

# Brève introduction à la génétique des populations

Christine Dillmann, Adrienne Ressayre

January 22, 2023

## 1 Fréquences alléliques, fréquences génotypiques et cycle de vie

### 1.1 Quelques définitions

Considérons une population d'une espèce diploïde, dans laquelle on observe un polymorphisme à un locus.

- **Locus** : position sur le chromosome.
- **Gène** : une séquence d'ADN codant pour une fonction. À un locus, un individu diploïde possède deux gènes, l'un hérité de son père et l'autre de sa mère
- **Allèle** : On peut regrouper les gènes en classes fonctionnelles appelées allèles, et le plus souvent représentées par des lettres. Par exemple,  $A$  et  $a$  sont les deux allèles présents au locus  $A$  dans la population.
- **Génotype** : Un individu d'une population est décrit par son génotype, c'est à dire les allèles qu'il possède à un locus. Pour un locus biallélique, il y a 3 génotypes possibles,  $AA$ ,  $Aa$ ,  $aa$ . Les individus possédant deux allèles différents au même locus sont appelés hétérozygotes.
- **Population** : Une population est un ensemble d'individus pouvant se reproduire entre eux, c'est à dire des individus connectés dans l'espace par la reproduction.
- **Fréquences génotypiques** : On représente une population comme la liste des génotypes présents en associant, à chaque génotype, une fréquence qui est la proportion d'individus de la population ayant ce génotype.

Génotype	$AA$	$Aa$	$aa$
Fréquence	$f_{AA}$	$f_{Aa}$	$f_{aa}$

Table 1: Génotypes et fréquences génotypiques dans une population à un locus biallélique

- **Fréquences alléliques** : La fréquence d'un allèle est la proportion de copies de cet allèle dans la population. Les fréquences alléliques se calculent à partir des fréquences génotypiques. Si l'on appelle  $p$  la fréquence de l'allèle  $A$  et  $q = 1 - p$  la fréquence de l'allèle  $a$ , on a

$$p = f_{AA} + \frac{1}{2}f_{Aa}$$

- **Polymorphisme** Un locus est polymorphe dans une population s'il existe plusieurs allèles différents, c'est à dire si  $p \neq 1$  et  $p \neq 0$ .
- **Variable aléatoire** : Une variable aléatoire est une variable à laquelle on peut associer une loi de probabilité. Le génotype  $G$  d'un individu pris au hasard dans la population est un variable aléatoire que peut prendre 3 valeurs possibles  $AA$ ,  $Aa$  ou  $aa$  avec les probabilités décrites dans la table 1.

## 1.2 Cycle de vie

Pour comprendre l'évolution de la composition génétique d'une population au cours du temps, il faut prendre en compte le cycle de vie des individus. Ce qui caractérise la plupart des êtres vivants est l'existence de deux phases, une phase haploïde ou gamétique, et une phase diploïde ou sporophytique de longueurs variables, séparées par la fécondation et la méiose. A la méiose, les individus diploïde produisent des gamètes, mâles et femelles. La fécondation est la rencontre d'un gamète mâle et d'un gamète femelle pour former un zygote diploïde.

Pool gamétique	$A$	$a$	Génération $g$
	$p$	$q$	<b>FECONDATION</b>
Zygotes	$AA$	$Aa$	$aa$
	$f_{AA}$	$f_{Aa}$	$f_{aa}$
	$w_{AA}$	$w_{Aa}$	$w_{aa}$
			<i>sélection</i>
Adultes reproducteurs	$f'_{AA}$	$f'_{Aa}$	$f'_{aa}$
Gamètes	$A$	$a$	
	$p'$	$q'$	
			<i>mutation</i>
			<i>dérive génétique</i>
	$\sigma_A$	$\sigma_a$	<i>sélection</i>
Pool gamétique	$p''$	$q''$	Génération $g + 1$

En outre, les évènements qui peuvent se produire durant chacune des phases vont conduire à une modification des fréquences génotypiques et des fréquences alléliques. On appelle **pression évolutive** tout évènement qui conduit à un changement des fréquences des allèles au cours du temps.

## 2 Pressions évolutives

### 2.1 Régime de reproduction

Les fréquences génotypiques dépendent des fréquences des allèles du pool gamétique fécondant, mais aussi de la façon dont les individus ou les gamètes se rencontrent, que l'on appelle régime de reproduction. Le cas le plus simple est la **panmixie** où les gamètes se rencontrent au hasard. En **autofécondation** stricte, un individu se reproduit exclusivement avec lui-même. Dans ce cas, la composition du pool de gamètes mâles est différente pour chaque individu et dépend du génotype de la mère.

En panmixie, les fréquences génotypiques dans les zygotes valent,

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

En panmixie, et en l'absence de pressions évolutives, la fréquence des allèles ne change pas :  $p'' = p$ .

## 2.2 Sélection sporophytique

La sélection se traduit par le fait que le génotype détermine la capacité des individus à participer à la génération suivante. On peut formaliser l'effet de la sélection en considérant qu'elle modifie la proportion des différents génotypes parmi les adultes reproducteurs. La **valeur sélective**  $w_G$  du génotype  $G$  mesure sa contribution au pool gamétique qui constitue la génération suivante. La fréquence du génotype  $AA$  parmi les adultes reproducteurs s'écrit, par exemple,

$$f'_{AA} = \frac{f_{AA}w_{AA}}{\bar{w}}$$

où  $\bar{w} = f_{AA}w_{AA} + f_{Aa}w_{Aa} + f_{aa}w_{aa}$  est la valeur sélective moyenne de la population.

On peut calculer  $\Delta p_{sel} = p' - p$  le changement de la fréquence de l'allèle  $A$  sous l'effet de la sélection,

$$\Delta p_{sel} = \frac{pq}{\bar{w}} (p(w_{AA} - w_{Aa}) - q(w_{AA} - w_{Aa})) = pq \frac{d \ln(\bar{w})}{dp} \quad (1)$$

Ainsi, en l'absence d'autre pressions évolutives que la sélection sporophytique, le devenir d'un allèle favorable dépend de la valeur sélective des hétérozygotes par rapport à celle des homozygotes. Il existe un état d'équilibre polymorphe uniquement lorsque la valeur sélective des hétérozygotes est supérieure à celle des deux homozygotes (**superdominance**). Dans tous les autres cas, un allèle est perdu. Ainsi, la sélection naturelle ne permet pas en général d'expliquer le polymorphisme génétique observé dans les populations. Il existe cependant des exceptions notables :

- Sélection fréquence dépendance.
- Sélection variable dans le temps ou l'espace.

L'équation (1) montre aussi que la sélection naturelle conduit au changement des fréquences des allèles dans le sens d'une augmentation de la valeur sélective moyenne de la population, mais ne garantit pas que la valeur sélective moyenne d'une population à l'équilibre soit la plus grande possible. Il n'y a pas de principe d'optimisation.

**La fitness est une mesure relative :** On voit bien d'après (1) que l'unité dans laquelle est exprimée la valeur sélective n'a pas d'importance, tant que l'on s'intéresse aux changements de la fréquences des allèles. On peut donc utiliser indifféremment les notations suivantes :

Valeurs sélectives		
$AA$	$Aa$	$aa$
$w_{AA}$	$w_{Aa}$	$w_{aa}$
$1 + s$	$1 + hs$	$1$
$1$	$1 - hs$	$1 - s$

$s$  est appelé le coefficient de sélection, et  $h$  le degré de dominance.  $s$  et  $h$  sont compris entre 0 et 1.

## 2.3 Mutation

Une mutation est une modification de la séquence d'ADN due à une erreur lors de la phase de réplication à la méiose. Elle se traduit par l'apparition d'un nouveau gène dans les gamètes produites par un individu. Ce nouveau gène peut appartenir à une classe allélique différente. Le taux de mutation peut être défini au niveau allélique comme la probabilité de passer d'une classe allélique dans une autre. Si l'on appelle  $u$  le taux de mutation de  $A$  vers  $a$ , et  $v$  le taux de mutation réverse, alors

$$p'' = (1 - u)p + vq$$

et le changement de fréquence allélique dû aux mutations vaut

$$\Delta p_{mut} = p'' - p' = -up + v(1 - p) \quad (2)$$

En l'absence d'autres pressions évolutives, il existe un état d'équilibre stable où la fréquence de l'allèle  $A$  vaut  $p^* = \frac{v}{u+v}$ .

### 2.3.1 Cryptopolymorphisme

Il existe un cas intéressant d'équilibre mutation-sélection pour des allèles défavorables récessifs ( $h=0$ ). Toutes les mutations conduisant à un gène non fonctionnel ont de bonnes chances d'être défavorables, et de réduire la valeur sélective des individus qui les portent. D'autre part, il existe de nombreuses cibles de mutations pour supprimer la fonctionnalité d'un gène (par exemple, création d'un codon stop dans la séquence d'ADN). On peut donc supposer que dans ce cas,  $v \approx 0$ . Un équilibre mutation-sélection sera atteint si  $\Delta p_{sel} + \Delta p_{mut} = 0$ , ce qui correspond à

$$q^2 = \frac{u}{s(1+u)} \approx \frac{u}{s}$$

Ainsi, des mutations défavorables récessives peuvent atteindre des fréquences relativement élevées dans une population, car elles sont "cachées" à l'état hétérozygote. On appelle ce phénomène le **fardeau de mutation**, que l'on retrouve chez toutes les espèces dont le régime de reproduction est la panmixie. Les conséquences sont la dépression de consanguinité et l'hétérosis.

## 2.4 Dérive génétique

La dérive génétique est la fluctuation aléatoire des fréquences des allèles dans une population de taille finie. Elle est due au fait qu'en l'absence d'autre pression évolutive, les  $N$  zygotes de la génération suivante sont le résultat de l'échantillonnage au hasard de  $2N$  gamètes dans le pool gamétique produit par les individus de la génération précédente.

Pour comprendre le mécanisme de la dérive génétique, on introduit une nouvelle variable aléatoire  $X_{g+1}$  qui est le nombre d'allèles  $A$  dans le pool gamétique fécondant de la génération  $g+1$ . Dans une population de taille finie  $N$ , la fréquence  $p''$  des allèles  $A$  s'écrit alors  $p'' = \frac{X_{g+1}}{2N}$ . En l'absence d'autre pression évolutive,  $p' = p$  et  $X_{g+1}$  est le résultat d'un tirage dans une loi Binomiale  $\mathcal{B}(N, p_g)$ .

### 2.4.1 Composition génétique de la population à la génération $g + 1$

On peut calculer l'espérance et la variance de  $X_{g+1}$  :

$$E(X_{g+1}) = \sum_{i=1}^{2N} iP(X_{g+1} = i) = 2Np_g$$

et

$$Var(X_{g+1}) = 2Np_g(1 - p_g)$$

et donc  $E(p_{g+1}) = E(p'') = p_g$  et  $Var(p_{g+1}) = p_g(1 - p_g)/2N$ . On peut en conclure les choses suivantes :

- $p_{g+1}$  est une variable aléatoire. Les fréquences des allèles changent d'une génération à l'autre sous l'effet de la dérive génétique avec des probabilités qui sont celles de la loi binomiale.
- La variance de la fréquence de l'allèle  $A$  est d'autant plus faible que la taille des populations est grande. L'amplitude des fluctuations de fréquences alléliques dépend de la taille des populations.
- La dérive génétique est un processus de Markov. En l'absence d'autres pressions évolutives, si l'on considère les valeurs successives  $\{X_0, X_1, \dots, X_g\}$  du nombre de copies de l'allèle  $A$ , on voit que

$$P(X_{g+1} = i/X_0, X_1, \dots, X_g) = P(X_{g+1} = i/X_g)$$

Ainsi, à un locus, la composition génétique d'une population à une génération  $g + 1$  ne dépend que de sa composition génétique à la génération précédente.

### 2.4.2 Dérive génétique et perte du polymorphisme

Pour comprendre comment la dérive modifie la composition génétique des populations, on peut considérer pour simplifier une population haploïde de  $2N$  gamètes, de taille constante. Les  $2N$  gamètes de la génération suivante sont constituées par un tirage avec remise parmi les gamètes de la génération précédente. C'est le modèle de **Whright-Fisher**. On s'intéresse au nombre de copies  $X_g$  de l'allèle  $A$  à une génération  $g$ , et on considère un état initial  $X_0$ . Soit  $z_{i,g} = P(X_g = i)$ . On peut décrire la chaîne de Markov par son état initial Soit  $P(X_0 = i)$  et sa matrice de transition  $T$  :

$$T = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ t_{2,1} & \dots & & & t_{2,2N} \\ & \dots & & & \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} \quad (3)$$

avec  $t_{i,j} = P(X_{g+1} = i/X_g = j)$ , qui correspond aux probabilités de la loi binomiale  $\mathcal{B}(2N, \frac{j}{2N})$ .

Cette chaîne de Markov a la propriété suivante :

$$P(X_0 = i_0, X_1 = i_1, \dots, X_g = i_g) = P(X_0 = i_0) \cdot t_{i_0, i_1} \cdot t_{i_1, i_2} \cdot \dots \cdot t_{i_{g-1}, i_g}$$

On voit aisément d'après (3) que les états intermédiaires polymorphes ( $X_g \neq 0, X_g \neq 2N$ ) communiquent avec l'ensemble des états possibles, et qu'il existe deux états absorbants

$X_g = 0$  et  $X_g = 2N$  qui correspondent à l'absence de polymorphisme génétique. Ainsi, quelle que soit la condition initiale, les seuls états stationnaires du système sont les états absorbants. La dérive génétique conduit forcément à la **disparition du polymorphisme génétique** dans une population.

Pour calculer la probabilité de fixation d'un allèle initialement présent en fréquence  $p_0$ , on peut utiliser une martingale pour montrer que  $E(X_{g+1}) = E(X_g)$ . En effet,

$$\begin{aligned}
 E(X_{g+1}) &= \sum_{i=1}^{2N} iP(X_{g+1} = i) \\
 &= \sum_i \sum_j iP(X_{g+1} = i / X_g = j)P(X_g = j) \\
 &= \sum_j P(X_g = j) \sum_i iP(X_{g+1} = i / X_g = j) \\
 &= \sum_j P(X_g = j)j \\
 &= E(X_g)
 \end{aligned} \tag{4}$$

Donc,  $E(X_\infty) = 0 \cdot z_{0,\infty} + 2N \cdot z_{2N,\infty} = X_0$ , d'où

$$P_{fix}(A) = P(X_\infty = 2N) = p_0 \tag{5}$$

Donc, la probabilité de fixation de l'allèle  $A$ , présent initialement à une fréquence  $p_0$ , est égale à sa fréquence.

## 2.5 Migration

La migration permet d'introduire de nouveaux allèles dans une population. Elle a un effet analogue à celui de la mutation. Pour traiter l'effet de la migration sur les fréquences alléliques, il faut traiter séparément les cas où les migrants arrivent sous forme de sporophytes (ce sont les individus qui migrent) ou sous forme de gamètes (ce sont les gamètes, par exemple le pollen chez les plantes, qui migrent).