Introduction à la génétique et à la génomique des populations

Fabien Halkett, Stéphane de Mita

Génétique des populations appliquée à la protection des végétaux

Stéphane De Mita
demita@gmail.com
INRA Nancy
UMR Interactions Arbres/Micro-organismes

sur la base d'un cours de Fabien Halkett

WILEY Evolutionary Applications Open Access

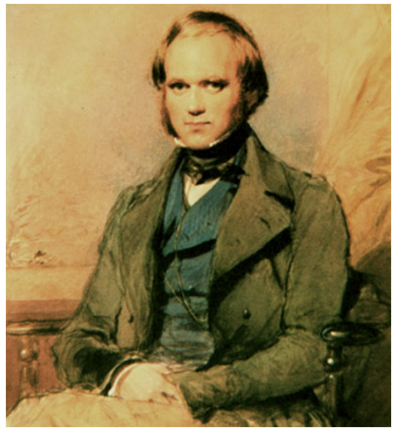
OBITUARY

Isabelle Olivieri (1957–2016)

Isabelle Olivieri passed away on Saturday 10 December at 4 a.m. She passed away quietly after having fought not only with an admirable courage against her cancer but also against a paraneoplastic syndrome, which heavily handicapped her for the last years of her life. Isabelle would have passed 60 years in March 2017. She entered the Agronomic College (AgroParisTech) in Paris in 1977, specialized in zoology and then oriented her research towards evolutionary genetics. Her PhD work was concerned with Mediterranean thistles. These plants were invasive in Australia, and one of the goals of the CSIRO where she was based in Montpellier was to find control agents. She would come back to the question of plant–insect interactions later in her life. She did her postdoc in Paul Ehrlich's laboratory in Stanford in 1983 and there further developed her ideas on the importance of the metapopulation concept for the understanding of the evolution of migration mechanisms. The paper which she published on this subject in American Naturalist (after a more than 4 years debate with the editors) provided her with a congratulation letter from Ernst Mayr

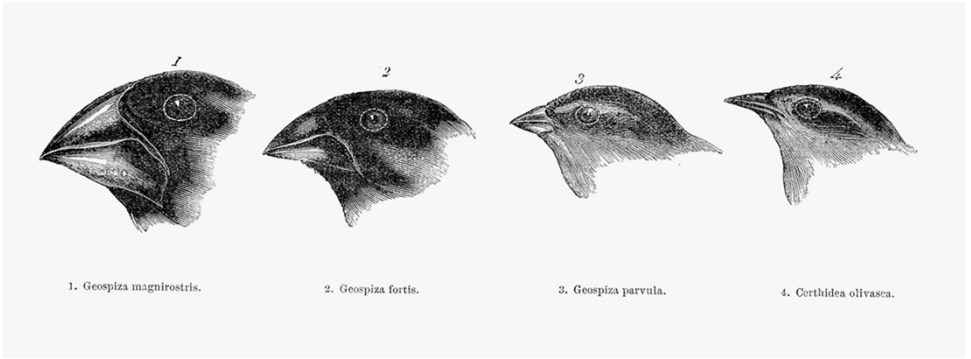
for the Study of Evolution, the Dutch Academy of sciences attributed her a comfortable grant allowing her to come whenever and wherever she wishes in any university to interact with local scientists, she was a

Synthèse "moderne" de l'évolution



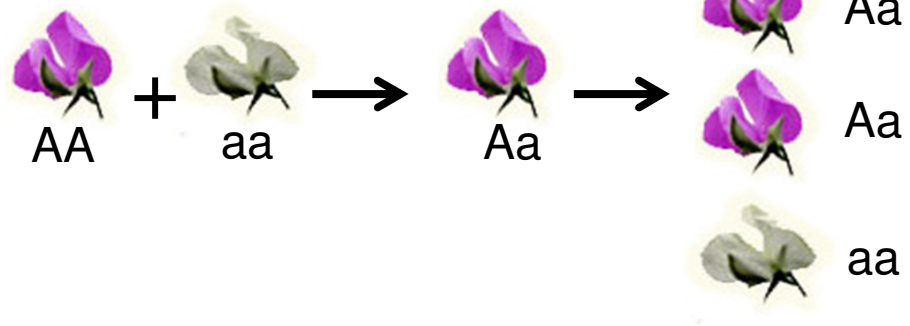
Charles Darwin : théorie de l'évolution des espèces au moyen de la sélection naturelle

1859



Gregor Mendel : lois de l'hérédité

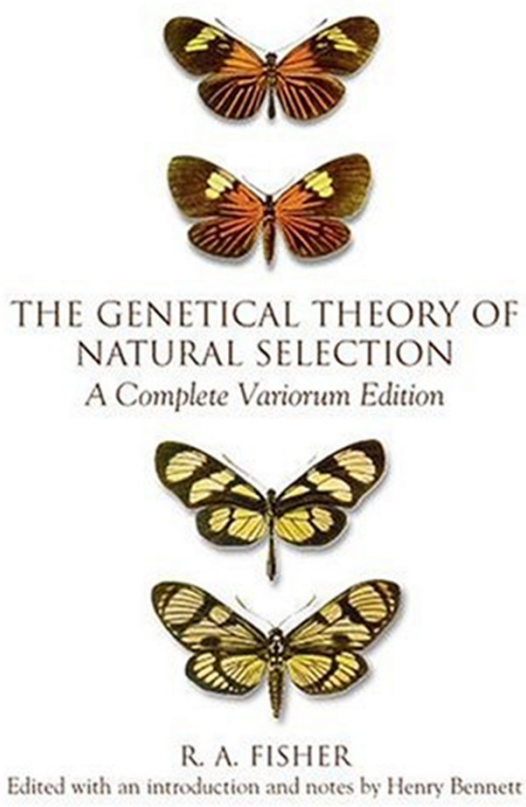
1866



Synthèse "moderne" de l'évolution

Ronald Fisher : mathématicien / statisticien / biologiste

1930



► Les caractères continus peuvent être expliqués par des unités discrètes (gènes)

► Concept de génétique quantitative (ou *biométrie*), eugénisme…

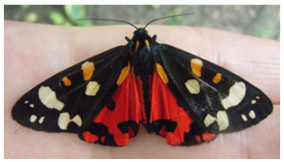
► Évolution des caractères complexes sous l'effet de la sélection et d'autres facteurs → **génétique des populations**

Objectifs des généticiens des populations

- Premier objectif : analyser les effets de la **sélection naturelle**
- Lien entre variabilité génétique (polymorphisme) et adaptation
- Limites à l'adaptation (dérive, mutation, migration)

Theodosius Dobzhansky

Genetics and the Origin of Species – 1937



Objectifs des généticiens des populations

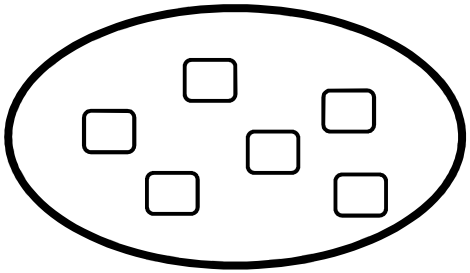
- Premier objectif : analyser les effets de la **sélection naturelle**
- Analyse de l'effet de **facteurs démographiques**
- Notamment : taille des populations, histoire, migration
- Thématiques de biologie des populations (connectées à l'écologie)
- Rendu possible par le développement des **marqueurs moléculaires**
- La sélection devient un facteur de nuisance

Définitions avant d'entrer dans le vif du sujet…

- **Espèce** : plusieurs définitions dont "ensemble d'individus interféconds"
  (parfois un peu arbitraire, dépend du groupe taxonomique)
- **Population** : groupe d'individus susceptibles de se reproduire de manière homogène et constante
  (même problème)
- **Gène** : portion du génome codant une fonction élémentaire
  (protéine ou régulateur)
- **Locus** : région physique du génome
  (de 1 base jusqu'à plusieurs kilobases selon la technique utilisée)
- **Allèle** : un des variants pour un locus polymorphe
- **Génotype** : ensemble des allèles portés par un individu

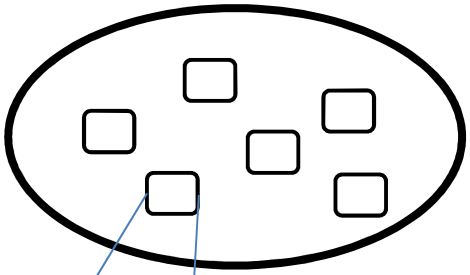
Niveaux d'organisation essentiels

Population

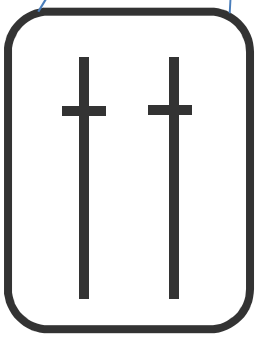


Niveaux d'organisation essentiels

Population



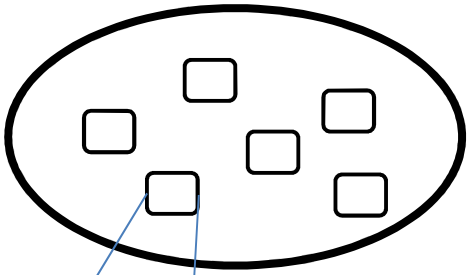
Individu



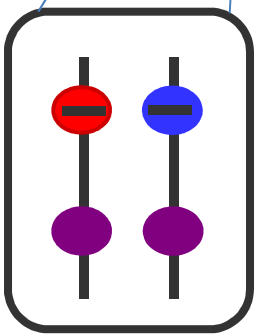
Locus

Niveaux d'organisation essentiels

Population



Individu

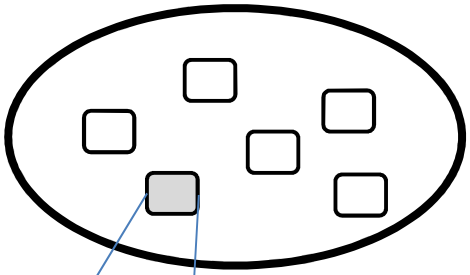


Locus hétérozygote

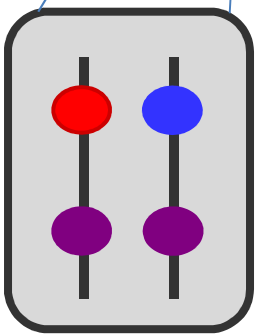
homozygote

Niveaux d'organisation essentiels

Population



Individu

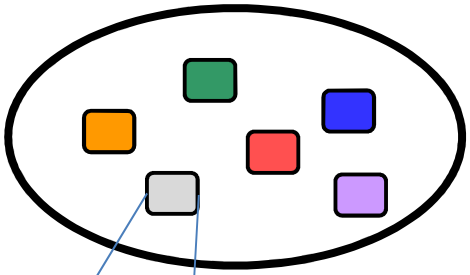


Locus hétérozygote

homozygote

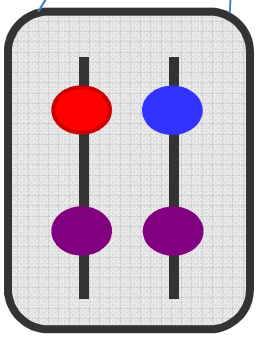
Niveaux d'organisation essentiels

Population



génotypes différents

Individu



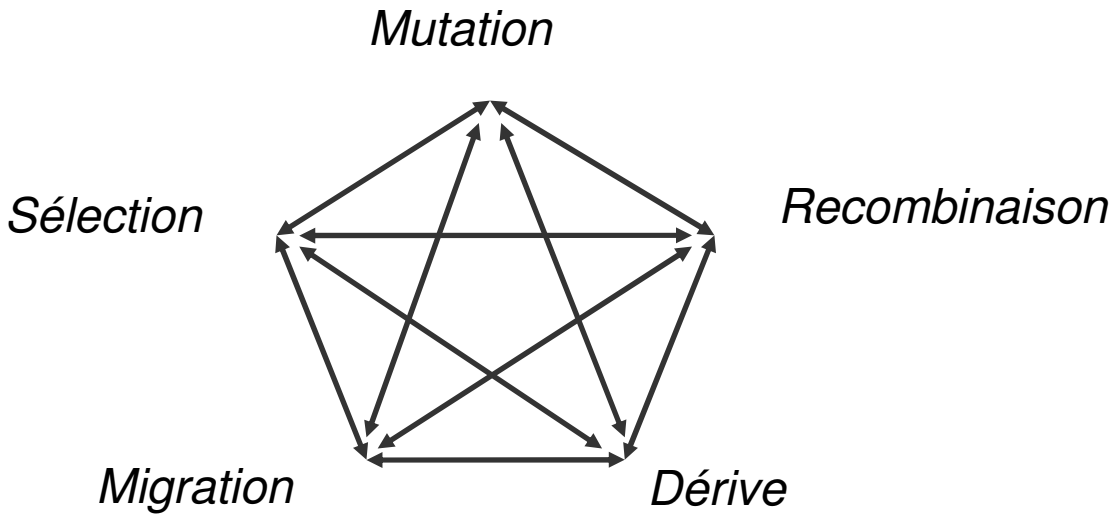
Locus hétérozygote

homozygote

Objet essentiel de la GdP : étudier les niveaux et le partitionnement de la variabilité génétique au sein et entre populations

Principe fondamental de la GdP

Étudier l'influence des différentes **forces évolutives** sur la **structuration génétique** des populations



Loi de **Hardy-Weinberg** :
- population isolée, de taille infinie,
- absence de mutation et de sélection
- reproduction panmictique

} les fréquences alléliques et génotypiques ne varient pas entre générations.

Phénomènes étudiés

- **Propres aux populations**
- Taille des populations
- Régime de reproduction
- Structure (sous-population) et patron de migration
- Histoire (variation de taille passée, anciennes structure et migration)

- **Sélectifs**

- **Moléculaires**
- Mutation
- Recombinaison et autres réarrangements génomiques

L'échantillonnage

- Il n'est pas possible (sauf exception) de caractériser des populations entières
- On se base sur des **échantillons**
- L'échantillon doit être :
- Représentatif (idéalement, aléatoire)
- Assez grand pour ne pas être *trop* aléatoire
- Assez petit pour rentrer dans le budget et le temps disponibles
- Prendre en compte la réalité du terrain

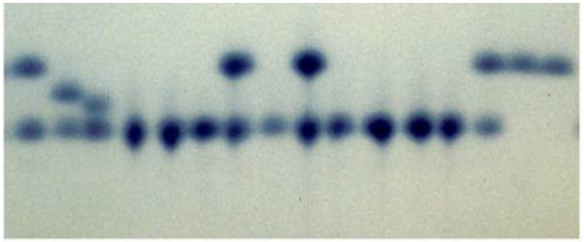
Marqueurs moléculaires

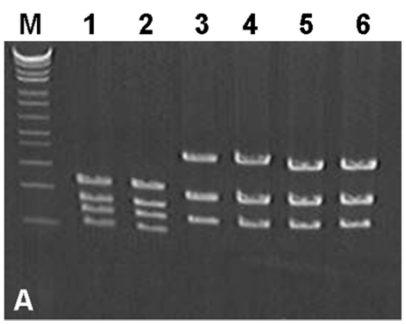
Différencier les lignages différents et détecter les individus proches évolutivement (ancêtre commun récent)

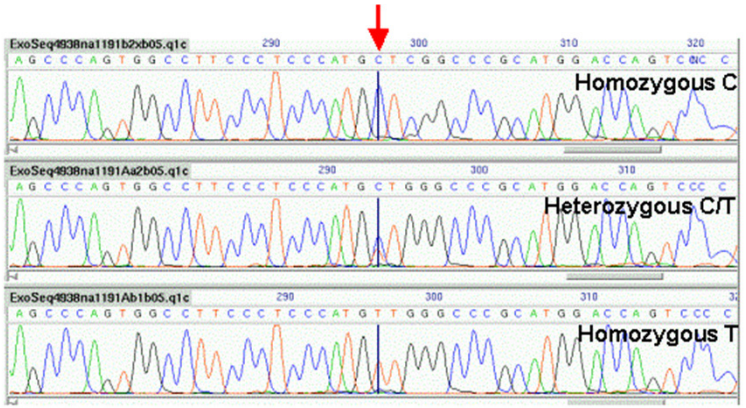
- **1953** Structure de l'ADN
- **1966** Allozymes (enzymes différents par leur pouvoir de migration)
- **1972** Séquençage du premier gène (ARN)
- **1974** Marqueurs basés sur l'ADN (RFLP)
- **1977** Séquençage de l'ADN par Fred Sanger
- **1992** Marqueurs microsatellites : ACGTG**GGAGGAGGAGGAGGAGGAGGAGGA**GCGAGTGT
- **1995** Premier génome complètement séquencé (Haemophilus influenzae)

Marqueurs moléculaires

- **Allozymes** : enzymes de masse et/ou charge variables
- **RFLP** : fragments d'ADN digérés et marqués
- **Microsatellites** : polymorphisme de longueur de motifs courts répétés
- **Génotypage SNP** : caractérisation de l'allèle pour une mutation ponctuelle
- **Séquençage** : lecture directe de la molécule (de plus en plus rapide)







Quel(s) marqueur(s) utiliser ?

Définition : caractère héréditaire possédant plusieurs états utilisés permettant de différentier les individus

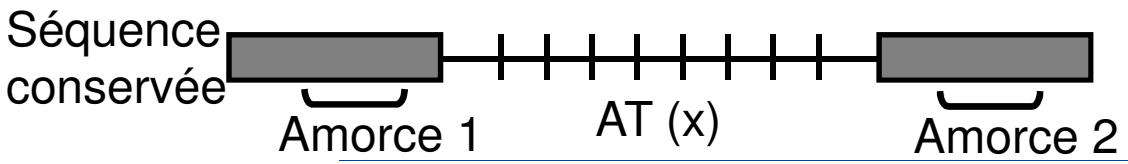
- Niveau de polymorphisme élevé

- Codominance (détection des hétérozygotes pour les diploïdes)

- Amplification par PCR (peu d'ADN requis)

- Fiabilité dans l'interprétation des données, répétabilité

Les marqueurs **microsatellites** sont les plus employés en génétique des populations



Démarche méthodologique

Echantillonnage des individus, conditionné à la question posée

Extraction de l'**ADN** des individus échantillonnés

PCR : amplification des loci d'intérêt

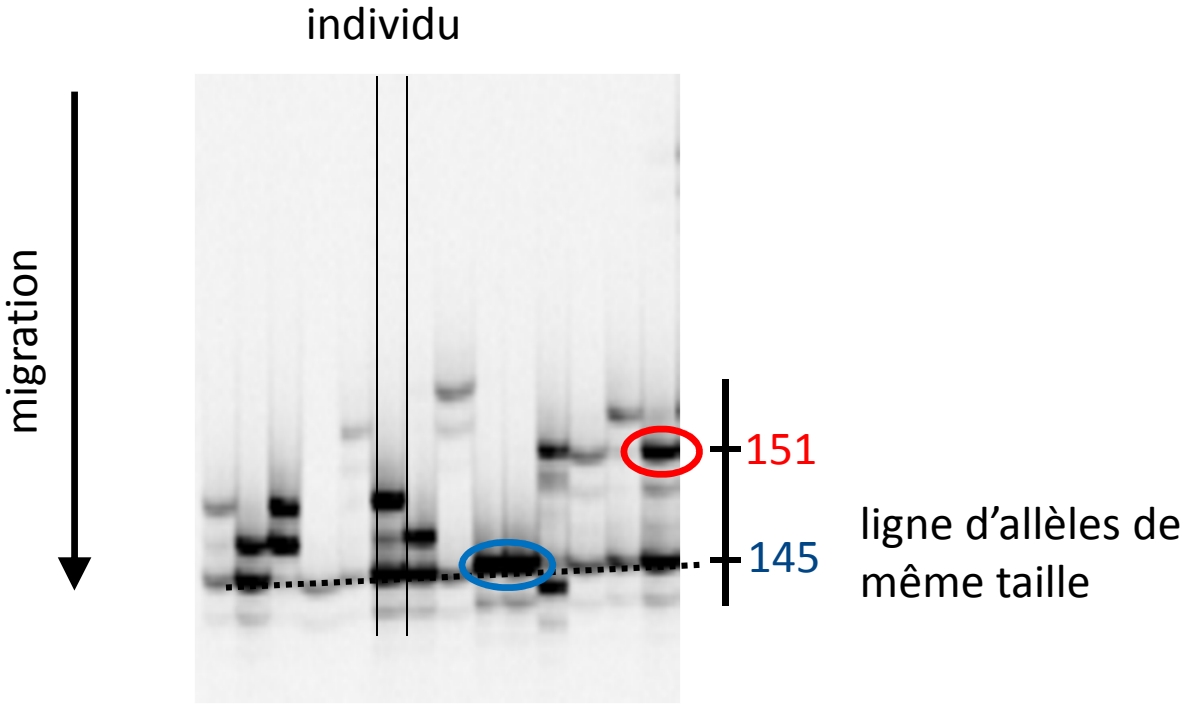
Révélation des allèles de chaque individu

séquençage

électrophorèse capillaire ou gel

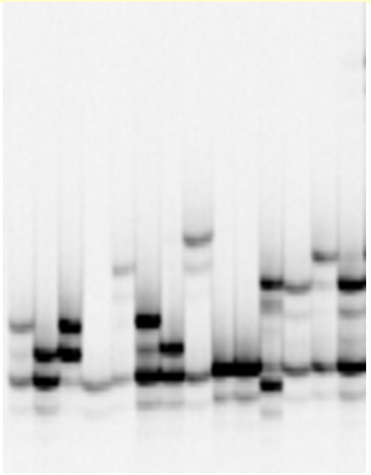
Exemple de résultat brut

Gel d'electrophorèse sur acrylamide révélant des amplifications microsatellites

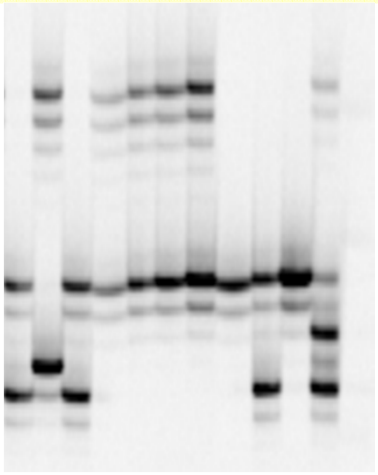


Variabilité génétique à un locus

Echantillon 1



Echantillon 2



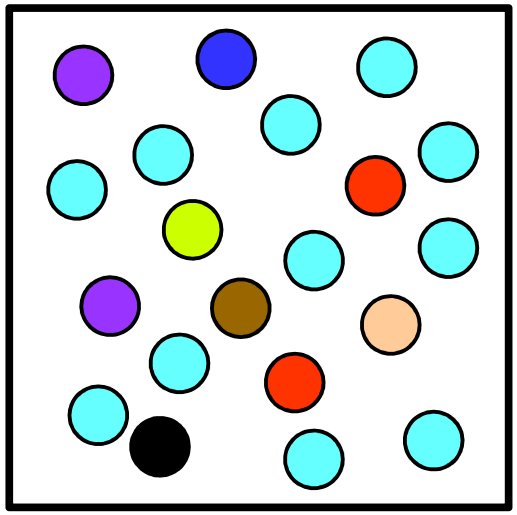
Caractères du polymorphisme

► **Nombre d'allèles** : nombre de séquences de gènes, de locus microsatellite, de profils RFLP, d'allozymes… différents les uns des autres

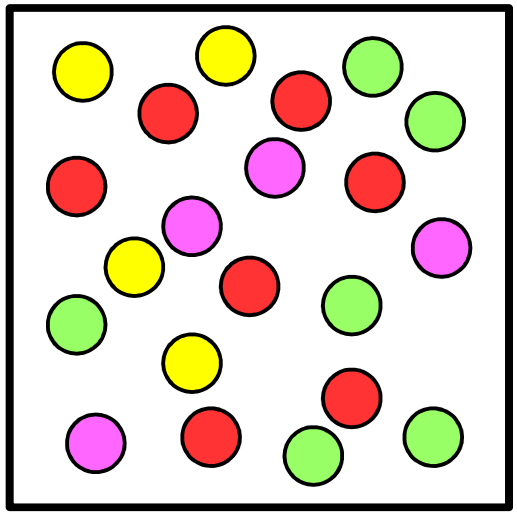
► **Fréquences alléliques** : proportion de chaque allèles dans l'échantillon

Le nombre d'allèles différents ne fait pas tout – leurs fréquences importent

Échantillon 1

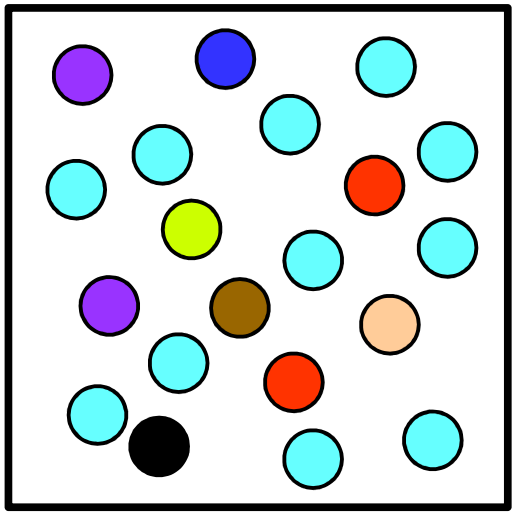


Échantillon 2



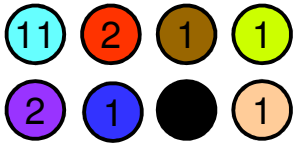
Caractères du polymorphisme

Échantillon 1

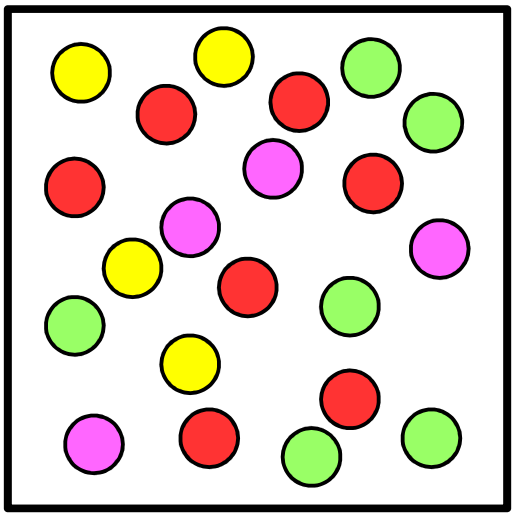


Nombre d'allèles

$A = 8$



Échantillon 2



Nombre d'allèles

$A = 4$

Quel échantillon est le plus divers génétiquement ?

Caractères du polymorphisme

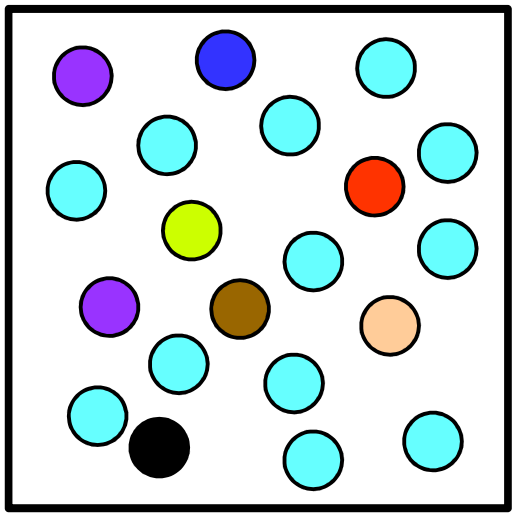
Diversité génétique

► H_E : probabilité de tirer deux allèles différents au hasard

Calcul :
H_E : probabilité de tirer deux allèles différents
= 1 – (probabilité de tirer deux fois le même allèle)
= 1 – (somme des carrés des fréquences alléliques)

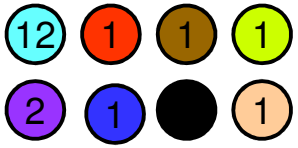
Caractères du polymorphisme

Échantillon 1



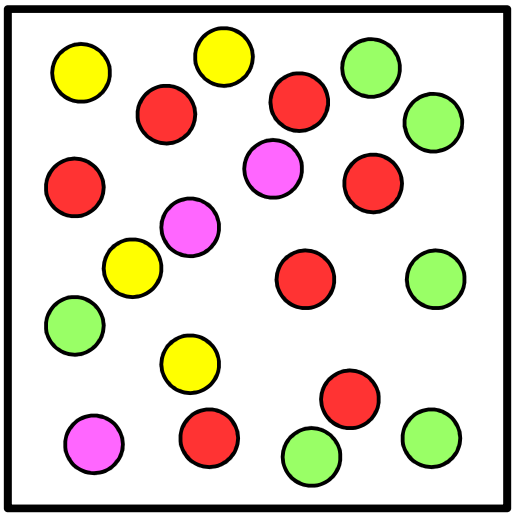
Nombre d'allèles

$A = 8$



$H_E = 0.615$

Échantillon 2

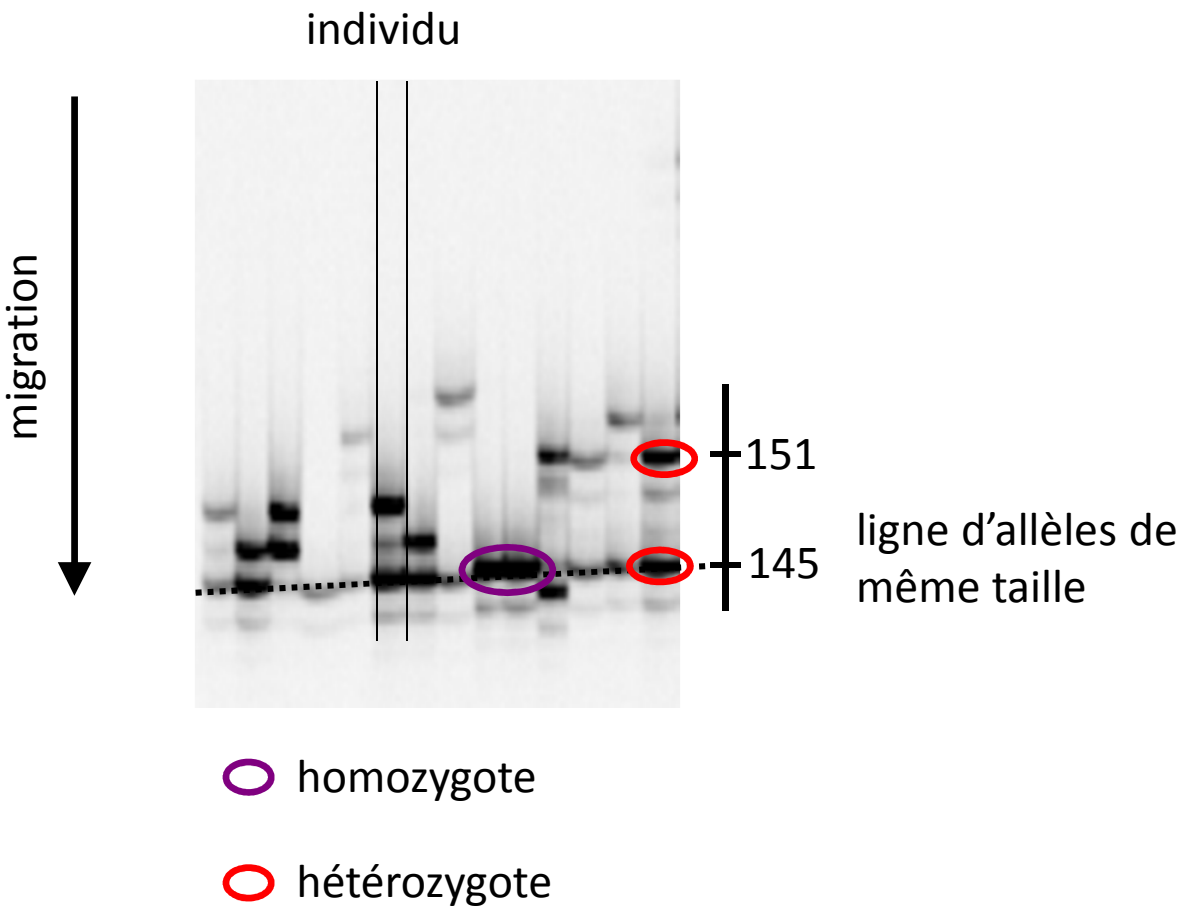


Nombre d'allèles

$A = 4$

$H_E = 0.725$

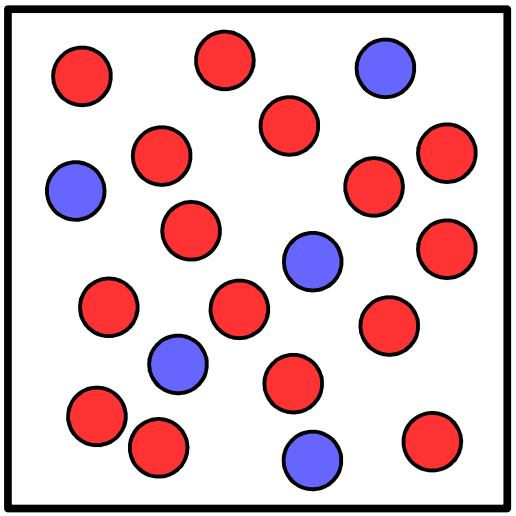
Structure du polymorphisme



Structure du polymorphisme

- Prise en compte du niveau des individus

$2n$ = 20 allèles (16) (4) H_E = 0.32



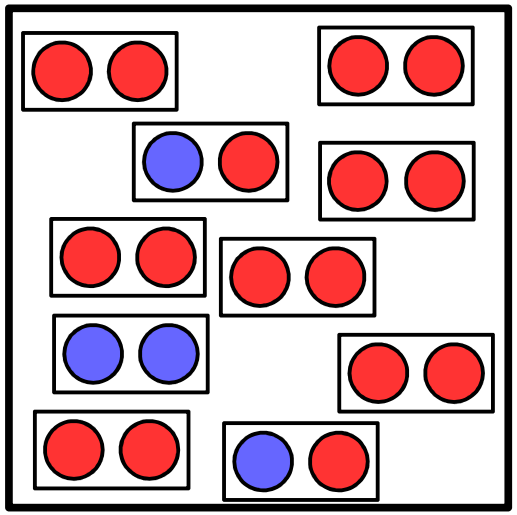
Structure du polymorphisme

- Prise en compte du niveau des individus

$2n$ = 20 allèles H_E = 0.32

H_E : probabilité de tirer deux allèles différents au hasard
= fréquence théorique (« expected ») des individus hétérozygotes

Comparaison avec la fréquence observée dans l'échantillon H_O



n = 10 individus H_O = 1 - 8/10 = 0.2

Structure du polymorphisme

- Prise en compte du niveau des individus

$2n = 20$ allèles $\quad H_E = 0.32$

$n = 10$ individus $\quad H_O = 0.2$

- **Coefficient de consanguinité**

$$F_{IS} = 1 - \frac{H_O}{H_E}$$

$F_{IS} = 0.375$

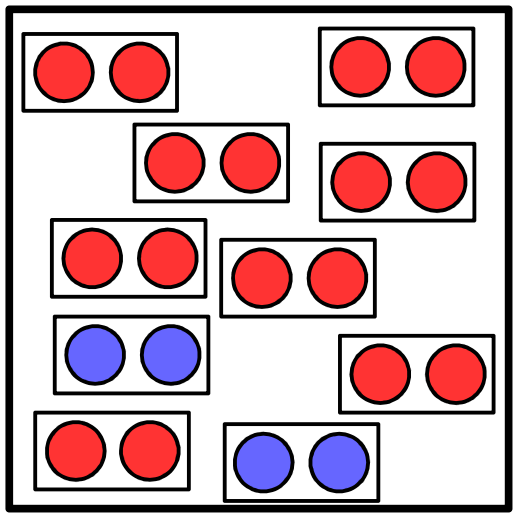
Structure du polymorphisme

- **Coefficient de consanguinité** $F_{IS} = 1 - \frac{H_O}{H_E}$

Structure du polymorphisme

► **Coefficient de consanguinité** $F_{IS} = 1 - \frac{H_O}{H_E}$

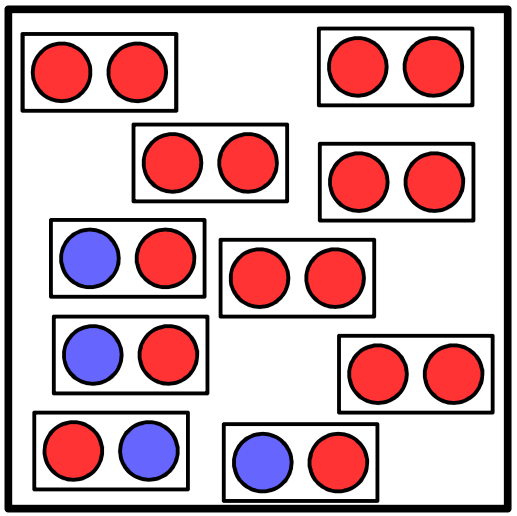
Déficit en hétérozygotes



$H_E = 0.32$
$H_O = 0$

$F_{IS} = 1$

Excès en hétérozygotes



$H_E = 0.32$
$H_O = 0.4$

$F_{IS} = -0.25$

De l'importance des modèles

- Les modèles sont essentiels en génétique des populations
- Simplification de la réalité
- Permettent d'élaborer des **prédictions** sur la base d'hypothèses données
- Permettent également de **tester les hypothèses** (tests statistiques)

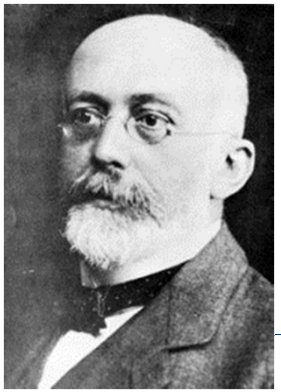
Le modèle de Hardy-Weinberg

- Un locus avec deux allèles A et B en fréquences p et $q = (1 - p)$

- Reproduction au hasard (panmixie) :
- AA = p^2
- AB = $2pq$
- $BB = q^2$

- Nombre infini d'individus et pas de sélection/mutation/migration :
- Les fréquences alléliques ne changent pas

1908

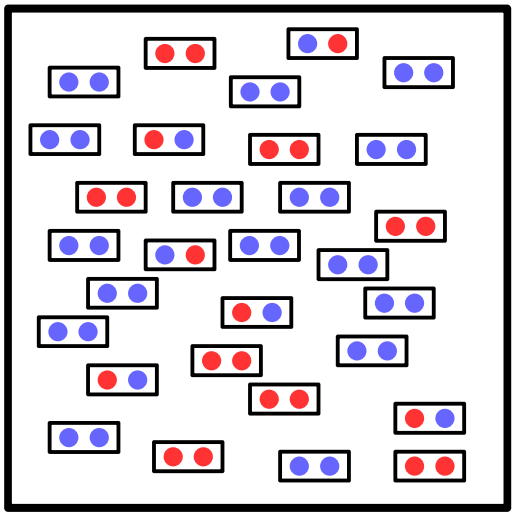


Sous les hypothèses de Hardy et Weinberg :

$F_{IS} = 0$ (nombre d'hétérozygotes = attendu)

Le modèle de Hardy-Weinberg

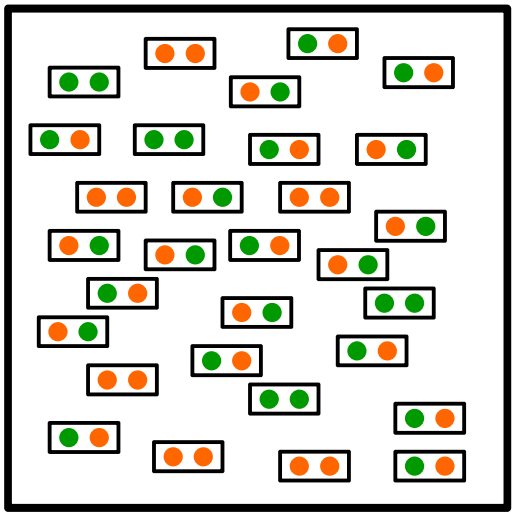
- Prédictions : les fréquences des génotypes sont connues et sont stables
- Test d'un jeu de données observé :



- Fréquences : ● 0.633 ● 0.367
- Attendus : ●● 12.03 ●● 13.93 ●● 4.03
- Observations : ●● 16 ●● 6 ●● 8
- $\chi^2 = 9.73$; $P < 0.005$
- Déficit significatif en hétérozygotes

Le modèle de Hardy-Weinberg

- Prédictions : les fréquences des génotypes sont connues et sont stables
- Test d'un jeu de données observé :



- Fréquences : ● 0.53 ● 0.47
- Attendus : ●● 8.53 ●● 14.93 ●● 6.53
- Observations : ●● 6 ●● 20 ●● 4
- $\chi^2 = 3.45$; $P > 0.05$
- Excès en hétérozygotes… mais non significatif

Causes des écarts à Hardy-Weinberg

- **Évolution des fréquences alléliques**
- Dérive génétique
- Sélection
- Mutation
- Migration

- **Déviation des fréquences génotypiques**
- Régime de reproduction
- Sélection
- Structure de population

Causes des écarts à Hardy-Weinberg

► Autogamie

De nombreuses plantes, animaux, champignons…
se reproduisent par autogamie (ou autofécondation)

Reproduction avec sexe mais sans partenaire (≠ clonal)

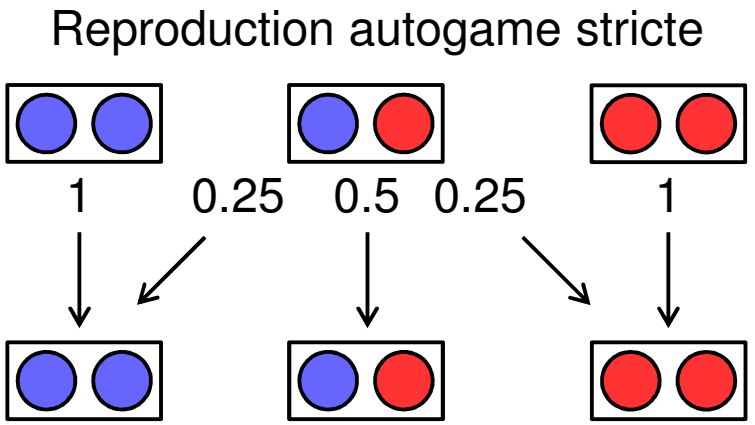
Souvent partielle (régimes mixtes)

Arabidopsis thaliana

Physa acuta

Causes des écarts à Hardy-Weinberg

► Autogamie



Le taux d'hétérozygotes diminuent de 50% à chaque génération

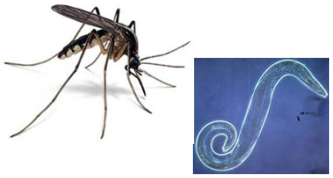
Forts déficits en hétérozygotes

À l'équilibre : $F_{IS} \rightarrow \frac{s}{2 - s}$ (s vaut 1, F_{IS} tend vers 1)

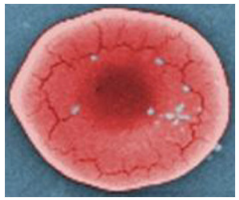
Causes des écarts à Hardy-Weinberg

► **Sélection**

Cas de la drépanocytose en région impaludée

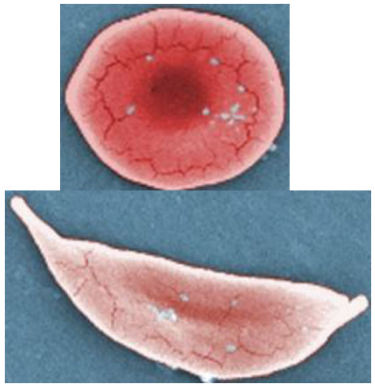


HbAA



Sensible
au paludisme

HbAS



Résistant au paludisme
Maladie bégnine

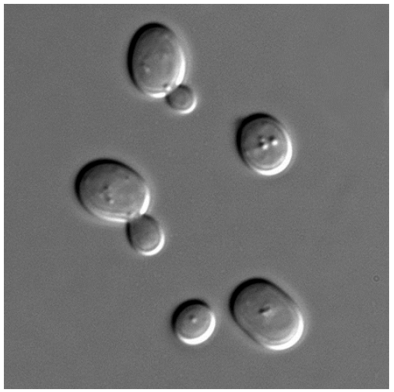
HbSS

Maladie
sévère

Superdominance : fort excès en hétérozygotes

Causes des écarts à Hardy-Weinberg

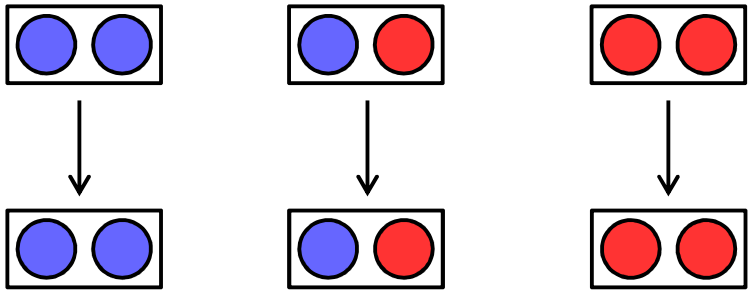
- **Clonalité**



Causes des écarts à Hardy-Weinberg

► **Clonalité**

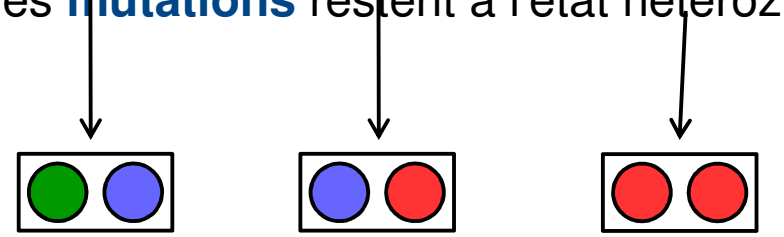
Reproduction clonale stricte



Les fréquences génotypiques ne changent pas

(quelles qu'elles soient…)

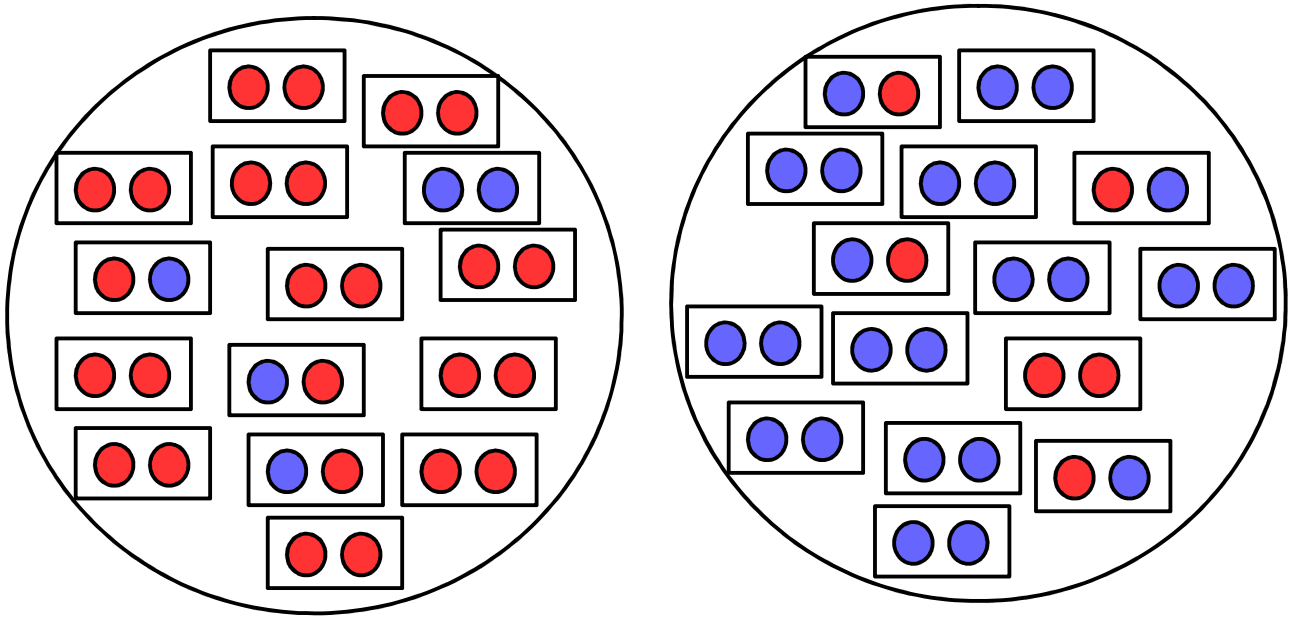
mais les **mutations** restent à l'état hétérozygote

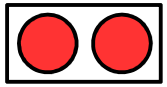
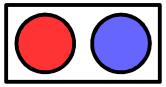
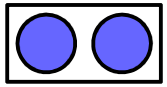


Excès en hétérozygotes

Causes des écarts à Hardy-Weinberg

- **Effet de la sous-structure de population**



						F_{IS}	
Sous-pop 1	0.83	0.16	0.73	0.2	0.07	0.28	$P > 0.05$
Sous-pop 2	0.2	0.8	0.07	0.27	0.67	0.17	$P > 0.05$
Total	0.52	0.48	0.4	0.23	0.37	0.53	**$P < 0.01$**

Variabilité génétique entre locus

► Equilibre gamétique

On considère **2 locus** possédant chacun 2 allèles

f(**A**) = *p* f(**B**) = *r*

f(a) = *q* f(**b**) = *s*

A l'équilibre, association au hasard des génotypes aux deux loci

Lors de la production des gamètes :

P(AB) = *pr*

P(Ab) = *ps*

P(aB) = *qr*

P(ab) = *qs*

gamètes en couplage : P(AB) x P(ab) = *prqs*

gamètes en répulsion : P(Ab) x P(aB) = *psqr*

P(AB) x P(ab) - P(Ab) x P(aB) = 0

Variabilité génétique entre locus

► Déséquilibre gamétique

On considère **une liaison statistique** entre 2 allèles de 2 locus

Par exemple A est lié à B, alors $P(AB) > pr$

On pose $D = P(AB) - pr$

D est appelé déséquilibre gamétique ou déséquilibre de liaison

$P(AB) > pr$

$P(Ab) < ps$

$P(aB) < qr$

$P(ab) > qs$

1

gamètes en couplage : $P(AB) \times P(ab) > prqs$

gamètes en répulsion : $P(Ab) \times P(aB) < psqr$

$P(AB) \times P(ab) \neq P(Ab) \times P(aB)$

Variabilité génétique entre locus

► Déséquilibre gamétique

Multiples causes du déséquilibre de liaison

- Sélection, si des associations entre allèles sont favorisées

- Modes de reproduction
  clonalité (absence de recombinaison)
  consanguinité (croisements entre proches)

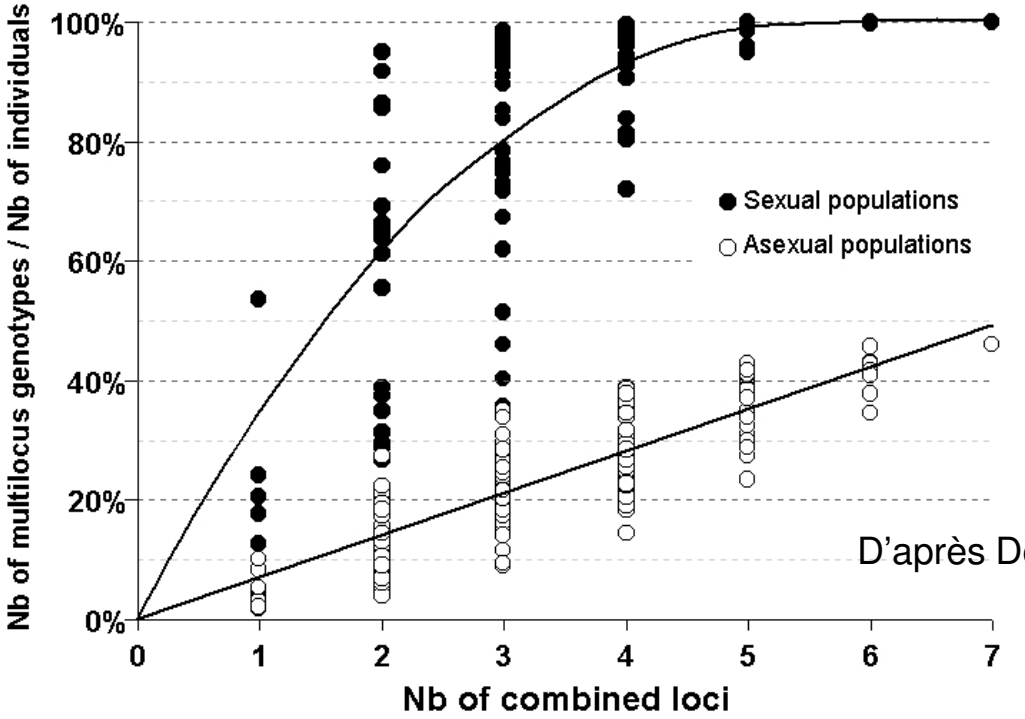
- Faible effectif dans la population (isolement reproducteur)

- Mélange de populations

Variabilité génotypique

Mesure simple Dg = ng/N

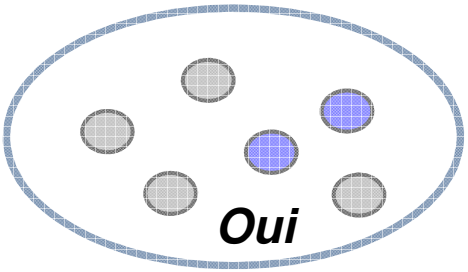
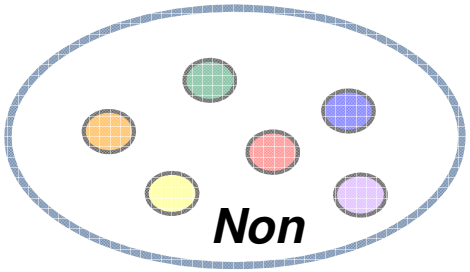
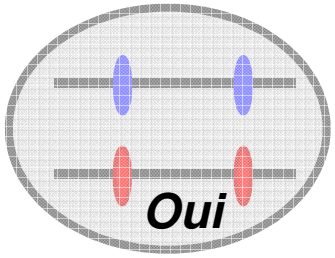
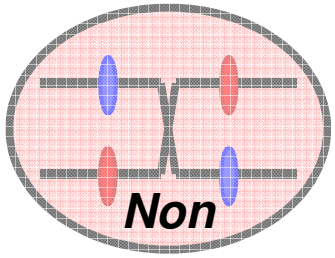
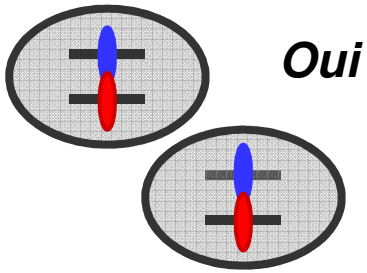
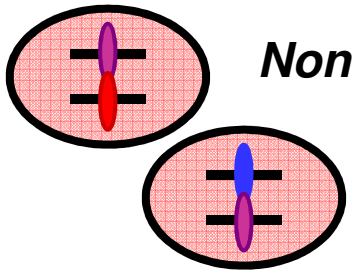
Mais problème d'interprétation :



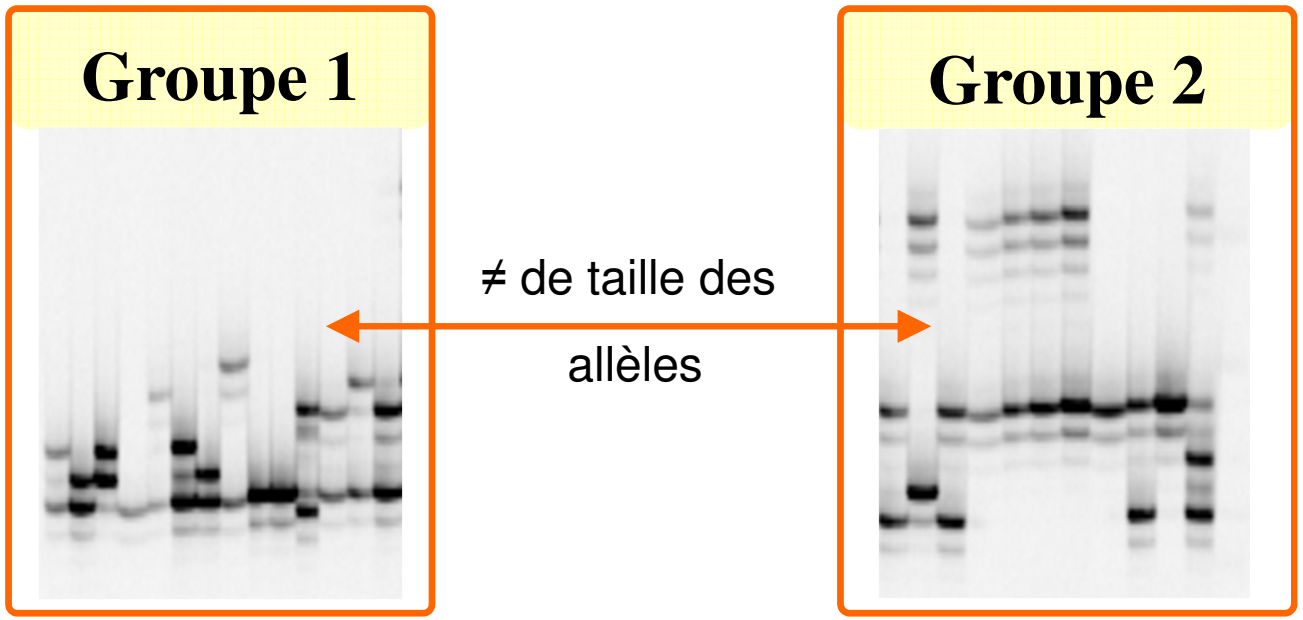
D'après Delmotte et al. 2002

Nécessite de définir la probabilité d'observer *n* fois le génotype et calculer un niveau de significativité

Caractéristiques génétiques et clonalité

	Clonalité	Sexualité	Indices
Génotypes *répétés ?*			n_G/N
Loci *liés ?*			D
Allèles *divergents ?*			$F_{IS} < 0$

Diversité vs différenciation génétique

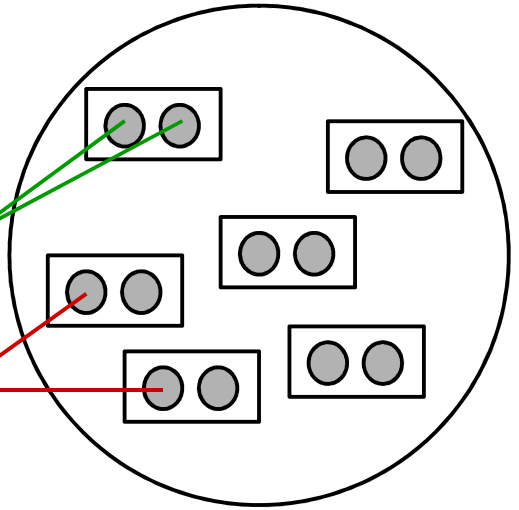


- Hétérozygotie observée H_O
- Diversité génétique $H_E = 1 - \sum p_i^2$

Indice de fixation entre populations

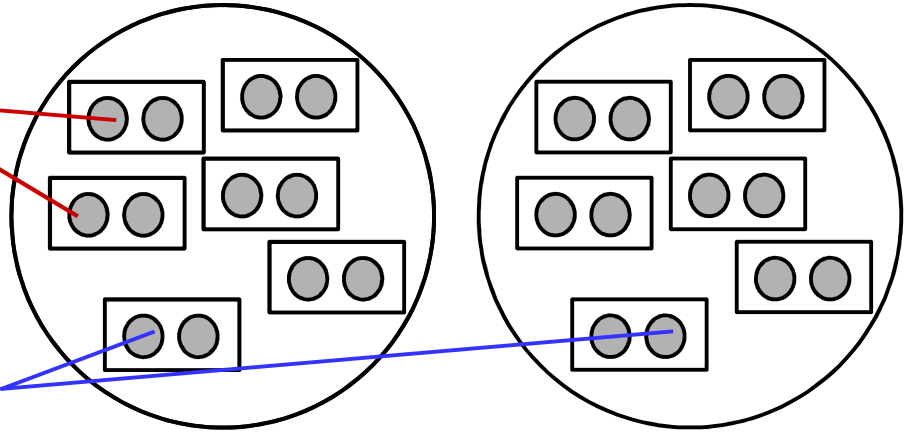
- **Fixation intra-individus**

$$F_{IS} = 1 - \frac{H_O}{H_E}$$



- **Fixation intra-population**

$$F_{ST} = 1 - \frac{H_S}{H_T}$$



Indice de fixation entre populations

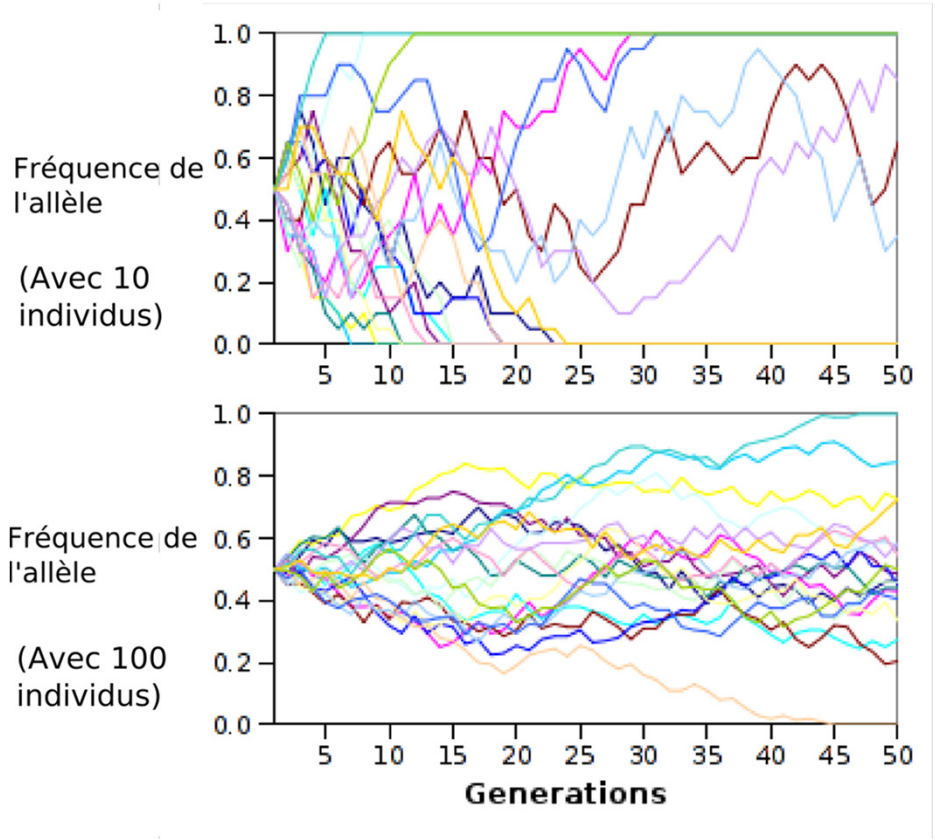
- **$F_{IS} = 0$** Intra-individus = ensemble de la population
- **$F_{IS} > 0$** Intra-individus moins variable que l'ensemble de la population
- **$F_{IS} = 1$** **Fixation** : pas de variation intra-individus

- **$F_{ST} = 0$** Intra-population = total (toutes les populations)
- **$F_{ST} > 0$** Intra-population moins variable que total (plus d'homozygotie)
- **$F_{ST} = 1$** **Fixation** : pas de variation intra-population

Valeur d'équilibre du $F_{ST} \approx \dfrac{1}{1 + 4Nm}$

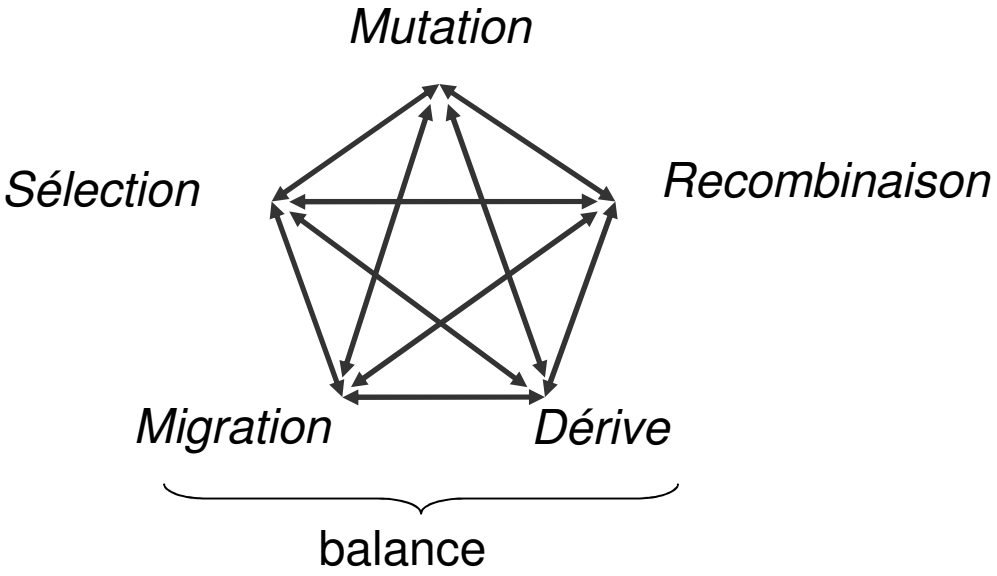
Modèle de Wright-Fisher – dérive génétique

► **Modèles en taille de population finie** (2*N* individus diploïdes)



Les allèles **dérivent**
et finissent par se fixer

Quelles forces différencient les populations ?

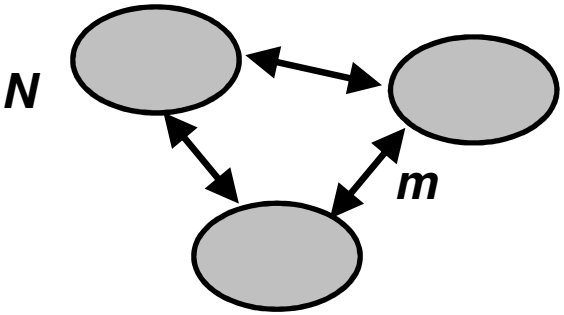


Principaux facteurs d'isolements reproducteurs

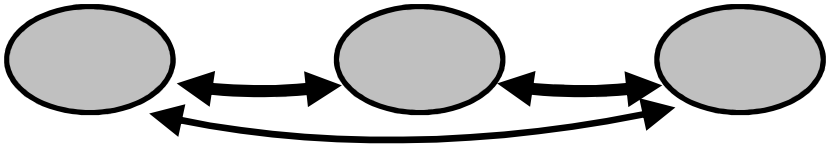
- espace
- temps
- hôtes

Estimation de la dispersion

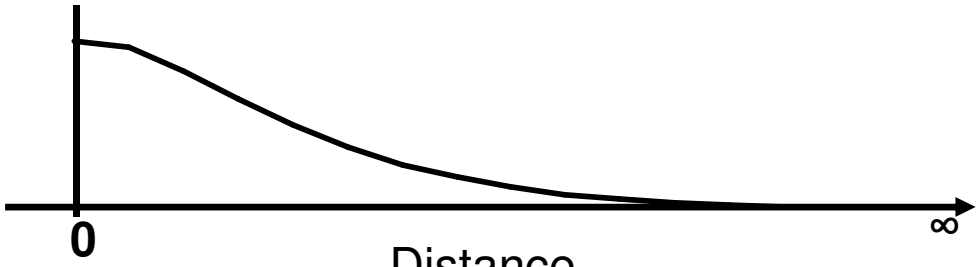
Modèle en îles ***vs*** **modèle d'Isolement par la distance**



$$F_{ST} \approx \frac{1}{1+4Nm}$$

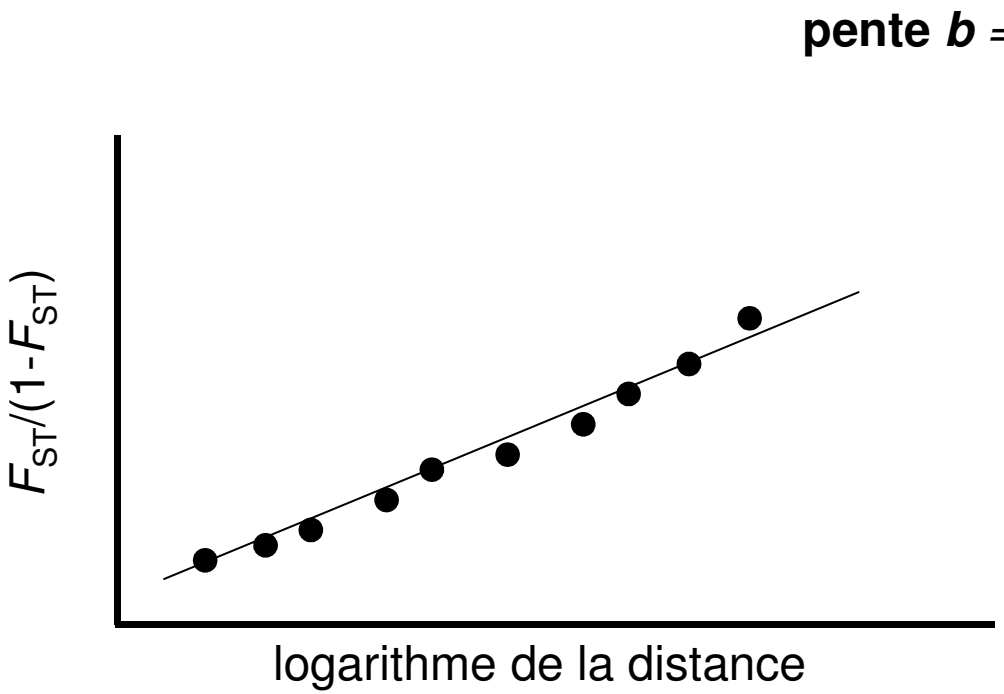


Dispersion graduelle



Modèle d'isolement par la distance

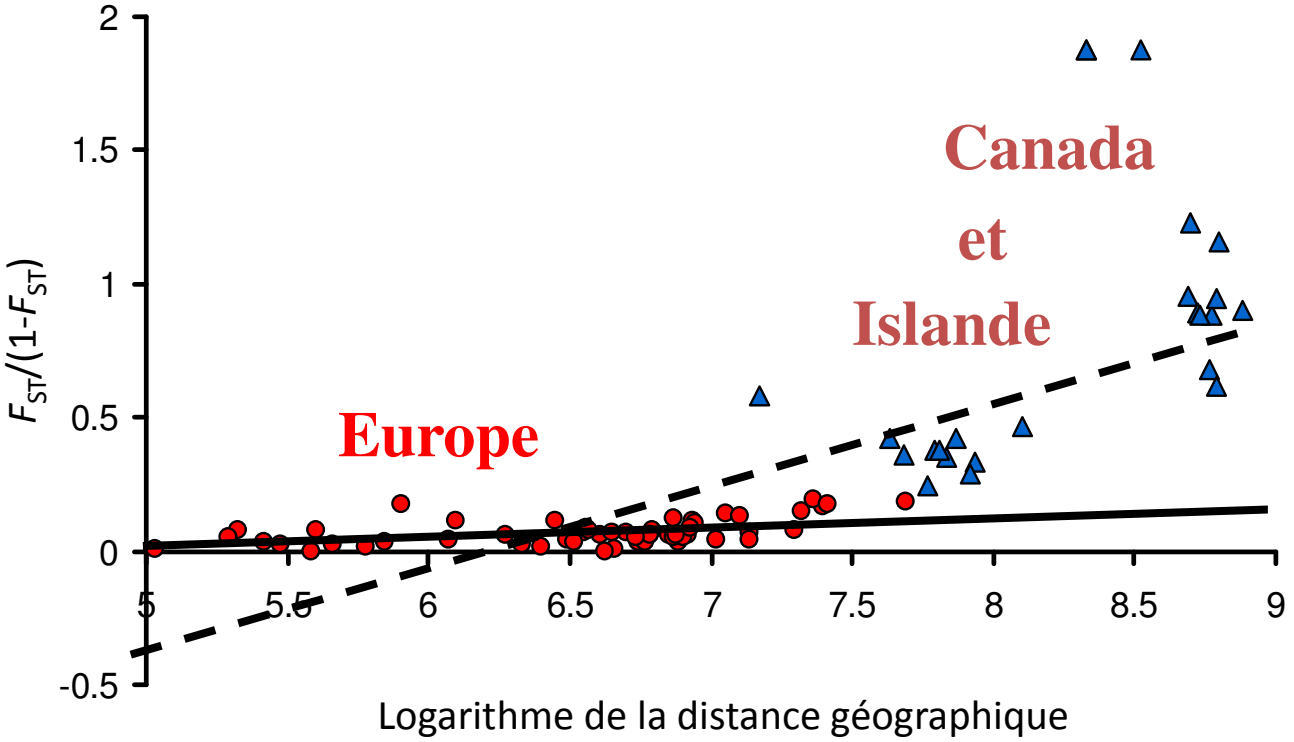
Représentation graphique



pente $b = 1 / (4\pi D\sigma^2)^2$

densité en individus **x** distance moyenne de dispersion

Un exemple d'isolement par la distance

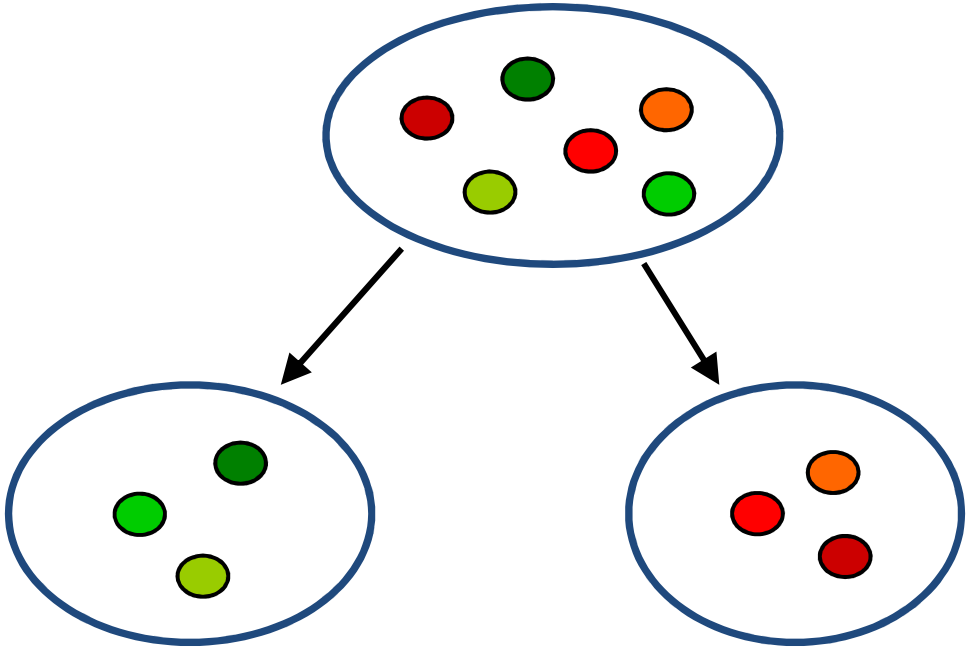


Rupture d'isolement par la distance

Qu’est ce qu’une population ?

Auparavant l’objet d’étude **‘population’** dépendait de l’échantillonnage

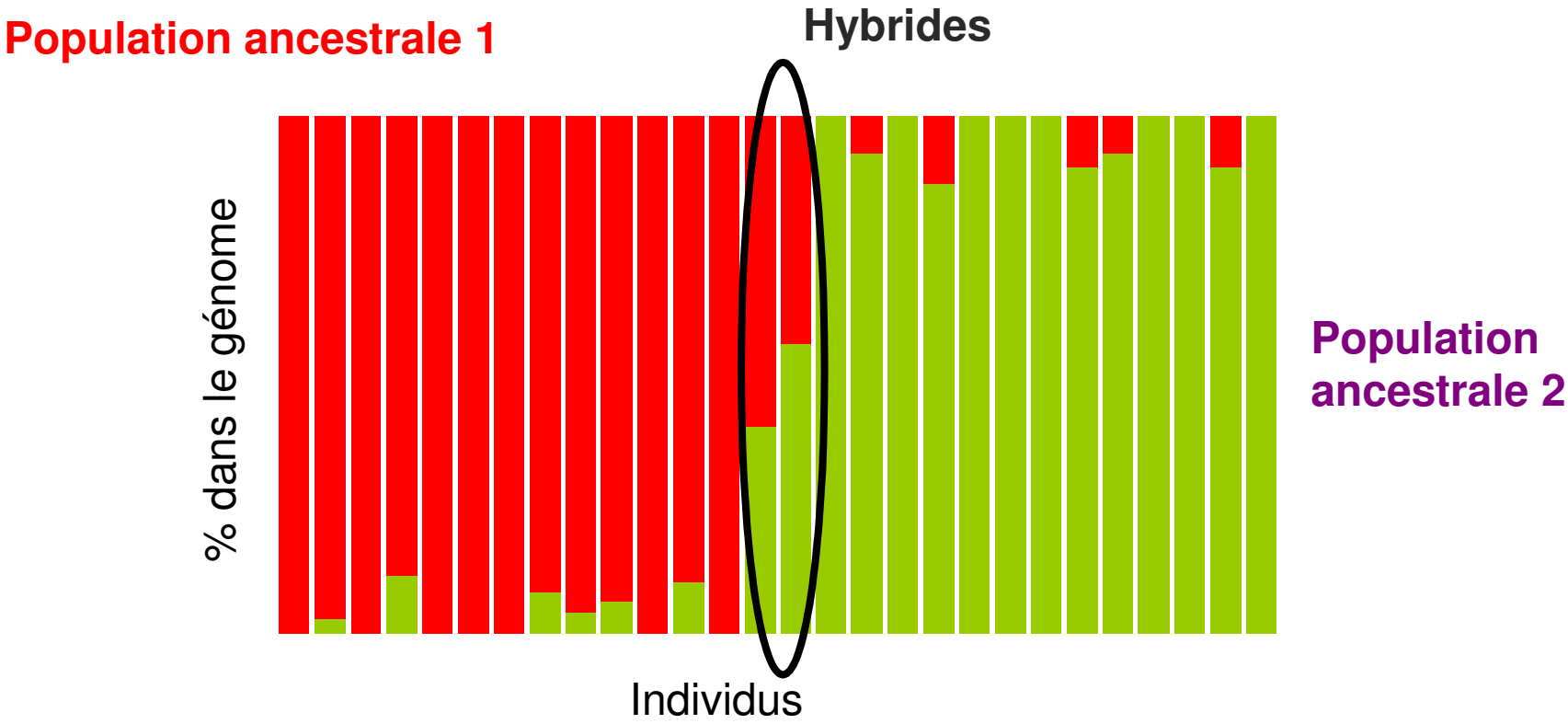
Plus récemment



Regroupement des individus selon leur **proximité génétique**

Intérêt des assignations Bayesiennes

Approche individu centrée



Modèle neutre de l'évolution moléculaire

1968

- Les mutations apparaissent à un taux μ
- Les mutations dérivent en fonction de la taille de la population *N*
- Calcul : nombre de mutations, fréquences, temps de fixation, taux d'évolution

- Hypothèse :

Le polymorphisme des populations naturelles est *majoritairement* expliqué par l'action de mutations neutres et de la dérive génétique

→ Mutations favorables rares
→ Mutations délétères rapidement éliminées

MOTOO KIMURA

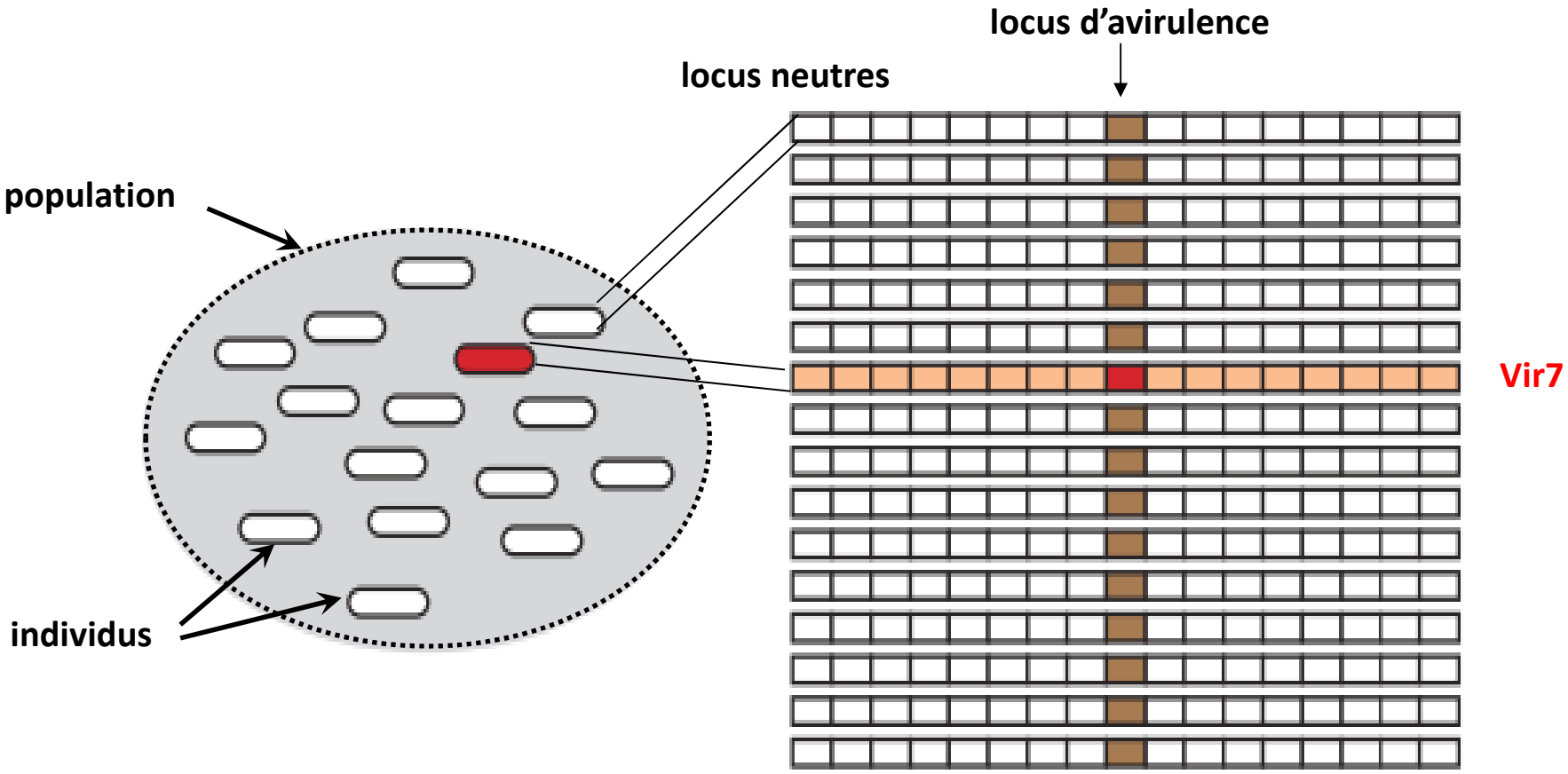
Le modèle neutre à l'ère génomique

- Données génétiques années 1980/90 : modèle neutre généralement vérifié
- Utilisé comme *hypothèse nulle* pour les tests de détection de la sélection
- Mais le modèle neutre fait aussi des hypothèses démographiques
  (population stable dans le temps, non structurée…)
- Ère génomique : données à grande échelle (grande résolution)
- Quantification des effets sélectifs
- Non négligeables à grande échelle
- Épisodes récurrents à effets faibles
- Épisodes intenses ponctuels affectant les régions voisines

Effets d'un épisode de sélection

Vue de la population :

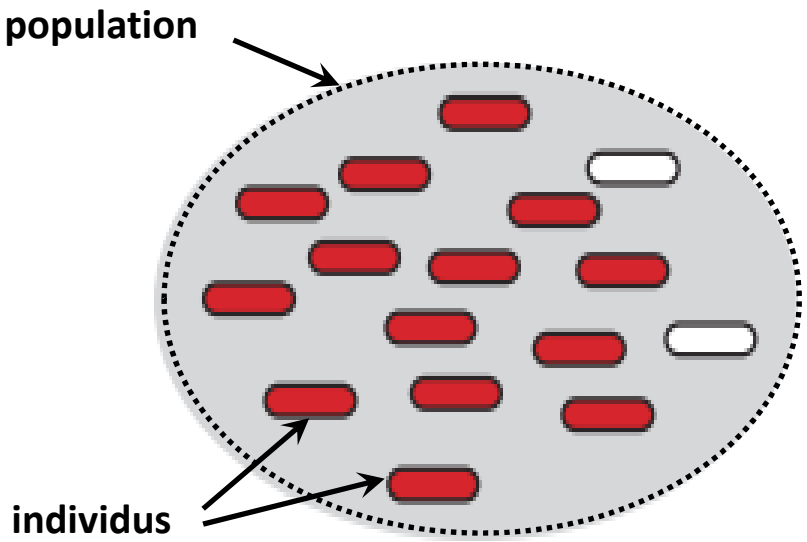
Organisation du génome :



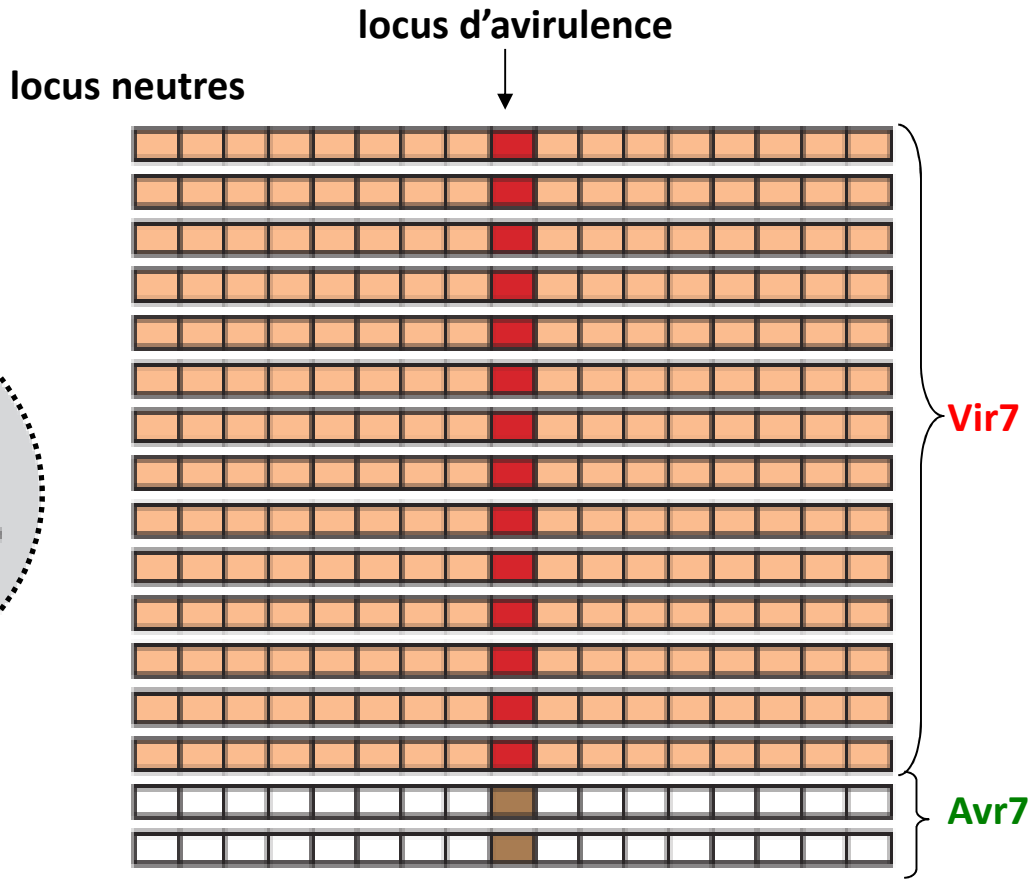
Au départ peu d'individus sélectionnés…

Effets d'un épisode de sélection

Vue de la population :



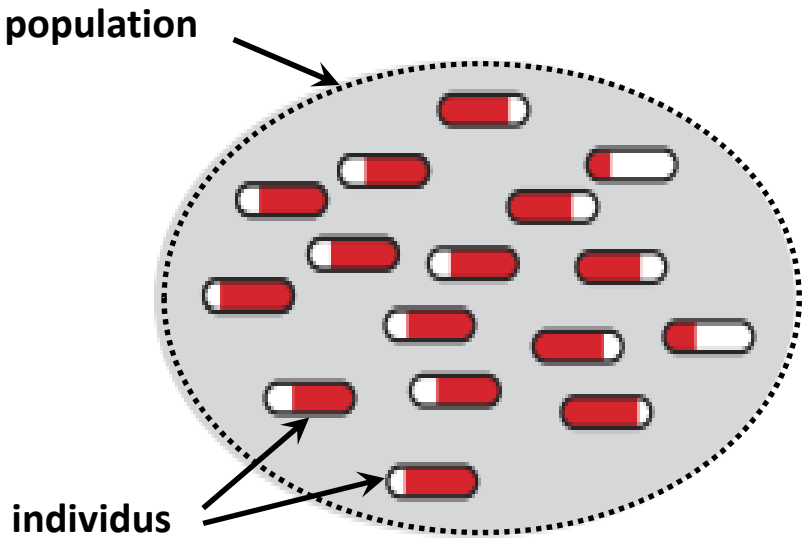
Organisation du génome :



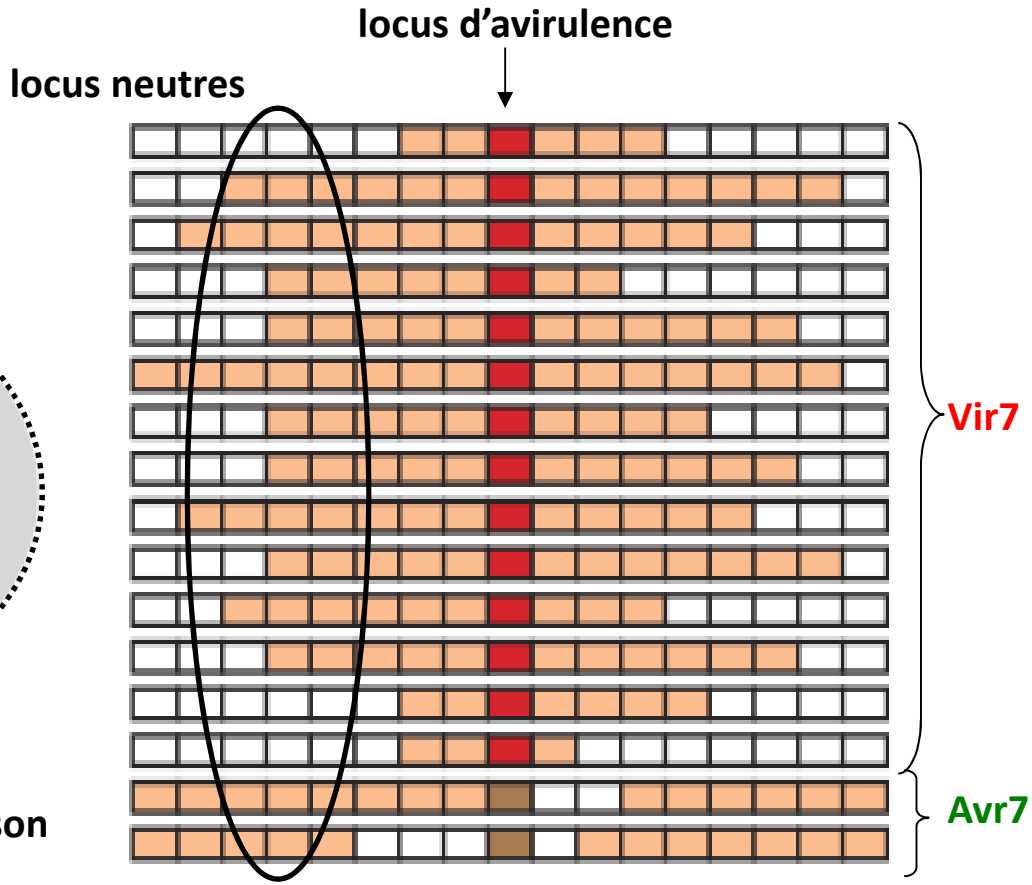
… qui se multiplient rapidement

Effets d'un épisode de sélection

Vue de la population :



Organisation du génome :

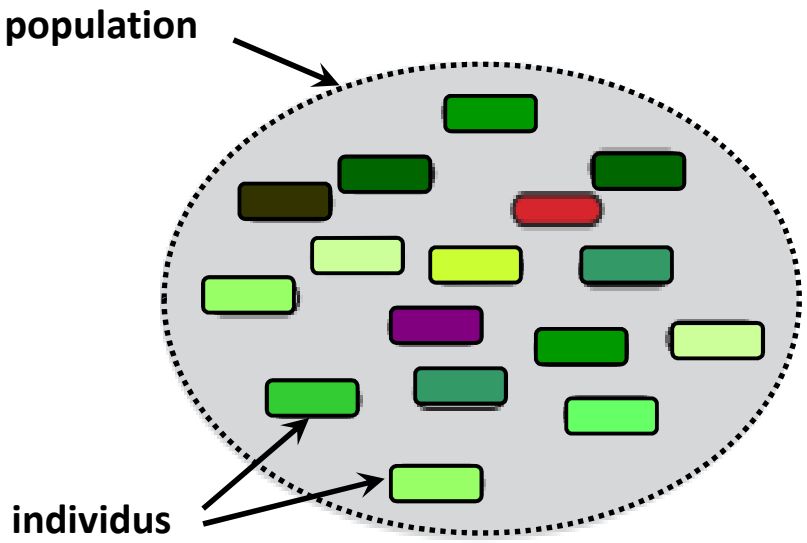


Évolution par recombinaison

Puis se mélangent aux autres individus

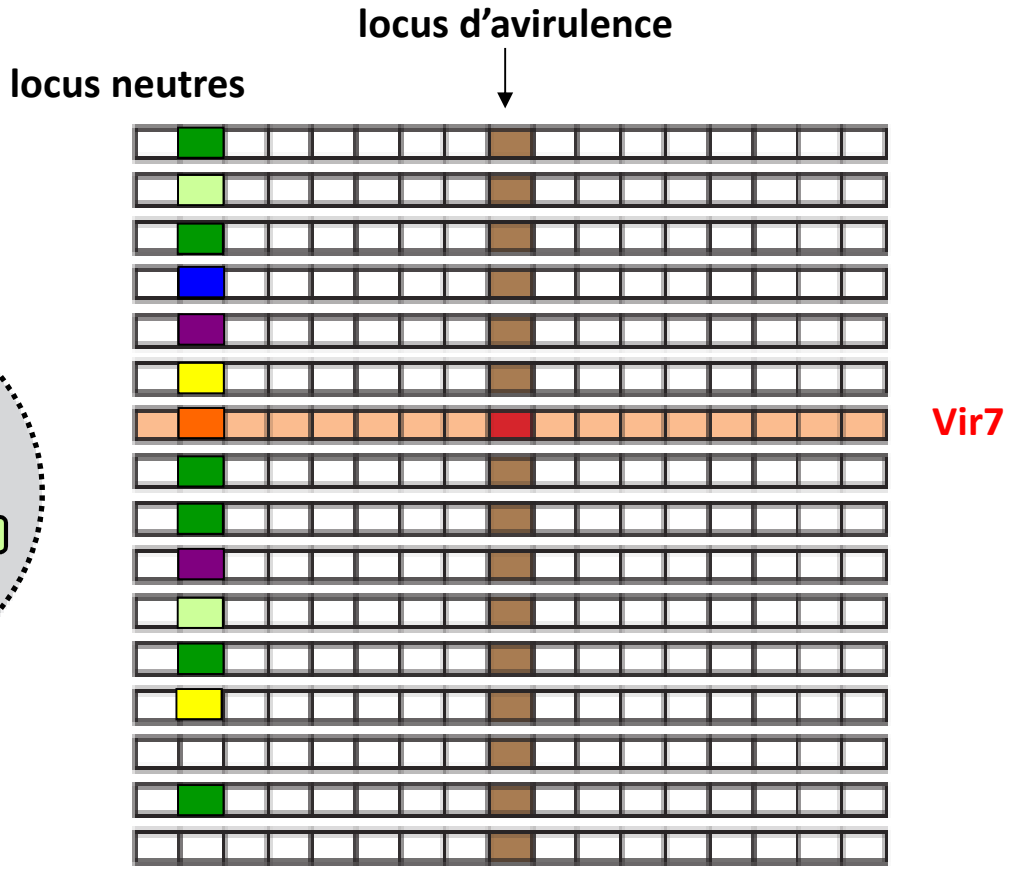
L'effet d'auto-stop

Vue de la population :



Population initiale diversifiée

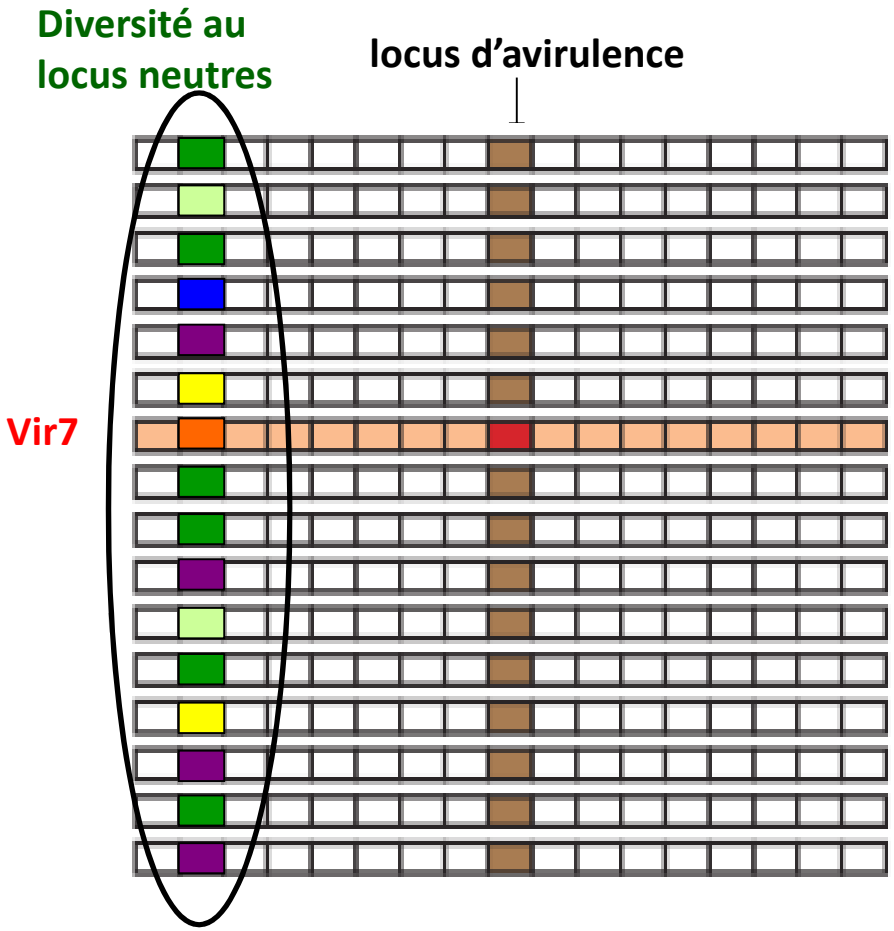
Organisation du génome :



L'effet d'auto-stop

Au tout début de l'évènement de sélection

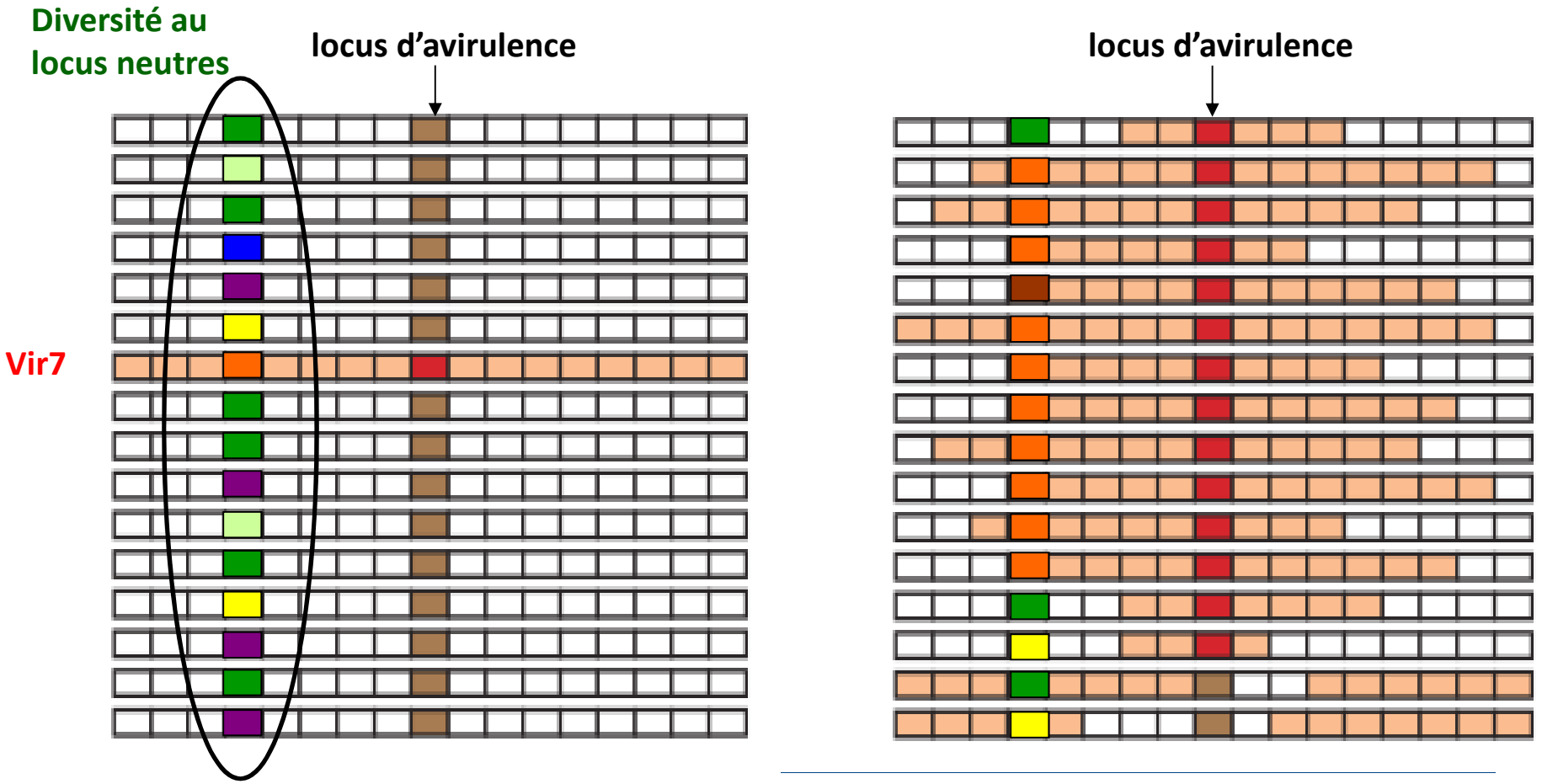
Après un certain temps



L'effet d'auto-stop

Au tout début du contournement

Après un certain temps



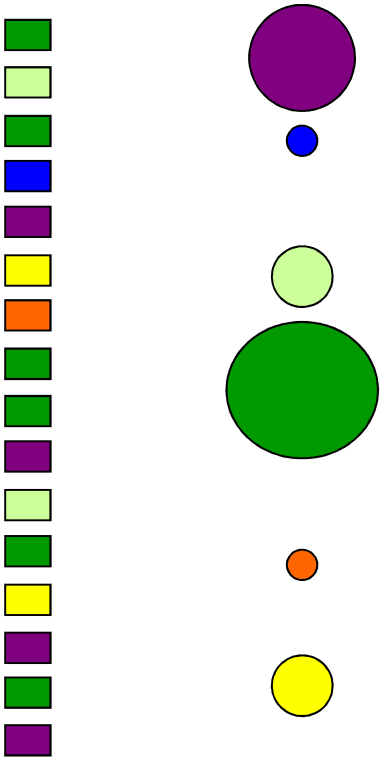
Evolution des fréquences alléliques

Au tout début du contournement

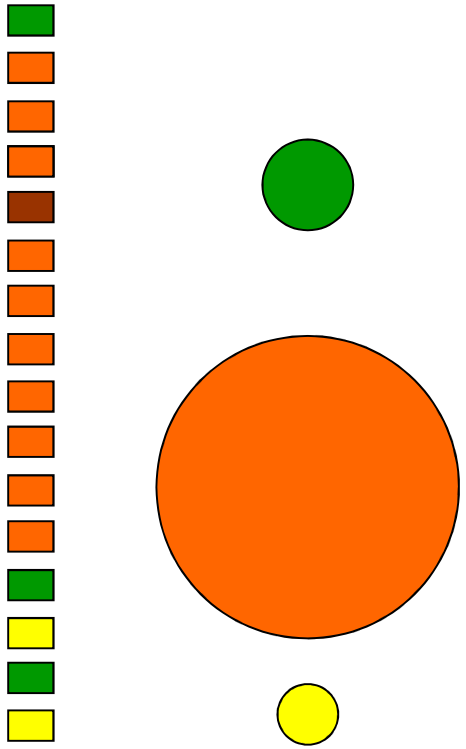
Après un certain temps

Diversité au locus neutres

Fréquence initiale

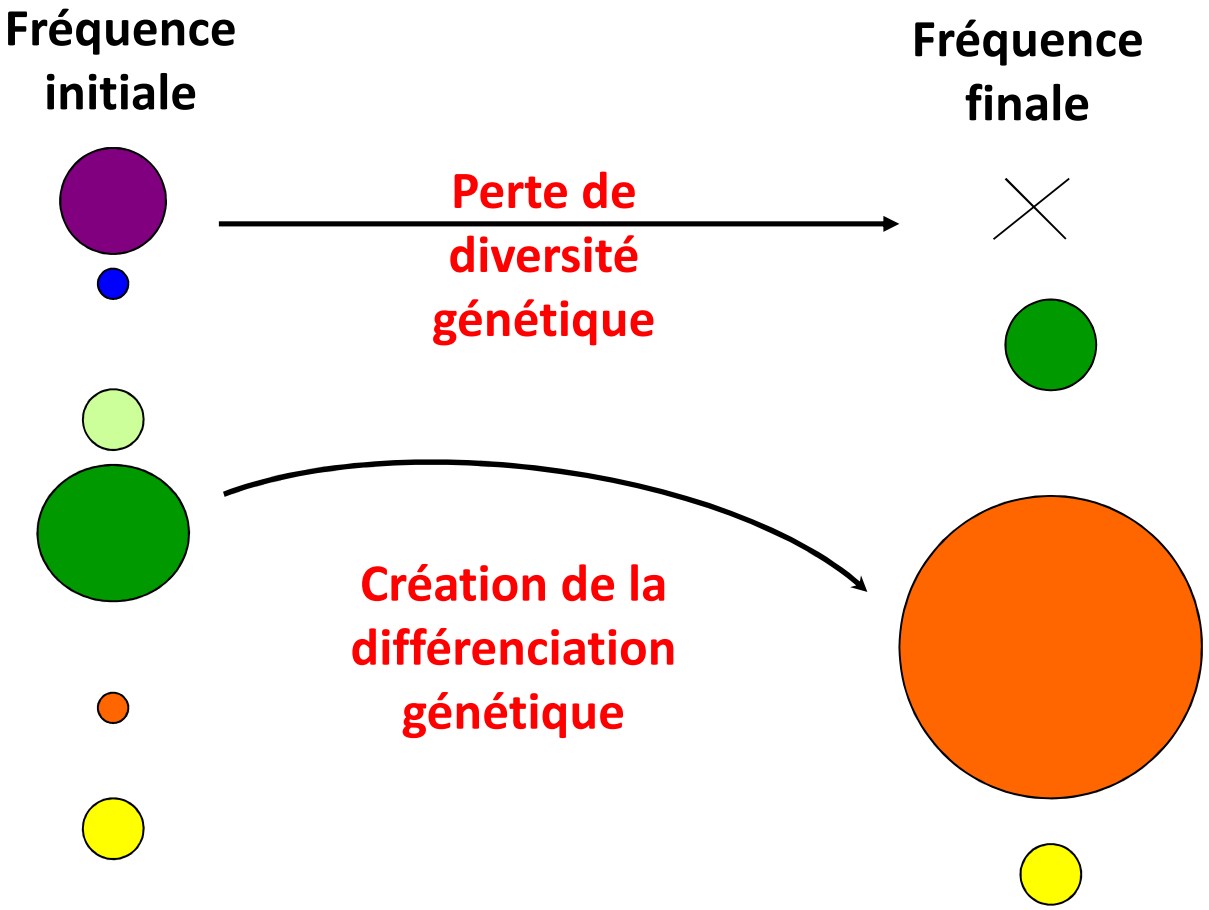


Fréquence finale

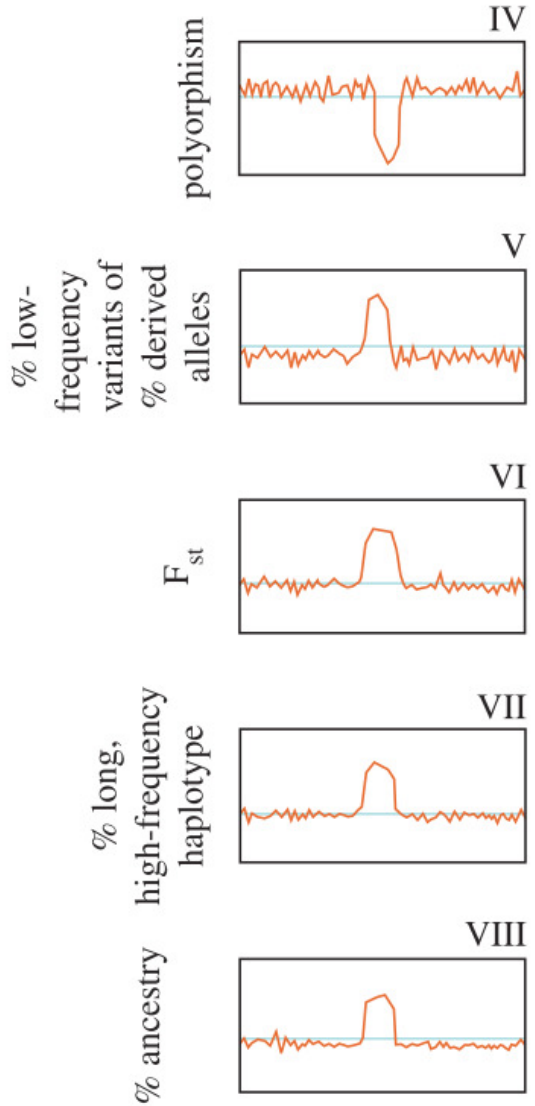


Evolution des fréquences alléliques

Au tout début du contournement **Après un certain temps**



Scan génomique



Variation des indices le long du génome :

- polymorphisme
- variants en faibles fréquences
- forte différenciation
- haplotype plus long
- différences d'ancestralité