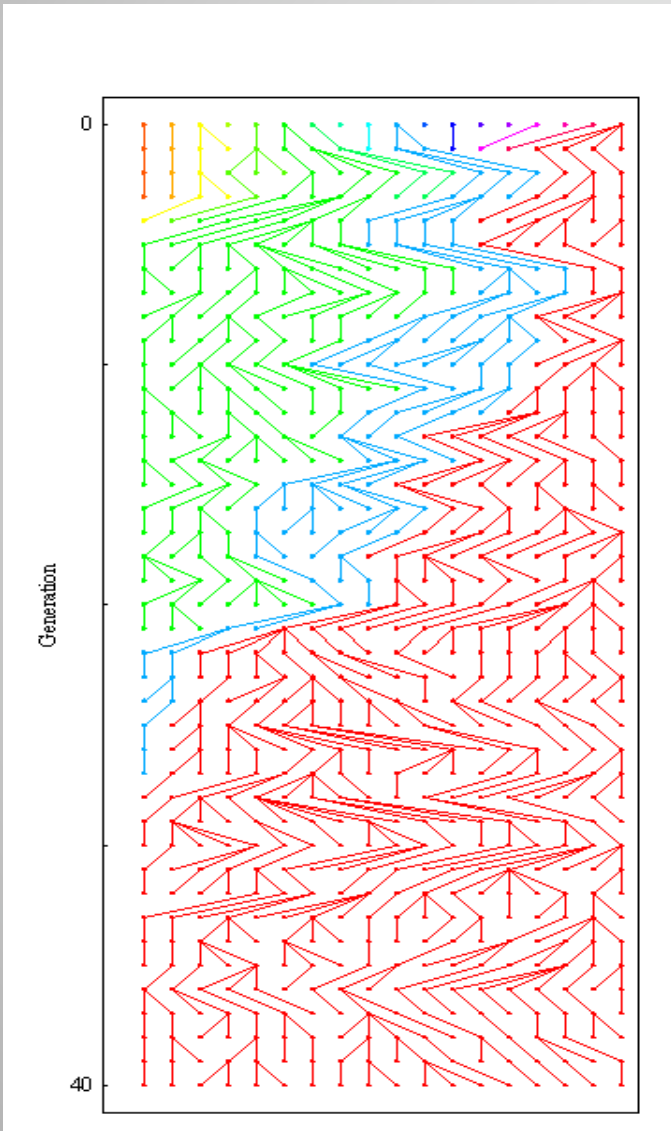


Ronald A Fisher



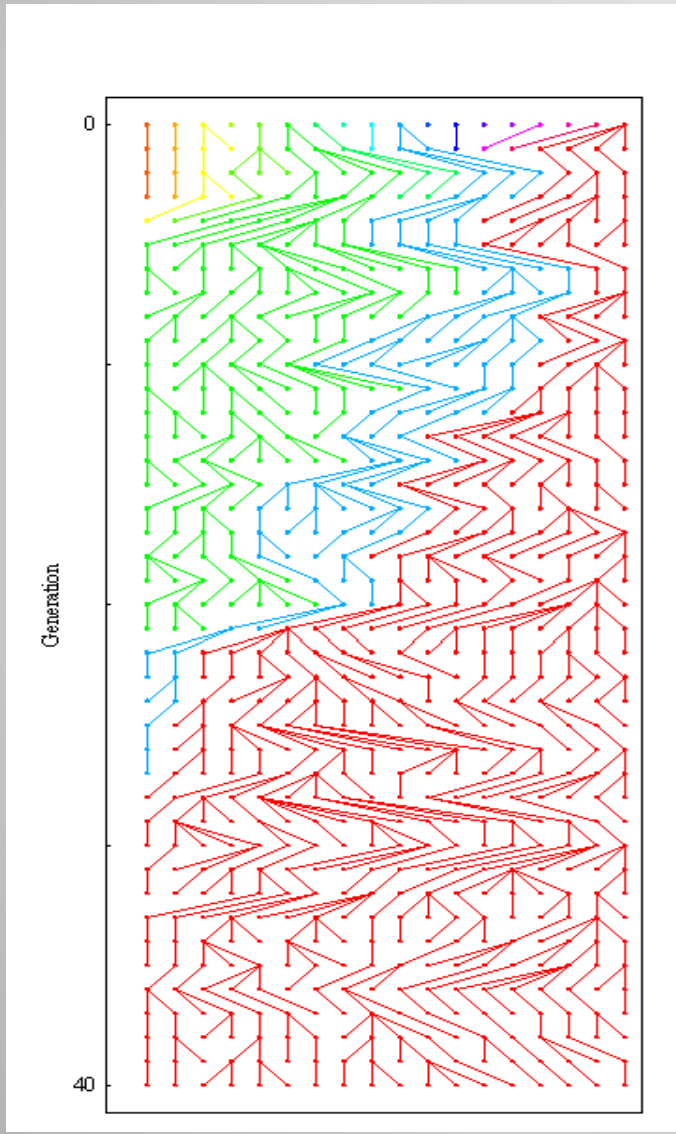
Sewall Wright

Le modèle de Wright-Fisher



- Une population de taille constante, dans laquelle les individus se reproduisent une unique fois et au même moment (générations non chevauchantes)
- Chaque gène à une génération est la copie d'un gène de la génération précédente

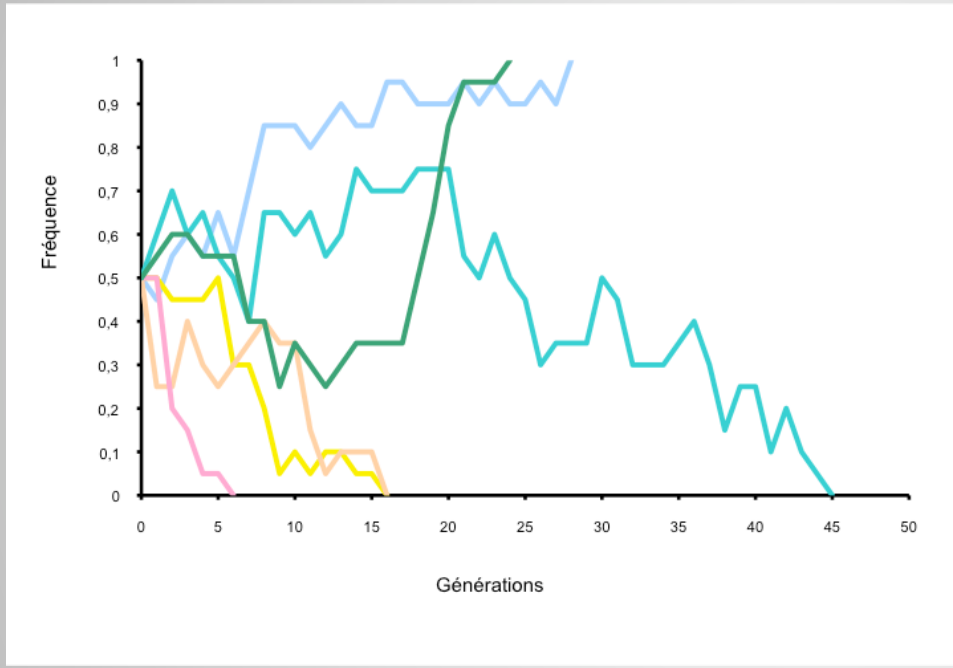
Le modèle de Wright-Fisher



- En l'absence de mutation et de sélection, les fréquences alléliques dérivent (augmentent et diminuent) inévitablement jusqu'à la fixation d'un allèle

- La dérive conduit donc à la perte de variation génétique à l'intérieur des populations

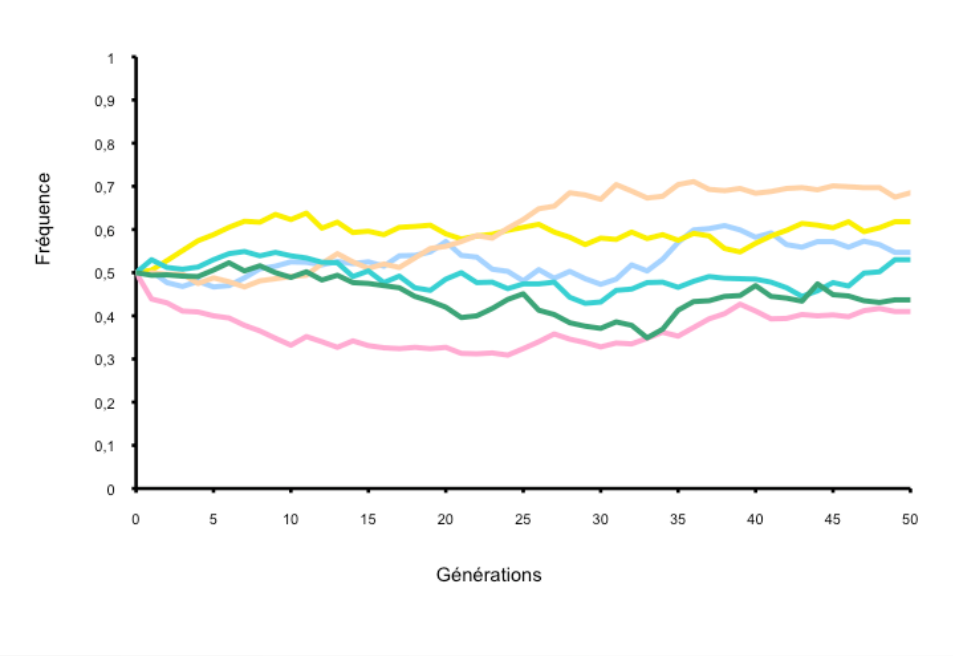
La dérive



- Évolution des fréquences alléliques dans 6 populations de ***N* = 10** individus

- Au bout de 50 générations, toutes les populations sont fixées pour un allèle

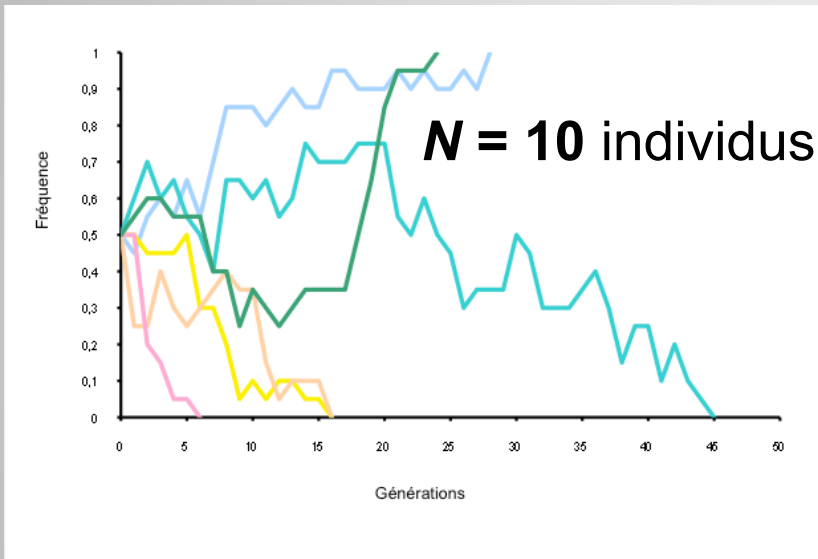
La dérive

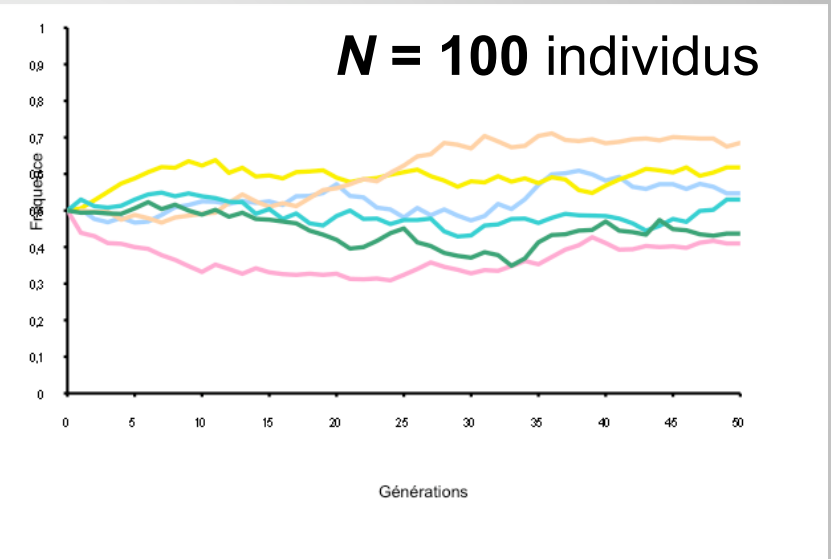


- Évolution des fréquences alléliques dans 6 populations de ***N* = 100** individus

- Au bout de 50 générations, aucune des populations n'est fixée pour un allèle

La dérive





- Les fréquences alléliques fluctuent d'autant plus que les populations sont de petite taille
- Des populations « filles » issues d'une même population « mère » divergent de façon indépendante. La variance augmente au fil du temps

Le modèle de Wright-Fisher

- On considère une population haploïde, isolée de taille N
- On considère un locus où 2 allèles (A et a) ségrégent. On note $p[t]$ la fréquence de A et $q[t] = (1 - p[t])$ la fréquence de a au temps t
- Il n'y a pas de mutation
- Chaque génération, chaque individu produit un grand nombre de gamètes, en espérance le même pour tous (neutralité sélective)
- On tire N gamètes pour constituer la génération suivante (tirage dans une urne gamétique de taille infinie)

Le modèle de Wright-Fisher

- Au temps (t+1), on réalise donc un tirage de N gènes avec remise dans une urne gamétique de taille infinie constituée d'allèles A en fréquence $p[t]$ et d'allèles a en fréquence $q[t]$

- Pour un tirage (disons celui correspondant à l'individu i), on définit la variable aléatoire indicatrice $X_i\,[t+1]$ telle que :

$$X_i\left[t+1\right] = \begin{cases} 1 & \text{si le gène } i \text{ est de type } A \\ 0 & \text{autrement} \end{cases}$$

- Selon les hypothèses du modèle, X_i suit une loi de Bernouilli de paramètre $p[t]$, et :

$$\begin{cases} E\left(X_i[t+1]\right) = p[t] \\ V\left(X_i[t+1]\right) = p[t]q[t] = p[t](1-p[t]) \end{cases}$$

Le modèle de Wright-Fisher

- Pour *N* tirages, on définit la variable aléatoire *X*[*t*+1] qui donne le nombre de copies de A, c'est-à-dire :

$$X[t+1] = \sum_{i=1}^{N} X_i[t+1]$$

- Quelles sont l'espérance et la variance de cette variable aléatoire ? (on fera l'hypothèse que les tirages sont indépendants)

Le modèle de Wright-Fisher

- Pour *N* tirages, on définit la variable aléatoire *X*[*t*+1] qui donne le nombre de copies de A, c'est-à-dire :

$$X[t+1] = \sum_{i=1}^{N} X_i[t+1]$$

- L'espérance de cette variable aléatoire est donnée par :

$$E(X[t+1]) = E\left(\sum_{i=1}^{N} X_i[t+1]\right) = \sum_{i=1}^{N} E(X_i[t+1]) = Np[t]$$

- Et sa variance (puisque les tirages sont indépendants) :

$$V(X[t+1]) = V\left(\sum_{i=1}^{N} X_i[t+1]\right) = \sum_{i=1}^{N} V(X_i[t+1]) = Np[t]q[t]$$

Le modèle de Wright-Fisher

- $X[t+1]$ suit donc une loi Binomiale, de paramètres N et $p[t] = X[t] / N$

$$\Pr\left(X\left[t+1\right]=k\right)=\binom{N}{k}p[t]^{k}\left(1-p[t]\right)^{N-k}$$

- Soit $Y[t+1] = X[t+1] / N = p[t+1]$, quelles sont l'espérance et la variance de la fréquence de A à la génération $(t+1)$?

Le modèle de Wright-Fisher

- $X[t+1]$ suit donc une loi Binomiale, de paramètres N et $p[t] = X[t] / N$

$$\Pr\left(X\left[t+1\right]=k\right)=\binom{N}{k}p[t]^{k}\left(1-p[t]\right)^{N-k}$$

- Soit $Y[t+1] = X[t+1] / N = p[t+1]$, la fréquence de A à la génération $(t+1)$, on a :

$$E\left(Y[t+1]\right)=E\left(p[t+1]\right)=E\left(\frac{X\left[t+1\right]}{N}\right)=\frac{E\left(X\left[t+1\right]\right)}{N}=p[t]$$

$$V\left(Y[t+1]\right)=V\left(p[t+1]\right)=V\left(\frac{X\left[t+1\right]}{N}\right)=\frac{V\left(X\left[t+1\right]\right)}{N^{2}}=\frac{p[t]q[t]}{N}$$

- En espérance, la fréquence ne change pas d'une génération à l'autre, mais la variance est d'autant plus grande que N est petit

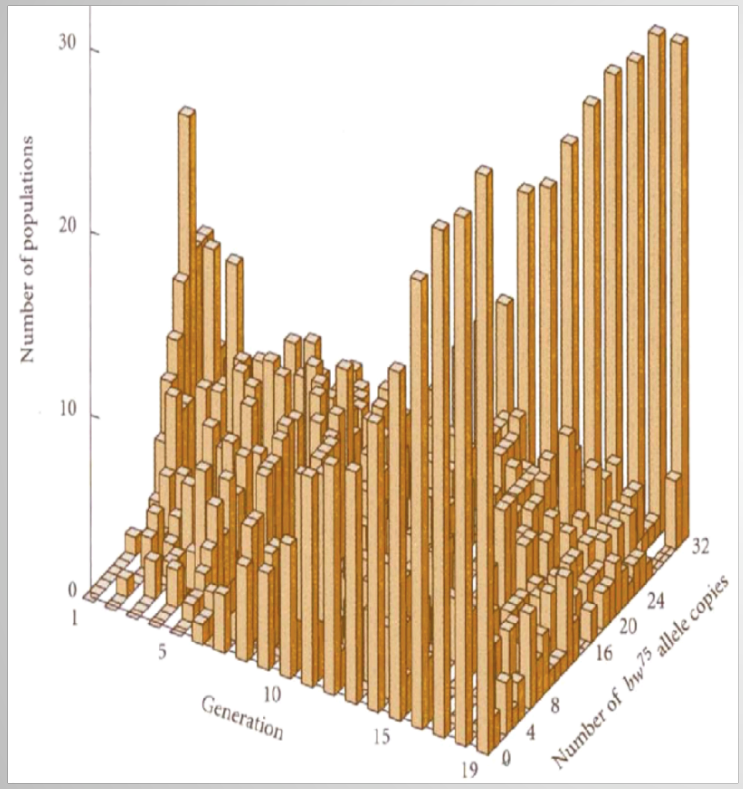
Le modèle de Wright-Fisher

- La variance représente la variation de la fréquence allélique de A à la génération suivante, si l'on répétait l'expérience un grand nombre de fois à partir d'une même population constituée initialement de A et a en fréquences *p* et *q*

- Au bout d'un grand nombre de générations, chaque population va nécessairement fixer un allèle A ou a ($p = 0$ et $p = 1$ sont des états absorbants)

- Mais si l'on considère un nombre infini de populations isolées, *p* populations vont fixer A, et (1 - *p*) populations vont fixer a. Dans ce cas, la fréquence totale de A reste *p*

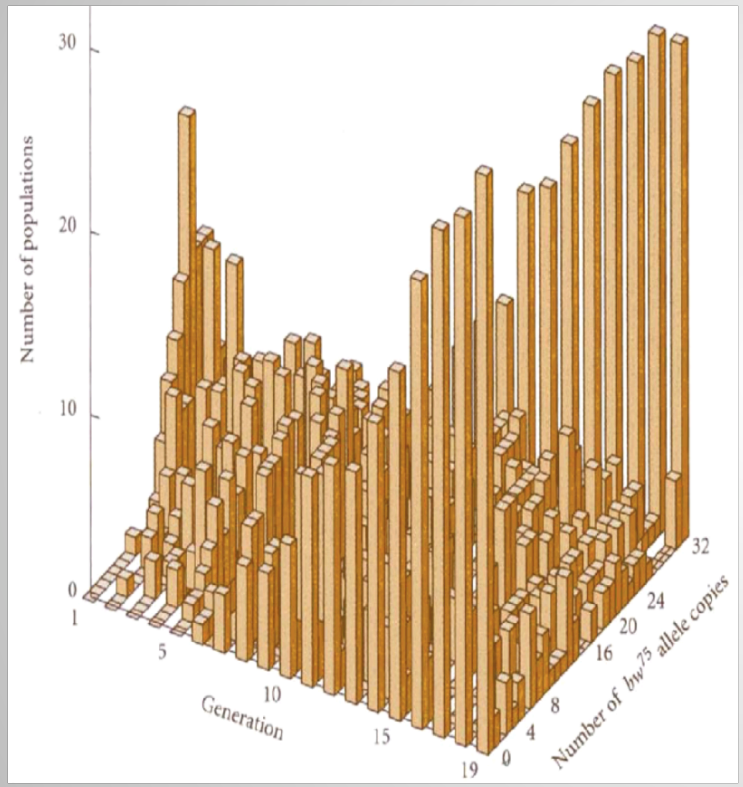
Évolution en populations finies

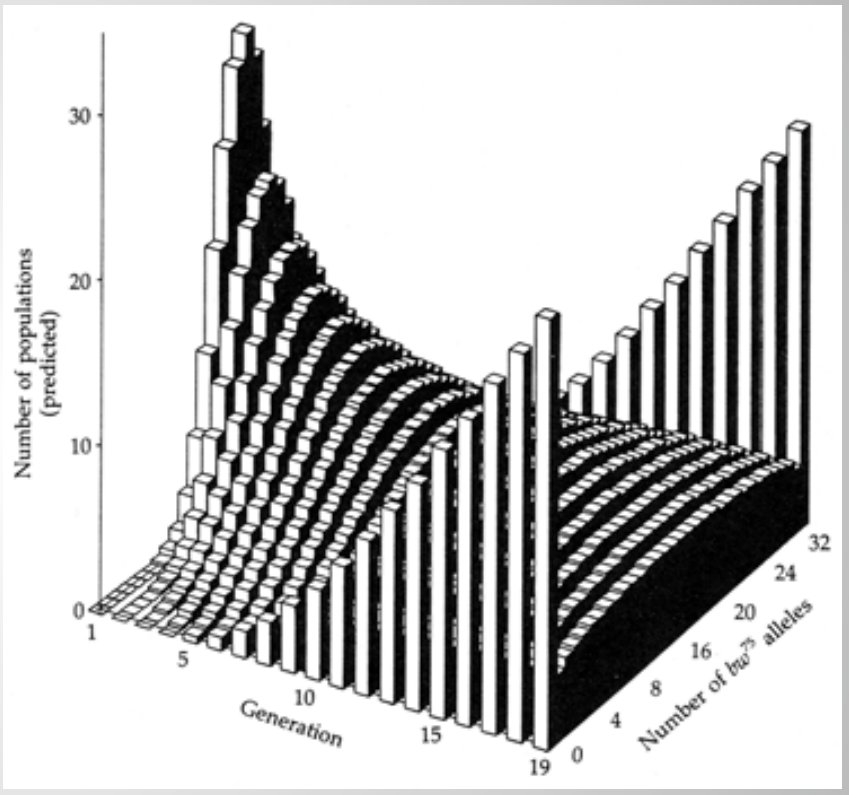


- 107 populations de drosophiles, chacune ayant été fondée avec 16 individus hétérozygotes pour la mutation ‘brown eye’ (*bw*^{75})

Les conséquences de la dérive : expérience de Buri (1956)

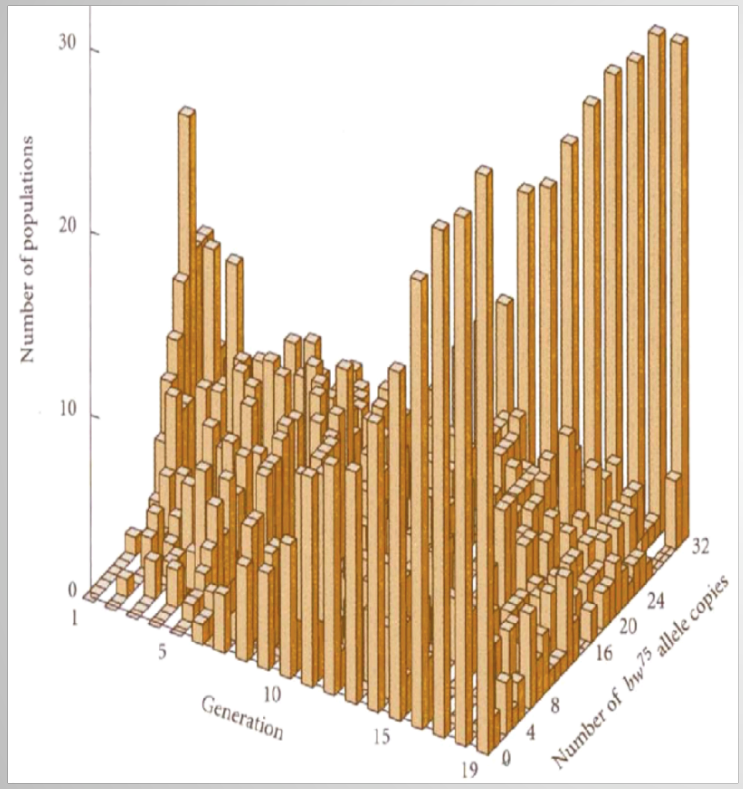
Évolution en populations finies

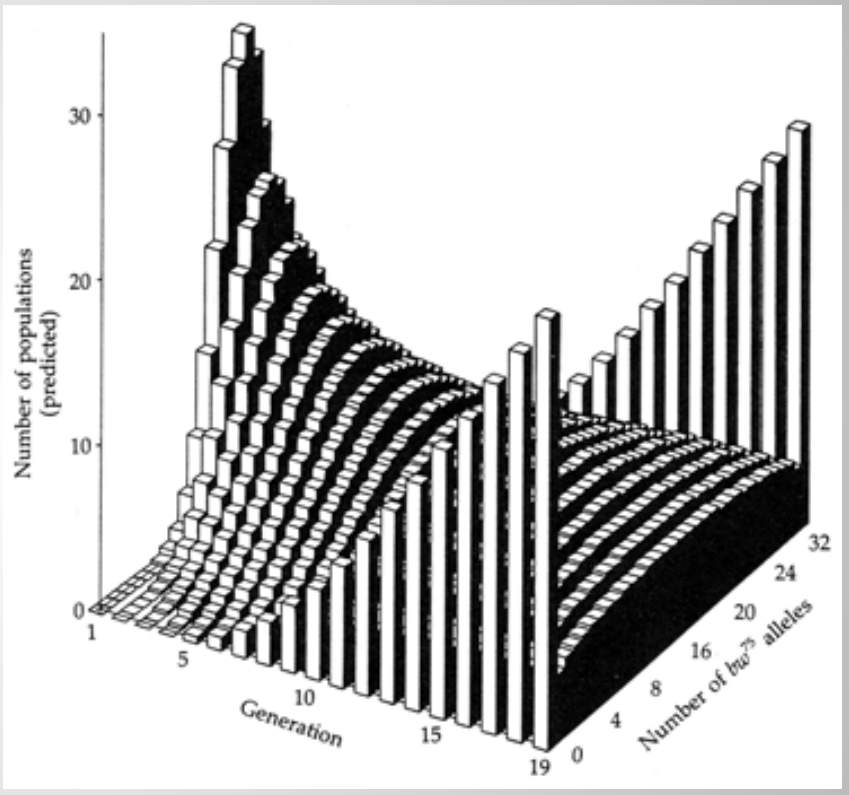




- Au fil du temps, la variation à l'intérieur des populations diminue
- La différenciation entre populations augmente
- Ceci s'exprime par une diminution de la proportion d'hétérozygotes au niveau global (effet Wahlund)

Évolution en populations finies





- Le modèle avec $2N = 32$ prédit moins de populations fixées que ce qui est observé au bout de 19 générations. Ceci parce que la variance du succès reproducteur est environ 70% plus élevé que ce qui est supposé dans le modèle de Wright-Fisher.

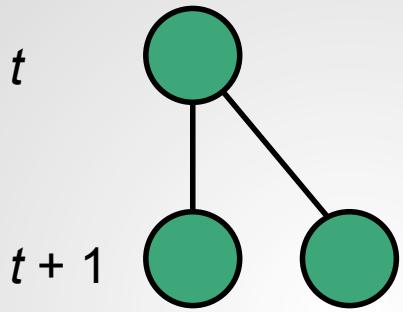
Dérive et consanguinité

- La perte de variation due à la dérive entraîne une augmentation de l'identité entre gènes (de façon ultime tous les gènes d'une population finie sont des copies d'un même gène ancêtre ; ils sont tous identiques par descendance ; la population est entièrement consanguine)

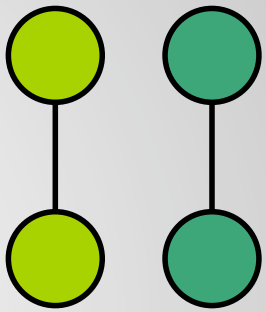
- On peut mesurer la perte de variation à l'intérieur d'une population par le coefficient de consanguinité

- Le coefficient de consanguinité (*F*), ou homozygotie, est la probabilité que deux gènes tirés au hasard sont des copies du même gène ancêtre

Dérive et consanguinité



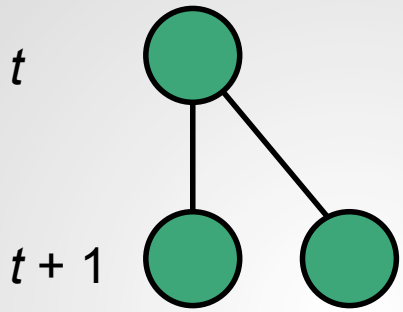
Deux gènes descendant d'un même parent

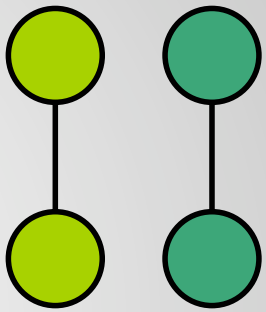


Deux gènes descendant de parents distincts

Dans une population diploïde, quelle est la valeur de *F* ?

Dérive et consanguinité





	Deux gènes descendant d'un même parent	Deux gènes descendant de parents distincts
Probabilité de l'évènement :	$1/(2N)$	$[1-1/(2N)]$
Probabilité d'identité :	1	$F[t]$

$$F[t+1] = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F[t]$$

Dérive et consanguinité

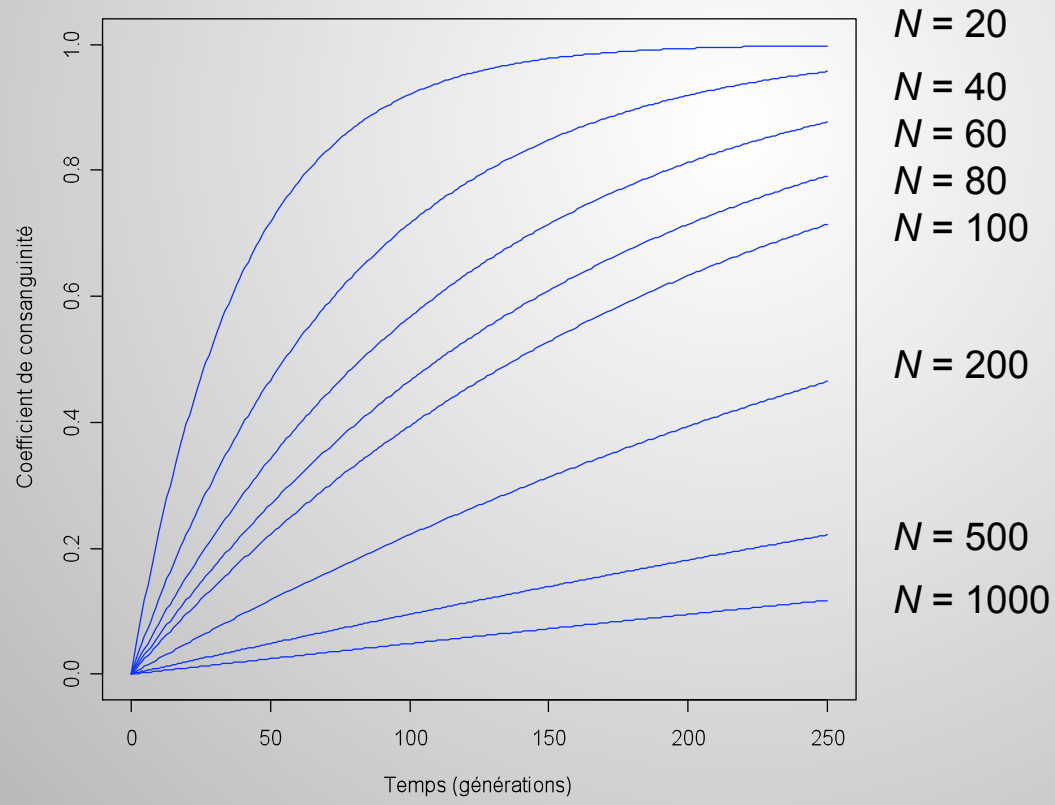
- Par récurrence, on obtient :

$$1 - F[t] = \left(1 - \frac{1}{2N}\right)^{t} (1 - F[0])$$

- $F[t]$ tend vers 1 lorsque t tend vers l'infini. La population tend à l'identité totale (fixation d'un allèle particulier). En l'absence de mutation, tous les gènes sont des copies identiques d'un gène ancêtre : identité par descendance

Dérive et consanguinité

- La population tend d'autant plus vite vers l'identité (homozygotie) qu'elle est de petite taille



Dérive et consanguinité

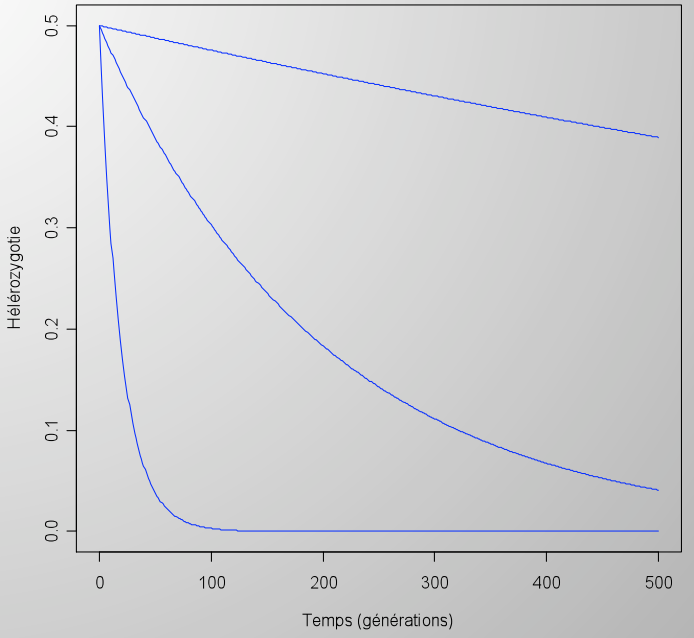
- Puisque l'homozygotie augmente, l'hétérozygotie diminue.
- On définit l'hétérozygotie attendue (*H*) comme la probabilité que deux gènes tirés au hasard dans la population soient différents allèles

$$H[t+1] = 1 - F[t+1] = \left(1 - \frac{1}{2N}\right) H[t]$$

- A l'équilibre :

$$H[t] = \left(1 - \frac{1}{2N}\right)^{t} H[0]$$

- Ce ne sont que des espérances !

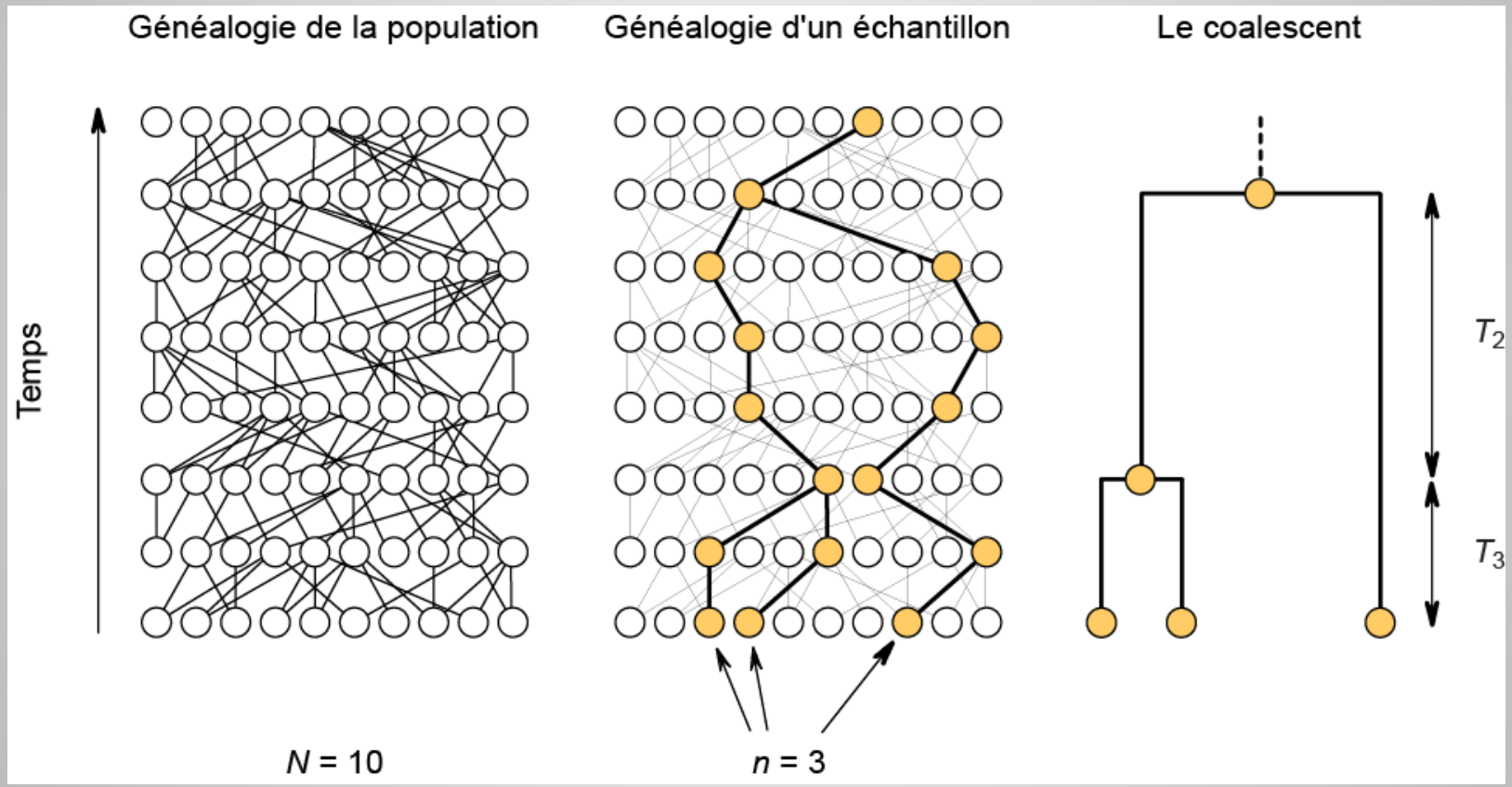


N = 1000

N = 100

N = 10

La coalescence



Regarder le processus de dérive en « remontant le temps » jusqu'à l'ancêtre commun d'un échantillon de gènes : à voir mardi 29 octobre

Effectif efficace

- On définit la taille efficace (notée N_e) d'une population comme étant la taille d'une population « idéale » de Wright-Fisher où la dérive génétique aurait la même intensité (*) que dans la population (ou bien le modèle de population) qui nous intéresse

- (*) *même taux de dérive, même augmentation de consanguinité, même augmentation de variance de fréquences alléliques entre populations, etc.*

Effectif efficace

- N_e ne représente donc le nombre d'individus reproducteurs de la population que dans le cas d'une population de Wright-Fisher, où un gène en particulier transmet un nombre de copies de lui-même tiré dans une loi binomiale de moyenne 1 et de variance (1 – 1/2*N*)

- De quel(s) facteur(s) dépend la taille efficace ?

Effectif efficace : sexe-ratio

- Sexe-ratio déséquilibrée $N_m \neq N_f$ (autosomes) :

- Quelle est la probabilité de 2 gènes pris au hasard soient identiques?
- Apres chaque événement de reproduction
- 1/4 (1/4) des paires de gènes proviennent de mâles (femelles), avec une probabilité $1/N_m$ ($1/N_f$) viennent d'un(e) même mâle (femelle), et ont une probabilité 1/2 de coalescer

$$2N_e = 2\frac{4N_mN_f}{N_m + N_f}$$

Effectif efficace : sexe-ratio

- Taille efficace des autosomes avec un sexe-ratio déséquilibrée $N_m \neq N_f$, $N = N_m + N_f$ (autosomes) :

- Quelle est la probabilité de 2 gènes pris au hasard aient le même gène parent?
- Ils doivent appartenir a 2 individus différents (sinon ils viennent d'un male et d'une femelle)
  - Prob = (N-1)/N
- Parmi les paires de gènes appartenant à 2 individus différents, ¼ de ces paires proviennent de 2 gènes parents présents chez un male à la génération précédente, ¼ à des femelles, le reste provient d'un male et d'une femelle.
- Prob = (N-1)/N*(1/4 + 1/4)
- Ils ont ensuite les probabilités respectives $1/N_m$ et $1/N_f$ de provenir d'un même individus parmi les males et femelles respectivement.
- Prob = $(N-1)/N*(1/(4N_m) + 1/(4N_f))$
- Et enfin une probabilité ½ d'avoir le même gène parent :
- Prob = $(N-1)/N*(1/(8N_m) + 1/(8N_f))$
- En considérant N suffisamment grand, on a donc Prob = $1/8 *(N_m + N_f)/N_mN_f$

Effectif efficace : sexe-ratio

- Par analogie avec la consanguinité dans une population de Wright-Fisher donnée par 1/(2*N*), on définie la taille efficace d'une population avec un sexe-ratio déséquilibrée comme le double de la moyenne harmonique des tailles de population de chaque sexe :

$$N_e = \frac{4 N_m N_f}{N_m + N_f}$$

- Exemple numérique : si $N_m = 100$ et $N_f = 250$, N_e = ? et N_e / N ?

Effectif efficace : sexe-ratio

- Sexe-ratio déséquilibrée (double de la moyenne harmonique) :

$$N_e = \frac{4N_m N_f}{N_m + N_f}$$

- Exemple numérique : si $N_m = 100$ et $N_f = 250$, $N_e = 285$, $N_e / N = 0.81$

Chromosomes sexuels : X

- Sur le chromosome X, (transmis des mères à leurs enfants et des pères à leurs filles) N_e est égale à :

- 1/9 (4/9) des paires de gènes proviennent de mâles (femelles), avec une probabilité $1/N_m$ ($1/N_f$) viennent d'un(e) même mâle (femelle), et ont une probabilité 1 (1/2) de coalescer chez les mâles (femelles)

$$2N_e = 2\frac{9N_mN_f}{4N_m + 2N_f}$$

- Si le sexe-ratio est équilibré, la taille efficace sur le chromosome X est égale à 3 /4 de celle sur les autosomes

Chromosomes sexuels : Y

- Sur le chromosome Y, (transmis des pères à leurs fils) N_e est égale à :

$$N_e = N_m$$

- 4 fois plus faible que sur les autosomes pour un sexe-ratio équilibré

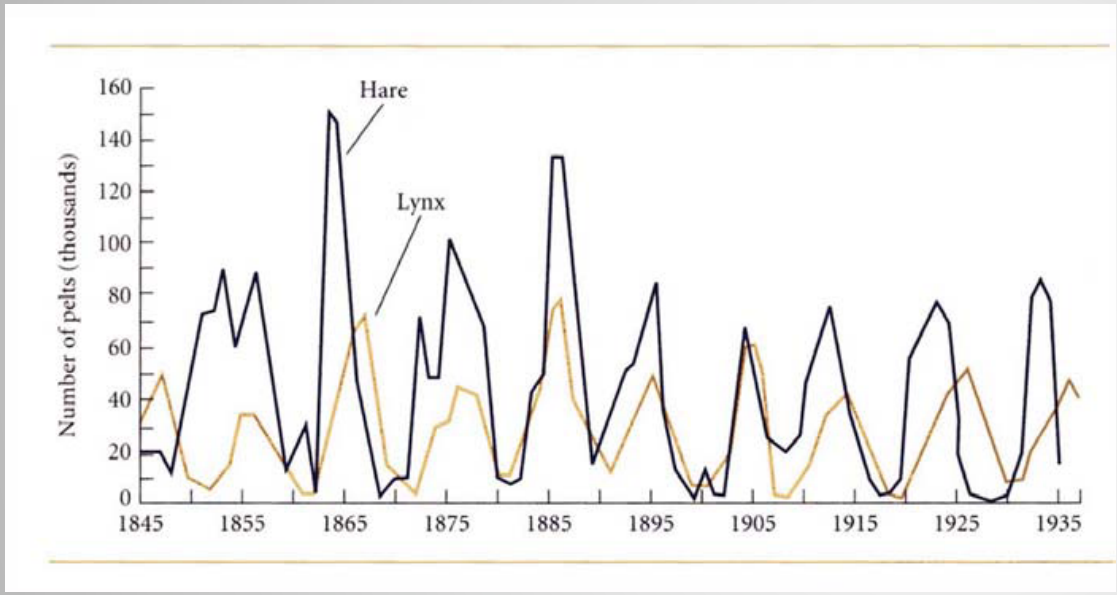
Succès reproducteur

- N_e dépend du nombre d'adultes reproducteurs mais aussi de la variance du nombre de descendants :
- La variance du succès reproducteur implique que certains individus ont plus de descendants et d'autres moins. Donc il y a une plus forte probabilité que des descendants aient le même parent…
- Pour des diploïdes :

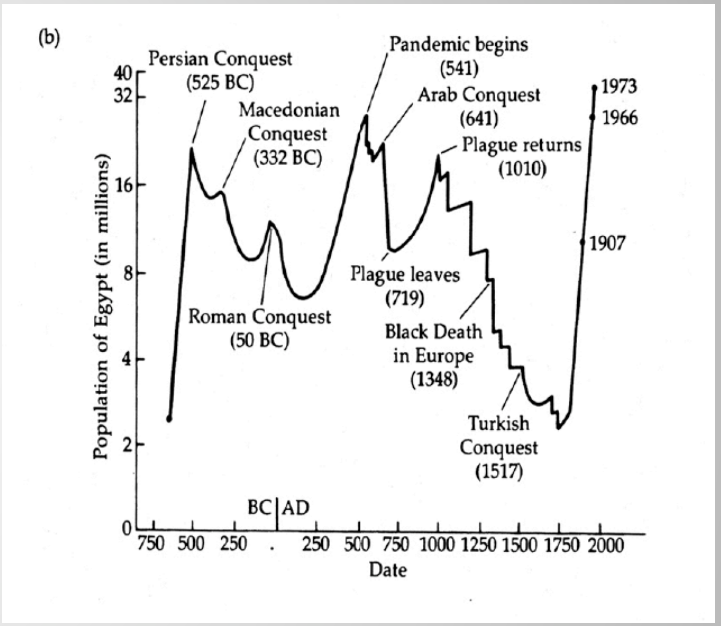
$$N_e \approx \frac{4N_{Adult.}}{2+\sigma^2_{desc.}}$$

- La variance du nombre de descendants diminue donc la taille efficace

Fluctuations démographiques



Effectifs du lynx canadien et du lièvre (basés sur le nombre de fourrures importées par la *Hudson Bay Company*)



Histoire démographique de l'Egypte, marquée par les invasions et la peste

Fluctuations démographiques

- Si l'on considère que $N = N(t)$ fluctue d'une génération à l'autre, alors à chaque génération la diversité décline ainsi :

$$1 - Q' = (1 - 1/N_t)(1 - Q)$$

- Sur T générations, on a : $\prod_{t=0}^{T-1} (1 - 1/N_t)$

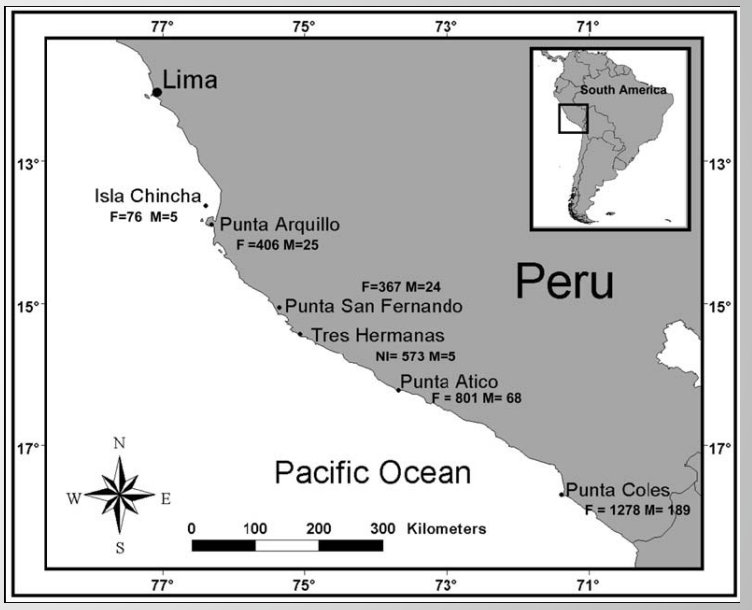
- S. Wright (1938) a défini implicitement une taille efficace telle que ce facteur soit égal à : $(1 - 1/N_e)^T$

- Si bien que la taille efficace peut être approximée par la moyenne harmonique des $N(t)$:

$$\frac{1}{N_e} \approx \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{N_t}$$

Fluctuations démographiques

Otarie à fourrure australe (Rosa de Oliveira *et al*. 2006)



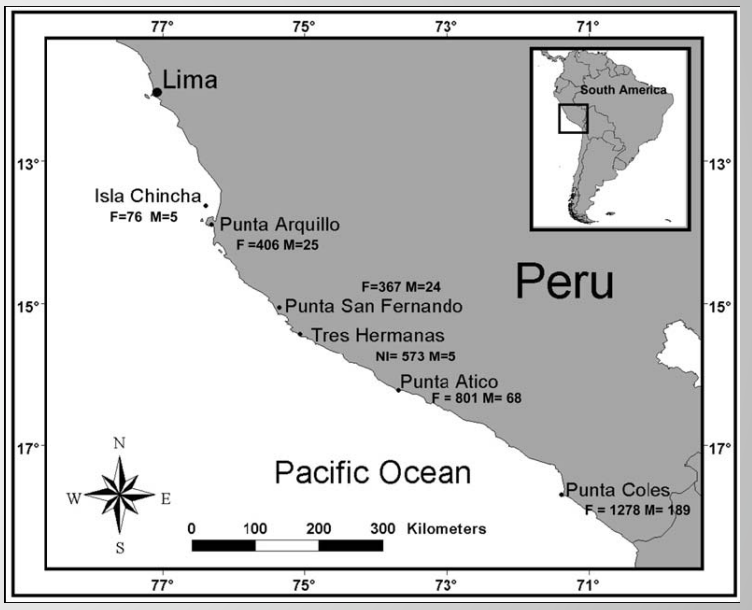
Avant El Niño (1996-97) : **24 481** individus (**10 720** femelles, **2 903** mâles)

Après El Niño (1999) : **8 223** individus (**3 215** femelles, **337** mâles)

Quelles sont les tailles efficaces dans chacun des cas, et globalement ?

Fluctuations démographiques

Otarie à fourrure australe (Rosa de Oliveira *et al*. 2006)



Avant El Niño (1996-97) : **9 138**

Après El Niño (1999) : **1 220**

Globalement : 1 076

Effectif efficace

Le génome des orangs-outans marqué par leur effondrement démographique(*) (Goossens *et al*. 2006)

(*) goulet d'étranglement, *bottleneck*

Paramètre important, notamment en génétique de la conservation

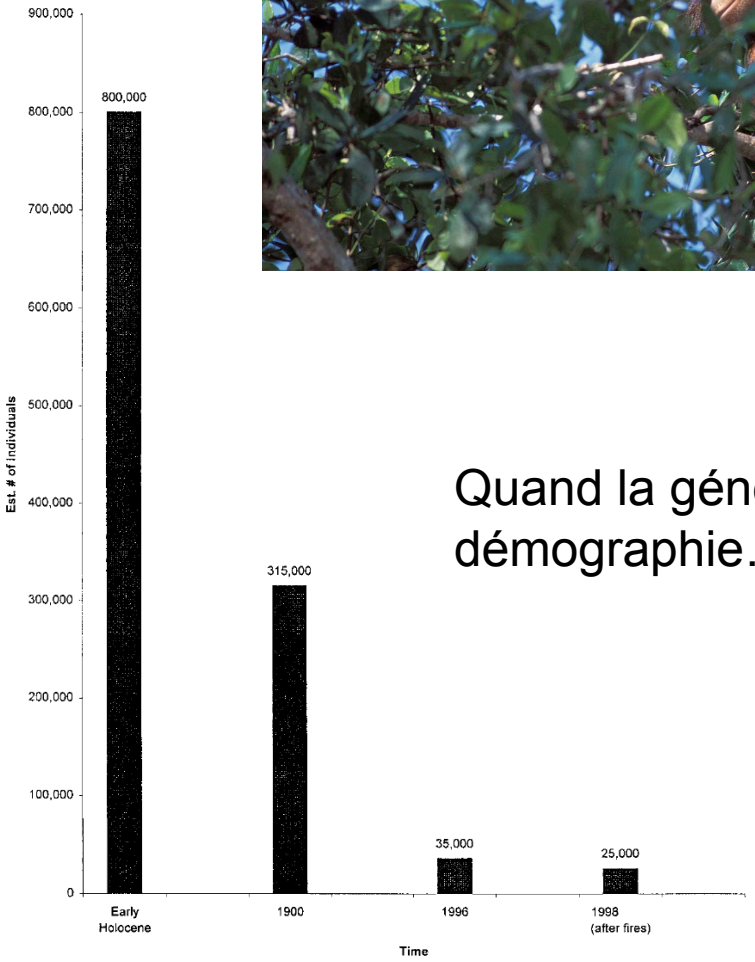


Figure 5. Total estimated orangutan populations since the early Holocene. Data are taken from Rijksen & Meijaard.[48]

Quand la génétique révèle la démographie…

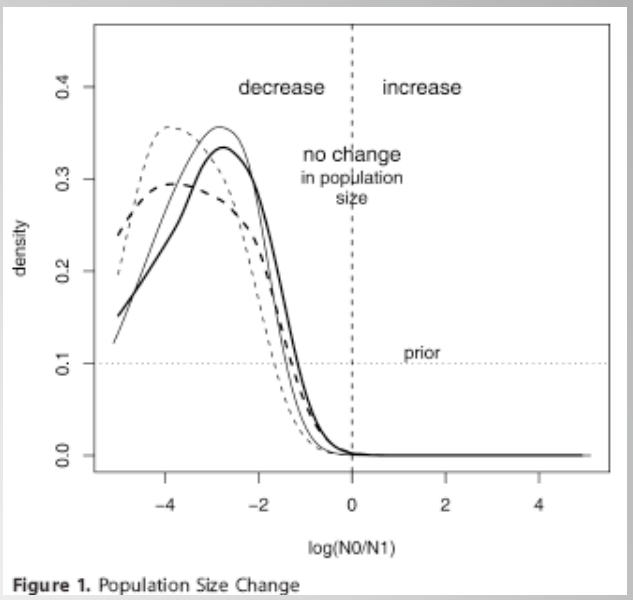


Figure 1. Population Size Change

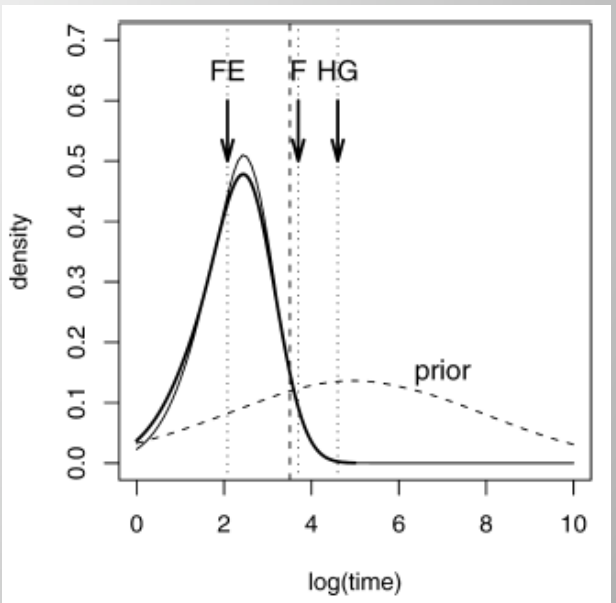


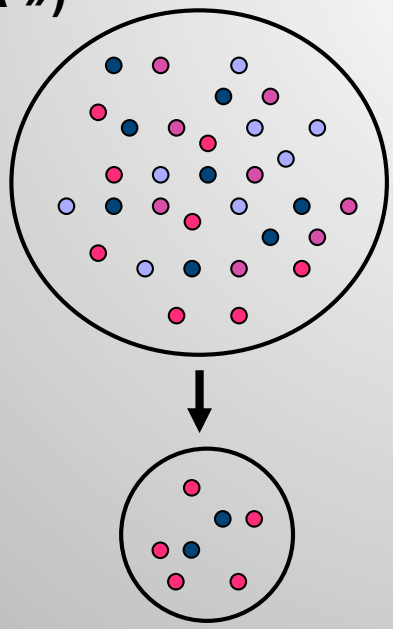
Figure 3. Time since the Population Collapse

Effets fondateurs

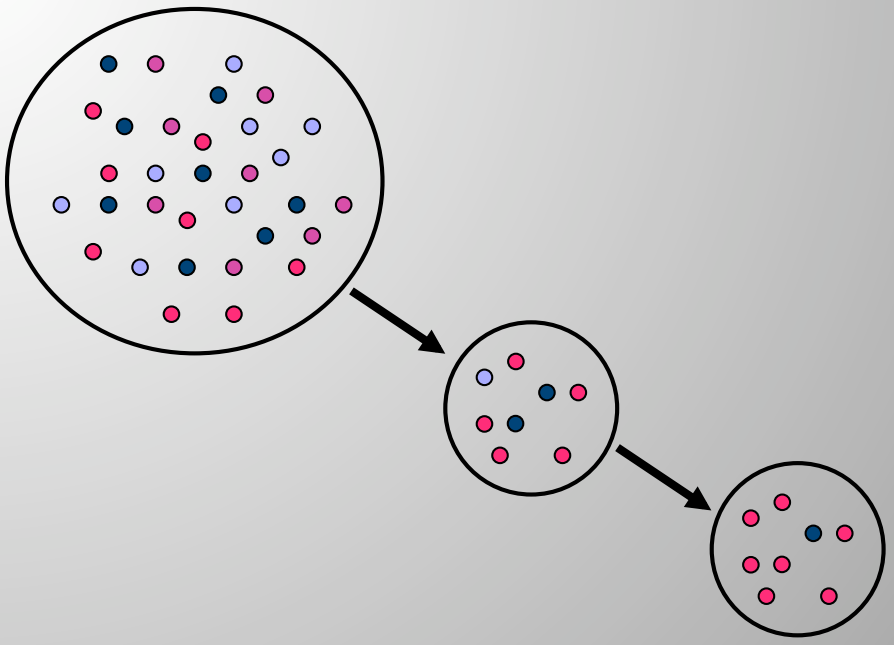
Ces réduction d'effectif efficace se traduisent par des « effets fondateurs », qui font que les populations nouvellement établies portent une fraction de la variabilité génétique de la population ancestrale.

On peut envisager un effet fondateur dans deux situations :

goulet d'étranglement (« *bottleneck* »)



Colonisations successives

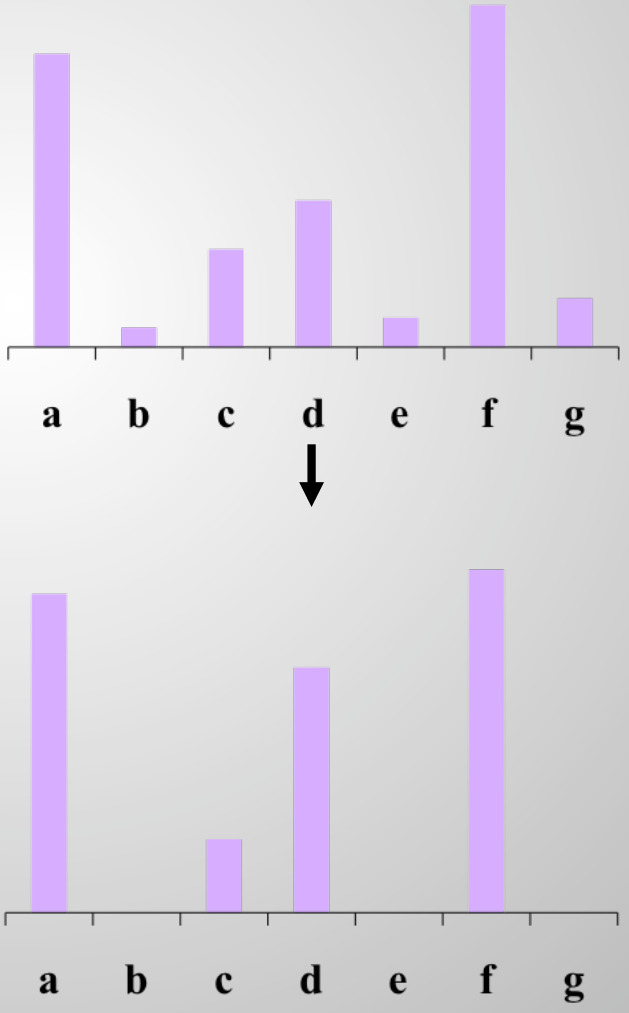


Effets fondateurs

Après un goulot d'étranglement

n_A

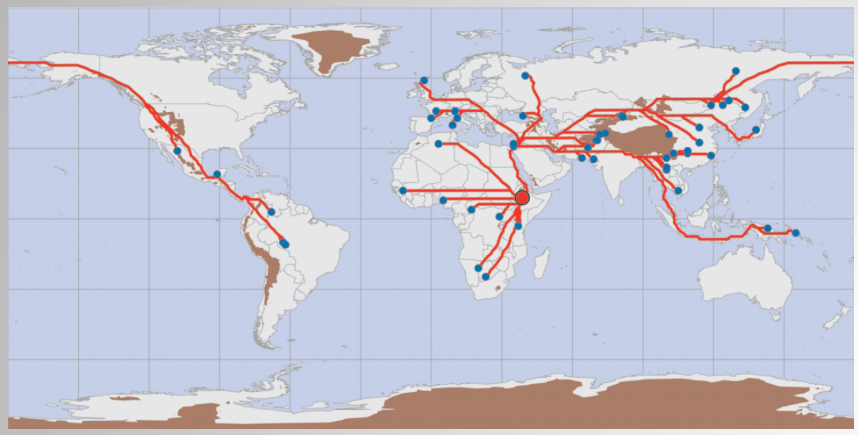
Diminue plus vite que H_e car les allèles rares sont éliminés les premiers.



$n_A = 7$
$H_e = 0.75$

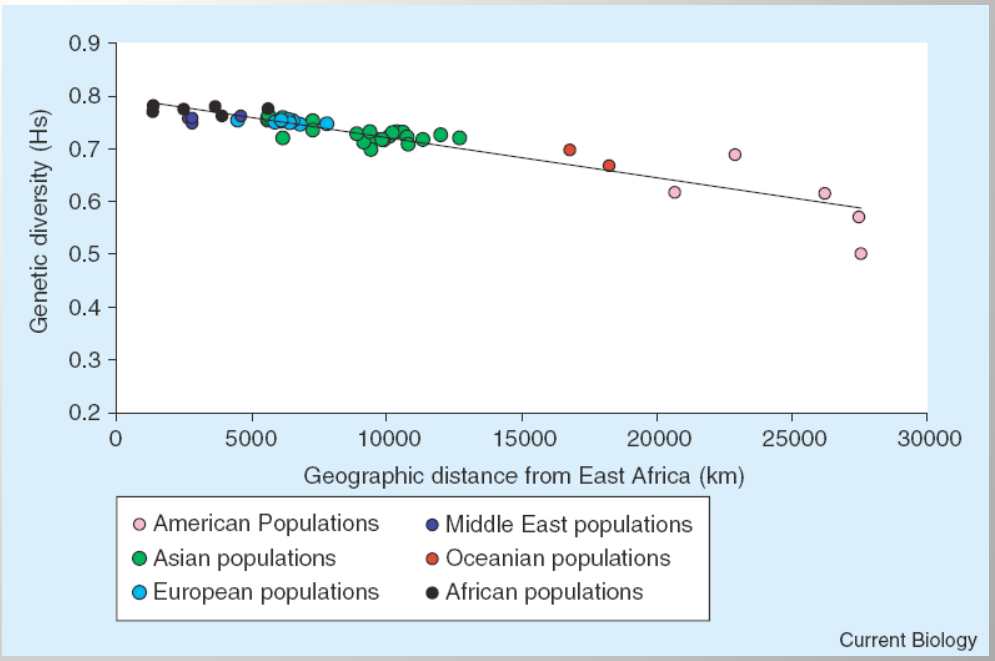
$n_A = 4$
$H_e = 0.70$

Effets fondateurs



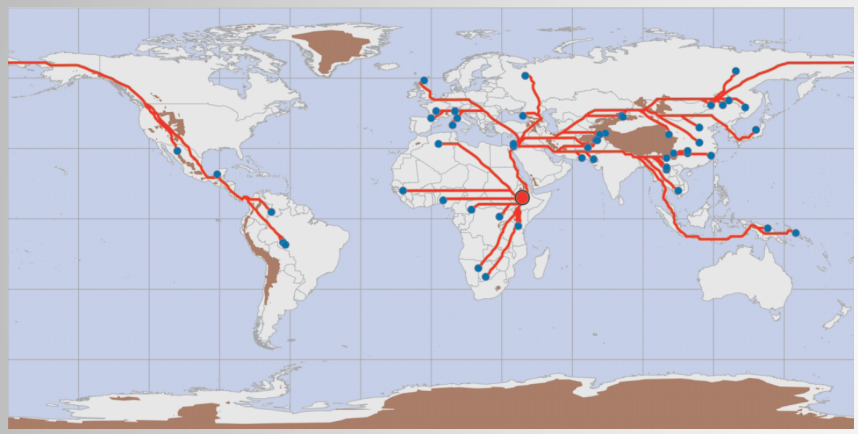
Distance à l'Afrique (Addis Abeba)

L'hétérozygotie diminue en s'éloignant de l'Afrique (Prugnolle *et al*. 2005)

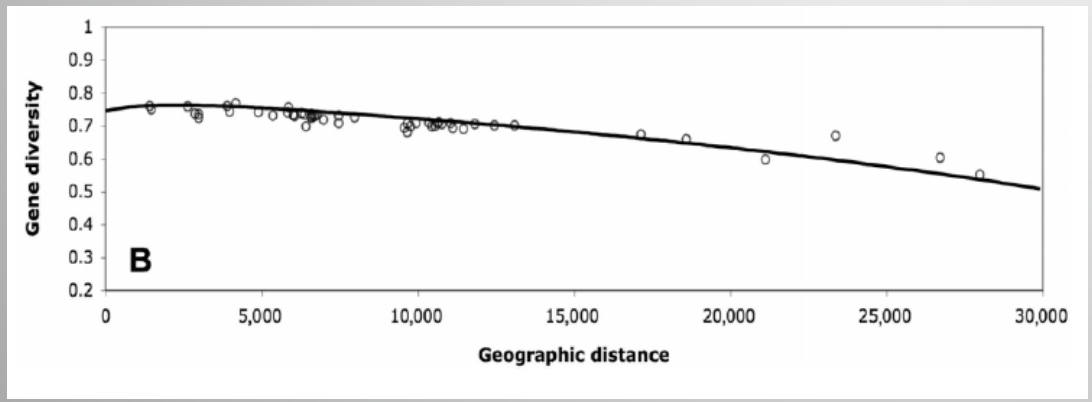


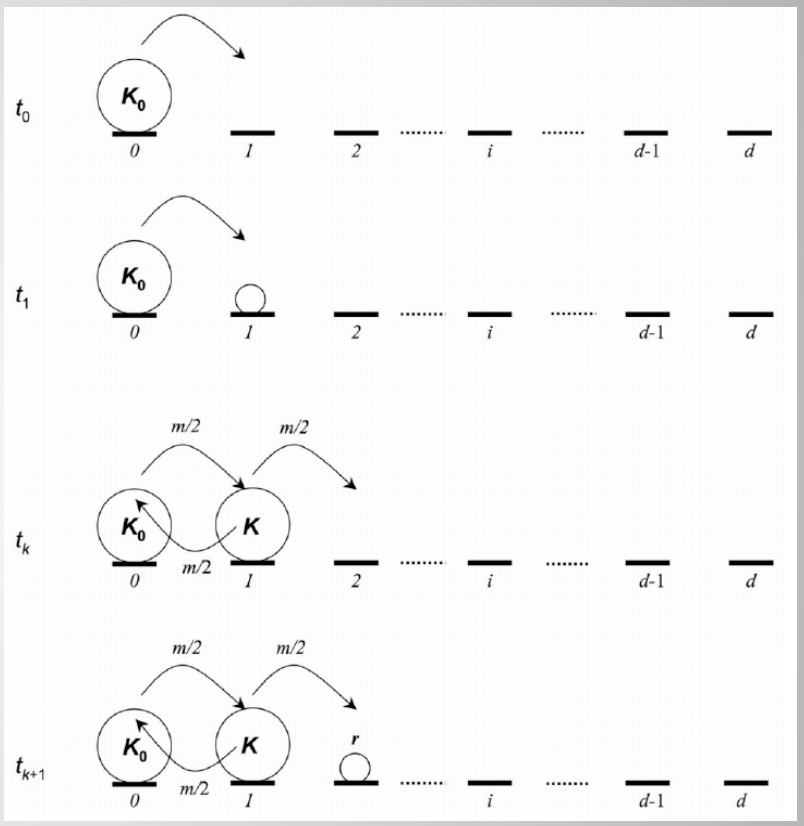
Effets fondateurs

A partir d'un modèle de colonisations successives (Liu *et al*. 2006)…



… on retrouve les résultats observés sur les données réelles :





Dérive et mutation

- Mais la perte de diversité génétique due à la dérive peut être compensée par l'arrivée de nouvelles mutations

- 2 gènes tirés au hasard ne sont pas identiques par descendance si l'un des deux (ou les deux) mute(nt) vers un nouvel état allélique

$$F[t+1] = \left[\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F[t]\right](1-\mu)^2$$

- A l'équilibre :

$$\hat{F} \approx \frac{1}{1+4N\mu}$$

Dérive et mutation

- Puisque :

$$\hat{H} = 1 - \hat{F}$$

- On a :

$$\hat{H} \approx \frac{4N\mu}{1 + 4N\mu}$$

- Que l'on écrit aussi, en posant $\theta = 4N\mu$:

$$\hat{H} \approx \frac{\theta}{1 + \theta}$$

Dérive et mutation

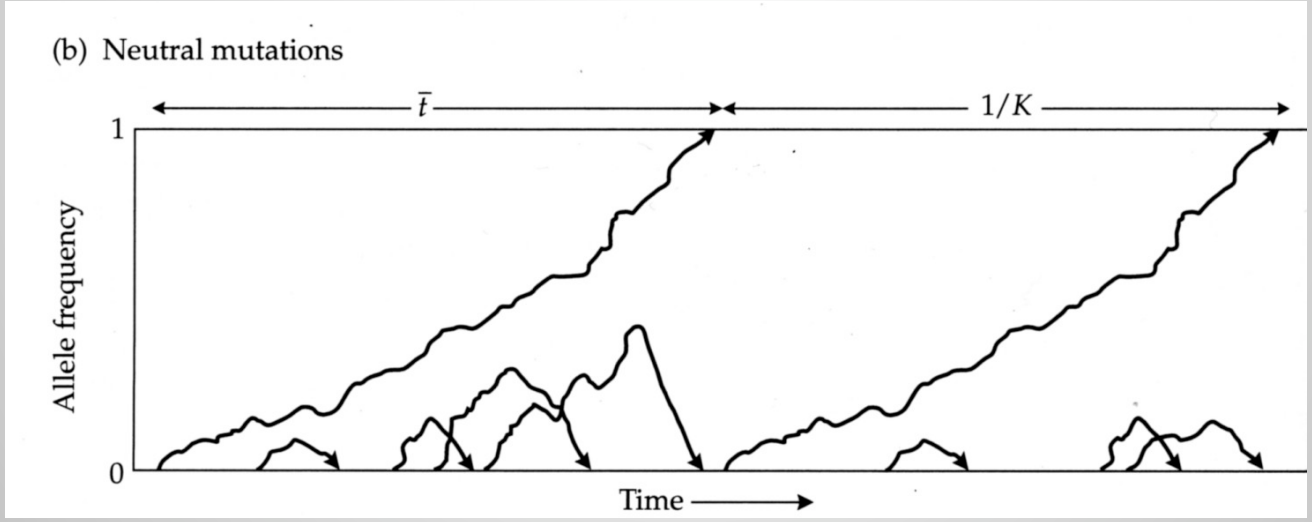
- En espérance, la dérive et la mutation maintiennent un niveau de variation intermédiaire à l'équilibre.
- Mais chaque mutation qui apparaît peut être fixée ultimement par dérive…
- Il y a donc un renouvellement constant des allèles

La théorie neutraliste

- Soit une population diploïde de taille N
- Soit un taux de mutation récurrent μ
- On a donc $2N\mu$ nouvelles mutations qui apparaissent par génération
- Chaque nouvelle mutation neutre a une fréquence initiale de $1/2N$, donc une probabilité de fixation de $1/2N$
- Par génération on a donc $K = 2N\mu / 2N = \mu$ mutations qui se fixeront (taux de substitution nucléotidique)
- En moyenne le temps de fixation est de $t = 4N_e$ générations (approximations de diffusion)

La théorie neutraliste

- Toutes les 1/μ générations, apparaît une mutation destinée à se fixer : les taux d'évolution sont constants



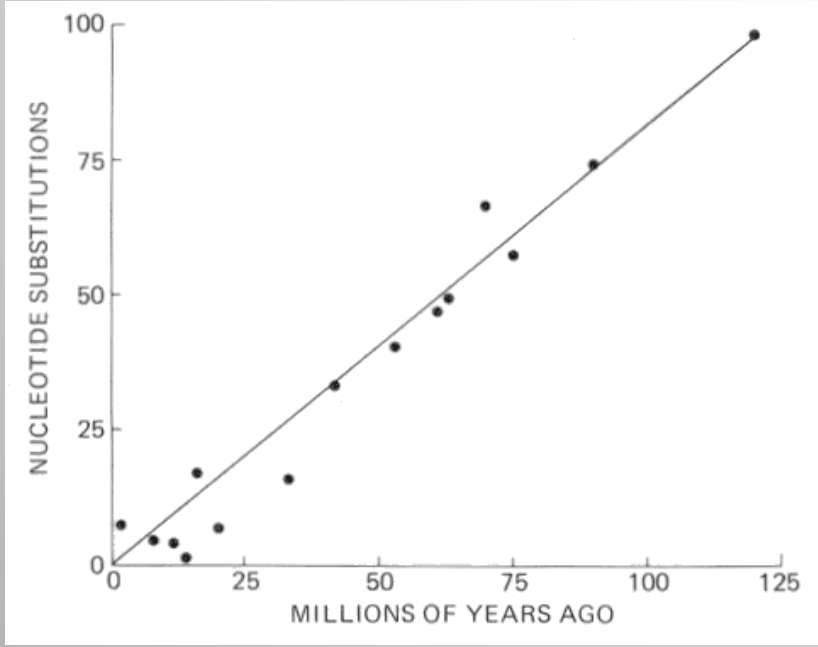
L’horloge moléculaire

Prédiction : le nombre de substitutions (mutations) dans le génome augmente linéairement avec le temps

Motoo Kimura

L’horloge moléculaire

- En comparant les gènes de l’α-globine chez des vertébrés, Motoo Kimura (1983) a montré que le nombre de différences (substitutions nucléotidiques) entre paires d’espèces vérifie cette prédiction

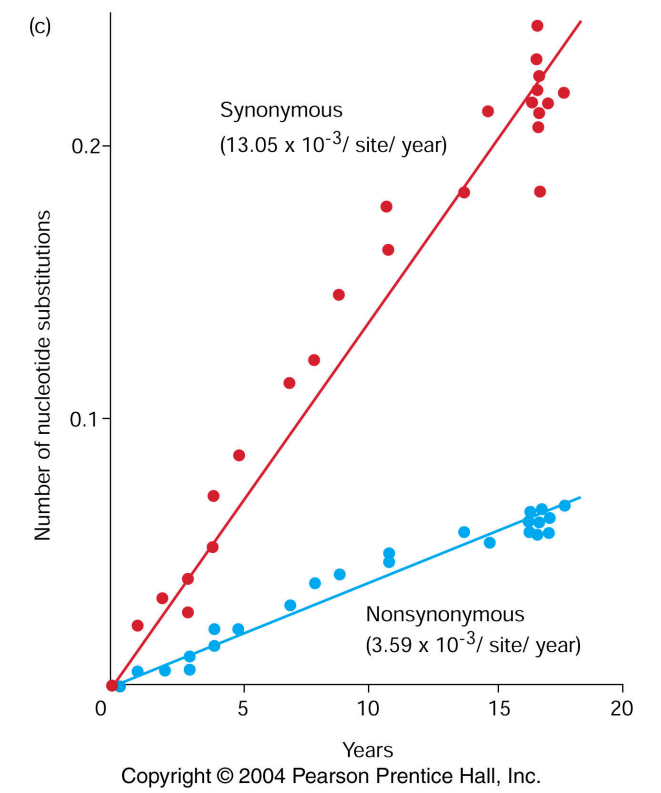


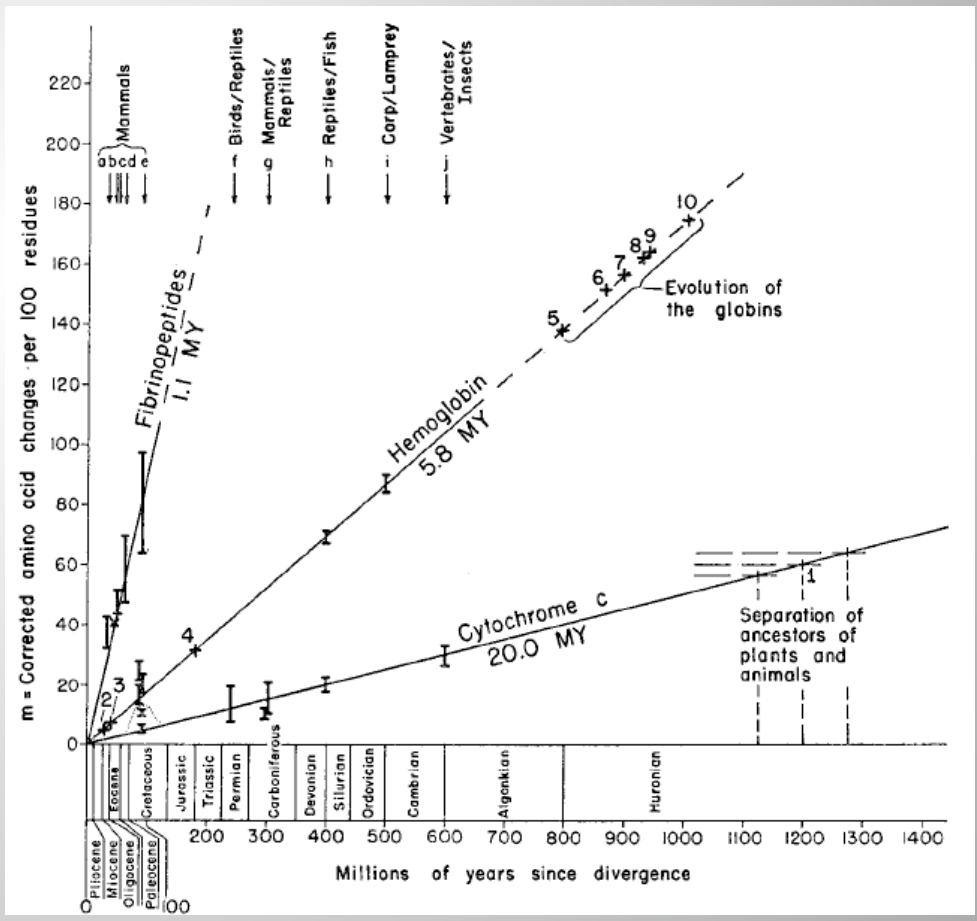
Le nombre de mutations qui séparent deux séquences fournit ainsi une mesure du temps depuis lequel elles divergent (Kimura 1983)

L'horloge moléculaire

Mais : l'horloge avance à un rythme différent selon les gènes, ou selon la nature des sites (taux de mutations différents, ou contraintes fonctionnelles)

(Données : BRCA 1)





(Dickerson 1971)

Evolution des fréquences alléliques en populations naturelles

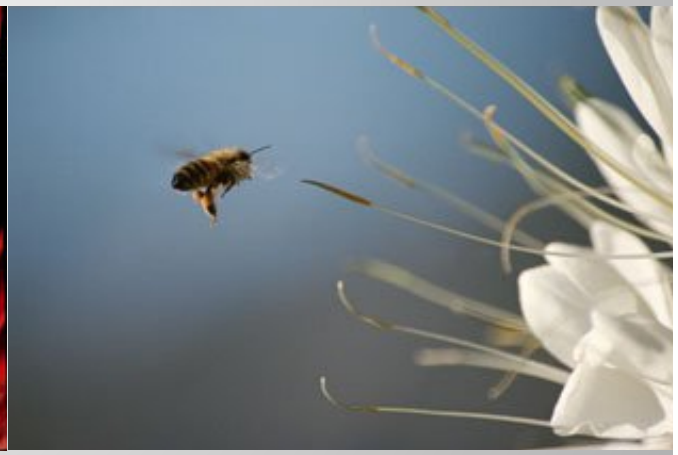
Il existe 4 "forces évolutives" qui agissent en interactions et font évoluer les fréquences alléliques en populations naturelles :

- **la mutation**
- **la dérive génétique (variations stochastiques des fréquences alléliques dues aux effets d'échantillonnages)**
- **la migration (ou flux de gènes)**
- **la sélection naturelle**

La migration

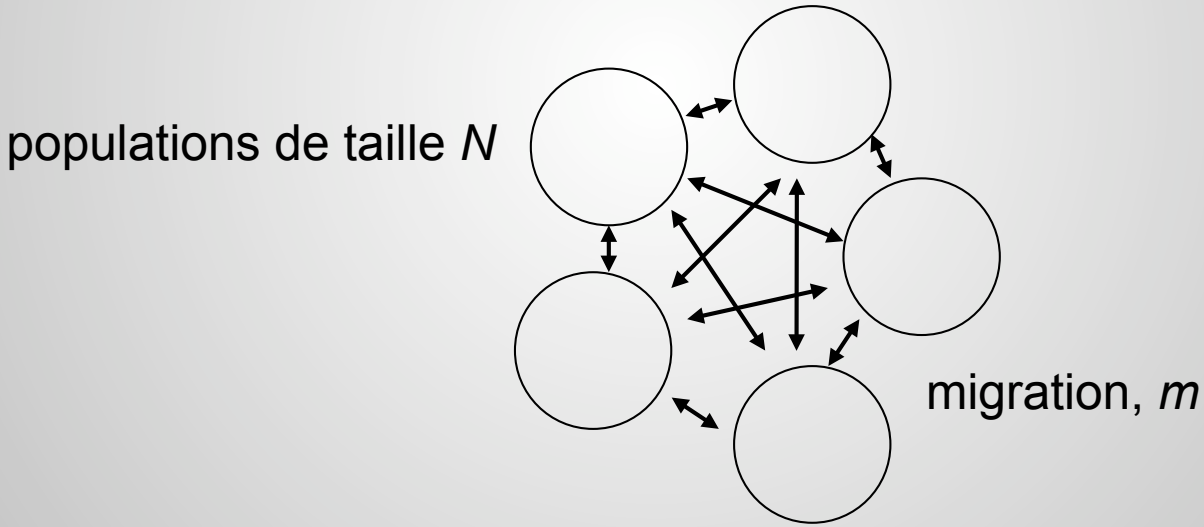
L’échange d’individus (ou de gamètes) entre sous populations permet les flux de gènes…





La migration

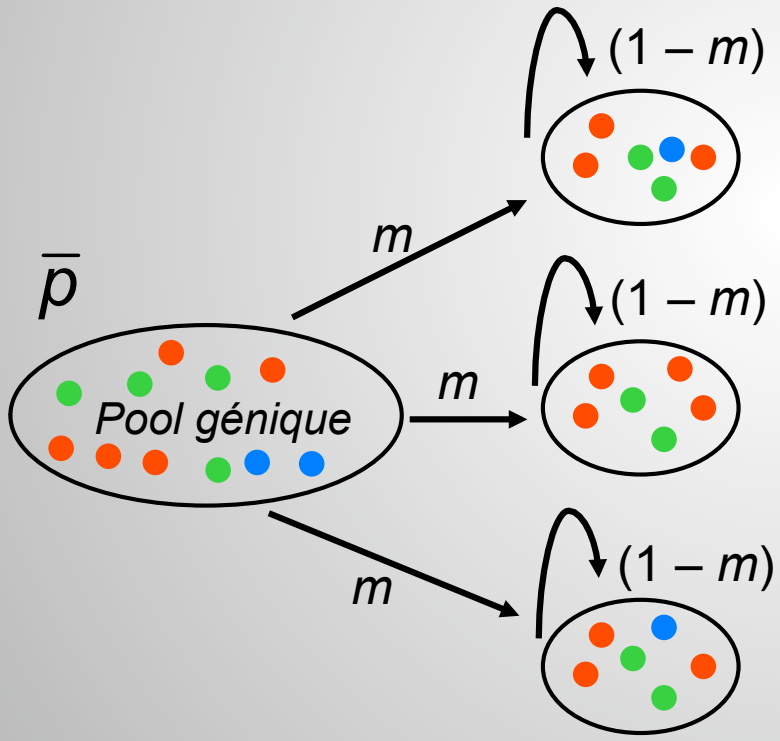
- Le modèle en îles (ou modèle de l'archipel) considère que la migration s'effectue entre toutes les sous-populations (dèmes) d'une population subdivisée.



- Ce modèle n'est pas spatialisé : tous les dèmes échangent des migrants au même taux, quelle soit leur « *position* »

La migration

- Si le nombre de dèmes est suffisamment grand (infini), la fréquence des allèles parmi les migrants est constante.

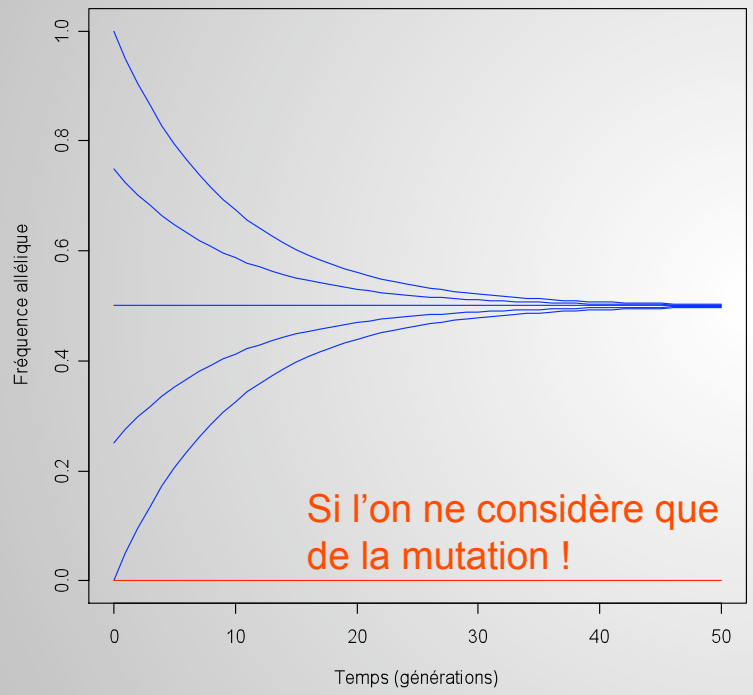


$$p[t+1] = (1-m)p[t] + m\overline{p}$$

Si l'on part de p[0], au bout de *t* générations :

$$p[t] = \overline{p} + (1-m)^t (p[0] - \overline{p})$$

La migration



$\bar{p}$

- Évolution des fréquences au cours du temps dans cinq dèmes qui échangent des migrants au taux m = 0.1 par génération
- Avec p[0]=0, combien de générations faut-il pour atteindre 90% de la valeur d'équilibre ? **22** (**1.15×10^6** pour la mutation !)

- En l'absence d'autres forces évolutives, la migration homogénéise complètement les fréquences alléliques entre dèmes.

La migration

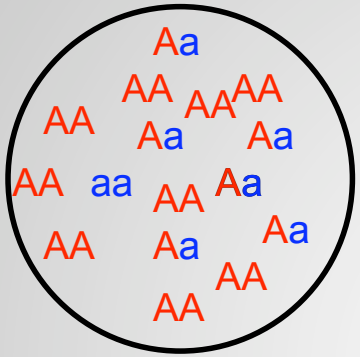
- C’est une force de plus forte intensité que la mutation
- La migration homogénéise les fréquences alléliques entre populations
- Le modèle de migration (la façon dont les individus se déplacent dans l’espace, la dispersion « en groupe » ou solitaire, etc.) influence beaucoup la distribution spatiale du polymorphisme

Mélanges de population panmictiques : l'effet wahlund (Rappel)

Pop1 (HW)

p_A=0.75

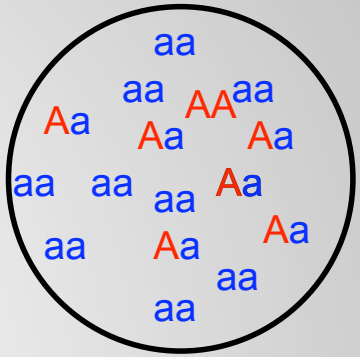
p_a=0.25



Pop2 (HW)

p_A=0.25

p_a=0.75



Fréquences observées

p_{AA} = 9 / 16 = 0.5625

p_{Aa} = 6 / 16 = 0.375

p_{aa} = 1 / 16 = 0.0625

p_{AA} = 1 / 16 = 0.0625

p_{Aa} = 6 / 16 = 0.375

p_{aa} = 9 / 16 = 0.5625

Fréquences attendues

p_{AA} = 0.75^2 = 0.5625

p_{Aa} = 2*0.75*0.25 = 0.375

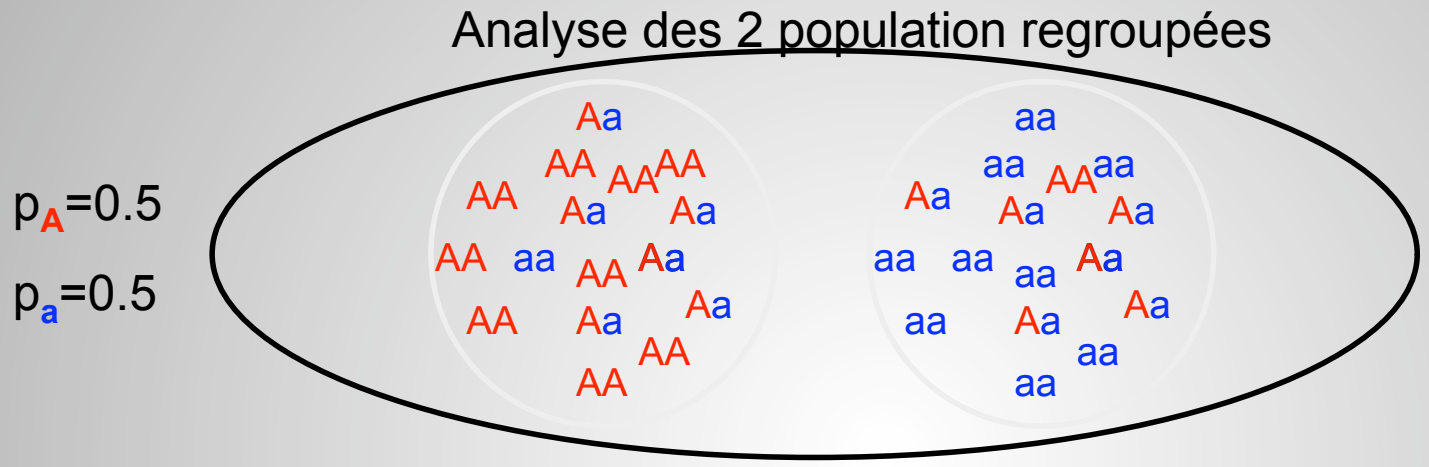
p_{aa} = 0.25^2 = 0.0625

p_{AA} = 0.25^2 = 0.0625

p_{Aa} = 2*0.75*0.25 = 0.375

p_{aa} = 0.75^2 = 0.5625

Mélanges de population panmictiques : l'effet wahlund (Rappel)



Fréquences observées

p_{AA} = 10 / 32 = 0.3125

p_{Aa} = 12 / 32 = 0.375

p_{aa} = 10 / 32 = 0.3125

Fréquences attendues

p_{AA} = 0.5^2 = 0.25

p_{Aa} = 2*0.5*0.5 = 0.5

p_{aa} = 0.5^2 = 0. 25

On observe un déficit d'hétérozygote (et donc un excés d'homozygote) par rapport à l'équilibre de Hardy-Weinberg : c'est l'effet Walhund (1923)

Un mélange de population (sous-populations) panmictiques n'est pas une population panmictique à cause de l'effet de la structuration et des flux de gènes (migration entre populations) limités

Formalisation de l'analyse de populations subdivisées (i.e. structurées)

Considérons *n* populations panmictiques et un locus bi-allélique avec Fréq[A]=p_i, et Fréq[a]=q_i dans chaque population *i*

les fréquences génotypiques dans chaque population sont à l'équilibre de Hardy-Weinberg :

AA	p_i^2
Aa	$2p_iq_i$
aa	q_i^2

Et en moyenne sur l'ensemble des populations :

AA	$E(p_i^2)$
Aa	$E(2p_iq_i)$
aa	$E(q_i^2)$

Formalisation de l'analyse de populations subdivisées (i.e. structurées)

Soit $p=E(p_i)=\frac{1}{n}\sum_{i=1}^{n} p_i$, si la population totale était panmictique on observerait sur l'ensemble des populations:

AA	p^2
Aa	$2pq$
aa	q^2

mais la fréquence d'hétérozygote réellement observée (H_o) est:

$H_o = E(2p_iq_i) = 2 * E(p_i - p_i^2) = 2 * E(p_i) - 2 * E(p_i^2)$

$= 2p - 2 * (Var(p) + p^2)$ car $Var(p) = E(p_i^2) - E(p_i)^2 = E(p_i^2) - p_i^2$

$= 2pq - 2 * Var(p) = 2pq (1 - 2Var(p)/2pq)$

Formalisation de l'analyse de populations subdivisées (i.e. structurées)

Si la population totale était panmictique on observerait sur l'ensemble des populations:

AA	p^2
Aa	$2pq$
aa	q^2

mais la fréquence d'hétérozygote réellement observée (H_o) est:

$H_o = 2pq\ (\ 1 - 2\mathrm{Var}(p)/2pq)$

On note $F_{ST} = \mathrm{Var}(\ p\ ) / pq$, où Var(p) est la variance de p entre population, on a alors

$$H_o = \mathrm{E}(\ 2p_iq_i\ ) = 2pq(\ 1 - F_{ST}\ )$$

Formalisation de l'analyse de populations subdivisées (i.e. structurées)

On observe donc la structure génotypique suivante, correspondant a Hardy-Weinberg généralisé à un ensemble de populations panmictiques :

- Fréq[AA] = E(p_i^2) = $p^2 - pq*F_{ST}$
- Fréq[Aa] = E($2p_iq_i$) = $pq(1 - F_{ST})$
- Fréq[aa] = E(q_i^2) = $q^2 - pq*F_{ST}$

F_{ST} peut donc être perçu comme le déficit en hétérozygote du aux échanges limités par flux de gènes/migration entre différentes populations (i.e. l'écart a la panmixie entre les populations), c'est l'effet wahlund

Formalisation de l'analyse de populations subdivisées (i.e. structurées)

$$F_{ST} = \mathrm{Var}(p)/pq = 1 - H_o/H_e$$

où H_o est l'hétérozygotie observée et H_e l'hétérozygotie attendue si l'on avait une seule population

On note que pq est la variance maximale des fréquence entre population ($\mathrm{Var}_{max}(p)$) obtenue si toute les populations sont fixées.

On a alors p populations fixés pour A et q populations fixées pour a,

d'où $\mathrm{Var}(p) = \mathrm{E}(p_i^2) - \mathrm{E}(p_i)^2 = p*1^2 + 0 - p^2 = p(1 - p) = pq$

F_{ST} est donc la variance des fréquences alléliques entre populations, standardisées par la variance maximale

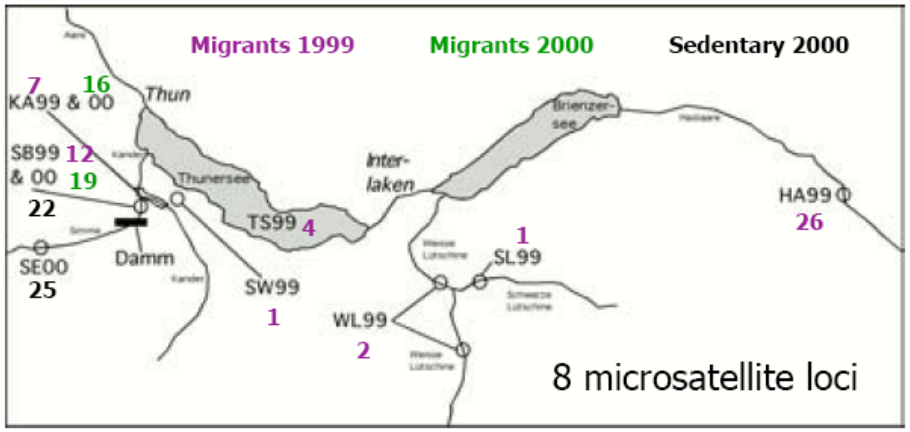
Formalisation de l'analyse de populations subdivisées (i.e. structurées)

On note aussi que *pq* est la variance totale (sur l'ensemble de toutes les populations) des fréquence alléliques ($Var_{tot}(p)$ obtenue quand on regroupe les populations). On a alors *p* allèles A et *q* allèles a, d'où $Var_{tot}(p) = p*1^2 + 0 - p^2 = pq$ (Bernoulli)

F_{ST} est donc aussi la proportion de la variance totale qui se trouve entre populations (analyse de variance)

Formalisation de l'analyse de populations subdivisées (i.e. structurées)

Genetic structure of the brown trout in Lake Brienz and Lake Thun



Pooled samples Lake Thun years 1999 and 2000

$F_{IS} = 0.05$

P-value = 0.009

Pooled samples Lake Thun and Lake Brienz 1999

$F_{IS} = 0.044$

P-value = 0.016

Wahlund effect

Comparison between Simme and Kander rivers in 2000

$F_{ST} = -0.001$

$F_{IS} = 0.06$

$F_{IT} = 0.053$

Differentiation between Lake Thun and Lake Brienz 1999

$F_{ST} = 0.049$

P-value = 0

About 5% of the total variance is due to differences beween the two lakes

Formalisation de l'analyse de populations subdivisées (i.e. structurées)

Considérons maintenant *n* populations ayant un régime de reproduction non panmictique (i.e. partiellement consanguin, $F_{IS}>0$), la structure génotypique d'une population est toujours données par :

AA	$p_i^2+p_iq_iF_{IS\,i}$
Aa	$2p_iq_i\,(1-F_{IS\,i})$
aa	$q_i^2+p_iq_iF_{IS\,i}$

Et sur toutes l'ensemble des populations, on observe :

AA	$E[\,p_i^2+p_iq_iF_{IS\,i}\,]$
Aa	$E[\,2p_iq_i\,(1-F_{IS\,i})\,]$
aa	$E[\,q_i^2+p_iq_iF_{IS\,i}\,]$

Formalisation de l'analyse de populations subdivisées (i.e. structurées)

On suppose l'indépendance entre F_{ISi} et p_i, on a alors

$$H_o = E\,[\,2p_iq_i\,(1-F_{IS\,i})\,] = E\,[\,2p_iq_i\,]^*\,E[\,(1-F_{IS\,i})\,]$$

Si on pose $F_{IS}=E\,[F_{ISi}]$ et, comme on a vu précédement, $E\,[\,2piqi\,]=2pq(1-F_{ST})$

On a donc

$$H_o = \text{Freq}[Aa] = 2pq(1-F_{ST})(1-F_{IS})\,]$$

La loi de Hardy-Weinberg généralisée à *n* populations de régime de reproduction consanguin donne les proportions génotypiques suivantes :

AA	$p^2+pq(F_{IS}+F_{ST}+F_{IS}F_{ST})$
Aa	$2pq\,(1-F_{IS})\,(1-F_{ST})$
aa	$q^2+pq(F_{IS}+F_{ST}+F_{IS}F_{ST})$

Formalisation de l'analyse de populations subdivisées (i.e. structurées)

F_{ST} est donc :

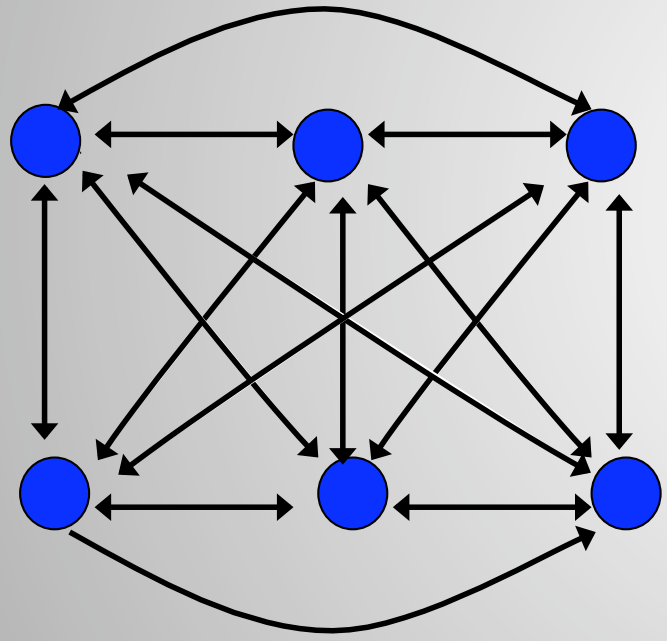
1. le déficit en hétérozygote du aux échanges limités par flux de gènes/migration entre différentes populations (i.e. l'écart a la panmixie entre les populations)
2. la variance des fréquences alléliques entre populations crée par la dérive et ou la migration faible, standardisées par la variance maximale
3. la proportion de la variance totale qui se trouve entre populations

FST mesure donc la différenciation entre les populations

Mais quelle sont l'influence des paramètres populationnels (tailles de populations (N), taux de migration (m), temps de divergence) sur les valeures de F_{ST}?

La migration : le modèle en îles

Le modèle en îles, ou modèle de l'archipel, considère que la migration se fait de façon homogène entre toutes les sous-populations d'une population subdivisée



Simple car homogèneité réduit à 3-5 le nombre de paramètres :

n_d= nombre de sous-populations (ou ∞)

N = taille des sous-populations

m= taux de migration

μ= taux de mutation

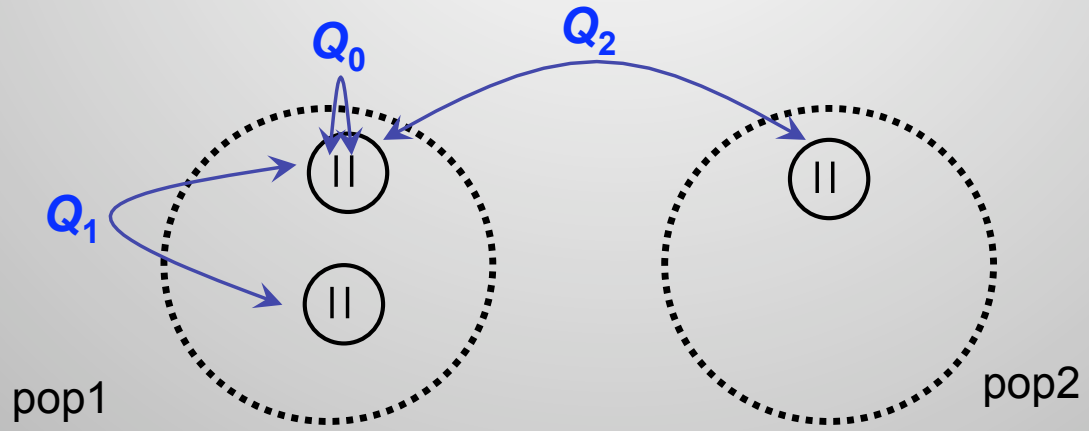
s= taux d'autofécondation (ou 1/N)

Avantage principal : Modèle homogène simple avec peu de paramètres -> analyse mathématique relativement simple

Probabilités d'identités : définitions

On défini des probabilité d'identité Q entre paires de gènes homologues (i.e. à un même locus) :

- Q_0 pour la probabilité que 2 gènes pris dans un même individu soient identiques
- Q_1 pour la probabilité que 2 gènes pris dans une même population soient identiques
- Q_2 pour la probabilité que 2 gènes pris dans deux populations différentes soient identiques



Probabilités d'identités et *F*-statistiques

On défini la relations entre *F*-statistiques et probabilités d'identités :

- $F_{IS} \equiv \frac{Q_0 - Q_1}{1 - Q_1}$
- $F_{ST} \equiv \frac{Q_1 - Q_2}{1 - Q_2}$
- $F_{IT} \equiv \frac{Q_0 - Q_2}{1 - Q_2}$

On retrouve bien la relation (Wright, 1943) :

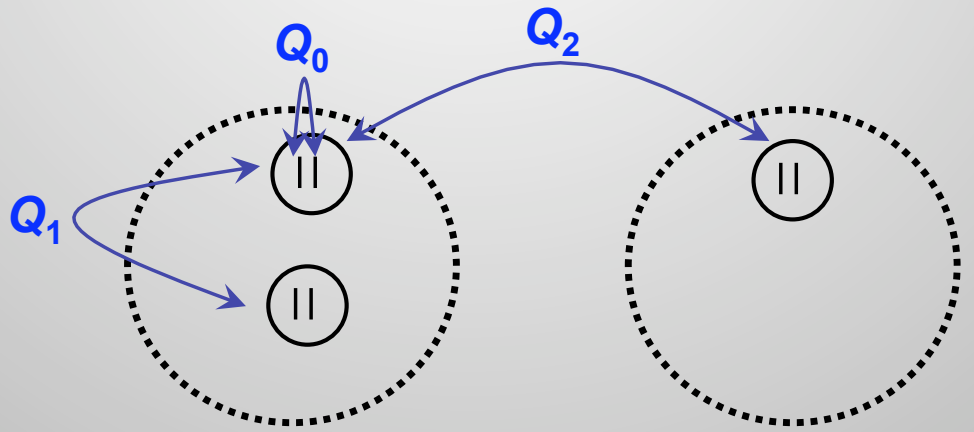
$$(1 - F_{IT}) \equiv (1 - F_{IS})(1 - F_{ST})$$

Probabilités d'identités et *F*-statistiques

Relation entre *F*-statistiques et probabilités d'identités : on retrouve les résultats précédents car

$$F_{ST} \equiv \frac{Q_1 - Q_2}{1 - Q_2} = \frac{(1 - Q_2) - (1 - Q_1)}{1 - Q_2}$$

$$F_{ST} \equiv 1 - \frac{E(2 p_i q_i)}{2 p q} = 1 - H_o / H_e$$



Calcul des Probabilités d'identités : formules de récurrences

On cherche a calculer l'évolution des probabilités d'identités dans le temps en fonction des paramètres démo-génétiques du modèle (e.g. migration, mutation, tailles de pops), afin d'en prendre ensuite les valeurs à l'équilibre.
On cherche donc a résoudre le système d'équation de récurrence suivant :

- $Q_0(t+1)=f(Q_0(t), Q_1(t), Q_2(t))$
- $Q_1(t+1)=f(Q0(t), Q1(t), Q2(t))$
- $Q_2(t+1)=f(Q0(t), Q1(t), Q2(t))$

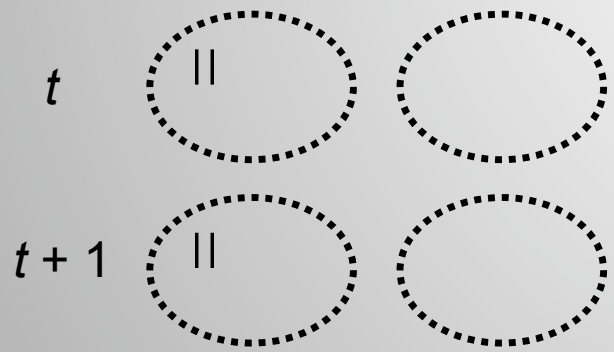
en évaluant tous les événements possibles en une génération pouvant agir sur ces probabilités

Calcul des Probabilités d'identités : principe des formules de récurrences

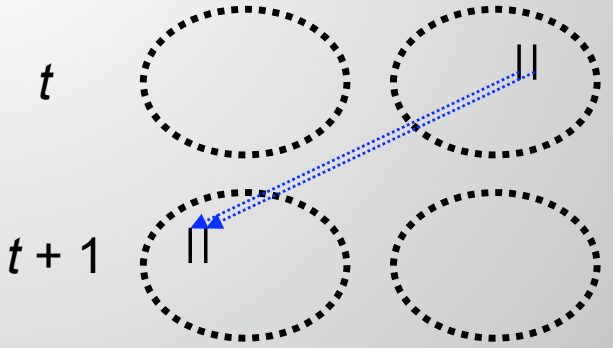
Étape 1 : on s'intéresse aux probabilité d'identité entre paires de gènes, il faut donc que les 2 gènes n'aient pas muté entre t et (t+1) (avec une probabilité $(1-\mu)^2$ que l'on définie comme γ)

Étape 2 : on regarde d'où ils viennent à la génération précédente :

a ≡ Probabilité que 2 gènes pris dans une population viennent d'une même population à la génération précédente



Prob(pas de migration)=$(1-m)^2$



Prob(migration de 2 gènes vers la même pop)=$m^2 /(n_d-1)$

Calcul des Probabilités d'identités : principe des formules de récurrences

Étape 1 : on s'intéresse aux probabilité d'identité entre paires de gènes, il faut donc que les 2 gènes n'aient pas muté entre t et (t+1) (avec une probabilité $(1-\mu)^2$ que l'on définie comme γ)

Étape 2 : on regarde d'où ils viennent à la génération précédente :

a ≡ Probabilité que 2 gènes pris dans une population soient les copies de 2 gènes provenant d'une même population à la génération précédente

$a = (1-m)^2 + m^2 /(n_d-1)$

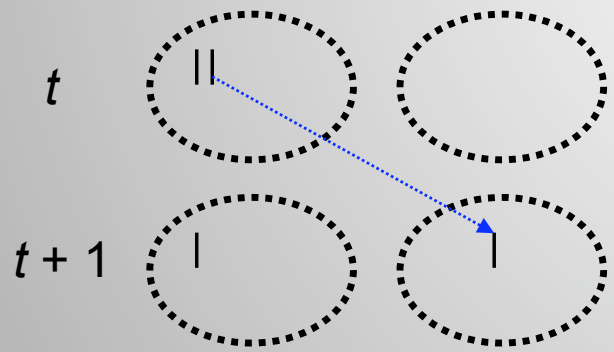
et (1-a) est alors la probabilité qu'ils soient les copies de gènes provenant de deux populations distinctes à la génération précédente

Calcul des Probabilités d'identités : principe des formules de récurrences

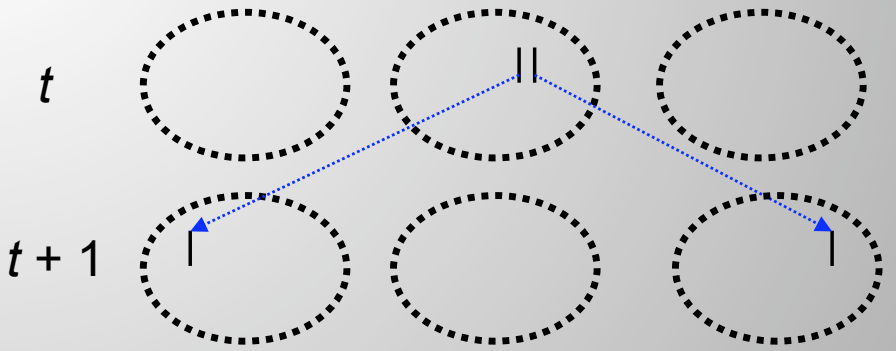
Étape 1 : on s'intéresse aux probabilité d'identité entre paires de gènes, il faut donc que les 2 gènes n'aient pas muté entre t et (t+1) (avec une probabilité $(1-\mu)^2$ que l'on définie comme γ)

Étape 2 : on regarde d'où ils viennent à la génération précédente :

b ≡ Probabilité que 2 gènes pris dans deux populations distinctes viennent d'une même population à la génération précédente



Prob(un des deux gène n'a pas migré)=2*(1-m)*m/(n_d-1)



Prob(migration de 2 gènes issus d'une troisième population)=(n_d-2)[m /(n_d-1)]²

Calcul des Probabilités d'identités : principe des formules de récurrences

Étape 1 : on s'intéresse aux probabilité d'identité entre paires de gènes, il faut donc que les 2 gènes n'aient pas muté entre t et (t+1) (avec une probabilité $(1-\mu)^2$ que l'on définie comme γ)

Étape 2 : on regarde d'où ils viennent à la génération précédente :

b ≡ Probabilité que 2 gènes pris dans deux populations distinctes soient des copies de 2 gènes provenant d'une même population à la génération précédente

$$b = \frac{2m(1-m)}{n_d - 1} + (n_d - 2)\left(\frac{m}{n_d - 1}\right)^2$$

et (1-b) est alors la probabilité que 2 gènes pris dans deux populations distinctes soient des copies de 2 gènes provenant de 2 populations différentes à la génération précédente

Calcul des Probabilités d'identités : principe des formules de récurrences

Étape 1 : on s'intéresse aux probabilité d'identité entre paires de gènes, il faut donc que les 2 gènes n'aient pas muté entre t et (t+1) (avec une probabilité $(1-\mu)^2$ que l'on définie comme γ)

Étape 2 : on regarde si les 2 gènes ont migré ou non à la génération précédente (probabilités a, $(1-a)$, b, $(1-b)$)

Étape 3 : s'ils proviennent d'une même population à la génération précédente, ils proviennent d'un même individu (cas d'une espèce diploïde) avec une probabilité s, le taux d'autofécondation

Calcul des Probabilités d'identités : principe des formules de récurrences

Étape 1 : la mutation ($\gamma=(1-\mu)^2$)

Étape 2 : la migration (a, $(1-a)$, b, $(1-b)$)

Étape 3 : l'autofécondation ($s, 1-s$)

$$Q_0(t+1) = \gamma\left[a\left(s\left(\frac{Q_0(t)+1}{2}\right)+(1-s)Q_1(t)\right)+(1-a)Q_2(t)\right]$$

$$Q_1(t+1) = \gamma\left[a\left(\frac{1}{N}\left(\frac{Q_0(t)+1}{2}\right)+(1-\frac{1}{N})Q_1(t)\right)+(1-a)Q_2(t)\right]$$

$$Q_2(t+1) = \gamma\left[b\left(\frac{1}{N}\left(\frac{Q_0(t)+1}{2}\right)+(1-\frac{1}{N})Q_1(t)\right)+(1-b)Q_2(t)\right]$$

Calcul des Probabilités d'identités : principe des formules de récurrences

A l'équilibre (migration-dérive-mutation), on a $Q_{\cdot}(t+1)=Q_{\cdot}(t)$ et on peut alors résoudre alors le système suivant pour exprimer Les probabilités d'identité en fonction des paramètres (n_d, N, m, s, μ) du modèle :

$$Q_0(t) = \gamma\left[a\left(s\left(\frac{Q_0(t)+1}{2}\right)+(1-s)Q_1(t)\right)+(1-a)Q_2(t)\right]$$

$$Q_1(t) = \gamma\left[a\left(\frac{1}{N}\left(\frac{Q_0(t)+1}{2}\right)+(1-\frac{1}{N})Q_1(t)\right)+(1-a)Q_2(t)\right]$$

$$Q_2(t) = \gamma\left[b\left(\frac{1}{N}\left(\frac{Q_0(t)+1}{2}\right)+(1-\frac{1}{N})Q_1(t)\right)+(1-b)Q_2(t)\right]$$

F-statistiques, Probabilités d'identités et paramètres des modèles

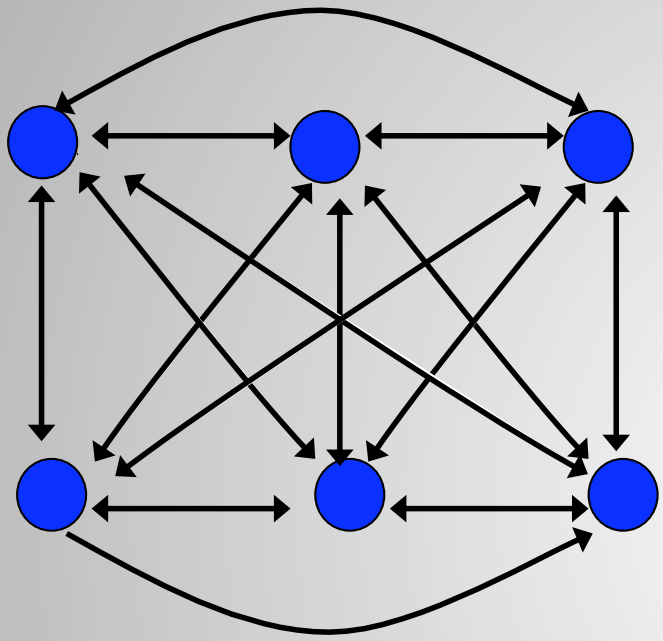
Puisque l'on a défini les relations entre *F*-statistiques et probabilités d'identités :

$$F_{IS} \equiv \frac{Q_0 - Q_1}{1 - Q_1} \qquad F_{ST} \equiv \frac{Q_1 - Q_2}{1 - Q_2} \qquad F_{IT} \equiv \frac{Q_0 - Q_2}{1 - Q_2}$$

On peut donc exprimer les *F*-statistiques en fonction des paramètres du modèle en faisant le rapport de expression de chaque probabilité d'identité trouvé précédemment.
Apres simplification, et sous panmixie au sein des pops (*s*=1/*N)*, on trouve :

$$F_{\mathrm{ST}} = \frac{1}{1 + 2N\left(2\mu + 2\frac{n_{\mathrm{d}}}{n_{\mathrm{d}} - 1} m\right)}$$

La migration : le modèle en îles et le F_{ST}



4 paramètres :

n_d= nombre de sous-populations (ou ∞)

N = taille des sous-populations

m= taux de migration

μ= taux de mutation

s=1/N (sous-populations panmictiques)

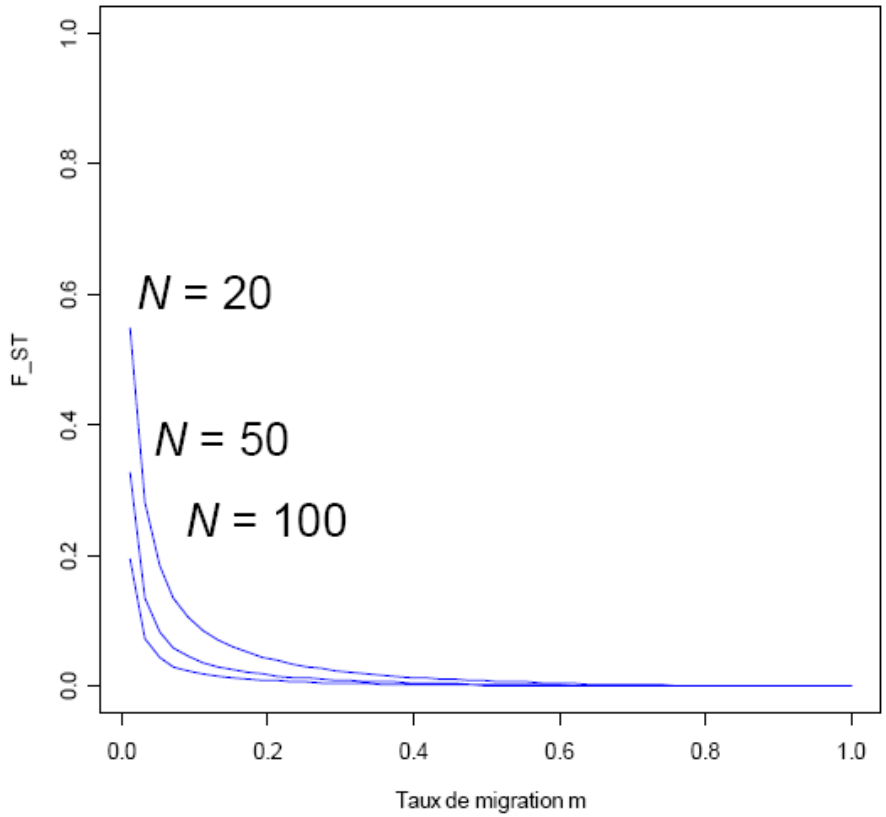
$$F_{\mathrm{ST}} = \frac{1}{1 + 2N(2\mu + 2\frac{n_{\mathrm{d}}}{n_{\mathrm{d}}-1}m)} \approx_{\mu \to 0} \frac{1}{1 + 4N\frac{n_{\mathrm{d}}}{n_{\mathrm{d}}-1}m}$$

$$F_{ST} \approx_{nd \to \infty} \frac{1}{1+4Nm}$$

La migration : le modèle en îles et le F_{ST}

$$F_{ST} \approx \frac{1}{1+4Nm} \text{ quand } \mu \to 0, n_d \to \infty$$

La différenciation décroît avec le taux de migration et augmente avec la dérive



F_{ST} et estimation sur des donnes réelles

Comment estimer le F_{ST} sur un jeu de données réel?

Données sur un locus allozymiques pour 3 populations de salamandre

Population	allèle1	allèle2
Konstanz	0.49	0.51
Bregenz	0.83	0.17
Schaffhausen	0.91	0.09

On peut partir de la définition *F*ST = 1 – *Ho*/*He* :

He, hétérozygosie attendue sous HW sur l'échantillon total = 2pq

Ho, hétérozygosie moyenne sous HW au sein des populations = $\frac{1}{n_{pop}} \sum_{pop} 2 p_i q_i$

On a donc besoin des fréquences moyennes sur l'échantillon total
Hyp : même taille d'échantillon pour chaque population
allèle1 p=(0.49+0.83+0.91)/3=0.74, et allèle2 q=(0.51+0.17+0.09)/3=0.26

On a donc, *He* = 2 pq = 2 x 0.74 x 0.26 = 0.384
D'autre part
Ho = [2x (0.49 x 0.51) + 2 x (0.83 x 0.17) + 2 x (0.91 x 0.09)] / 3 = 0.315

F_{ST} = 1 - 0.315/ 0.384 = 0.180

Il existe un certain nombre de logiciels :

Genepop	Rousset (2007)	http://www.isem.cnrs.fr/spip.php?article389
FSTAT	Goudet (1995)	http://www2.unil.ch/popgen/softwares/fstat.htm
Arlequin	Schneider et al. (2000)	http://cmpg.unibe.ch/software/arlequin3/
SPAGeDi	Hardy et Vekemans (2002)	http://www.ulb.ac.be/sciences/ecoevol/spagedi.html
GENETIX	Belkhir et al. (1996)	http://www.genetix.univ-montp2.fr/genetix/intro.htm

(liste non exhaustive…)

F_{ST} et estimation sur des donnes réelles

Comment estimer le F_{ST} sur un jeu de données réel?

Données sur un locus allozymiques pour 3 populations de salamandre

Population	allèle1	allèle2
Konstanz	0.49	0.51
Bregenz	0.83	0.17
Schaffhausen	0.91	0.09

F_{ST} = 1 - 0.315/ 0.384 = 0.180

Un FST de 0.18 indique que 18% de la variance génétique est due la différenciation entre population (et que 82% est due à la variabilité intrapopulation!)

Si l'on fait l'hypothèse d'un modèle en îles à l'équilibre migration-dérive, les flux de gènes dans le système correspondent à :

Nm = (1-FST) / 4 FST*2/3 = 0.76 migrants/génération

D'après la formule $$F_{ST} \approx \frac{1}{1 + 4N \frac{n_d}{n_d - 1} m}$$

(4) Estimations des taux de migration

-Estimations indirectes: comme le Fst est estimé à partir des fréquences alléliques (ou des hétérozygoties observées et attendues), on peut l'utiliser pour estimer le produit Nm:

$$Nm = \frac{1}{4}\left[\frac{1}{F_{ST}} - 1\right]$$

Exemples ci-contre:
-à une extrême, le bivalve Mytilus edulis a des larves planctoniques qui peuvent être transportées sur de longues distances par des courants. Nm = 42.
-À l'autre extrême, certaines salamandres sont extrêmement sédentaires. Nm=0.10.

Table 5. Estimates of *Nm* and $\hat{F}_{ST}$.

Species	Type of organism	Estimated *Nm*	Estimated $\hat{F}_{ST}$
Stephanomeria exigua	Annual plant	1.4	0.152
Mytilus edulis	Mollusc	42.0	0.006
Drosophila willistoni	Insect	9.9	0.025
Drosophila pseudoobscura	Insect	1.0	0.200
Chanos chanos	Fish	4.2	0.056
Hyla regilla	Frog	1.4	0.152
Plethodon ouachitae	Salamander	2.1	0.106
Plethodon cinereus	Salamander	0.22	0.532
Plethodon dorsalis	Salamander	0.10	0.714
Batrachoseps pacifica ssp. 1	Salamander	0.64	0.281
Batrachoseps pacifica ssp. 2	Salamander	0.20	0.556
Batrachoseps campi	Salamander	0.16	0.610
Lacerta melisellensis	Lizard	1.9	0.116
Peromyscus californicus	Mouse	2.2	0.102
Peromyscus polionotus	Mouse	0.31	0.446
Thomomys bottae	Gopher	0.86	0.225

(Data from Slatkin 1985a.)

-Estimations directes: la méthode la plus utilisée est celle du marquage-recapture par les écologistes (par ex. pour les lézards). Mais on peut suivre aussi des individus marqués par un trait génétique par exemple (yeux colorés chez la drosophile).

La migration : le modèle en îles et le F_{ST}

$$F_{ST} \approx \frac{1}{1+4Nm} \quad \Rightarrow Nm \approx \frac{1}{4}\left(\frac{1}{F_{ST}} - 1\right)$$

Cette formule a trop souvent été utilisée pour estimer un nombre de migrant entre populations par génération mais :

- Modèles peu réalistes, mauvaise description de la dispersion
- Hypothèses de stabilité démographiques dans le temps et dans l'espace
- Hypothèses associées aux taux de mutation et processus mutationnels
- Hypothèses de neutralité des marqueurs utilisés

La migration : le modèle en îles et le F_{ST}

$$F_{ST} \approx \frac{1}{1+4Nm} \quad \Rightarrow Nm \approx \frac{1}{4}\left(\frac{1}{F_{ST}} - 1\right)$$

Indirect measures of gene flow and migration: $F_{ST} \neq 1/(4Nm+1)$

MICHAEL C. WHITLOCK*† & DAVID E. MCCAULEY‡

†Department of Zoology, University of British Columbia, Vancouver, BC V6T 1Z4 Canada and ‡Department of Biology, Vanderbilt University, Nashville, Tennessee 37235, U.S.A.

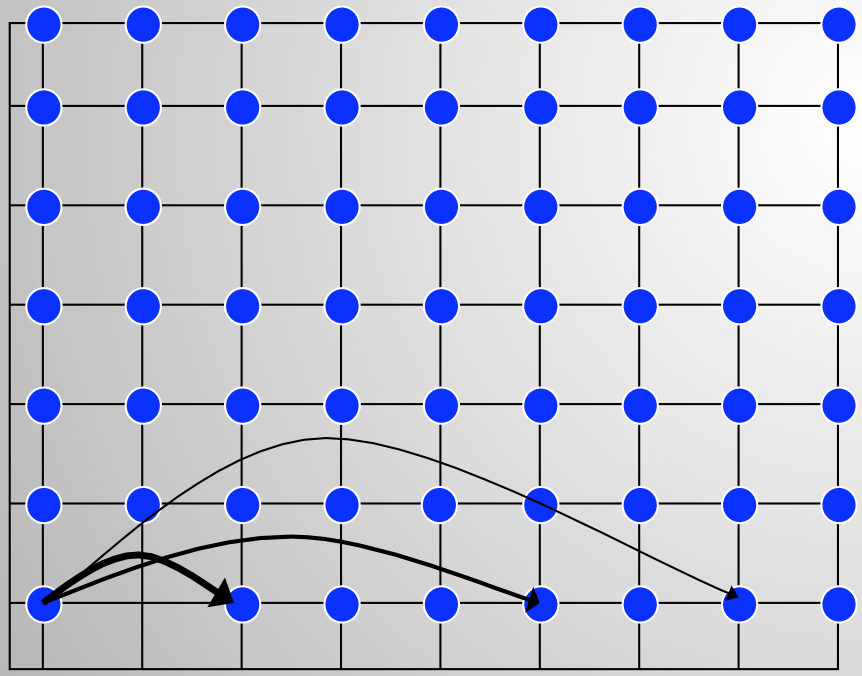
The difficulty of directly measuring gene flow has lead to the common use of indirect measures extrapolated from genetic frequency data. These measures are variants of F_{ST}, a standardized measure of the genetic variance among populations, and are used to solve for Nm, the number of migrants successfully entering a population per generation. Unfortunately, the mathematical model underlying this translation makes many biologically unrealistic assumptions; real populations are very likely to violate these assumptions, such that there is often limited quantitative information to be gained about dispersal from using gene frequency data. While studies of genetic structure *per se* are often worthwhile, and F_{ST} is an excellent measure of the extent of this population structure, it is rare that F_{ST} can be translated into an accurate estimate of Nm.

Keywords: allozymes, dispersal, F_{ST}, gene flow, indirect measures, migration.

Un modèle plus réaliste pour une meilleure estimation de la migration en populations structurées : **Le modèle d'isolement par la distance**

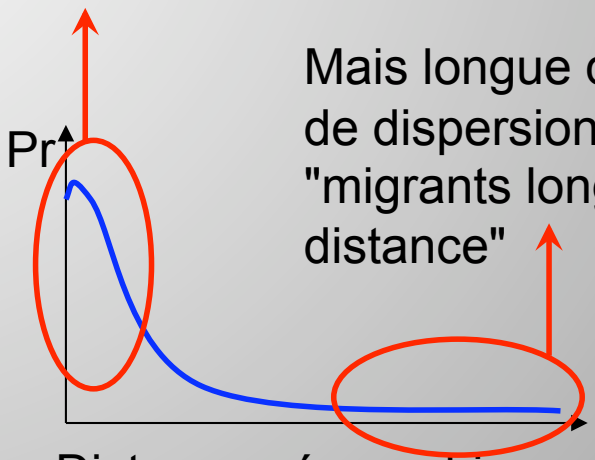
Dispersion limitée dans l'espace ↔ 2 individus ont plus de chance de se reproduire ensemble si ils sont proches géographiquement

Endler 1977 (revue biblio): la majorité des espèces ont une dispersion localisé



Migration fonction de la distribution de dispersion :

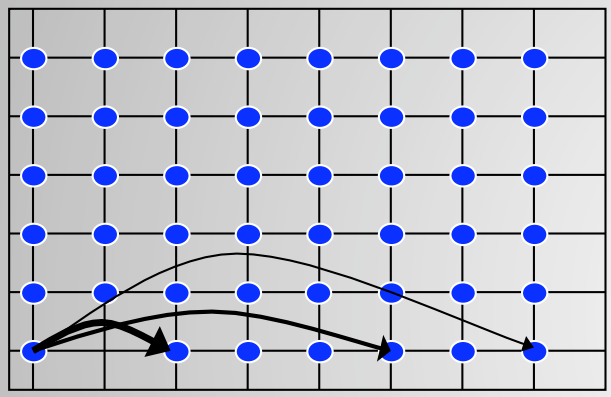
Majorité de la dispersion à très courte distance



Mais longue queue de dispersion = "migrants longue distance"

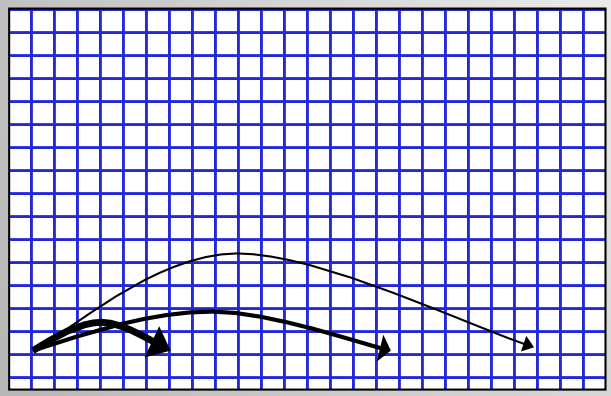
Les modèle d'isolement par la distance

2 modèles en fonction du type de distribution des organismes dans le paysage :



Population en dèmes

Chaque nœud du réseau correspond à une sous population panmictique



Population "continue" en réseaux
Chaque nœud du réseau correspond à 1 individu

L'isolement par la distance

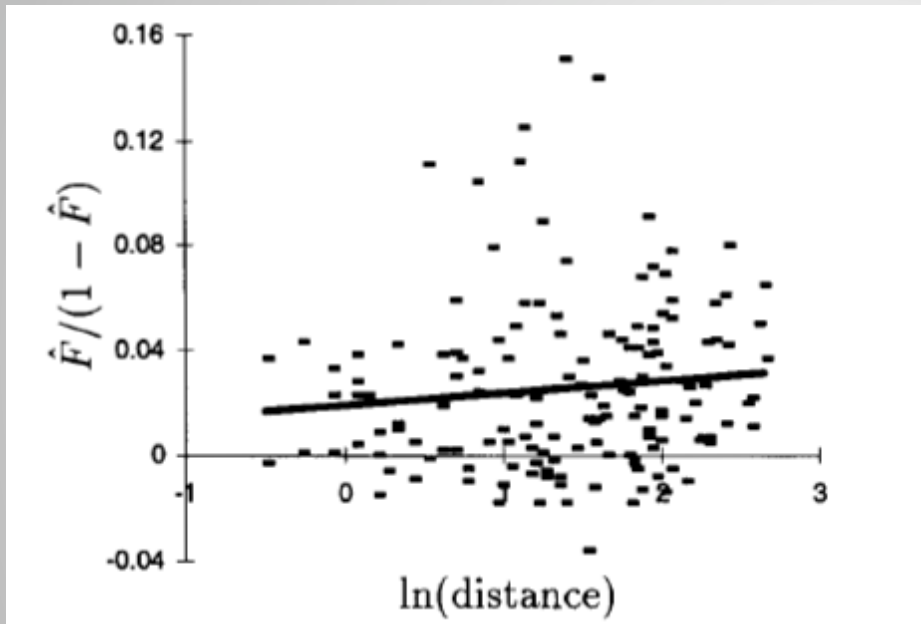


FIGURE 5.—Differentiation among Gainj- and Kalam-speaking peoples. Multilocus estimates of pairwise differentiation are plotted against logarithm of map distances (in Km). The regression is $y = 0.0047x + 0.0191$ and the maximum distance between two subpopulations is 14 km. Genotypic data appear in LONG *et al.* (1986). F_{ST} was estimated according to WEIR and COCKERHAM (1984).

Rousset 1997

- Le modèle d'isolement par la distance (Malécot 1956) prédit une relation linéaire entre la distance génétique et le logarithme de la distance géographique
- La pente de la droite de régression donne un estimateur de la distance de dispersion

L’isolement par la distance

Coenagrion mercuriale : données démographique (capture / marquage / recapture) : Watt *et al*. 2006

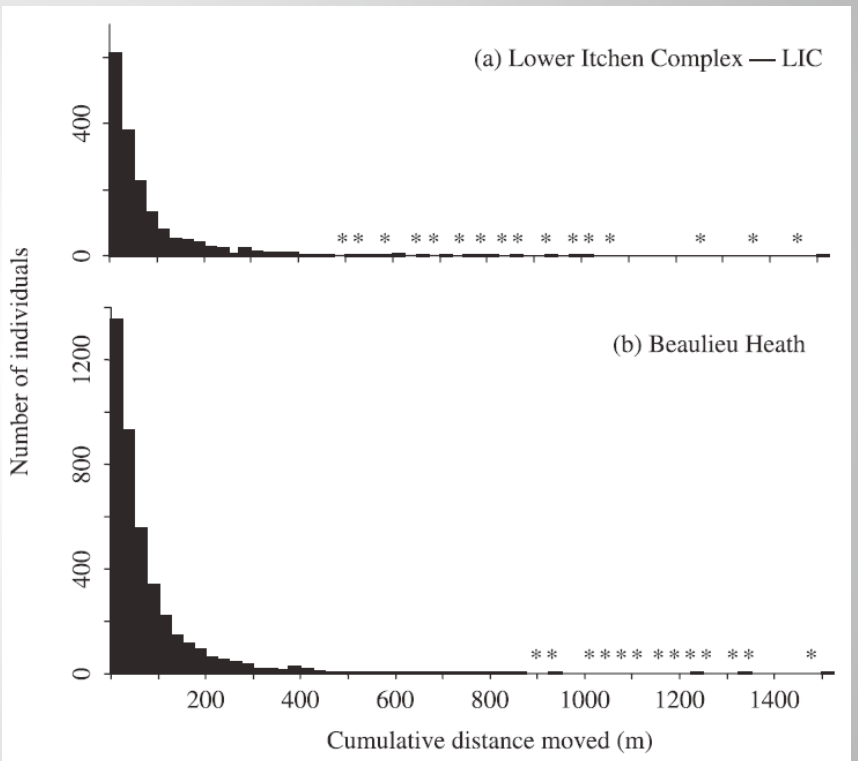
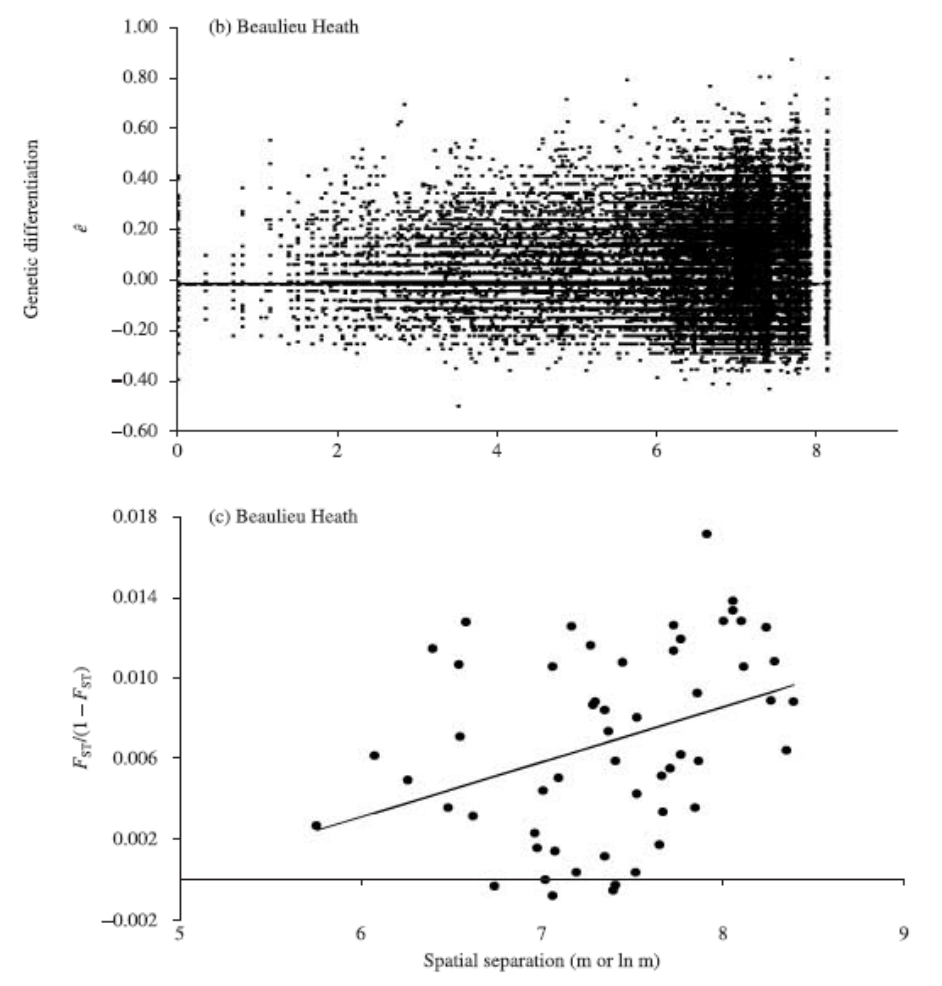


Fig. 3 Frequency of cumulative lifetime movement of adult *Coenagrion mercuriale* in 25-m distance categories for (a) the Lower Itchen Complex (LIC) and (b) Beaulieu Heath, both to the same scale. *n*, number of recaptured individuals; *highlights infrequent ($n = 1$ or 2) movement events.

L'isolement par la distance



Données génétiques (marqueurs microsatellites) : estimation de la « taille de voisinage » ($D\sigma^2$)

L’isolement par la distance

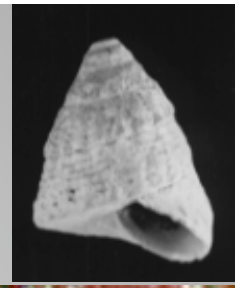
Coenagrion mercuriale : excellente concordance entre estimations directes et indirectes

	Estimation de $D\sigma^2$	
	démographie	génétique
Site 1	277	222
Site 2	249	259
Site 3	555	606

Les modèle d'isolement par la distance

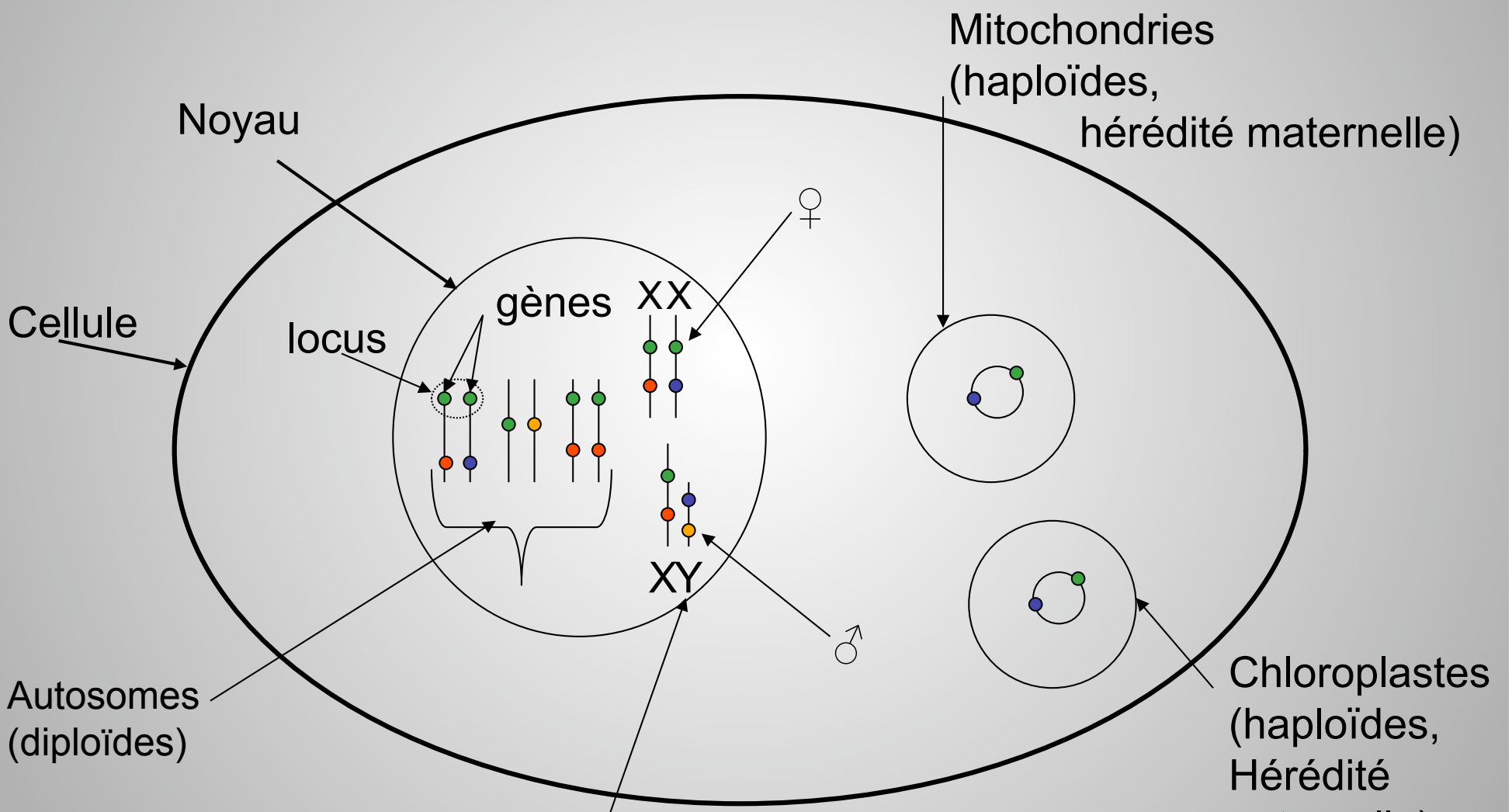




	Direct (Demography)	Indirect (genetic)
American Marten (Martes americana)	7.5	3.8
Kangaroo rats (Dipodomys)	1.43	2.58
intertidal snails (Bembicium vittatum)	2.4	3.6
Forest lizards (Gnypetoscincus queenslandiae)	11.5	5.5
Humans in the rainforest (Papous)	29.3	21.1
Legumin (Chamaecrista fasciculata)	9.6	13.9

Différents génomes chez un même individu



Différents types de marqueurs

- Pour un gène mitochondrial (haploïde, transmis de mère à enfant)

$$F_{ST\,mito} = \frac{1}{1 + 2Nm_f}$$

- Pour un gène sur le chromosome Y (haploïde, transmis de père à fils)

$$F_{ST\,Y} = \frac{1}{1 + 2Nm_m}$$

- Chez l'Homme, les F_{ST} sur le Y sont plus forts que sur les mitochondries : dispersion biaisée en faveur des femmes

Chromosome Y *vs.* mtDNA

Table 1

Estimates of sex-biased migration based on partitioning of genetic variance.

Region	mtDNA F_{ST}	NRY F_{ST}	Male:female *Nm*	NRY Data	Notes	References
Global	0.186	0.645	0.13	STR		[49]
Global	0.401	0.357	1.21	Seq	Overall	[50••]
	0.261	0.209	1.34		Within continents	
	0.189	0.187	1.01		Among continents	
Thailand	0.290	0.131	2.71	STR	Matrilocal	[29]
	0.118	0.450	0.16		Patrilocal	
	0.038	0.130	0.26		Food-producers	
Sub-Saharan Africa	0.431	0.072	9.76	SNP	Hunter-gatherer	[31••]
	0.025	0.174	0.12		Overall	
Caucasus	0.025	0.174	0.12	SNP	Overall	[25]
	0.008	0.060	0.13		Within groups	
	0.018	0.121	0.13		Among groups	
Sub-Saharan Africa	0.16	0.33	0.39	SNP	Overall	[40••]
	0.13	0.28	0.38		Within groups	
	0.04	0.06	0.65		Among groups	

This table lists the reported F_{ST} values for mtDNA and NRY data from a number of recent studies, with the geographic region studied listed in the first column. With the exception of the first study, by Seilstad *et al* [49], each of these studies is based on mtDNA and NRY samples drawn from the same individuals. The male:female *Nm* ratio was calculated assuming that $Nm = (1/F_{ST})-1$. Ratio values greater than 1 correspond to higher male migration (i.e. higher male effective population size), whereas ratios less than 1 correspond to higher female migration (i.e. higher female effective population size). The NRY data column indicates what type of data was collected from the Y chromosome: Seq, direct sequencing; SNP, single nucleotide polymorphisms; STR, microsatellite repeats. The mtDNA was all collected by direct sequencing. For each study in which the authors performed analyses on different subsets of the data, or at different levels of resolution, the estimated *Nm* ratio for each sub-analysis is shown, with the partitioning of the data indicated in the Notes column.

- Chez l'homme il y a une forte influence de la structure sociale et de l'échelle !

Différents types de marqueurs

- Peu de polymorphisme sur le Y chez de nombreuses espèces : comparaison des F_{ST} nucléaires aux F_{ST} mitochondriaux : philopatrie des femelles si F_{ST} mitochondriaux sont plus forts

Physeter macrocephalus (cachalot) et *Carcharodon carcharias* (grand requin blanc) : absence de différenciation entre océans sur marqueurs microsatellites mais différenciation significative sur ADNmt

Différents types de marqueurs

- Chez les plantes, deux possibilités de migration :
  - par les graines (taux m_S).
  - par le pollen (taux m_P).

- Pour un gène nucléaire :

$$F_{STn} = \frac{1}{1 + 4N(m_S + m_P/2)}$$

- Pour un gène cytoplasmique à hérédité maternelle :

$$F_{STm} = \frac{1}{1 + 2Nm_S}$$

- Pour un gène cytoplasmique à hérédité paternelle :

$$F_{STp} = \frac{1}{1 + 2N(m_S + m_P)}$$

- Obtention de ratios, par ex. :

$$\frac{m_P}{m_S} = \frac{1/F_{STn} - 1}{1/F_{STm} - 1} - 2$$

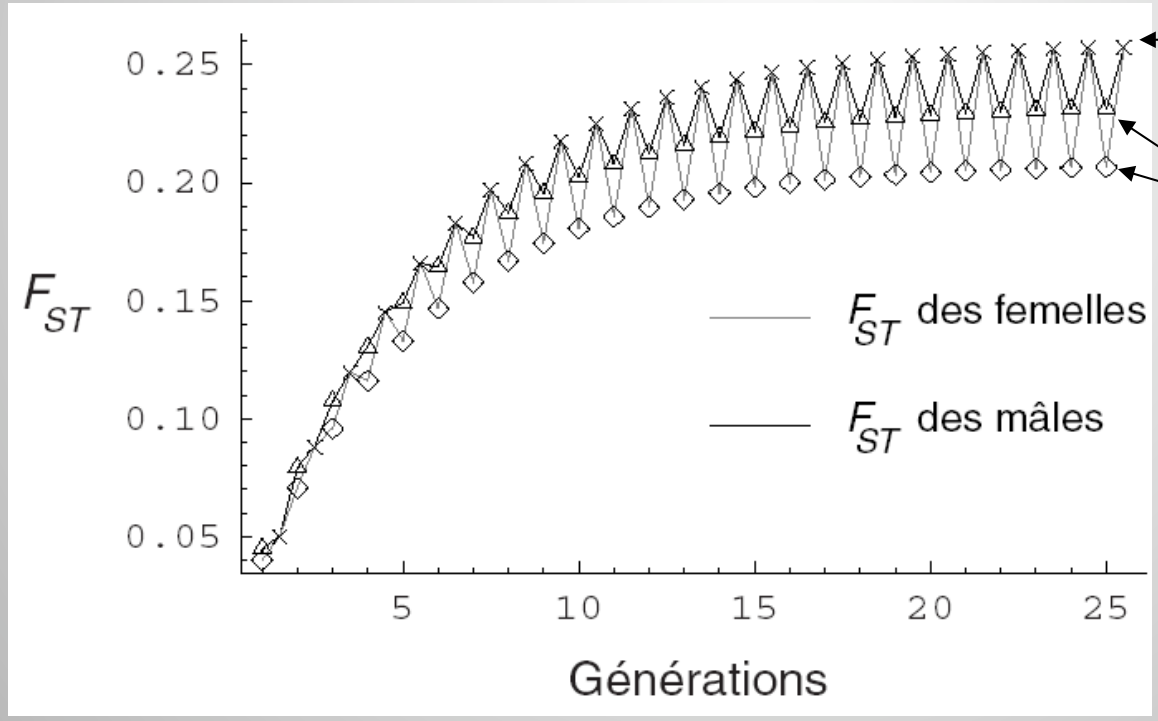
Différents types de marqueurs

Espèces	F_{STn}	F_{STm}	m_p / m_s	
Quercus robur / Q. petraea	0.037	0.884	196	Graînes lourdes Disp. pollen anémophile
Pinus contorta	0.0608	0.66	28	Graînes + légères
Pinus attenuata	0.12	0.863	44	
Pinus muricata	0.22	0.882	24	
Sorbus torminalis	0.11	0.34	1.18	Graînes ingérées Disp. pollen entomophile

- Résultats cohérents avec le cycle de vie des espèces :
  - Graines lourdes ou légères, ingérées ou non.
  - Pollinisation entomophile ou anémophile.

Dispersion sexe-spécifique

Si l'on mesure les F_{ST} chez des mâles et des femelles avant et après dispersion (ici les mâles dispersent moins que les femelles) :



Dispersion sexe-spécifique

On peut montrer que :

$$\frac{F_{ST}^{XX}}{F_{ST}^{*}} \simeq (1 - m_{XX})^2$$

Ce qui suggère un estimateur du taux de migration sexe-spécifique de la forme :

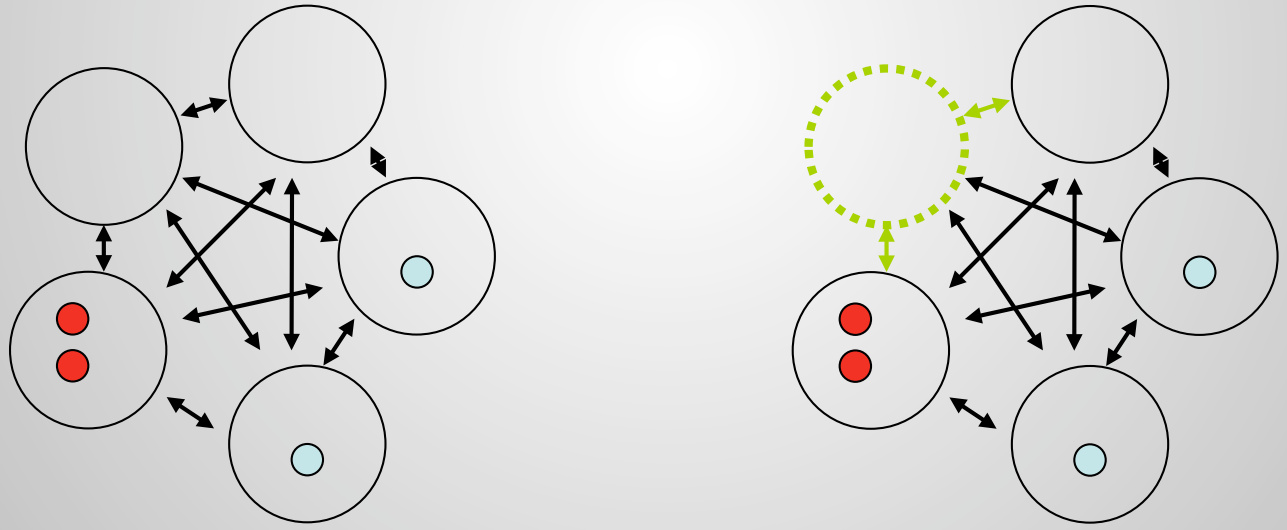
$$\hat{m}_{XX} \simeq 1 - \sqrt{\hat{F}_{ST}^{XX} / \hat{F}_{ST}^{*}}$$

Exemple chez *Crocidura russula* :

		Non-hierarchical analysis	
	n	F_{ST}	$\hat{a}_2$
Juveniles	170	0.090 [0.077; 0.104]	
Adult females	181	0.053 [0.047; 0.059]	**0.24** [0.18; 0.29]
Adult males	185	0.065 [0.052; 0.080]	**0.15** [0.05; 0.25]

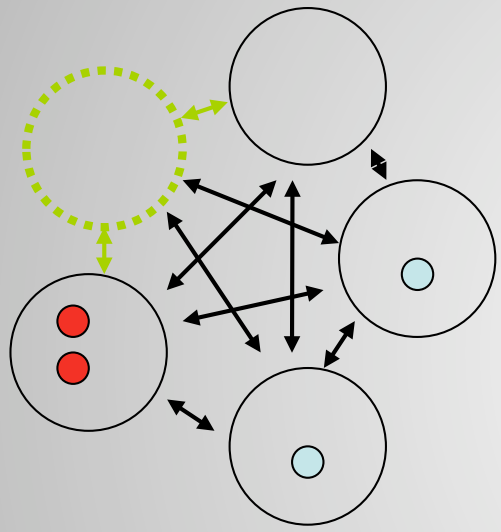
F_{ST} et métapopulations

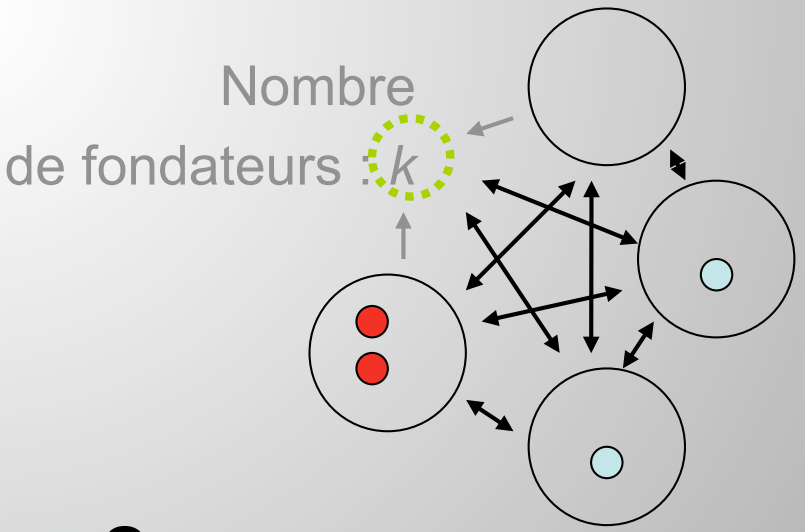
- Une métapopulation est une population subdivisée où chaque dème/sous-population peut s'éteindre avec une certaine probabilité *e* à chaque génération…



Rôle des extinction ?

F_{ST} et métapopulations





… et des recolonisations ?

F_{ST} et métapopulations

$$F_{ST} \approx \frac{(1-e)k(1-m)^2 + e[N(N-1)-\phi]+ek\phi}{(N-1)[k(N-(1-e)(1-m)^2)-e(k-1)\phi]}$$

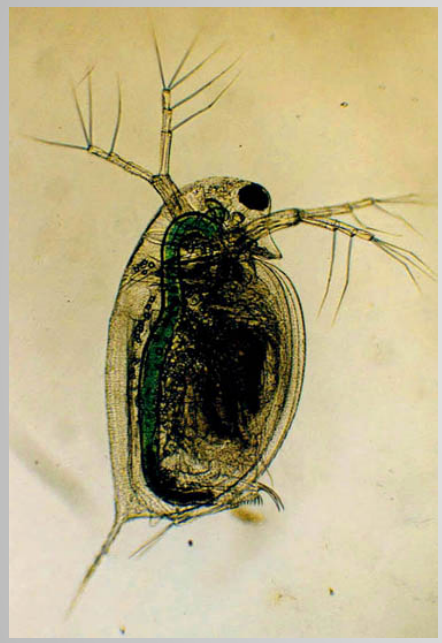
- La différenciation est plus forte qu'en l'absence d'extinctions,
- et d'autant plus si la diversité des populations nouvellement colonisée est faible, c'est-à-dire si la colonisation crée des goulets d'étranglement forts (effets de fondation forts avec *k* faible)

F_{ST} et métapopulations

- Les extinctions conduisent à la mise en place d'une structure en âges dans la métapopulation : des populations « jeunes » (nouvellement colonisées) et des populations « anciennes »

- Les populations jeunes subissent un fort effet de fondation, les plus âgées atteignent un équilibre (modèle en îles)

F_{ST} et métapopulations



Daphnia longisperma

Daphnia magna

Haag et al. (2005) *Genetics 170*: 1809-1820

- Suivi écologique de 507 patches sur 20 ans
- 17% des patches occupés (âge max. : 16-17 ans)
- Grandes populations isolées

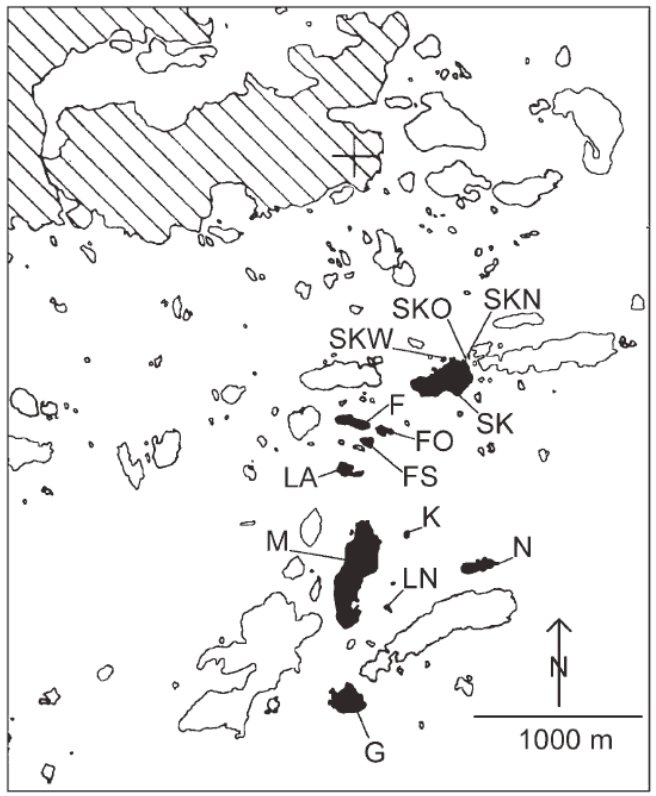
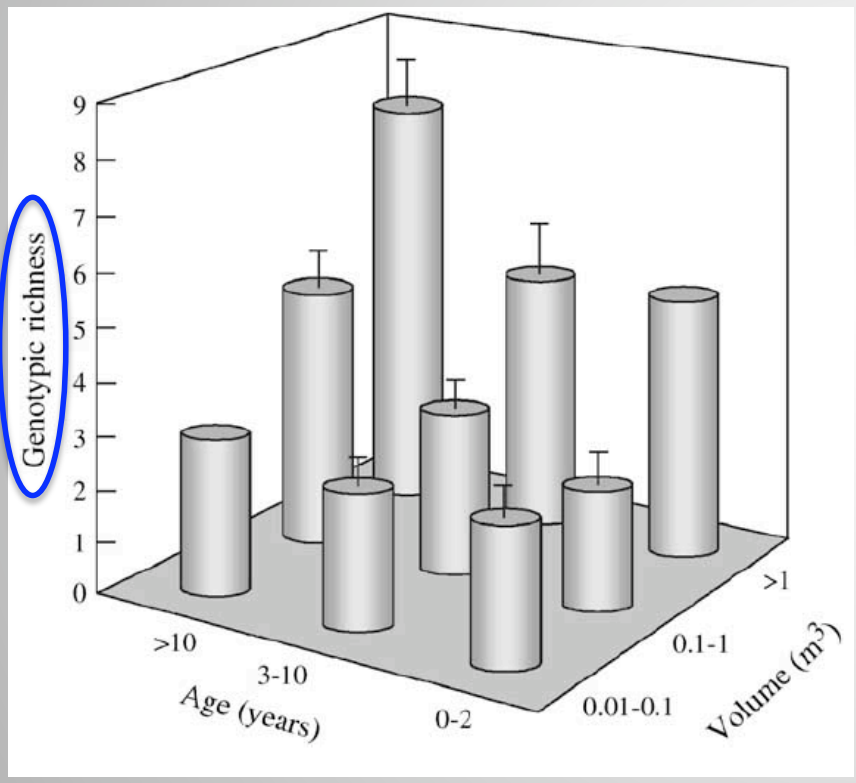


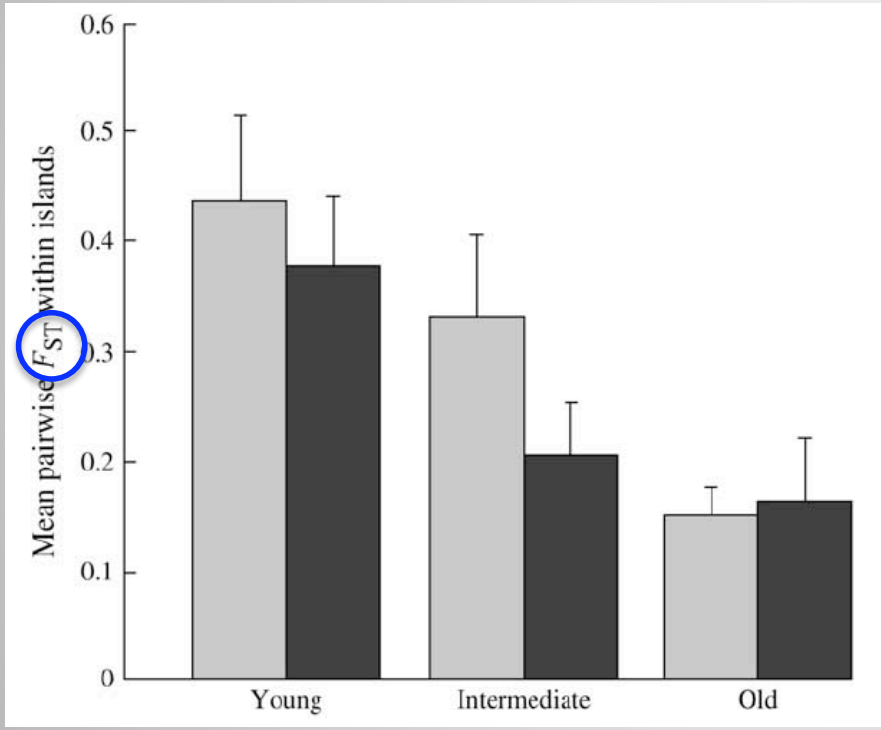
FIGURE 1.—Map of the study area on the Baltic coast of southern Finland. Islands included in this study are solid and labeled with abbreviations of island names as in Table 1. To our knowledge, there are no populations of either of the two species present on the mainland (hatched area) within 20 km, but populations are present on many other islands. The cross within the hatched area indicates the location of the Tvärminne Zoological Station at 59°50′N and 23°15′E.

F_{ST} et métapopulations



- La diversité génétique en fonction de l'âge des populations et de leur volume…

F_{ST} et métapopulations



- La différenciation entre paires de populations en fonction de leur âge…

F_{ST} et métapopulations

- L'âge (+), le volume (+), la distance à la mer (+) et la distance au plus proche voisin (-) sont corrélés aux mesures de diversité génétique

- L'âge (-), le volume (-), la distance (+) sont corrélés aux mesures de différenciation génétique