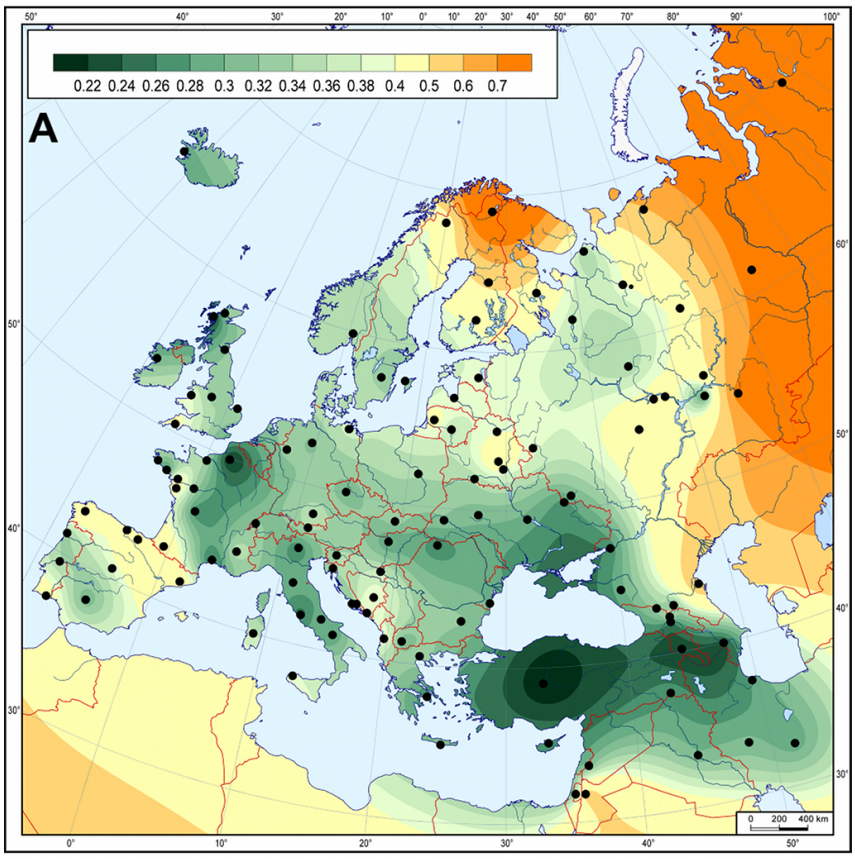
Génétique des Populations



Master GEB

Illustration de couverture : Haak et coll, *Ancient DNA from European Early Neolithic Farmers Reveals Their Near Eastern Affinities*, PLOS Biology (2010)

1 Équilibre de Hardy-Weinberg : modèle à un locus génétique

1.1 Population

Une population est un groupe d'individus qui vivent dans une aire géographique assez restreinte pour que chacun des membres de cette population ait la possibilité de se reproduire avec un autre membre de sexe opposé ; les échanges reproductifs avec d'autres populations d'individus de la même espèce sont supposés rares ou inexistants.

Cette définition est parfois bien adaptée aux espèces végétales ou animales : les animaux domestiques d'un même élevage forment une population ; un groupe d'hirondelles qui se reproduit d'année en année dans la même vallée également. Il peut être nécessaire de diviser une population en sous-populations (ou « dèmes »), en petites unités reproductives qui ne sont pas isolées les unes des autres. Les animaux élevés par les hommes sont parfois divisés en groupes totalement isolés (du point de vue de la reproduction) les uns des autres.

Quand on en vient à la génétique humaine, la définition est plus problématique, les groupes humains habituellement désignés sous le nom de « population » n'étant généralement ni assez homogènes ni assez isolés pour que le concept soit parfaitement adapté. Il existe cependant de petites populations humaines relativement isolées, pour des raisons géographiques (îles, hauts plateaux ou vallées isolées) ou sociales (juifs Ashkénazes, Huttérites, Amish).

1.2 Modèle de l'urne gamétique

On considère une population « idéale » d'individus diploïdes avec une reproduction sexuée (émission de gamètes qui fusionnent pour produire un nouvel individu). Les générations sont séparées et non-chevauchantes, c'est-à-dire que les individus arrivés à maturité sexuelle se reproduisent avec les individus de leur propre génération exclusivement.

On considère le cas d'un locus autosomal di-allélique A/a.

On suppose que les gamètes produits représentent fidèlement la composition génétique de la population, et qu'à la reproduction, les gamètes sont émis dans l'environnement et qu'ils s'apparient au hasard : c'est le modèle de l'urne gamétique, qui est réalisé pour des espèces végétales (émission du pollen dispersé au hasard dans l'environnement), certains animaux (coquillages).

Dans ce cas, si p est la proportion des gamètes portant l'allèle A et $q = 1 - p$ est celle des gamètes portant l'allèle a, la fusion de deux gamètes produit

- un individu AA avec probabilité $f_{\mathrm{AA}} = p^2$
- un individu Aa avec probabilité $f_{\mathrm{A}a} = 2pq$
- un individu aa avec probabilité $f_{aa} = q^2$

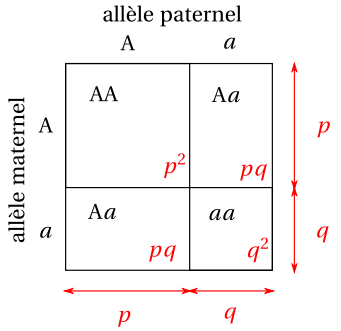


FIGURE 1.1 – Les proportions de Hardy-Weinberg illustrées par un pseudo carré de Punnett

À la génération suivante, les individus AA sont donc en proportion p^2 (environ), les Aa (resp. aa) en proportion $2pq$ (resp. q^2).

Dans les gamètes émis par cette nouvelle génération, les gamètes A sont à nouveau présents en proportion $f_{AA} + \frac{1}{2} f_{Aa} = p^2 + pq = p$, et les gamètes a en proportion q : la composition de l'urne gamétique ne change pas, et les proportions génotypiques données ci-dessus sont donc constantes de génération en génération : c'est l'*équilibre de Hardy-Weinberg*.

Notons qu'on a bien $f_{AA} + f_{Aa} + f_{aa} = p^2 + 2pq + q^2 = (p+q)^2 = 1$.

Revenons au cas d'une population ayant une reproduction avec accouplement. Cette situation est équivalente au modèle de l'urne gamétique, à deux conditions : le choix du partenaire se fait au hasard (on parle de « panmixie » ; la population est dite « panmictique ») et que les gamètes émis par ces partenaires s'apparient ensuite au hasard (pangamie).

1.2.1 Les conditions de validité

Nous avons utilisé plusieurs hypothèses de façon implicite.

Tout d'abord, pour que les génotypes des individus présents à une génération soient en proportions parfaitement égales aux probabilités f_{AA}, f_{Aa} et f_{aa} (calcul encadré ci-dessus), il faut supposer la condition suivante :

— La population est infinie (!)

Pour que la composition de la population ne change pas entre la fécondation et la reproduction :

— Pas de sélection (pas d'individus qui meurent avant la reproduction)
— Pas de migration

Pour que la composition de l'urne gamétique reflète fidèlement la composition allélique des individus :

— Pas de sélection (tous les individus contribuent de la même façon)
— Pas de mutation
— Pas de distorsion de ségrégation méiotique (quand deux allèles sont présents chez un individu, il émet moitié de gamètes avec un allèle, moitié avec l'autre allèle)

Pour que les nouveaux individus soient formés par tirage au hasard des gamètes dans l'urne :

— Panmixie : formation des couples au hasard

— Pangamie : lors de la fécondation, les gamètes s'unissent au hasard ; pas de sélection gamétique.

Il est possible que cette liste ne soit pas complète. Bien sûr, ces conditions ne sont jamais toutes vérifiées. L'équilibre de Hardy-Weinberg concerne une population idéale, une population réelle n'est jamais tout à fait à l'équilibre. Cependant, dans beaucoup de situations et à des échelles de temps suffisament faibles (quelques générations), cela reste une description satisfaisante de la réalité.

Même quand on étudie des gènes soumis à une forte sélection, comme dans le cas d'une maladie récessive létale, si on considère la population avant l'âge où la sélection s'opère, pourvu que la population parentale soit panmictique on observera les proportions de Hardy-Weinberg (voir également section 1.5).

Remarque : écarts à la panmixie

Les écarts à la panmixie sont de deux types : homogamie (en anglais : *assortative mating*) et hétérogamie (en anglais : *dissortative mating*). Il y a homogamie quand les partenaires se choisissent pour leur ressemblance phénotypique ou génétique (c'est peut-être le cas pour la stature, au niveau phénotypique) ; il y a hétérogamie quand les partenaires se choisissent pour leur dissemblance (c'est peut-être le cas pour HLA...).

1.3 Généralisation à un locus multi-allélique

Dans le cas d'un locus multi-allélique avec allèles $\mathrm{A}_1, \ldots, \mathrm{A}_n$ de fréquences respectives $p_1, \ldots, p_n$, le même raisonnement dans le modèle de l'urne gamétique montre que les fréquences génotypiques sont

$$\boxed{\begin{aligned} f_{\mathrm{A}_i\mathrm{A}_j} &= 2p_i p_j \text{ (avec } i \neq j) \\ f_{\mathrm{A}_i\mathrm{A}_i} &= p_i^2 \end{aligned}}$$

La somme de ces fréquences vaut bien 1 ; comme dans le cas di-allélique, les valeurs de fréquences génotypiques correspondent au développement du carré $(p_1 + \cdots + p_n)^2$.

1.4 Cas des chromosomes sexuels

Les chromosomes sexuels, appelés également hétérosomes ou gonosomes sont, chez l'homme et chez la plupart des mammifères, les chromosomes X et Y, les hommes étant XY (on dit que les mâles sont hétérogamétiques) et les femmes XX (on dit que les femelles sont homogamétiques).

Pour les locus situés sur le chromosome Y, la fréquence des allèles portés par les spermatozoïdes Y dans l'urne gamétique est évidemment égale à leur fréquence chez les mâles, qui reste inchangée au fil des générations.

Reste le cas du chromosome X. On se restreint au cas d'un locus di-allélique A/a. Notons $p_{m,t}$ et $p_{f,t}$ les fréquences de l'allèle A chez les mâles et les femelles à la génération t. On note $q_{m,t} = 1 - p_{m,t}$ et $q_{f,t} = 1 - p_{f,t}$.

Un mâle de la génération $t+1$ n'a qu'un X, reçu de sa mère, donc

$$p_{m,t+1} = p_{f,t} \tag{1.1}$$

la fréquence allélique chez les mâles d'une génération est la fréquence allélique chez les femelles de la génération précédente. Les fréquences des deux génotypiques A et a sont égales aux fréquences alléliques.

Une femelle de la génération $t+1$ reçoit un X de chacun de ses parents, donc

$$p_{f,t+1} = \frac{1}{2}(p_{m,t} + p_{f,t}), \tag{1.2}$$

c'est-à-dire que la fréquence allélique chez les femelles d'une génération est la moyenne des fréquences alléliques chez les mâles et les femelles de la génération précédente. Pour ce qui est des fréquences génotypiques, chez les femelles de la génération $t+1$, le génotype AA a pour fréquence $p_{m_t} p_{f,t}$, le génotype Aa a pour fréquence $p_{m_t} q_{f,t} + q_{m_t} p_{f,t}$ et le génotype aa $q_{m_t} q_{f,t}$.

Si la fréquence allélique est la même chez les mâles et les femelles

Si on note $p_{m,0} = p_{f,0} = p$, alors à toutes les générations suivantes, $p_{m,t} = p_{f,t} = p$, et chez les femelles les trois génotypes sont dans les proportions d'Hardy-Weinberg, $p^2, 2pq, q^2$.

Cas général : évolution vers l'équilibre

Supposons qu'au temps $t = 0$, on n'a pas $p_{m,0} = p_{f,0}$.

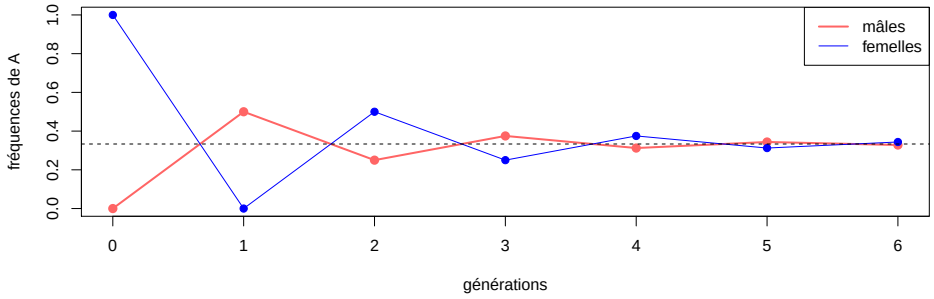


FIGURE 1.2 – Évolution des fréquences alléliques pour un locus situé sur le chromosome X, dans le cas extrême où $p_m = 0$ et $p_f = 1$ à la génération 0.

S'il y a un équilibre avec fréquence p_m chez les mâles et p_f chez les femelles, l'équation 1.1 donne

$$p_m = p_f$$

et l'équation 1.2 donne

$$p_f = \frac{1}{2}(p_m + p_f).$$

On en déduit qu'à l'équilibre la fréquence est la même chez les mâles et les femelles.

La différence des fréquences alléliques chez les deux sexes évolue selon l'équation

$$p_{f,t+1} - p_{m,t+1} = \frac{1}{2}(p_{f,t} - p_{m,t}),$$

obtenue en retranchant l'équation 1.1 de l'équation 1.2. Cette différence est divisée par deux à chaque génération, elle s'amenuise très vite.

D'autre part, toujours à partir des équations 1.1 et 1.2 on obtient

$$2p_{f,t+1} + p_{m,t+1} = 2p_{f,t} + p_{m,t}$$

et donc, pour tout t on a $2p_{f,t} + p_{m,t} = 2p_{f,0} + p_{m,0}$.

On en déduit que la limite commune de $p_{f,t}$ et $p_{m,t}$ est $p = \frac{2}{3}p_{f,0} + \frac{1}{3}p_{m,0}$.

Ceci reflète le fait que, dans l'urne gamétique émise par la génération 0, les deux tiers des gamètes qui portent un X sont émis par les femelles, et un tiers par les mâles.

1.5 Applications

1.5.1 Dominance et codominance

Il y a des exemples de gènes di-alléliques dont les allèles sont codominants : ainsi, le système de groupes sanguins MN, déterminé par un gène d'allèles m et n [1] selon la table 1.1.

Génotype	mm	mn	nn
Groupe	M	MN	N

TABLE 1.1 – Déterminisme des groupes sanguins MN

Un autre exemple est donné par la couleur de certaines fleurs, déterminée par un gène di-allélique A/a, où les fleurs AA sont blanches, les fleurs Aa sont roses, et les aa sont rouges.

Pour d'autres gènes, un allèle est dominant sur l'autre : c'est le cas de la drépanocytose, ou anémie falciforme (en anglais : *drepanocytosis* ou *sickle cell disease*), une maladie du sang due à une mutation du gène de la β-globine, un des constituants de l'hémoglobine [2]. On considèrera deux allèles de la β-globine, β^S et β^A. La drépanocytose est une maladie récessive : les individus atteints sont tous de génotype $\beta^S\beta^S$. Chez les individus atteints, la β-globine crée de longues chaînes de polymère, donnant aux globules rouges une forme caractéristique de faucille ; les symptômes principaux sont une anémie chronique et des crises vaso-occlusives (obstruction des vaisseaux sanguins capillaires par les globules rouges anormaux). Les hétérozygotes $\beta^S\beta^A$ ne présentent pas les symptômes de la maladie, la majorité de leurs globules rouges étant normaux.

Cependant l'observation du sang des hétéorygotes au microscope met en évidence la présence de quelques globules rouges en faucille. On parle dans ce cas de phénotype SA (en anglais, *sickle cell trait* ou *sicklemia*). Ainsi, selon la manière dont on décide d'observer le phénotype, on a dominance de β^A sur β^S, ou codominance des deux allèles.

1.5.2 Estimation des fréquences alléliques

Reprenons l'exemple des groupes sanguins MN. Si on observe le groupe sanguin dans un échantillon de 100 personnes, comme dans la table 1.2, on peut directement estimer les fréquences alléliques par comptage des allèles m et n. En effet, chacun des 21 individus de groupe M porte deux allèles m, et chacun des individus

Groupe	M	MN	N
Effectifs	21	60	19

TABLE 1.2 – Effectifs pour le groupe MN

1. La notation standard pour ces allèles est L^M et L^N.
2. L'hémoglobine est un « tétramère », c'est-à-dire qu'elle est formée par l'union de 4 molécules, deux α-globines et deux β-globines.

MN en porte un ; cela fait en tout $2 \times 21 + 60 = 102$ allèles m, sur un total de 200 allèles observés (deux par individu), donc une fréquence allélique $f_m = \frac{102}{200} = 0.51$.

Cependant, dans le cas dominant ou récessif, on ne peut plus procéder ainsi. Par exemple, dans la table 1.3 on dénombre les cas de mucoviscidose dans un effectif de 25000 personnes. La mucoviscidose est une

Phénotype	Sain	Atteint
Effectifs	24989	11

TABLE 1.3 – Effectifs pour la mucoviscidose

maladie récessive, due à une mutation du gène CFTR. Si on note A l'allèle normal et a l'allèle mutant, tous les individus atteints sont de génotype aa ; les individus sains sont AA ou Aa. Faute de pouvoir séparer les hétérozygotes des homozygotes AA, on ne peut pas compter les allèles.

On peut obtenir une estimation des fréquences alléliques en supposant que les fréquences génotypiques sont dans les proportions de Hardy-Weinberg. Cette hypothèse paraît audacieuse, car le trait est soumis à sélection ; cela se traduira pas un changement des fréquences alléliques au fil des générations. Cependant, il suffit que la population soit panmictique pour qu'à la naissance (avant que la sélection n'opère) les proportions de Hardy-Weinberg soient respectées. On peut donc estimer les fréquences p et q de A et a en supposant que les individus sains sont en proportion $p^2 + 2pq$ et les atteints en proportion q^2. On a donc $q^2 = 11/25000$, d'où on tire $q \simeq 0.021$.

Les tables 1.4 et 1.5 donnent des exemples de calculs de fréquences des allèles morbides (allèles causant la maladie) pour quelques maladies dominantes et récessives.

Maladie	Fréquence f	Fréquence de l'allèle morbide $q = \sqrt{f}$
Mucoviscidose	1/2500	0,02
Phénylcétonurie	1/16000	0,008
Drépanocytose	1/25 (Afrique)	0,2

TABLE 1.4 – Maladies récessives. Les individus atteints sont aa, les individus sains sont AA ou aa. La fréquence de AA est q^2, donc $q = \sqrt{f}$

Maladie	Fréquence f	Fréquence de l'allèle morbide $q = 1 - \sqrt{1-f}$
Huntington	1/10000	$5 \cdot 10^{-5}$
Achondroplasie	1/20000	$2{,}5 \cdot 10^{-5}$
Rétinoblastome	1/30000	$1{,}7 \cdot 10^{-5}$

TABLE 1.5 – Maladies dominantes. Les individus atteints sont Aa ou aa, les individus sains sont AA. La fréquence de AA est p^2, donc $p = \sqrt{1-f}$ et $p = 1 - q = 1 - \sqrt{1-f}$.

1.5.3 Contrôle qualité en épidémiologie génétique

L'équilibre de Hardy-Weinberg est utilisé à des fins de « contrôle qualité » des données génétiques : quand le contexte le permet, on estime que si les génotypes observés en un locus donné sont répartis selon des proportions trop différentes de celles qu'on doit observer à l'équilibre de Hardy-Weinberg, il est probable que la cause en est un problème technique lors du génotypage.

Le test utilisé peut être un simple χ^2 ; un « test exact » a également été proposé.

1.6 Exercices

Exercice 1 On considère un locus autosomal di-allélique A/a. À la génération 0, la fréquence de l'allèle A chez les hommes est $p_h = 0{,}1$, et chez les femmes, $p_f = 0{,}8$. Sous les hypothèses du modèle de Hardy-Weinberg, quelles sont les proportions génotypiques à la génération 1 ? à la génération 2 ? aux générations suivantes ?

Exercice 2 On note X_1 et X_2 deux variables de Bernoulli indépendantes de paramètre q. On pose $X = X_1 + X_2$. Quelle est la loi de X, son espérance, sa variance ?

Montrer que $\mathbb{P}(X = 0) = p^2$, $\mathbb{P}(X = 1) = 2pq$ et $\mathbb{P}(X = 2) = q^2$, où on a noté $p = 1-q$. Quel rapport avec l'équilibre de Hardy-Weinberg ?

Exercice 3 On considère le groupe sanguin ABO, déterminé par un gène à trois allèles notés a, b, o ; a et b sont co-dominants, et o est récessif.

On suppose que leurs fréquences sont $f_a = 0{,}4$, $f_b = 0{,}3$, $f_o = 0{,}3$. Quelles sont les proportions des groupes sanguins A, B, AB et O dans la population ?

Exercice 4 On a génotypé 100 individus en un locus di-allélique A/a. Les effectifs génotypiques sont résumés dans la table suivante.

	AA	Aa	aa
n	7	32	61

Estimer la fréquence des deux allèles A et a. Quelles sont les proportions génotypiques attendues sous l'équilibre de Hardy-Weinberg ? Est-on à l'équilibre ? (faire un test du χ^2)

Exercice 5 : un modèle simple en épidémiologie génétique On considère un locus di-allélique A/a, qu'on impliqué dans une maladie humaine, selon le modèle des risques multiplicatifs :

$$\mathbb{P}(\text{Att}|\text{AA}) = \varphi_0 \qquad \mathbb{P}(\text{Att}|\text{A}a) = r\varphi_0 \qquad \mathbb{P}(\text{Att}|aa) = r^2\varphi_0$$

où Att est l'événement « être atteint ».

1. On note p la fréquence de l'allèle A et q celle de l'allèle a. Montrer que dans ce modèle, la probabilité d'être atteint vaut

$$\mathbb{P}(\text{Att}) = \varphi_0(p + qr)^2.$$

2. En utilisant la formule de Bayes, calculez la probabilité qu'un individu atteint ait pour génotype AA, Aa et aa. Est-ce que les proportions d'Hardy-Weinberg sont vérifiées chez les individus atteints ?

2 Distance génétique

2.1 Les lois de Mendel à l'épreuve

Au début du XX^{e} siècle, en soumettant des drosophiles à des agents mutagènes, Thomas Hunt Morgan et ses collaborateurs ont obtenu des caractères récessifs « anormaux », gouvernés par des gènes di-alléliques, d'allèles A/a : le caractère n'apparaît que chez les individus aa. On appelle a l'allèle muté, et A l'allèle sauvage (on rencontre parfois ce vocabulaire en génétique humaine).

2.1.1 Backcross

Morgan a réalisé l'expérience suivante (appelée rétrocroisement « backcross ») : on croise une souche sauvage AA avec un mutant aa. On obtient un individu hybride de génotype Aa au locus considéré ; son phénotype normal (ou « phénotype sauvage »), ce qui correspond au faut que le caractère A est a est dominant (on parlera aujourd'hui également d'allèle dominant).

On croise cet individu avec son parent mutant (d'où le nom de « backcross »). Quels seront les phénotypes des descendants ?

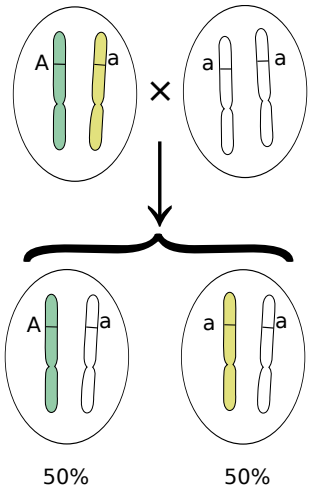


FIGURE 2.1 – Après un simple backcross

La réponse est simple (et correspond à la deuxième loi de Mendel) : les allèles A et a de chacun des deux parents de l'hybride se retrouve en proportions égales dans les gamètes qu'il émet ; le mutant n'émet que des gamètes portant l'allèle a ; on obtient donc 50% d'invididus de génotype Aa (phénotype sauvage) et 50% d'individus de génotype aa (phénotype mutant, ou ici, récessif, car il *revient*). Ceci est illustré figure 2.1. On parle de « ségrégation » des allèles ancestraux (du latin *segregatio*, séparation) : l'individu hybride ne transmet pas un mélange des deux caractères, mais l'un ou l'autre.

2.1.2 Double backcross

Considérons maintenant deux caractères récessifs, correspondant à des locus génétiques distincts, d'allèles A/a et B/b.

Le « double backcross » consiste à croiser une souche sauvage AA, BB avec un mutant aa, bb. On obtient un individu dihybride de phénotype normal (sauvage), de génotype Aa au premier locus et Bb au second. On croise cet individu avec son parent mutant. On pose la même question que précédemment : quels seront les phénotypes des descendants ?

La troisième loi de Mendel postule la « ségrégation indépendante » des caractères (ou des allèles) ancestraux : l'individu dihybride doit émettre 4 types de gamètes en quantité égales, des gamètes AB, Ab, ab, et aB. Le mutant n'émettant que des gamètes ab, on devrait observer chez les descendants les quatre combinaisons de phénotypes possibles, dans des proportions égales (25% de chaque).

Cependant une « théorie chromosomique naïve » (figure 2.2), postulant (à juste titre) que les chromosomes sont le support physique de l'hérédité, laisse envisager la co-transmission (ou co-ségrégation) par le dihybride des caractères reçus de chacun des deux parents : auquel cas les gamètes émis seront pour moitié AB (on transmets les deux allèles sauvages ensemble), et pour moitié ab (ou les deux allèles mutants ensemble). On ne devrait observer chez les descendants que deux combinaisons de phénotypes (soit les deux caractères mutants, soit les deux caractères sauvages), dans des proportions égales.

L'observation du phénotype des descendants doit permettre de trancher : le phénotype de chacun d'entre eux permet de connaître le type du gamète qu'il a hérité de l'individu di-hybride. On observe dans certains cas une situation intermédiaire (figure 2.3) : majoritairement, on a des gamètes AB et ab, donc co-transmission des caractères reçus de chacun des deux parents ; cependant, certains descendants ont reçu des gamètes Ab ou aB.

Ce phénomène est appelé *recombinaison*. On parle également d'individus ou de gamètes recombinants. On note θ la proportion de gamètes recombinant, ou *taux de recombinaison* (10% sur la figure 2.3). La figure 2.4 illustre les proportions de types gamétiques attendus pour un taux de recombinaison θ donné.

Quand le taux de recombinaison entre deux locus est inférieur à $\frac{1}{2}$, on dit que les locus sont *liés*.

Ces observations ont contribué à faire admettre la théorie chromosomique, selon laquelle les chromosomes sont les supports matériels de l'hérédité, tout en mettant en évidence le phénomène de la recombinaison, qui s'explique par l'existence d'enjambements ou *cross-overs* lors de la méïose.

2.2 Distance génétique

2.2.1 Définition de la distance génétique

Plus la distance est grande entre deux locus situés sur un même chromosome, plus il est probable qu'un enjambement ait lieu entre eux lors d'une méïose. Ceci conduit à définir une unité de distance génétique, le Morgan.

> Une distance d'un Morgan entre deux locus correspond à une moyenne d'un enjambement par méïose entre ces deux locus.

On utilisera le plus souvent le centiMorgan (abrégé cM), qui correspond à un enjambement toutes les 100 méïoses, en moyenne.

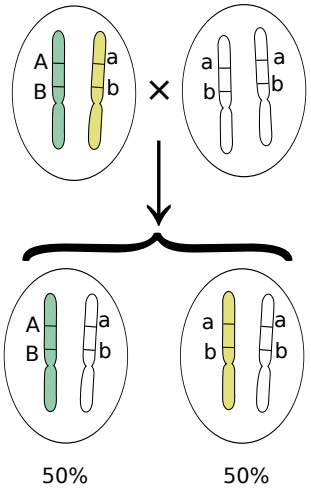


FIGURE 2.2 – Théorie chromosomique naïve : coségrégation des caractères ancestraux

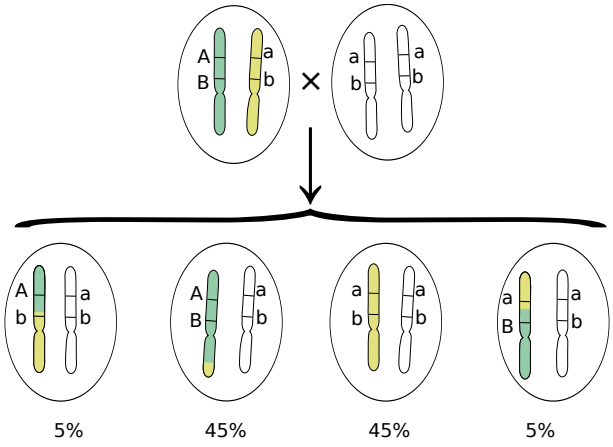


FIGURE 2.3 – Recombinaison partielle des caractès ancestraux

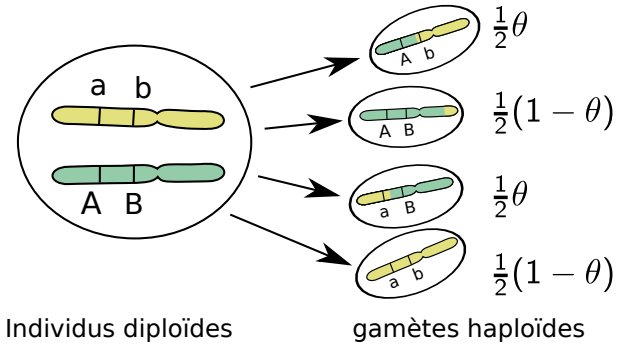


FIGURE 2.4 – Gamètes recombinants

Notons qu'en supposant que les enjambements sont indépendants les uns des autres, cette définition conduit à une distance additive : pour trois locus polymorphes A/a, B/b et C/c se trouvant dans cet ordre sur le génome, le nombre moyen d'enjambements entre a et c est égal à la somme du nombre moyen d'enjambements entre a et b, et du nombre moyen d'enjambements entre b et c :

$$d(a,c) = d(a,b) + d(b,c).$$

2.2.2 Taux de recombinaison

En pratique, on ne peut pas mesurer directement la distance génétique entre deux locus, parce qu'on ne peut pas compter les enjambements. La seule chose qu'on peut estimer c'est le taux de recombinaison, qui est est la probabilité qu'il y ait eu *un nombre impair* d'enjambements. En effet, si il y a eu un double enjambement entre les locus, on n'observe pas de recombinaison !

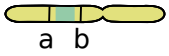


FIGURE 2.5 – Double enjambement : pas de recombinaison !

On note θ_{ab} ou simplement θ le taux de recombinaison entre deux locus A/a et B/b. C'est une mesure de la distance génétique entre locus ; plus θ est élevé, plus les locus considérés sont éloignés.

Si $\theta = 0$, les deux locus sont si proches qu'aucune recombinaison n'a lieu entre eux (ils ne sont pas pour autant physiquement confondus). À l'autre extrême, il y a le cas où les deux locus sont sur des chromosomes différents. Cette fois, les caractères ségrègent indépendamment. On n'a pas, au niveau moléculaire, d'enjambement, mais du point de vue de la transmission du matériel génétique reçu du parent on a une probabilité de recombinaison $\theta = \frac{1}{2}$.

C'est la plus grande probabilité possible. Une valeur plus élevée signifierait que la recombinaison est plus probable que la co-transmission, ce qu'on n'observe pas en pratique.

On peut tenter de donner une interprétation intuitive de ce fait : quand la distance entre les deux locus grandit, la probabilité qu'il y ait un enjambement grandit, mais la probabilité qu'il y en ait deux (et qu'on n'ait donc pas de recombinaison des allèles) grandit également. C'est ce qui empêche le taux de recombinaison de dépasser $\frac{1}{2}$. Nous allons voir ci-après que cette interprétation est correcte, pourvu qu'on suppose l'indépendance mutuelle des enjambements.

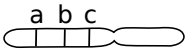
Pour les petites distances, la probabilité de faire plus d'un enjambement est très faible : par exemple, si on

observe une recombinaison entre deux locus toutes les 100 méïoses, on en déduit qu'on a un enjambement toutes les 100 méïoses, et que ces deux locus sont à 1 cM l'un de l'autre ; il est assez intuitif qu'il est si rare qu'il y ait un double enjambement entre ces deux locus qu'on peut négliger cette possibilité. Ceci revient à poser $d(a,b) \simeq \theta_{ab}$.

Dans leurs travaux, Morgan et ses collaborateurs ont utilisé les taux de recombinaison comme unité de distance. Il supposaient en particulier l'additivé : pour trois locus polymorphes A/a, B/b et C/c se trouvant dans cet ordre sur le génome, ils considéraient $\theta_{ac} = \theta_{ab} + \theta_{bc}$.

Cette égalité est à peu près vraie pour des petits taux de recombinaison. Ceci permet la cartographie des chromosomes, l'observation des valeurs de θ_{ac}, θ_{ab} et θ_{bc} permettant de savoir dans quel ordre les trois locus sont disposés sur un chromosome.

Dès que les taux de recombinaison sont importants (autrement dit, dès que les locus sont éloignés), cette propriété d'additivité est en défaut : en effet, si il y a eu recombinaison entre a et b, puis entre b et c, on n'observe pas de recombinaison entre a et c ! (les doubles enjambements « ne comptent pas ») (voir figure 2.6).



$$\theta_{ac} \neq \theta_{ab} + \theta_{bc}$$

FIGURE 2.6 – Non-additivité des taux de recombinaison

Il est donc nécessaire de fournir un effort conceptuel pour relier le taux de recombinaison et la distance génétique.

2.2.3 Distance de Haldane

Le généticien John Burdon Sanderson Haldane [1] a fait l'hypothèse que les enjambements sont indépendants les uns des autres ; ils surviennent au hasard avec une probabilité égale (très faible) en tout point d'un chromosome. Dans ce cas, le nombre N d'enjambements entre deux locus à distance d (exprimée en Morgans) suit une loi de Poisson de paramètre d. On a bien $E(N) = d$.

Le taux de recombinaison est alors la probabilité que N soit impair. En utilisant les propriétés de la loi de Poisson, on montre que

$$\boxed{\begin{aligned}\theta &= \mathbb{P}(\text{N impair}) \\ &= \frac{1}{2}\left(1 - e^{-2d}\right)\end{aligned}}$$

On en déduit le résultat comme annoncé plus haut que le taux de recombinaison est bien toujours plus petit que $\frac{1}{2}$:

$$0 \leq \theta \leq \frac{1}{2}.$$

On peut également reformuler le lien entre la distance et le taux de recombinaison comme ceci :

$$\boxed{d(a,b) = -\frac{1}{2}\log(1 - 2\theta_{ab}).}$$

1. JBS Haldane est un des pionniers de la génétique des populations, avec Fisher et Wright. C'était également un personnage fascinant ; très engagé politiquement, il quitta l'Angleterre en 1956 à l'âge de 64 ans pour s'installer en Inde, à la recherche d'une alternative aux deux premiers mondes. Il y mourut en 1964, ayant acquis la nationalité indienne.

On parle de « distance de Haldane ». Notons qu'un taux de combinaison de $\frac{1}{2}$ (locus non liés) correspond à une distance infinie.

> Pour θ_{ab} petit, on a $d(a,b) \simeq \theta_{ab}$; une distance d'un centiMorgan ($d(a,b) = 0{,}01$) correspond à un enjambement en moyenne toutes les 100 méïoses, et à (environ) un recombinant en moyenne toutes les 100 méïoses ($\theta_{ab} \simeq 0{,}01$).

La distance génétique est en première approximation proportionnelle à la distance physique : pour le génome humain, on utilise généralement l'approximation « un centimorgan ≃ un million de paires de bases », soit

> 1cM ≃ 1Mb.

Chr	cM	Mb	Chr	cM	Mb	Chr	cM	Mb
1	284	247	9	166	140	17	128	79
2	269	243	10	181	135	18	117	76
3	223	199	11	158	134	19	108	64
4	214	191	12	175	132	20	108	62
5	204	181	13	126	114	21	63	47
6	192	171	14	119	106	22	73	50
7	187	159	15	141	100	X	182	155
8	168	146	16	134	89	Y		58

TABLE 2.1 – Longueurs des chromosomes humains

2.2.4 Distances de Kosambi et de Felsenstein

La distance de Haldane suppose les enjambements indépendants, selon un processus de Poisson. Il est cependant possible qu'il existe des « interférences » : juste après un enjambement, il y a une légère diminution de la probabilité d'en observer un nouveau. C'est ce que modèlise la distance de Kosambi, proposée par Damodar Dharmanada Kosambi, d'après des données empiriques :

$$d_{\mathrm{K}}(a,b) = -\frac{1}{4}\log\left(\frac{1-2\theta_{ab}}{1+2\theta_{ab}}\right).$$

Comme pour la distance de Haldane, on a $d_{\mathrm{K}}(a,b) \simeq \theta_{ab}$ pour θ_{ab} petit. Cette distance peut être utilisée en analyse de liaison, c'est-à-dire quand on cherche à localiser des gènes impliqués dans une maladie sur le génome.

Une autre distance a été proposée par Felsenstein :

$$d_{\mathrm{F}}(a,b) = -\frac{1}{2(2-k)}\log\left(\frac{1-2\theta_{ab}}{1+2\theta_{ab}(1-k)}\right),$$

où k est entre 0 et 2. La constante k est une mesure d'interférence, $k > 1$ correspondant à la présence d'interférences positives, et $k < 1$ à la présence d'interférences négatives. Si $k = 0$, on retrouve la distance de Kosambi, si $k = 1$ la distance de Haldane.

Il faut également signaler que les recombinaisons sont plus fréquentes chez les femmes que chez les hommes : en toute rigueur on devrait utiliser en analyse de liaison des cartes génétiques différentes selon le sexe.

2.3 Implications en épidémiologie génétique

Il est à noter que pour des raisons éthiques, on ne peut pas procéder aux mêmes expériences sur les humains que sur les drosophiles. Cependant, l'observation de caractères mendéliens (dominant/récessifs) et de polymorphismes de sites de restriction a permis d'estimer la distance génétique entre divers locus, et d'établir les premières cartes du génome humain.

L'objet de l'analyse de liaison permet d'utiliser des données familiales (généalogies avec de multiples atteints) pour placer un gène dont les variations sont impliquées dans une maladie locus morbides sur une telle carte.

2.4 Exercices

Exercice 1 On considère trois locus a, b et c, dans cet ordre sur le chromomosome. Le taux de recombinaison entre a et b est $\theta_{ab} = 0{,}1$; et le taux de recombinaison entre b et c est $\theta_{bc} = 0{,}2$.

1. Calculer $d(a,b)$ et $d(b,c)$ dans la carte de Haldane (donner les résultats en Morgan ou en centiMorgan).

2. Quelle est la valeur de θ_{ac} ?

Exercice 2 : additivité de la distance de Haldane

Une façon de valider la distance de Haldane est de montrer qu'elle est bien additive, au contraire des taux de recombinaison. Supposons trois locus di-alléliques A/a, B/b et C/c, dans cet ordre sur un chromosome (figure 2.6).

1. Montrer qu'en supposant que les enjambements entre a et b sont indépendants de ceux qui surviennent entre b et c, on a

$$\theta_{ac} = \theta_{ab}(1-\theta_{bc}) + (1-\theta_{ab})\theta_{bc}.$$

2. En déduire

$$(1-2\theta_{ac}) = (1-2\theta_{ab})(1-2\theta_{bc}).$$

3. En déduire que la distance de Haldane est additive.

3 Équilibre et déséquilibre gamétique

3.1 Définitions

On considère toujours deux locus diallélіques A/a et B/b. Notons f_A (respectivement f_B) la fréquence de l'allèle A (respectivement B) et $f_a = 1 - f_A$ (respectivement $f_b = 1 - f_B$) la fréquence de l'allèle a (respectivement b).

Si on tire au hasard un gamète émis par la population, il peut être de type AB, Ab, aB ou ab ; quelles sont les fréquences f_{AB}, f_{Ab}, f_{aB} ou f_{ab} de ces 4 types d'allèles ?

Si on a

$$\begin{array}{lcllcl} f_{AB} & = & f_A f_B & f_{Ab} & = & f_A f_b \\ f_{aB} & = & f_a f_B & f_{ab} & = & f_a f_b \end{array}$$

alors on dit qu'on a *équilibre gamétique* entre les locus considérés. Dans le cas contraire, on est en présence d'un *déséquilibre gamétique.*

Si on observe sur un gamète, au premier locus, un allèle A, alors le second locus porte l'allèle B avec la probabilité $\frac{f_{AB}}{f_A}$. En l'absence de déséquilibre gamétique, on voit que cette probabilité est égale à f_B, la probabilité a priori qu'un gamète porte l'allèle B ; l'observation du premier locus n'a apporté aucune information sur le second. Si au contraire il y a déséquilibre gamétique, l'observation du premier locus apporte une information sur le second.

On parlera de *déséquilibre maximal* quand seulement trois des quatre gamètes possibles sont présents : par exemple, $f_{ab} = 0$. Dans ce cas, l'observation de l'allèle a au premier locus sera toujours accompagnée de l'observation de l'allèle B au second locus ; inversement, l'observation de l'allèle b au second locus sera toujours accompagnée de l'observation de l'allèle A au premier locus.

On parlera de *déséquilibre complet* quand seulement deux des quatres gamètes possibles sont présents : par exemple, $f_{Ab} = f_{aB} = 0$. Dans ce cas, l'observation de l'allèle présent à un des deux locus suffit à déterminer entièrement l'allèle présent à l'autre locus.

3.2 Mesures usuelles du déséquilibre gamétique

Il y a trois mesures utilisées, toujours notées D, D′ et r^2.

3.2.1 Le déséquilibre D

On pose

$$\boxed{D = f_{AB} - f_A f_B.}$$

On a D = 0 si, et seulement si, on a équilibre gamétique.

Remarque 1 La définition de D semble dépendre du choix qu'on a fait pour les nommer les allèles A/a et B/b. En fait, on a

$$\begin{cases} f_{aB} - f_a f_B &= -D \\ f_{Ab} - f_A f_b &= -D \\ f_{ab} - f_a f_b &= D, \end{cases}$$

donc D est bien défini – au signe près : si on échange A et a, on change simplement le signe de D. On notera que D est également la covariance des variables aléatoires X et Y définies par X = 1 (respectivement X = 0) si le gamète porte l'allèle A (respectivement l'allèle a), et Y = 1 (respectivement Y = 0) si le gamète porte l'allèle B (respectivement l'allèle b).

Une fois la référénce choisie, connaître f_A, f_B et D suffit à retrouver toutes les fréquences gamétiques :

$$\boxed{\begin{aligned} f_{AB} &= f_A f_B + D \\ f_{Ab} &= f_A f_b - D \\ f_{aB} &= f_a f_B - D \\ f_{ab} &= f_a f_b + D \end{aligned}}$$

Remarque 2 En utilisant $f_A = f_{AB} + f_{Ab}$ et $f_B = f_{AB} + f_{aB}$, on obtient également

$$\boxed{D = f_{AB} f_{ab} - f_{Ab} f_{aB}.}$$

3.2.2 Le déséquilibre D′ de Lewontin

À valeurs fixées pour f_A, f_B, les valeurs minimales et maximales que peut prendre D sont

$$\begin{aligned} D_{min} &= \max(-f_A f_B, -f_a f_b), \\ D_{max} &= \min(f_A f_b, f_a f_B). \end{aligned}$$

En effet, puisque $f_{AB} = f_A f_B + D \geqslant 0$ et $f_{ab} = f_a f_b + D \geqslant 0$, on a $D \geqslant -f_A f_B$, et $D \geqslant -f_a f_b$, d'où la valeur de D_{min}. La valeur de D_{max} s'obtient de façon similaire en considérant les fréquences f_{Ab} et f_{aB}.

Lewontin a suggéré de renormaliser D ainsi :

$$\boxed{D' = \begin{cases} \dfrac{D}{D_{max}} & \text{si } D \geq 0, \\ \dfrac{D}{D_{min}} & \text{si } D \leq 0. \end{cases}}$$

L'avantage de D′, c'est qu'il prend des valeurs entre 0 et 1 ; il prend la valeur 1 quand le déséquilibre est maximal au sens défini plus haut (un des quatre gamètes possibles n'est jamais observé).

3.2.3 Le coefficient de corrélation r^2

Enfin, on utilisera très souvent la mesure r^2, qui est un coefficient de corrélation :

$$\boxed{r^2 = \frac{D^2}{f_A f_a f_B f_b}.}$$

Les valeurs prises par r^2 sont également entre 0 et 1. On a $r^2 = 1$ quand le déséquilibre est complet.

Le déséquilibre gamétique entre locus est souvent représenté dans des diagrammes comme celui de la figure 3.1, où chaque carré représente le déséquilibre entre deux locus ; ici la couleur est d'autant plus sombre que la valeur de r^2 est élevée ; la valeur inscrite est 100 fois la valeur de r^2. On lit par exemple que le qu'entre les SNP rs4512434 et rs4740848 on a $r^2 = 0{,}32$, entre rs4512434 et rs2274874, $r^2 = 0{,}05$, etc. Les carrés noirs correspondent à $r^2 = 1$.

3.2.4 Calcul pratique du déséquilibre gamétique

En pratique, on n'échantillonne que rarement des gamètes : on observe la plupart du temps des données génotypiques, qui peuvent être résumés dans une table comme celle-ci :

	BB	Bb	bb	
AA	91	74	7	172
Aa	50	103	31	184
aa	3	15	26	44
	144	192	64	400

Sur un total de 400 individus, on a observé par exemple 172 individus de génotype AA parmi lesquels 91 individus de génotypes AA et BB, etc.

Peut-on se servir de ces données pour estimer les fréquences gamétiques ? Quand on est face à ce problème, on parle également des fréquences haplotypiques, un haplotype étant l'ensemble des allèles portés par un chromosome.

Si un individu est, par exemple, de génotypes AA et Bb, il a été formé d'un gamète AB et d'un gamète Ab ; un de ces chromosomes porte l'haplotype AB et l'autre l'haplotype Ab. On pourra utiliser la notation abrégée : un individu AA, Bb est AB + Ab.

Ainsi, les individus AA, BB sont AB + AB, on compte donc 182 haplotypes AB ; les 74 individus AA, Bb sont AB + Ab, on compte donc 74 haplotypes AB et 74 haplotypes Ab, etc. Si on tente de procéder ainsi à un dénombrement complet des haplotypes, on a un problème avec case centrale de la table, qui contient les effectifs des « doubles hétérozygotes » Aa, Bb : sont-ils AB + ab ou Ab + aB ?

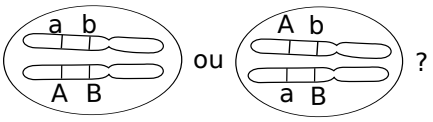


FIGURE 3.2 – Ambiguïté de la phase des doubles hétérozygotes

Cette ambiguïté (illustrée par la figure 3.2) rend impossible l'estimation des fréquences gamétiques et du déséquilibre gamétique par simple comptage : il faut utiliser une estimation par maximum de vraisemblance ; voir détails en encadré.

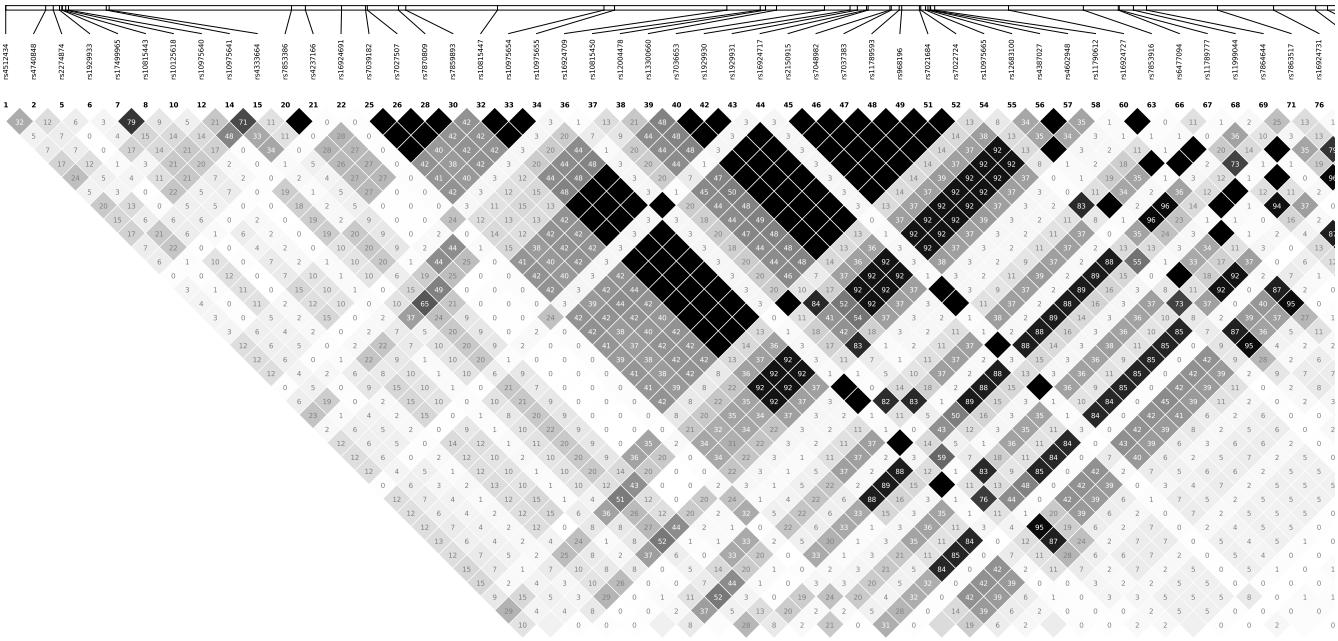
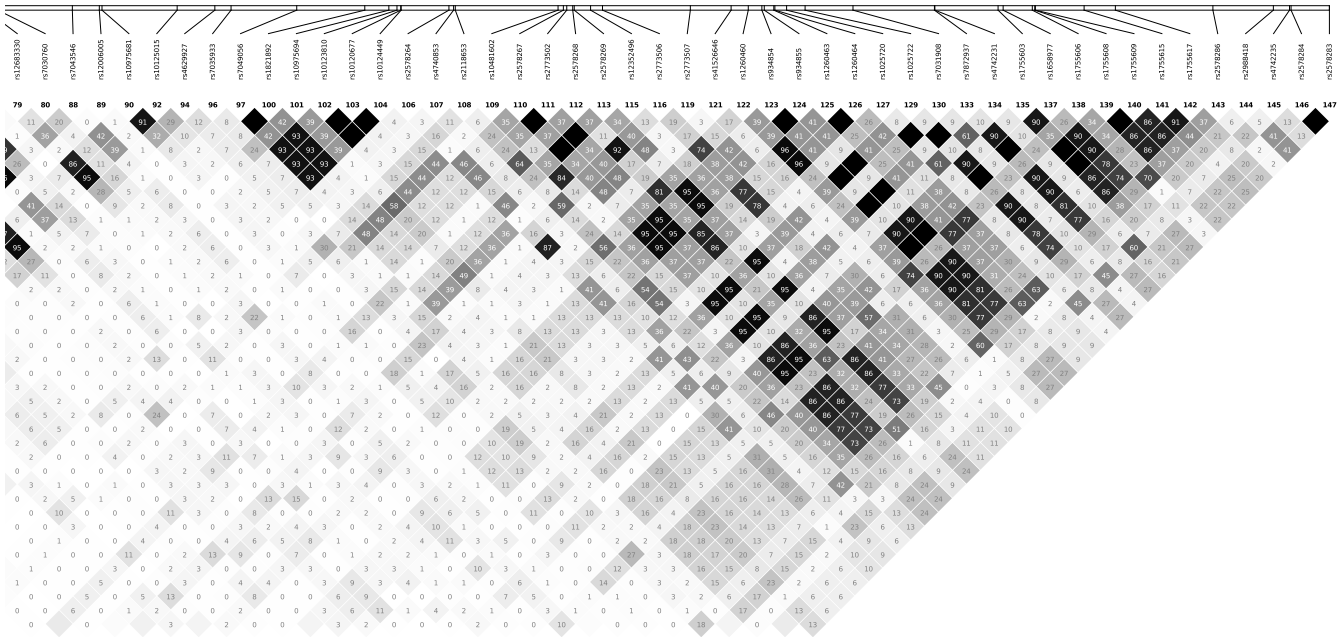


FIGURE 3.1 – Représentation du déséquilibre gamétique pour 96 SNP dans le gène GLDC



(figure réalisée avec Haploview, données HapMap pour la population européenne)

Estimation du déséquilibre par maximum de vraisemblance

Avec des observations comme celles-ci :

	BB	Bb	bb	
AA	91	74	7	172
Aa	50	103	31	184
aa	3	15	26	44
	144	192	64	400

Notons d'abord que f_A, f_B, f_a et f_b peuvent être estimés à partir des effectifs de chacun des génotypes :

$$\begin{aligned} f_A &= \tfrac{1}{800}(2\times 172+184) = 0{,}66 \qquad f_a = 1-f_A = 0{,}34 \\ f_B &= \tfrac{1}{800}(2\times 144+192) = 0{,}6 \qquad f_b = 1-f_B = 0{,}4. \end{aligned}$$

On peut maintenant écrire la fréquence de chacun des haplotypes en fonction du déséquilibre gamétique, qui est un paramètre inconnu que nous noterons ici d :

$$\begin{aligned} f_{AB} &= f_A f_B + d = 0{,}396 + d \qquad f_{Ab} = f_A f_b - d = 0{,}264 - d \\ f_{aB} &= f_a f_B - d = 0{,}204 - d \qquad f_{ab} = f_a f_b + d = 0{,}136 + d \end{aligned}$$

On utilise l'hypothèse de panmixie de la population : les proportions génotypiques sont les proportions de Hardy-Weinberg, et sont donc :

	BB	Bb	bb
AA	f_{AB}^2	$2f_{AB}f_{Ab}$	f_{Ab}^2
Aa	$2f_{AB}f_{aB}$	$2f_{AB}f_{ab}+2f_{aB}f_{Ab}$	$2f_{Ab}f_{ab}$
aa	f_{aB}^2	$2f_{aB}f_{ab}$	f_{ab}^2

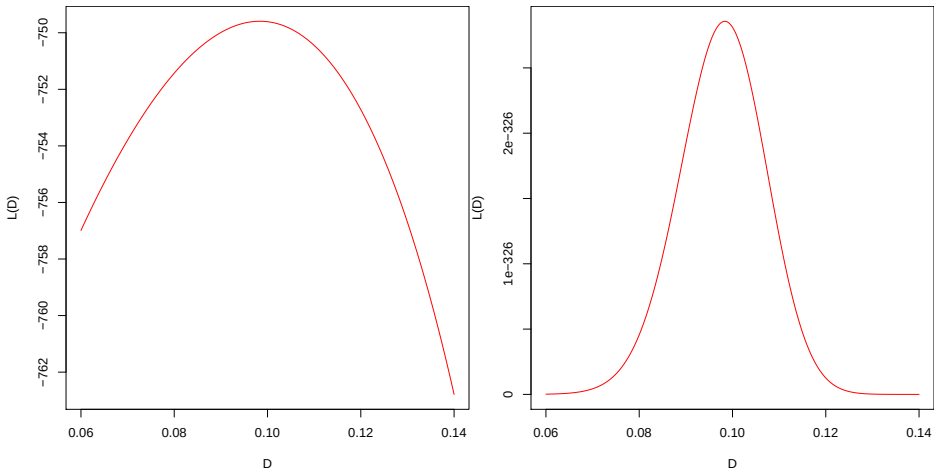
La vraisemblance d'observer un individu de génotypes AA, BB est donc f_{AB}^2, celle d'observer un individu de génoype AA Bb est $2f_{AB}f_{Ab}$, etc. La vraisemblance de la totalité de nos observations est le produit des vraisemblances de chacune d'elles ; et la log-vraisemblance est la somme des log-vraisemblances, qui valent $2\log(f_{AB})$, $\log(2f_{AB}f_{Ab})$, etc.
En reprenant les effectifs qui sont dans notre table d'observations, on écrit que la log-vraisemblance est

$$\begin{aligned} \ell(d) = {} & 91\times 2\times\log\big(f_{AB}\big) + 74\times\log\big(2f_{AB}f_{Ab}\big) + 7\times 2\times\log\big(f_{Ab}\big) \\ & + 50\times\log\big(2f_{AB}f_{aB}\big) + 103\times\log\big(2f_{AB}f_{ab}+2f_{aB}f_{Ab}\big) + 31\times\log\big(2f_{Ab}f_{ab}\big) \\ & + 3\times 2\times\log\big(f_{aB}\big) + 15\times\log\big(2f_{aB}f_{ab}\big) + 26\times 2\times\log\big(f_{ab}\big). \end{aligned}$$

Reste à remplacer les fréquences par leur valeur en fonction de d ; on obtient

$$\begin{aligned} \ell(d) = {} & 91\times 2\times\log((0{,}396+d)) + 74\times\log(2(0{,}396+d)(0{,}264-d)) + 7\times 2\times\log((0{,}264-d)) \\ & + 50\times\log(2(0{,}396+d)(0{,}204-d)) + 103\times\log(2(0{,}396+d)(0{,}136+d)+2(0{,}204-d)(0{,}264-d)) + 31\times\log(2(0{,}264-d)(0{,}136+d)) \\ & + 3\times 2\times\log((0{,}204-d)) + 15\times\log(2(0{,}204-d)(0{,}126+d)) + 26\times 2\times\log((0{,}136+d)). \end{aligned}$$

Il reste à chercher le maximum de la fonction $\ell(d)$. Notez qu'elle n'est définie que pour $d \in]-0{,}136; 0{,}204[$. Voici les graphes de la log-vraisemblance $\ell(d)$ et de la vraisemblance $\mathrm{L}(d) = \exp \ell(d)$.



Graphes de $\ell(d)$ et de $\mathrm{L}(d) = \exp \ell(d)$.

Les deux fonctions atteignent leur maximum en $d = 0{,}098$. Cette valeur du déséquilibre gamétique permet de calculer les fréquences haplotypiques :

Haplotype	Fréquence
AB	$f_\mathrm{A} f_\mathrm{B} + d = 0{,}494$
Ab	$f_\mathrm{A} f_\mathrm{b} - d = 0{,}166$
aB	$f_\mathrm{a} f_\mathrm{B} - d = 0{,}106$
ab	$f_\mathrm{a} f_\mathrm{b} + d = 0{,}234$

On peut également s'intéresser à la proportion de doubles hétérozygotes Aa, Bb qui sont AB + ab ; c'est

$$\frac{2 f_{\mathrm{AB}} f_{\mathrm{ab}}}{2 f_{\mathrm{AB}} f_{\mathrm{ab}} + 2 f_{\mathrm{Ab}} f_{\mathrm{aB}}} = 0{,}87,$$

soit 87% des doubles hétérozygotes.

3.3 Évolution du déséquilibre au fil des générations

On a considéré jusqu'à présent les fréquences gamétiques dans une population à un temps t donné.

On va voir comment ces fréquences varient au fil des générations. On considère deux locus en déséquilibre gamétique dans les gamètes émis par la génération 0 ; on note $D_0 = f_{AB} - f_A f_B$. On va s'intéresser aux gamètes émis par la génération suivante.

On note comme au chapitre précédent $\theta = \theta_{ab}$ le taux de recombinaison entre les deux locus considérés. Rappelons qu'on est à l'équilibre de Hardy-Weinberg, donc les fréquences f_A et f_B sont constantes d'une génération à l'autre.

Tirons un gamète au hasard parmi les gamètes émis par la génération 1. Quelle est la probabilité f'_{AB} qu'il soit AB ? Il y a deux façon d'obtenir un tel gamète :

— c'est un gamète AB de la génération précédente (probabilité f_{AB}) qui n'a pas recombiné (probabilité $1-\theta$) ;
— c'est un gamète formé après une recombinaison (probabilité θ) entre un d'une part gamète portant l'allèle A (probabilité f_A) et d'autre part un gamète portant l'allèle B (probabilité f_B).

On a donc

$$f'_{AB} = (1-\theta) f_{AB} + \theta f_A f_B.$$

On en déduit immédiatement que le déséquilibre gamétique à la génération suivante, noté D_1, est

$$\begin{aligned} D_1 &= f'_{AB} - f_A f_B \\ &= (1-\theta) f_{AB} + \theta f_A f_B - f_A f_B \\ &= (1-\theta)(f_{AB} - f_A f_B) \\ &= (1-\theta) D_0 \end{aligned}$$

Ainsi, après n générations, le déséquilibre gamétique est

$$\boxed{D_n = (1-\theta)^n D_0.}$$

On voit qu'entre deux locus non liés, c'est-à-dire quand $\theta = \frac{1}{2}$, le déséquilibre gamétique est divisé par 2 à chaque génération. Ainsi, dans une population panmictique, un déséquilibre gamétique qui existerait à un moment donné dans la population entre deux locus non liés s'estomperait en quelques générations.

Au contraire, si θ est petit, par exemple $\theta = 0{,}01$ (les locus sont à une distance d'un centimorgan l'un de l'autre, soit environ un million de paires de bases), le déséquilibre gamétique peut perdurer pendant des générations ; on a $0{,}99^{110} = 0{,}33$, c'est-à-dire que pour des locus distants d'un centimorgan il faut 110 générations, et donc environ 2750 ans pour diviser le déséquilibre par 3.

On parle de déséquilibre de liaison (DL), en anglais *linkage disequilibrium (LD)* quand il y a déséquilibre gamétique entre deux locus génétiquement liés.

Malheureusement la rigueur dans le vocabulaire n'est pas toujours au rendez-vous : on rencontrera souvent « déséquilibre de liaison » pour des locus qui ne sont pas liés, avec parfois la précision « pseudo déséquilibre de liaison » (en anglais, *spurious linkage disequilibrium*).

3.4 Genèse d'un déséquilibre gamétique

Voyons maintenant quels événements peuvent créer un déséquilibre gamétique.

3.4.1 Mélange de populations

On considère une population dont tous les individus sont de génotypes AA et BB, qu'on mélange avec une autre population dont tous les individus sont de génotypes aa et bb.

Dans la nouvelle population obtenue, à la première génération, tous les gamètes sont soit AB, soit ab : on a un déséquilibre complet ! Si la nouvelle population est panmictique, il s'estompera au fil des générations, plus ou moins rapidement selon la distance génétique qui sépare des deux locis.

De façon plus générale, si on mélange deux populations chez lesquelles aucun déséquilibre gamétique n'existe entre les locus considérés, il suffit qu'il existe une différence de fréquences alléliques f_A et f_B entre ces deux populations pour que dans la nouvelle population formée il y ait déséquilibre gamétique. Le cas considéré plus haut est celui où $f_A = 1, f_B = 1$ dans une population, et $f_A = 0, f_B = 0$ dans l'autre.

3.4.2 Mutation

Considérons une population dans laquelle existe un polymorphisme di-allélique A/a ; chez un individu donné, en un locus jusqu'alors monomorphe B, apparaît une mutation b. Si cet individu est de génotype AA au premier locus, il contribuera à l'urne gamétique par des gamètes AB et Ab. Aucun gamète aB ne sera observé dans l'urne gamétique ; on est présence d'un déséquilibre maximal.

Ce déséquilibre s'estompera à nouveau selon la formule $D_n = (1-\theta)^n D$.

Notons que l'allèle muté b pourrait disparaître très vite, ou du moins rester très rare ; le mécanisme décrit sera d'autant plus efficace pour créer un déséquilibre que la population sera petite, que l'allèle b sera soumis à sélection positive, etc.

3.4.3 Sélection, dérive, effet fondateur...

De façon générale, les mécanismes de « pression évolutive », qui tendent à modifier les fréquences alléliques en population, peuvent créer ou amplifier un déséquilibre gamétique.

3.5 Implications en épidémiologie génétique

Le déséquilibre gamétique permet d'obtenir une information sur un locus polymorphe X simplement en considérant un ou plusieurs locus en déséquilibre avec X. Même si le polymorphisme en X est inconnu de l'expérimentateur, il peut espérer détecter son existence en observant, dans une population de cas et de témoins, des locus en déséquilibre de liaison avec X. Ceci peut permettre de localiser avec précision la région du génome, voire précisément le gène au sein duquel existe un polymorphisme dont un ou plusieurs allèles sont impliqués dans une maladie.

Le projet HapMap a établi une carte des SNP les plus fréquents et du DL qui existe entre eux. On peut l'utiliser pour choisir un ensemble de SNP, le plus petit possible, tel que tous les SNP cartographiés sont en DL avec $r^2 > 0{,}80$ (ou un autre seuil arbitraire) avec un des SNP choisis ; on parle alors de tag SNP. Sur la figure 3.3, on

peut voir un choix de tag SNP pour les premiers SNP du gène GLDC ; par exemple les SNP 38 et 55 sont en $r^2 = 0{,}84$, on peut donc ne conserver que le SNP 55, qui « tague » le SNP 38.

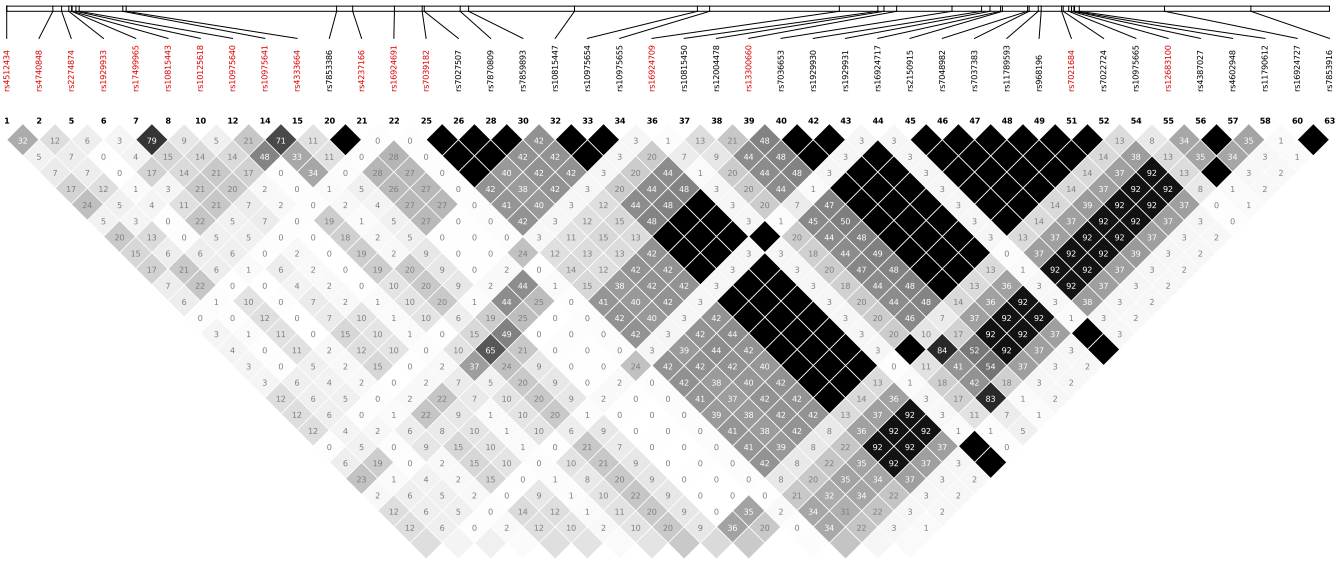


FIGURE 3.3 – Tag SNP pour les premiers SNP du gène GLDC (en rouge)

Cependant, nous l'avons vu, le déséquilibre gamétique peut exister entre des locus non liés, par exemple s'il y a un mélange de population dans la population considérée. Dans ce cas, le déséquilibre peut créer de fausses associations.

3.6 Exercices

Exercice 1 On échantillonne des gamètes au hasard dans une urne gamétique, et on considère deux locus A/a et B/b. On utilise ici toutes les notations du chapitre pour les fréquences alléliques, gamétiques, et les mesures du déséquilibre gamétique.

On note X (respectivement Y) la variable aléatoire qui vaut 1 si le gamète tiré porte l'allèle a (respectivement l'allèle b), et 0 sinon.

1. Les variables X et Y sont des variables de Bernoulli ; quel est leur paramètre ?

2. Montrer que la covariance entre X et Y est égale à D.

3. Montrer que la corrélation entre X et Y est égale à r.

Exercice 2 On considère deux locus di-alléiques A/a et B/b, avec fréquences $f_A = 0{,}6$, $f_a = 0{,}4$, $f_B = 0{,}3$, $f_b = 0{,}7$. En supposant qu'il y a équilibre gamétique, calculer la fréquence des gamètes AB, Ab, aB et ab. Quelle est la valeur minimale que peut prendre le déséquilibre D ? Et la valeur maximale ?

Exercice 3 On considère 1000 individus génotypés en deux locus di-alléliques A/a et B/b. On a les effectifs suivants :

Génotype	AA	Aa	aa
Effectifs	300	490	210

Génotype	BB	Bb	bb
Effectifs	100	420	480

1. S'il y a équilibre gamétique, quelles sont les fréquences des gamètes AB, Ab, aB, et ab ?

2. Les fréquences gamétiques sont données dans la table suivante. Calculer D, D′, r^2.

Gamète	AB	Ab	aB	ab
Fréquences	0,04895	0,49605	0,26105	0,19395

Exercice 4 On considère une population formée par un mélange à 50% de la sous-population S_1 et à 50% de la sous-population S_2. On a deux locus di-alléliques A/a et B/b, en équilibre gamétique dans les deux sous-populations. Les fréquences des allèles dans les sous-population sont données par la table suivante :

Population	f_A	f_a	f_B	f_b
S_1	0,2	0,8	0,5	0,5
S_2	0,6	0,4	0,3	0,7

1. Quelles sont les fréquences alléliques dans la population totale ?

2. La population globale est elle à l'équilibre gamétique ? Calculez D, D′, r^2.

Exercice 5 On considère une petite population et un locus di-allélique A/a, avec $f_A = 0{,}6$. En un locus pour lequel n'existait qu'un allèle B, un allèle « mutant » b apparaît chez un individu de génotype AA, qui émet donc des gamètes Ab.

On suppose que dans l'urne gamétique de cette génération, on a les fréquences suivantes :

f_{AB}	f_{Ab}	f_{aB}	f_{ab}
0,59	0,01	0,40	0,00

1. Pourquoi n'y a-t-il aucun gamète ab dans cette urne ?

2. Calculez les valeurs de D, D′ et r^2 pour ces deux locus.

3. On suppose que le taux de recombinaison entre les deux locus est $\theta = 0{,}1$. Quelle est la valeur de D attendue après 5 générations ? Quelles sont alors les fréquences gamétiques attendues ?

Exercice 6

On considère un locus di-allélique A/a, avec fréquences alléliques $f_A = 0{,}4$ et $f_a = 0{,}6$, et un locus di-allélique B/b, en déséquilibre de liaison avec A/a, avec $f_B = f_b = 0{,}5$.

1. En prenant $D = 0{,}1$, calculez les fréquences f_{AB}, f_{Ab}, f_{aB} et f_{ab}.

2. On note « AB + Ab » pour « porter les haplotypes AB et Ab », etc. Calculez les probabilités suivantes :

$$\begin{array}{lll} \mathbb{P}(AB+AB) & \mathbb{P}(AB+aB) & \mathbb{P}(aB+aB) \\ \mathbb{P}(AB+Ab) & \mathbb{P}(AB+ab) & \mathbb{P}(aB+ab) \\ & \mathbb{P}(aB+Ab) & \\ \mathbb{P}(Ab+Ab) & \mathbb{P}(Ab+ab) & \mathbb{P}(ab+ab) \end{array}$$

Quelle est la probabilité qu'un individu de génotypes Aa, Bb, porte les haplotypes AB + ab ?

3. Même question pour D = 0,2.

Exercice 7 : un modèle simple en épidémiologie génétique (suite)

On considère un locus di-allélique A/a, avec fréquences alléliques $f_A = 0{,}4$ et $f_a = 0{,}6$. On le suppose impliqué dans une maladie humaine, selon le modèle des risques multiplicatifs :

$$\mathbb{P}(\text{Att}|\text{AA}) = \varphi_0 \qquad \mathbb{P}(\text{Att}|\text{A}a) = r\varphi_0 \qquad \mathbb{P}(\text{Att}|aa) = r^2\varphi_0.$$

On prendra $r = 2$.

1. Calculez, en fonction de φ_0, la probabilité d'être atteint, et les proportions attendues des trois génotypes chez individus atteints.

2. On considère un locus di-allélique B/b, en déséquilibre de liaison D = 0,2 avec A/a, avec $f_B = f_b = 0{,}5$. On suppose que le second locus ne joue aucun rôle dans la maladie, de sorte que la probabilité d'être atteint pour une combinaison d'haplotypes ne dépend que du nombre d'allèles A et a présents :

$$\begin{array}{lll}
\mathbb{P}(\text{Att}|\text{AB}+\text{AB}) = \varphi_0 & \mathbb{P}(\text{Att}|\text{AB}+a\text{B}) = r\varphi_0 & \mathbb{P}(\text{Att}|a\text{B}+a\text{B}) = r^2\varphi_0 \\
\mathbb{P}(\text{Att}|\text{AB}+\text{A}b) = \varphi_0 & \mathbb{P}(\text{Att}|\text{AB}+ab) = r\varphi_0 & \mathbb{P}(\text{Att}|a\text{B}+ab) = r^2\varphi_0 \\
 & \mathbb{P}(\text{Att}|a\text{B}+\text{A}b) = r\varphi_0 & \\
\mathbb{P}(\text{Att}|\text{A}b+\text{A}b) = \varphi_0 & \mathbb{P}(\text{Att}|\text{A}b+ab) = r\varphi_0 & \mathbb{P}(\text{Att}|ab+ab) = r^2\varphi_0
\end{array}$$

En utilisant les calculs réalisés à l'exercice précédent et la formule de Bayes, calculez

$$\begin{array}{lll}
\mathbb{P}(\text{AB}+\text{AB}|\text{Att}) & \mathbb{P}(\text{AB}+a\text{B}|\text{Att}) & \mathbb{P}(a\text{B}+a\text{B}|\text{Att}) \\
\mathbb{P}(\text{AB}+\text{A}b|\text{Att}) & \mathbb{P}(\text{AB}+ab|\text{Att}) & \mathbb{P}(a\text{B}+ab|\text{Att}) \\
 & \mathbb{P}(a\text{B}+\text{A}b|\text{Att}) & \\
\mathbb{P}(\text{A}b+\text{A}b|\text{Att}) & \mathbb{P}(\text{A}b+ab|\text{Att}) & \mathbb{P}(ab+ab|\text{Att})
\end{array}$$

3. En déduire les fréquences des génotypes BB, Bb et bb chez les individus atteints. Que remarquez-vous ?

Les plus courageux referont ce calcul sans utiliser de valeur numériques pour f_A, f_a, f_B, f_b, D et r.

4 Apparentement et consanguinité

4.1 Introduction

Deux individus apparentés partagent un ou plusieurs ancêtres. On réserve souvent le terme au cas où ces ancêtres communs ne sont pas trop éloignés (en terme de nombre de générations), car remarquer que nous sommes tous apparentés (ainsi qu'aux mouches et aux algues bleues) ne nous mène pas très loin. Quand deux individus sont apparentés, il est possible qu'ils partagent deux allèles identiques parce que reçu d'un ancêtre commun : on parle alors en anglais d'allèles *Identical By Descent* (IBD) et en français d'*identiques par descendance*[1]. L'abréviation IBD est très largement utilisée.

Les individus consanguins sont nés d'une union entre apparentés (cousins, oncle/nièce, etc). La figure 4.1 montre l'exemple de l'union entre cousins germains.

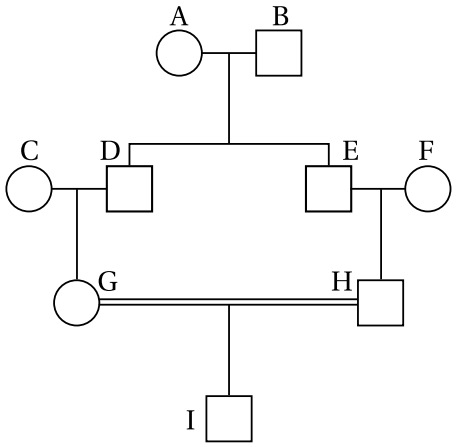


FIGURE 4.1 – Enfant issu de cousins germains

La probabilité pour un individu consanguin d'être homozygote en un locus polymorphe quelconque sur le génome est supérieure à ce qui est prédit par les proportions d'Hardy-Weinberg. Considérons la figure 4.1 : au locus considéré l'enfant I peut avoir reçu deux copies d'un même allèle ancestral ; par exemple, une même copie d'un allèle de la grand-mère A peut lui être transmise via les individus D et G, et via les individus E et H ; les deux allèles sont donc IBD. On dit que I est *Homozygous By Descent (HBD)*, en français *homozygote par descendance*.

Il est important de faire la différence entre IBD et IBS, *Identity By State*, identité par état : chez un individu homozygote en un locus considéré, les deux allèles sont les mêmes (ils sont IBS) mais cela n'implique pas qu'ils dérivent d'un ancêtre commun. Une différence notable est que quand deux individus partagent deux allèles IBD en un locus précis du génome, alors ils partegeront tout un segment de chromosome s'étendant de part et d'autre de ce locus (voir également la figure 4.3 à la fin du chapitre).

1. Il s'agit d'une mauvaise traduction : *descent* a pour définition *derivation from an ancestor :* on traduirait mieux par *identique par origine* voire *par ascendance*

4.2 Coefficients de parenté et de consanguinité

4.2.1 Définitions

Définissons d'abord quelques termes.

- Deux individus sont apparentés s'ils ont un ancêtre commun ;
- un individu est consanguin si ses parents sont apparentés ;
- le coefficient de parenté de deux individus A et B est la probabilité Φ_{AB} que deux allèles tirés au hasard, l'un chez A, l'autre chez B (au même locus) soient identiques par descendance (IBD) ;
- le coefficient de consanguinité d'un individu I est la probabilité f_I que les deux allèles portés par l'individu (en un locus donné) soient identiques par descendance (IBD).

Une conséquence immédiate de la définition est que le coefficient de consanguinité d'un individu I est égal au coefficient de parenté de ses parents P et M :

$$\boxed{f_I = \Phi_{PM}.}$$

Le coefficient de parenté d'un individu I avec lui-même est lié à son coefficient de consanguinité par

$$\boxed{\Phi_{II} = \frac{1}{2} + \frac{1}{2} f_I = \frac{1}{2}(1 + f_I);}$$

en effet, deux allèles tirés au hasard chez l'individu I sont, soit le même allèle (proba $\frac{1}{2}$), soit les deux allèles distincts portés par l'individu (proba $\frac{1}{2}$) qui sont alors IBD avec probabilité f_I.

4.2.2 Calcul par la méthode des boucles

Reprenons l'exemple de l'enfant de cousins germains. Nous supposons ci-dessous que les individus C et F ne sont pas apparentés entre eux ni avec A et B, et, dans un premier temps, que A et B ne sont pas consanguins.

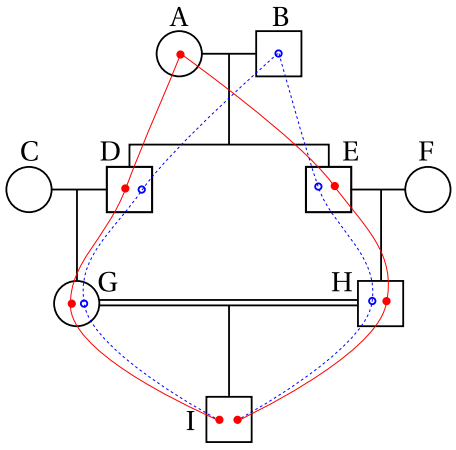


FIGURE 4.2 – Calcul du coefficient de consanguinité

Considérons les transmissions à partir de l'ancêtre A. Il y a une probabilité $\frac{1}{2}$ que A transmette un même allèle à D et E. Si D et E ont reçu le même allèle de A, D le transmet à G avec probabilité $\frac{1}{2}$ et E le transmet à H avec

probabilité $\frac{1}{2}$. Si enfin G et H ont reçu de la sorte le même allèle de A (ce qui arrive avec probabilité $\frac{1}{8}$), ils le transmettent à I **chacun** avec probabilité $\frac{1}{2}$; I reçoit donc deux allèles IBD provenant de A avec probabilité $\frac{1}{32}$.

Le même résultat vaut pour les transmissions à partir de B : I reçoit un allèle IBD de B avec probabilité $\frac{1}{32}$. Pour finir, I pouvant recevoir ses deux allèles IBD de A ou de B, le coefficient de consanguinité de I vaut $\frac{1}{32} + \frac{1}{32} = \frac{1}{16}$.

Si A et B sont consanguins, il faut tenir compte du fait que même s'ils ont transmis deux allèles différents, ces deux allèles peuvent être IBD. Reprenons la première « boucle de consanguinité », c'est-à-dire les transmissions à partir de A. Si A transmet chacun de ses deux allèles au locus considéré à D et E (probabilité $\frac{1}{2}$), ces deux allèles peuvent quand-même être IBD avec probabilité f_A ; ils sont ensuite transmis à I avec probabilité $\frac{1}{16}$. Donc au final la probabilité pour I d'avoir un allèle IBD reçu de A est $\frac{1}{32}(1+f_A)$. C'est la même chose pour l'autre boucle, et pour finir

$$f_I = \frac{1}{32}(1+f_A) + \frac{1}{32}(1+f_B).$$

Cet exemple permet de comprendre la formule générale de Wright :

$$f_I = \sum_{\text{A: ancêtre commun}} \left(\frac{1}{2}\right)^{n_A+m_A+1} (1+f_A)$$

où la somme se fait sur toutes les boucles de consanguinté, n_A et m_A sont les nombres de méioses entre d'une part le père de I et l'ancêtre A, et d'autre part la mère de I et l'ancêtre A.

La longueur totale de la boucle est donc $n_A + m_A + 2$, et le terme $n_A + m_A + 1$ peut également être écrit comme « longueur de la boucle -1 ».

Relation	Φ
Identité	$\frac{1}{2}$
Parent/enfant	$\frac{1}{4}$
Germains	$\frac{1}{4}$
Demi-germains	$\frac{1}{8}$
Oncle/nièce	$\frac{1}{8}$
Cousins germains	$\frac{1}{16}$

TABLE 4.1 – Coefficients de parenté classiques

La table 4.1 récapitule quelques coefficients de parenté classiques, quand les ancêtres ne sont ni consanguins, ni apparentés entre eux. La vérification de ces valeurs est à faire en exercice. Ces valeurs donnent également les coefficients de consanguinité des enfants nés d'union entre individus ayant cet apparentement. Notez l'emploi de « germain », qui veut dire frère ou soeur (en anglais : *sib* ou *sibling*).

4.2.3 Calcul par récurrence

Il y a une méthode plus mécanique, peut-être un peu fastidieuse mais sans doute plus sûre quand les généalogies sont compliquées. Il s'agit de calculer façon systématique les coefficients de parenté, en « remontant » l'arbre. On a les relations suivantes :

$$\Phi_{II} = \frac{1}{2}(1 + f_I)$$

Si P est le père de I et M est la mère de I,

$$\boxed{\begin{aligned}\Phi_{PI} &= \frac{1}{4} + \frac{1}{4}f_P + \frac{1}{2}\Phi_{PM} = \frac{1}{2}\Phi_{PP} + \frac{1}{2}\Phi_{PM}\\ \Phi_{MI} &= \frac{1}{4} + \frac{1}{4}f_M + \frac{1}{2}\Phi_{PM} = \frac{1}{2}\Phi_{MM} + \frac{1}{2}\Phi_{PM}\end{aligned}}$$

Et enfin, si I et J ne sont pas parent l'un de l'autre, et si P_I, M_I, P_J, M_J sont leurs parents respectifs,

$$\boxed{\begin{aligned}\Phi_{IJ} &= \frac{1}{2}\left(\Phi_{P_I J} + \Phi_{M_I J}\right)\\ &= \frac{1}{2}\left(\Phi_{I P_J} + \Phi_{I M_J}\right)\\ &= \frac{1}{4}\left(\Phi_{P_I P_J} + \Phi_{P_I M_J} + \Phi_{M_I P_J} + \Phi_{M_I M_J}\right)\end{aligned}}$$

Cette dernière relation reste valable si $P_I = P_J$ ou $M_I = M_J$ ou les deux (auquel cas I et J sont germains).

En appliquant ces relations de façon répétitive, on arrive à calculer les coefficients de parentés entre deux personnes quelconques de l'arbre, et donc les coefficients de consanguinité.

Exemple des cousins germains

Reprenons l'exemple de la figure 4.2. On a $f_I = \Phi_{GH}$; pour calculer Φ_{GH}, on écrit

$$\begin{aligned}\Phi_{GH} &= \frac{1}{4}\left(\Phi_{CE} + \Phi_{CF} + \Phi_{DE} + \Phi_{DF}\right)\\ &= \frac{1}{4}\Phi_{DE},\end{aligned}$$

car C et E ne sont pas apparentés, ni C et F, ni D et F. On a ensuite

$$\begin{aligned}\Phi_{DE} &= \frac{1}{4}\left(\Phi_{AA} + \Phi_{AB} + \Phi_{BA} + \Phi_{BB}\right)\\ &= \frac{1}{4}\left(\Phi_{AA} + \Phi_{BB}\right)\\ &= \frac{1}{8}\left(2 + f_A + f_B\right)\end{aligned}$$

et enfin $f_I = \Phi_{GH} = \frac{1}{16} + \frac{1}{32}(f_A + f_B)$.

4.2.4 Probabilité des génotypes chez un individu consanguin

On considère un individu dont le coefficient de consanguinité f est connu. À un locus donné, l'individu a deux allèles IBD avec probabilité f, et deux allèles ayant une provenance distincte avec probabilité $(1 - f)$.

Dans le premier cas, il a génotype AA avec probabilité p, et aa avec probabilité q ; dans le second cas, il a génotype AA avec probabilité p^2, Aa avec probabilité $2pq$ et aa avec probabilité q^2. Pour finir,

$$\begin{array}{ll} \mathbb{P}(\mathrm{AA}) = (1-f)p^2 + fp & = p^2 + fpq \\ \mathbb{P}(\mathrm{A}a) = (1-f)2pq & = 2pq - 2fpq \\ \mathbb{P}(aa) = (1-f)q^2 + fq & = q^2 + fpq \end{array}$$

Cette écriture peut servir pour tout écart à l'équilibre de Hardy-Weinberg, quelle qu'en soit la cause ; f est simplement un coefficient de corrélation allélique.

4.3 Corrélation allélique et défaut d'hétérozygotie

Le coefficient de consanguinité f s'interprète comme une corrélation. Considérons un individu avec un coefficient de consanguinité f, et un locus di-allélique A/a. On note X_p la variable qui vaut 1 si l'individu a reçu l'allèle a de son père, 0 sinon ; et X_m la variable homologue pour l'allèle maternel. Les variables X_p et X_m sont donc des variables de Bernoulli de paramètre q, la fréquence de l'allèle a dans la population. Leur covariance est

$$\begin{aligned} \mathrm{cov}(\mathrm{X}_p, \mathrm{X}_m) &= \mathrm{E}(\mathrm{X}_p\mathrm{X}_m) - \mathrm{E}(\mathrm{X}_p)\mathrm{E}(\mathrm{X}_m) \\ &= \mathbb{P}(\mathrm{X}_p = 1 \text{ et } \mathrm{X}_m = 1) - q^2 \\ &= q^2 + fpq - q^2 \\ &= fpq \end{aligned}$$

L'écart-type de X_p comme celui de X_m valent $\sqrt{pq}$, donc

> le coefficient de corrélation entre X_p et X_m est f.

Inversement,
- si X_p et X_m suivent une loi de Bernoulli de paramètre q (c.-à-d. chacun des deux allèles reçus des parents est a avec probabilité q),
- si $\mathrm{cor}(\mathrm{X}_p, \mathrm{X}_m) = f$ (peu importe la cause de cette corrélation),

alors l'individu considéré est de génotype aa avec probabilité égale à $\mathrm{E}(\mathrm{X}_p\mathrm{X}_m) = \mathrm{E}(\mathrm{X}_p)\mathrm{E}(\mathrm{X}_m) + \mathrm{cov}(\mathrm{X}_p, \mathrm{X}_m) = q^2 + fpq = (1-f)q^2 + fp$. De même, son génotype est Aa avec probabilité $\mathrm{E}((1-\mathrm{X}_p)\mathrm{X}_m) = 2pq - 2fpq = (1-f)2pq$ et de génotype AA avec probabilité $\mathrm{E}((1-\mathrm{X}_p)(1-\mathrm{X}_m)) = p^2 + fpq = (1-f)p^2 + fp$.

Enfin, une autre caractérisation possible de f est son lien avec le taux d'hétérozygotie :

$$1 - f = \frac{\mathrm{H}}{2pq}.$$

4.4 Consanguinité dans une population

On considère une population où une partie des unions se fait entre apparentés (mariage préférentiel entre cousins, entre oncle et nièce).

On peut définir un coefficient F de consanguinité moyen de la population, qui sera la probabilité que les deux allèles portés par un individu quelconque soient IBD.

On a alors, dans cette population, les fréquences génotypiques suivantes :

$$\boxed{\begin{aligned} f_{\mathrm{AA}} &= (1-\mathrm{F})p^2 + \mathrm{F}p \\ f_{\mathrm{A}a} &= (1-\mathrm{F})2pq \\ f_{aa} &= (1-\mathrm{F})q^2 + \mathrm{F}q \end{aligned}}$$

4.4.1 Exemple : population partiellement autogame

Cet exemple est pertinent pour des plantes telles que le pois, qui se reproduiront partiellement en régime autogame, partiellement en régime panmictique. Notons α la probabilité qu'un individu soit issu d'une autofécondation.

On calcule le taux d'hétérozygotie à la génération $t+1$. Si l'individu est issu d'une autofécondation (proba α), pour qu'il soit hétérozygote il faut que son géniteur soit lui-même hétérozygote (proba H_t) et qu'il transmette les deux allèles distincts qu'il porte (proba 1/2) ; si il est issu d'une union panmictique, il est hétérozygote avec probabilité $2pq$ (on sait que les fréquences alléliques ne varient pas).

On a donc

$$\mathrm{H}_{t+1} = \alpha \times \frac{1}{2} \times \mathrm{H}_t + (1-\alpha) \times 2pq.$$

Quand la population est à l'équilibre, le taux d'hétérozygotie vérifie

$$\mathrm{H} = \frac{1}{2}\alpha\mathrm{H} + (1-\alpha)2pq,$$

et donc $\mathrm{H} = \frac{2(1-\alpha)}{2-\alpha}2pq$. On calcule la consanguinité moyenne à l'équilibre :

$$\begin{aligned} \mathrm{F} &= 1 - \frac{\mathrm{H}}{2pq} \\ &= \frac{\alpha}{2-\alpha} \end{aligned}$$

Dans le cas particulier où $\alpha = 1$ (autogamie totale), on a $\mathrm{H}_{t+1} = \frac{1}{2}\mathrm{H}_t$: le taux d'hétérozygotie décroît rapidement et à l'équilibre, $\mathrm{F} = 1$ et tous les individus sont homozygotes.

4.5 Implications en épidémiologie génétique

4.5.1 Maladies récessives et consanguinité

Considérons une maladie récessive dans une population globalement panmictique, avec de rares cas d'union entre apparentés, par exemple entre cousins germains ($\Phi = \frac{1}{16}$). On suppose que la fréquence de l'allèle morbide est q.

Un enfant qui naît d'une union panmicitique a une probabilité q^2 d'être atteint. Si ses parents sont apparentés avec un coefficient de parenté Φ, cette probabilité devient $q^2 + \Phi pq > q^2$.

Prenons l'exemple de la mucoviscidose, où $q = 0{,}02$. Le risque en population est $q^2 = 1/2500$; $\Phi = \frac{1}{16}$, on calcule $q^2 + \Phi pq = 1/615$, soit une multiplication du risque par 4 environ.

Dans le cas de la phénylcétonurie, $q = 0{,}008$ et $q^2 = 1/16000$, et le risque pour un enfant issu de cousins germains est de $q^2 + \Phi pq = 1/1800$, soit un risque multiplié par 9.

Le risque relatif est

$$\frac{q^2 + \Phi pq}{q^2} = 1 + \Phi \frac{p}{q};$$

plus q est petit, plus ce risque relatif est important. Dans le cas de certaines maladies récessives rares, la plupart des atteints seront issus de familles consanguines[2].

Remarque Ces considérations remettent naturellement en cause le calcul des fréquences alléliques sur la seule base des prévalences, en supposant l'équilibre de Hardy-Weinberg ; il faudrait distinguer les enfants issus d'une union entre apparenté des autres enfants, et disposer d'une estimation de la fréquence des unions entre apparentés dans la population.

4.5.2 Homozygosity mapping

Comme on l'a dit, quand un individu consanguin a reçu deux allèles IBD en un locus, il est IBD sur tout un segment de chromosome de part et d'autre de ce locus. La figure 4.3 illustre ce fait en faisant apparaître les recombinaisons successives sur un chromosome (de longueur approximativement 1 cM).

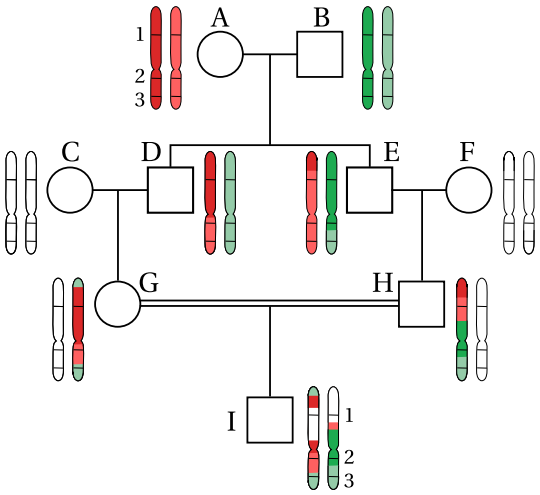


FIGURE 4.3 – Homozygosity mapping

On voit que l'individu a reçu deux allèles IBD au locus 3, et sur tout un segment de chromosome. On dit qu'il est *Homozygous By Descent,* HDB.

Ceci peut permettre de localiser un gène impliqué dans une maladie récessive rare, qui sera essentiellement présente chez des individus consanguins, l'individu ayant reçu deux allèles morbides du gène impliqués d'un ancêtre commun. Ce gène se trouve donc dans une région HBD du génome. On génotype les atteints en une famille de marqueurs couvrant le génome ; les segments HBD sont détectables car tous les marqueurs dans ces régions sont homozygotes. Une longue série de marqueurs homozygotes (un *run of homozygosity*) est donc l'indice d'une région HBD.

Cette méthode appelée *homozygosity mapping* a été proposée par Lander et Botstein en 1987 ; elle fait encore l'objet de développements méthodologiques aujourd'hui.

2. Garrod avait déjà remarqué en 1902 que c'était le cas pour les enfants atteints de phénylcétonurie.

4.6 Exercices

Exercice 1 Calculez le coefficient de consanguinité de l'individu A dans le pedigree de la figure 4.4.

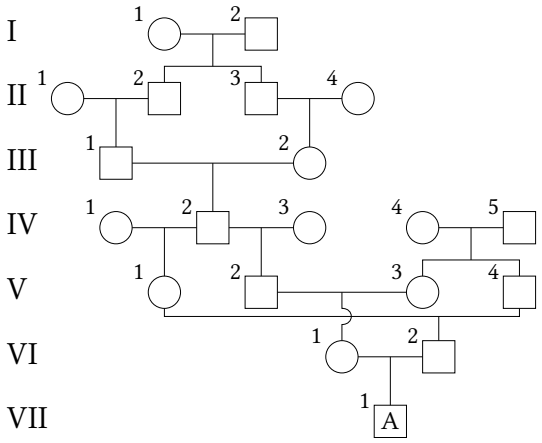


FIGURE 4.4 – Un pedigree complexe

Exercice 2 On reprend ici les notations de la section 4.3. Si on note $X = X_p + X_m$, le génotype de l'individu considéré et les valeurs de X sont en correspondance univoque : AA pour X = 0, Aa pour X = 1 et aa pour X = 2. Calculez la variance X.

Exercice 3 Le syndrôme d'Hergé est une maladie récessive. On suppose qu'en Syldavie, un petit royaume d'Europe centrale, on observe un cas de syndrôme d'Hergé pour 625 naissances (ou 16 pour 10 000).

a. En supposant que la population syldave est panmictique, estimer la fréquence q de l'allèle morbide.

b. En fait les syldaves se marient souvent entre cousins germains, de sorte que 17,8% des naissances sont des enfants de cousins germains ; on négligera les autres mariages consanguins, très rares. Dans ces conditions, quelle proportion de syndrôme d'Hergé attend-on parmi les nouveaux nés si $q = 0{,}03$? si $q = 0{,}035$? et si $q = 0{,}04$?

Laquelle de ces valeurs de q produit la proportion de syndrômes d'Hergé la plus proche de la valeur observée (1/625) ? On prendra cette valeur de q pour répondre à la question suivante.

c. Quelle proportion d'enfants atteints sont issus d'une union entre cousins germains ?

Exercice 4 (pour programmeurs aguerris) Écrire un programme qui calcule les coefficients d'apparentement entre tous les individus d'une famille. On utilisera bien sûr la méthode par récurrence. L'entrée du programme pourra être un table dont les colonnes seront `id`, `pere`, `mere`, `f` : chaque individu a un identifiant `id` non nul, on donne l'identifiant de ses parents (0 si c'est un fondateur), son coefficient de consanguinité si c'est un fondateur (on ignorera la valeur de cette colonne sinon). Il est naturel que le résultat du proramme soit rendu sous forme matricielle.

Par exemple, la famille de la figure 4.1 sera codée par une table

```
A 0 0 0
B 0 0 0
C 0 0 0
D A B -
E A B -
F 0 0 0
G C D -
H E F -
I G H -
```

5 Populations structurées

Nous considérons à présent une population structurée en sous-populations isolées.

5.1 Effet Wahlund : mélange de populations panmictiques

Commençons par un exemple numérique. Supposons qu'on a deux sous-populations panmictiques S_1 et S_2 de sorte que l'équilibre de Hardy-Weinberg y est réalisé. La fréquence de A dans la première population est 0,2, et 0,6 dans la seconde. La population totale est constituée d'un mélange de S_1 et S_2 en proportions égales. La table suivante donne les fréquences génotypiques dans les deux sous-populations et dans la population totale.

Population	p	$f(\mathrm{AA})$	$f(\mathrm{A}a)$	$f(aa)$
S_1	0,2	0,04	0,32	0,64
S_2	0,6	0,36	0,48	0,16
Globale	0,4	0,20	0,40	0,40

TABLE 5.1 – Mélange de deux populations à l'équilibre de Hardy-Weinberg

On constate que dans la population totale, on n'est pas à l'équilibre de Hardy-Weinberg ; il y a un déficit d'hétérozygotes.

5.2 Formalisation

5.2.1 Corrélation allélique

De même que la consanguinité, cette structure de population induit une corrélation positive entre les allèles portés par un individu pris dans la population totale : ses deux parents étant issus de la même sous-population, les allèles qu'il a reçu de chacun d'eux sont plus souvent identiques que si la population totale était panmictique.

Nous calculerons dans un moment la valeur de cette corrélation en fonction des fréquences alléliques dans chacune des populations ; en attendant, notons la simplement F.

En vertu des calculs effectués à la section 4.3, on a

$$\begin{aligned} f_{\mathrm{AA}} &= p^2 + pq\mathrm{F} \\ f_{\mathrm{A}a} &= 2pq - 2pq\mathrm{F} \\ f_{aa} &= q^2 + pq\mathrm{F} \end{aligned}$$

où p et $q = 1 - p$ sont les fréquences des allèles A et a dans la population totale.

5.2.2 Calcul de $F = F_{ST}$

On considère n sous-populations $S_1,\dots,S_n$, en proportion $\alpha_1,\dots,\alpha_n$ dans la population totale. On suppose que chacune des sous-populations est à l'équilibre de Hardy-Weinberg, et que l'allèle A a fréquence p_i dans la sous-population S_i ; on pose $q_i = 1 - p_i$.

On a $p = \sum_i \alpha_i p_i$ et $q = \sum_i \alpha_i q_i = 1 - p$. Pour calculer F, on considère la probabilité qu'un individu pris au hasard ait génotype AA : c'est $f_{AA} = \sum_i \alpha_i p_i^2$. On a alors

$$\begin{aligned} F &= \frac{1}{pq}(f_{AA} - p^2) \\ &= \frac{1}{pq}\left(\sum_i \alpha_i p_i^2 - p^2\right) \qquad\qquad = \frac{1}{pq}\sum_i \alpha_i \left(p_i^2 - p^2\right) \\ &= \frac{1}{pq}\sum_i \alpha_i (p_i - p)^2 \end{aligned} \tag{5.1}$$

Certaines des égalités ci-dessus nécessitent de remarquer que $\sum_i \alpha_i = 1$. La dernière montre que F_{ST} est positif.

Une autre façon de faire le calcul est de considérer la probabilité qu'un individu pris au hasard ait génotype Aa : c'est $f_{Aa} = \sum_i \alpha_i 2 p_i q_i$. On a alors

$$\begin{aligned} 1 - F &= \frac{1}{2pq} f_{aA} \\ &= \frac{1}{2pq}\sum_i \alpha_i \cdot 2 p_i q_i \\ &= \frac{1}{pq}\sum_i \alpha_i p_i q_i. \end{aligned}$$

On peine à reconnaître l'expression précédente. Cependant on a $\sum_i \alpha_i p_i q_i = \sum_i \alpha_i \left(p_i - p_i^2\right) = p - \sum_i \alpha_i p_i^2$ et on retombe pour F sur l'expression écrite à l'équation 5.1 :

$$\begin{aligned} F &= 1 - \frac{1}{pq}\sum_i \alpha_i p_i q_i \\ &= \frac{1}{pq}\left(pq - p + \sum_i \alpha_i p_i^2\right) \\ &= \frac{1}{pq}\left(-p^2 + \sum_i \alpha_i p_i^2\right). \end{aligned}$$

5.2.3 Indice de fixation

On note F_{ST} cette quantité ; on l'appelle *indice de fixation de Wright*. On vient de voir plusieurs façons équivalentes de le calculer, notamment

$$\boxed{\begin{aligned} F_{ST} &= \frac{1}{pq}\left(\sum_i \alpha_i p_i^2 - p^2\right) \\ &= \frac{1}{pq}\sum_i \alpha_i \left(p_i^2 - p^2\right) \\ &= \frac{1}{pq}\sum_i \alpha_i (p_i - p)^2 \end{aligned}}$$

$$\boxed{1 - \mathrm{F_{ST}} = \frac{1}{pq} \sum_i \alpha_i p_i q_i.}$$

L'emploi du terme de « fixation » s'explique ainsi : on dit que dans une population, un allèle est *fixé* si c'est le seul qui est observé, l'autre allèle ayant disparu. Si dans chacune des sous-populations considérées, un des deux allèles A ou a est *fixé* (c'est-à-dire $p_i = 0$ ou 1 pour $i = 1, \ldots, n$), alors on a $\mathrm{F_{ST}} = 1$.

5.3 La magie des statistiques F de Wright

On ne suppose plus que chacune des sous-populations est à l'équilibre de Hardy-Weinberg ; il existe dans chacune d'elle une corrélation allélique notée $\mathrm{F_{IS}^{(i)}}$ (pour $i = 1, \ldots, n$).

On a dans chaque population des fréquences alléliques p_i, q_i. On pose toujours

$$\mathrm{F_{ST}} = \frac{1}{pq} \sum_i \alpha_i \left(p_i^2 - p^2 \right).$$

Cet indice de fixation représente la corrélation allélique induite par les différences de fréquences alléliques entre les sous-population. Si chacune des sous-population est à l'équilibre de Hardy-Weinberg, l'histoire s'arrête là, comme nous l'avons vu plus haut.

Quand ça n'est pas le cas, il faut calculer la valeur de F induite par l'effet cumulé des différences alléliques entre les sous-populations ($\mathrm{F_{ST}}$) et par les corrélations alléliques intra-populations ($\mathrm{F_{IS}^{(1)}}, \ldots, \mathrm{F_{IS}^{(n)}}$).

On note $\mathrm{F_{IT}}$ cette valeur.

5.3.1 Même écart à l'équilibre dans chaque sous-population

On suppose pour simplifier le problème qu'on a la même corrélation allélique dans chaque sous-population : $\mathrm{F_{IS}^{(1)}} = \cdots = \mathrm{F_{IS}^{(n)}} = \mathrm{F_{IS}}$.

Calculons $\mathrm{F_{IT}}$ en utilisant la caractérisation par le taux d'hétérozygotie. Le taux d'hétérozygotie dans la sous-population i est $(1 - \mathrm{F_{IS}}) \cdot 2 p_i q_i$. Le taux d'hétérozygotie dans la population totale est donc

$$\sum_i \alpha_i (1 - \mathrm{F_{IS}}) \cdot 2 p_i q_i = (1 - \mathrm{F_{IS}}) \sum_i \alpha_i \cdot 2 p_i q_i.$$

On a alors

$$\begin{aligned} 1 - \mathrm{F_{IT}} &= \frac{1}{2pq} (1 - \mathrm{F_{IS}}) \sum_i \alpha_i \cdot 2 p_i q_i \\ &= (1 - \mathrm{F_{IS}})(1 - \mathrm{F_{ST}}) \end{aligned}$$

grâce à la deuxième caractérisation de $\mathrm{F_{ST}} = \frac{1}{pq} \sum_i \alpha_i p_i q_i$.

5.3.2 Écarts à l'équilibre spécifiques à chaque sous-population

Reprenons le calcul ci-dessus ; on obtient

$$1 - \mathrm{F_{IT}} = \frac{1}{2pq} \sum_i \alpha_i \left(1 - \mathrm{F_{IS}^{(i)}} \right) 2 p_i q_i$$

et on ne peut plus metre $\left(1-\mathrm{F}_{\mathrm{IS}}^{(i)}\right)$ en facteur dans la somme comme on l'a fait précédemment.

Posons

$$\overline{\mathrm{F}_{\mathrm{IS}}} = \frac{\sum_i \alpha_i p_i q_i \mathrm{F}_{\mathrm{IS}}^{(i)}}{\sum_i \alpha_i p_i q_i}.$$

C'est un indice de fixation moyen, où chaque $\mathrm{F}_{\mathrm{IS}}^{(i)}$ est pondéré par $\alpha_i p_i q_i$. Cette définition est posée de sorte que l'égalité suivante soit vraie :

$$\sum_i \alpha_i \left(1-\mathrm{F}_{\mathrm{IS}}^{(i)}\right) p_i q_i = \left(1-\overline{\mathrm{F}_{\mathrm{IS}}}\right) \sum_i \alpha_i p_i q_i.$$

La mise en facteur voulue est faite. On peut finir le calcul de $1-\mathrm{F}_{\mathrm{IT}}$:

$$\begin{aligned} 1-\mathrm{F}_{\mathrm{IT}} &= \left(1-\overline{\mathrm{F}_{\mathrm{IS}}}\right) \times \frac{1}{pq} \sum_i \alpha_i p_i q_i \\ &= \left(1-\overline{\mathrm{F}_{\mathrm{IS}}}\right) (1-\mathrm{F}_{\mathrm{ST}}) \end{aligned}$$

On a montré comme précédemment la formule magique

$$\boxed{(1-\mathrm{F}_{\mathrm{IT}}) = \left(1-\overline{\mathrm{F}_{\mathrm{IS}}}\right)(1-\mathrm{F}_{\mathrm{ST}})\,.}$$

Les indices qui décorent les divers indices de fixation sont des abréviations :

— F_{IS}, Individu dans la Sous-population
— F_{ST}, Sous-population dans le Total
— F_{IT}, Individu dans le Total

Cette formule permet de séparer, dans l'indice de fixation global, ce qui provient de la structure spatiale en sous-population (F_{ST}), et ce qui provient d'effets internes propres aux sous-populations (F_{IS}). Les corrélations F_{IS} internes aux sous-populations peuvent venir par exemple d'une consanguinité moyenne, ou d'une subdivision en sous-populations... On peut ajouter des « étages » à volonté, on obtient des formules analogues.

5.4 Implications en épidémiologie génétique

Les indices de Wright permettent de mesurer la différentiation entre des populations en un locus donné. Cela peut permettre de détecter des régions génomiques qui ont été soumises à une pression de sélection différente selon les populations, et donc la présence de polymorphismes fonctionnels dans ces régions. La connaissance des régions qui varient beaucoup d'une population à l'autre est essentielle en analyse d'association sur le génome entier, car un signal d'association dans une de ces régions peut être liée à un problème de « stratification de population ».

5.5 Exercices

Exercice 1 Montrer que si un allèle est fixé dans chaque population ($p_i = 0$ ou 1), alors $\mathrm{F}_{\mathrm{ST}} = 1$ (on fera le calcul avec les deux caractérisations de F_{ST}).

Exercice 2 Calculer la valeur de F_{ST} pour le mélange de population décrit en section 5.1.

Exercice 3 Le SNP rs4988235 est situé dans un intron du gène MCM6, un gène impliqué dans la réplication du génome. Cet intron est une région régulatrice d'un gène proche, LCT, qui encode la lactase. Les deux allèles de rs4988235 sont C et T, et l'allèle T est associé à la persistance de la lactase dans la population européenne. La table ci-dessous donne les génotypes observés dans les échantillons de population européennes du projet 1000 génomes.

Population	TT	TC	CC	–
CEU	54	38	7	99
FIN	32	53	14	99
GBR	49	33	9	91
IBS	25	48	34	107
TSI	2	15	90	107
Total	162	187	154	503

Calculez les indices de fixation population par population, dans la population totale, et vérifiez que la relation de Wright est exacte (on doit trouver $F_{IT} = 0{,}256$, $F_{ST} = 0{,}231$, $\overline{F_{IS}} = 0{,}0328$).

6 Sélection

Introduction

On va s'intéresser dans ce court chapitre aux conséquences de l'abandon de l'hypothèse d'absence de sélection dans le modèle de Hardy-Weinberg. On conserve les autres hypothèses, notament la panmixie, l'absence de sélection gamétique, les générations séparées, et la population de taille infinie.

Ici cependant les individus qui composent une génération données ont des chances inégales de transmettre leurs gamètes à la génération suivante ; ceci peut être dû à des différences de mortalité avant l'âge reproductif, ou encore à une fécondité réduite.

On ne va considérer qu'un seul locus génétique di-allélique soumis à sélection. On s'intéresse à l'évolution des fréquences alléliques en ce locus. On ne traitera pas, parmis plusieurs sujets intéressants, aux effets que la sélection peut induire sur le déséquilibre gamétique dans la région où ce trouve ce locus.

6.1 Modèle

6.1.1 Valeur sélective

On considère un locus diallélique d'allèles A et a, soumis à sélection, c'est-à-dire que la fécondité des individus dépend de leur génotype en ce locus.

On notera p_t la fréquence de l'allèle A à la génération t (pour pallier toute ambiguïté, on conviendra qu'il s'agit de la fréquence à la naissance, voire à la formation des zygotes – en tout cas avant que la sélection fasse son œuvre). On note $q_t = 1 - p_t$ la fréquence de l'allèle a.

Notons u, v et w les valeurs sélectives[1] des trois génotypes :

Génotype :	AA	Aa	aa
Valeur sélective :	u	v	w

On peut définir les valeurs sélectives comme la probabilité qu'un individu de génotype donné contribue à l'urne gamétique. En fait, il suffit que u, v et w soient proportionnels à ce nombre ; seules les valeurs sélectives relatives importent, c'est-à-dire les proportions $u : v : w$.

En l'absence de sélection, on a $u = v = w$.

6.1.2 L'urne gamétique

L'hypothèse de panmixie restant valide, on peut utiliser le modèle de l'urne gamétique. Quelle est la probabilité qu'un gamète tiré au hasard dans l'urne contenant les gamètes émis par la génération t soit A ?

1. En anglais, *fitness*. Il ne s'agit pas des activités pratiquées dans les salles de sport, mais de la viabilité ou de la fertilité des individus — de leur succès reproducteur.

Notons tout d'abord que la génération parentale considérée étant elle-même issue d'une population panmictique, les génotypes des individus qui la composent sont dans les proportions de Hardy-Weinberg (p_t^2, $2p_tq_t$, q_t^2) ; seule la population des *reproducteurs* en dévie.

Posons

$$\mathrm{T} = p_t^2 u + 2p_t q_t v + q_t^2 w;$$

c'est la valeur sélective moyenne de la population.

La probabilité qu'un individu qui contribue à l'urne gamétique soit AA se calcule ainsi

$$\begin{aligned}\mathbb{P}(\mathrm{AA}|\text{reprod.}) &= \frac{\mathbb{P}(\text{reprod.}|\mathrm{AA})\mathbb{P}(\mathrm{AA})}{\mathbb{P}(\text{reprod.}|\mathrm{AA})\mathbb{P}(\mathrm{AA}) + \mathbb{P}(\text{reprod.}|\mathrm{A}a)\mathbb{P}(\mathrm{A}a) + \mathbb{P}(\text{reprod.}|aa)\mathbb{P}(aa)} \\ &= \frac{u \times p_t^2}{u \times p_t^2 + v \times 2p_t q_t + w \times q_t^2} \\ &= \frac{1}{\mathrm{T}} p_t^2 u\end{aligned}$$

De même on a

$$\begin{aligned}\mathbb{P}(\mathrm{A}a|\text{reprod.}) &= \frac{1}{\mathrm{T}} 2p_t q_t v \\ \mathbb{P}(aa|\text{reprod.}) &= \frac{1}{\mathrm{T}} q_t^2 w\end{aligned}$$

La probabilité qu'un gamète tiré dans l'urne soit A est 1 quand l'individu qui l'a émis est AA, $\frac{1}{2}$ quand il est Aa, et 0 quand il est aa. Finalement, la probabilité qu'un allèle tiré au hasard soit A, qui est aussi la fréquence de A dans la génération suivante, est

$$\boxed{p_{t+1} = \frac{p_t^2 u + p_t q_t v}{p_t^2 u + 2p_t q_t v + q_t^2 w}}$$

6.1.3 Équation d'évolution

Posons

$$f(p) = \frac{p^2 u + p(1-p)v}{p^2 u + 2p(1-p)v + (1-p)^2 w},$$

pour $p \in [0,1]$.

La fréquence de A à la t$^{\mathrm{e}}$ génération s'obtient par la relation $p_t = f(p_{t-1})$. L'évolution de cette fréquence dépend entièrement de f, que nous devons donc étudier. Notons tout d'abord quelques résultats issus de calculs simples.

Les points fixes de f (c'est-à-dire les valeurs de p pour lesquelles $f(p) = p$) sont les états d'équilibre du système. Si p_0 est un tel point fixe, ou aura pour tout t, $p_t = p_0$: la fréquence n'évolue pas au fil du temps. On a les résultats suivants :

- si $w \neq 0$, f admet 0 comme point fixe, et $f'(0) = \frac{v}{w}$;
- de même, si $u \neq 0$, f admet 1 comme point fixe, et $f'(1) = \frac{v}{u}$;
- f peut admettre un troisième point fixe p_e :

$$\boxed{p_e = \frac{w - v}{u - 2v + w},}$$

si cette quantité est entre 0 et 1. La dérivée en ce point fixe est

$$f'(p_e) = 1 - \frac{(v-u)(v-w)}{v^2 - uw}.$$

Au fil des sections suivantes, nous allons considérer un à un tous les cas possibles. Auparavant, éliminons le cas simple où il n'y a pas de sélection : $u = v = w$. Il est facile de vérifier qu'alors $f(p) = p$. La fréquence de A reste constante au cours du temps, comme on s'y attendait.

6.2 Sélection directionnelle

Il s'agit du cas où un des allèles est favorable, par exemple l'allèle A : on a alors $u > v \geq w$ ou $u \geq v > w$. On parle aussi de sélection purificatrice, car l'allèle défavorable tend à disparaître.

6.2.1 Pas de dominance de l'allèle A : $u > v \geq w > 0$

Ici, le génotype AA a une meilleure valeur sélective que Aa, qui lui-même a soit une meilleure valeur sélective que aa (co-dominance), soit une valeur sélective égale à celle de aa (A est récessif).

Dans ce cas, f a deux points fixes, 0 et 1. L'allure du graphe de f est visible figure 6.1, et l'évolution de p_t au fil du temps figure 6.2.

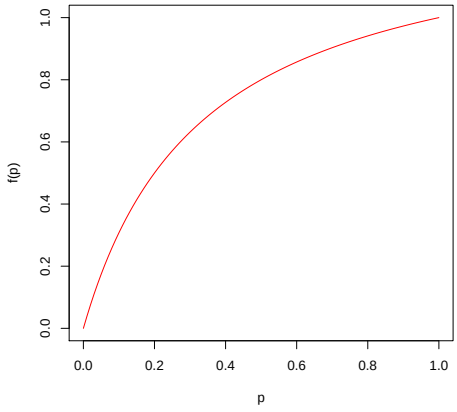


FIGURE 6.1 – Graphe de f

On voit que p_t tend vers 1 et q_t vers 0 : l'allèle favorable A va se fixer dans la population. On peut estimer la vitesse de convergence : passées les premières générations, la suite q_t tend vers 0 à peu près comme $\left(\frac{v}{u}\right)^t$, c'est-à-dire que toutes les demi-périodes

$$\boxed{\mathrm{T} = \frac{\log(2)}{\log(u) - \log(v)} \simeq \frac{0{,}7}{\log(u) - \log(v)}},$$

la fréquence de l'allèle a est divisée par deux.

Le point d'équilibre $p = 0$ est possible, mais il est instable : dès qu'une petite proportion d'allèles A apparaît (par mutation, migration), la suite des p_t va tendre vers 1.

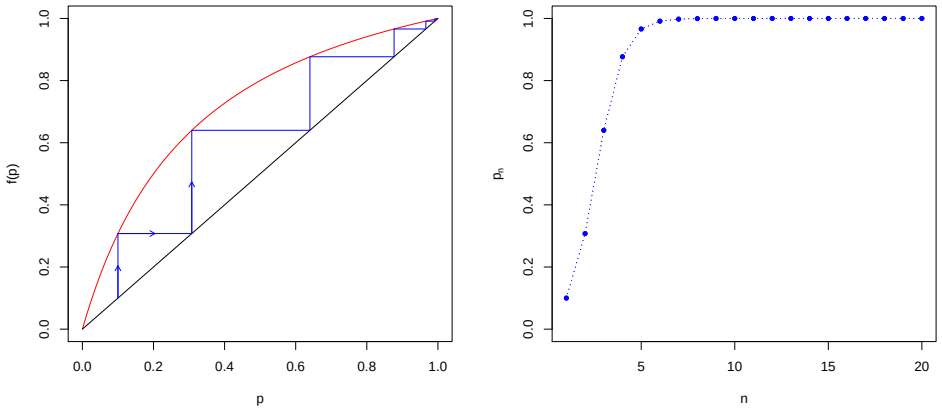


FIGURE 6.2 – Graphe de f et évolution de p_t

Cas où le génotype aa est létal : $u > v$ et $w = 0$

Il s'agit aussi du cas où les individus de génotype aa sont stériles... Ce cas est peu différent du précédent, à part pour l'allure du graphe de f en 0 ; voir figure 6.3.

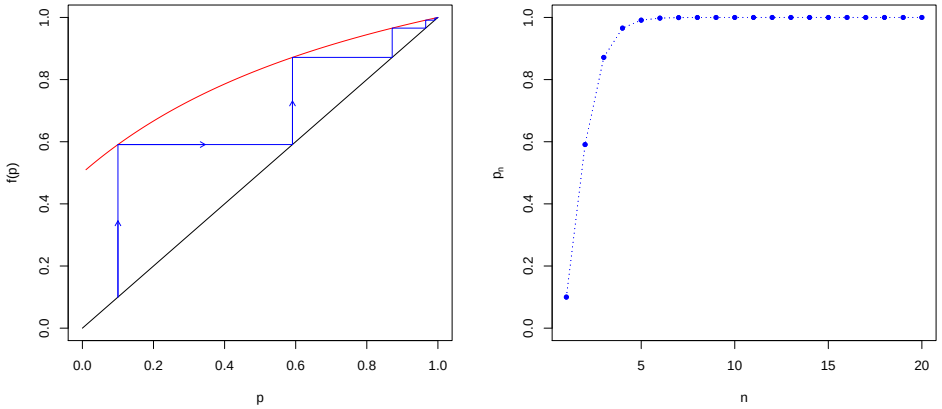


FIGURE 6.3 – Graphe de f et évolution de p_t

6.2.2 Si l'allèle A est dominant : $u = v > w > 0$

Dans ce cas, les génotypes AA et Aa ont la même valeur sélective : A est dominant. Le comportement général est le similaire, comme illustré figure 6.4.

Cependant, comme on le voit sur la figure, du fait qu'on a $f'(1) = 1$, la vitesse à laquelle l'allèle récessif a disparaît est plus lente : la fréquence de l'allèle a, $q_t = 1 - p_t$, se comporte à peu près comme $\frac{u}{(u-w)t}$.

Cas où le génotype aa est létal : $u = v$ et $w = 0$

Ici, on a dominance de A, et aa est létal : c'est par exemple le cas de certaines maladies récessives. Ce cas est donc particulièrement intéressant. Son traitement mathématique est peu différent du précédent, à part pour l'allure du graphe de f en 0 (figure 6.5).

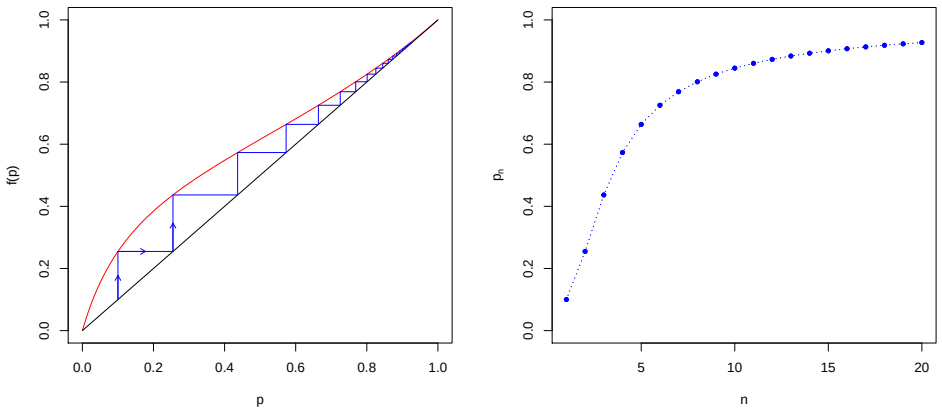


FIGURE 6.4 – Graphe de f et évolution de p_t

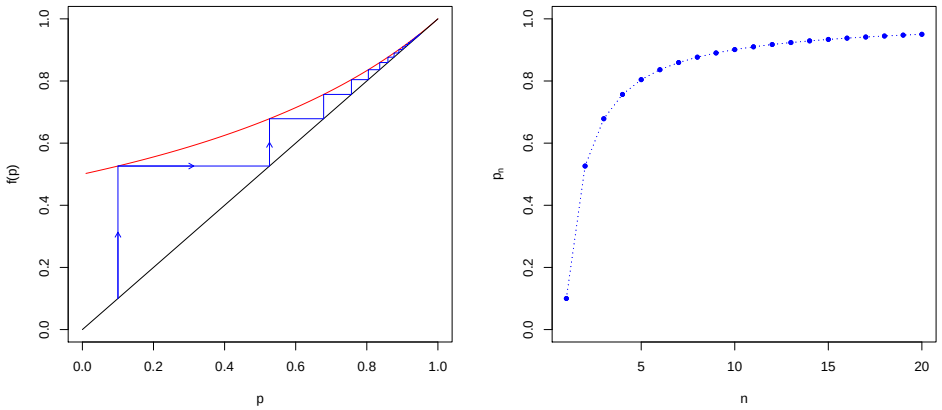


FIGURE 6.5 – Graphe de f et évolution de p_t

Ici la récurrence prend une forme simple :

$$p_{t+1} = \frac{p_t}{p_t(p_t+2q_t)} = \frac{1}{1+q_t},$$

et donc

$$q_{t+1} = 1 - \frac{1}{1+q_t} = \frac{q_t}{1+q_t}.$$

On peut calculer la valeur de p_t et q_t en fonction de n : on a $q_t = \frac{q_0}{1+tq_0}$.

Exercice : Si $q_0 = 0{,}01$, après combien de générations a-t-on $q_t = 0{,}005$?

6.2.3 Si l'allèle A est défavorable : $u < v \leq w$ ou $u \leq v < w$

Il suffit d'échanger les rôles des allèles A et a dans ce qu'on vient de voir. Pour cela, il suffit de permuter u et w. On aura dans tous les cas fixation de l'allèle favorable a, avec la même discussion sur les vitesses d'évolution... La figure 6.6 illustre le cas $0 < u < v < w$.

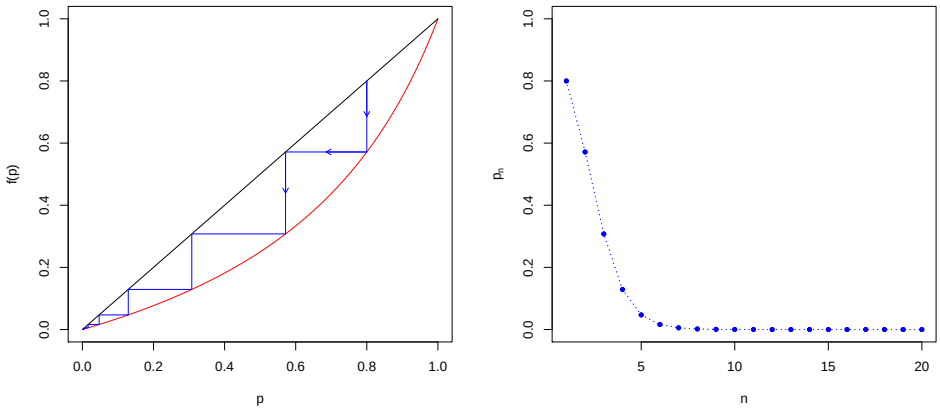


FIGURE 6.6 – Graphe de f et évolution de p_t

.

6.3 Sélection balancée : $u < v > w$

Ici, l'hétéroygote a une meilleur valeur sélective que les deux homozygotes. On parle aussi de sélection stabilisante.

Dans ce cas, on a, en plus des points fixes éventuels en 0 et 1 (selon que v et w sont nuls ou non), un point fixe en

$$p_e = \frac{w-v}{u-2v+w}.$$

La figure 6.7 montre l'évolution de p_t, selon que $p_0 < p_e$, en bleu, ou $p_0 > p_e$, en vert :

Dans tous les cas, p_t tend vers cette valeur d'équilibre p_e.

Notons en passant que la pente de f en p_e est toujours plus petite que 1, et qu'on pourrait là encore calculer une demi-période pour l'évolution des fréquences dès lors qu'on est assez proche de l'équilibre.

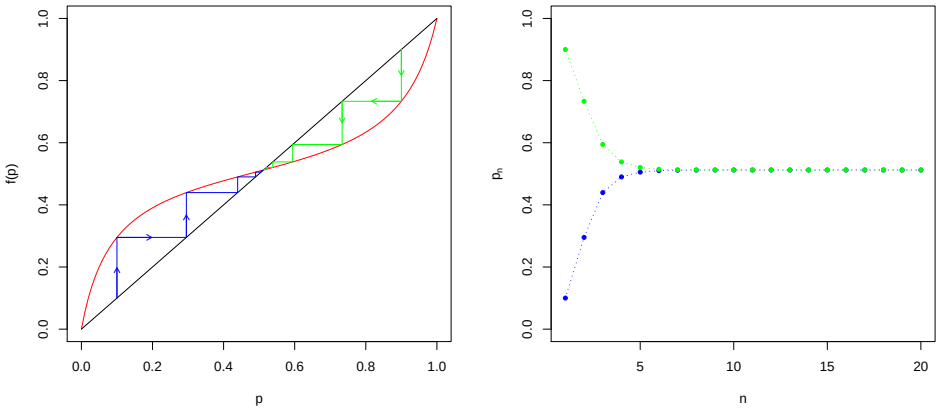


FIGURE 6.7 – Graphe de f et évolution de p_t

L'exemple classique de ce cas est la drépanocytose, ou anémie falciforme. Cette maladie récessive (déjà décrite au premier chapitre) est due à une mutation β^S du gène β de l'hémoglobine, dont l'allèle normal est noté β^A. Les individus $\beta^S\beta^S$, porteurs de deux copies de cette mutation ont une maladie très grave qui diminue considérablement leur espérance de vie. Cependant, les hétérozygotes $\beta^S\beta^A$ ne sont pas touchés avec la même gravité par la maladie. De plus, sont moins vulnérables au paludisme que les homozygotes $\beta^A\beta^A$ – la paludisme est une maladie causée par le plasmodium, un parasite qui colonise les globules rouges.

C'est cet avantage de l'hétérozygote qui explique que l'allèle β^A se maintienne à une fréquence élevée (jusqu'à 20% dans certaines régions).

On a également évoqué un possible avantage de l'hétérozygote pour expliquer la fréquence élevée de la mutation responsable de la mucoviscidose, qui permettrait, selon les auteurs, de mieux résister au choléra, à la tuberculose, etc. Aucune hypothèse ne fait l'unanimité.

6.4 Sélection disruptive : $u > v < w$

Ici, l'hétérozygote est l'individu ayant la plus mauvaise valeur sélective.

On a de nouveau un point d'équilibre

$$p_e = \frac{w - v}{u - 2v + w},$$

mais cette fois il s'agit d'un équilibre répulsif : toute déviation de ce point entraine la fréquence vers 0 ou 1. La figure 6.8 montre l'évolution de p_t, selon que $p_0 < p_e$, en bleu, ou $p_0 > p_e$, en vert.

Cette fois, le processus est divergent ; on parle de sélection diversifiante ou disruptive.

Cela peut conduire des populations séparées, issues d'une même population, à fixer des allèles différents, ou même contribuer à couper une population vivant dans une même aire reproductive en deux sous-populations de phénotypes différents. L'exemple classique est celle des différences de caryotype. Les individus porteurs d'un caryotype équilibré mais « hybride »émettent en effet une certaine proportion de gamètes déséquilibrés.

Certains auteurs ont émis l'hypothèse qu'un tel processus a pu être à l'œuvre dans la population ancestrale des humains (46 chromosomes) et des chimpanzés (48 chromosomes). Cependant à l'heure actuelle, des différences de caryotypes existent par exemple au sein des populations de porcs, de cercopithèques, sans empêcher l'interfécondité des individus.

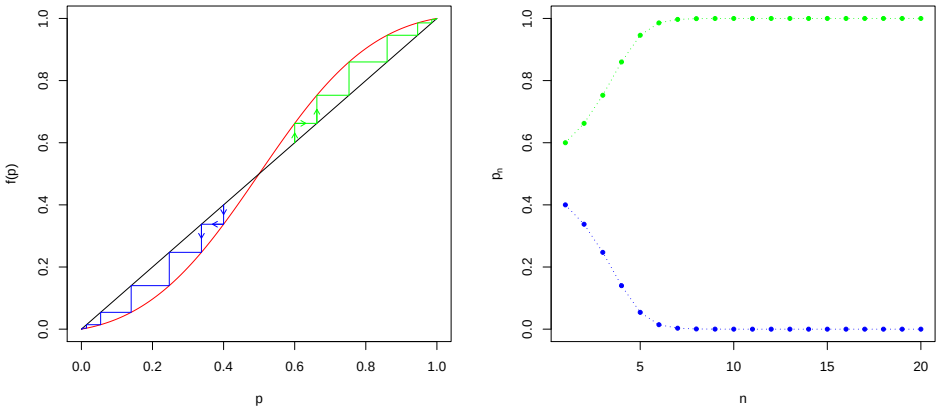


FIGURE 6.8 – Graphe de f et évolution de p_t

6.5 Des modèles plus généraux

Le modèle présenté ici est très rudimentaire. Il est possible de l'enrichir de plusieurs façons.

On pourrait notament envisager des environnements multiples : la valeur sélective d'un génotype peut dépendre de l'environnement. Ceci peut contribuer à maintenir la diversité dans une population qui occupe un espace géographique important.

Un autre phénomène à prendre en compte est « l'avantage des variants rares ». Prenons par exemple le cas d'un papillon dont les ailes peuvent avoir des couleurs différentes ; les prédateurs prennent l'habitude de repérer les papillons qui ont la couleur la plus fréquente, et les papillons qui ont une couleur rare sont avantagés, même si par ailleurs elle n'offre pas de meilleures possibilités de camouflage. La même chose vaut pour la résistance aux agents pathogènes, qui s'adaptent de façon à infester les individus porteur des génotypes les plus fréquents.

Pour tenir compte de ce phénomène, on est conduit à introduire des valeurs sélectives qui dépendent de la valeur de p, ce qui peut faire apparaître de nouveaux points d'équilibres.

7 Mutation et migration

7.1 Mutation

Ici on abandonne une autre des hypothèses du modèle de Hardy-Weinberg : l'absence de mutation. Nous ne considérerons qu'un modèle très simple, à un seul locus. On considère un gène diallélique d'allèles A et a, et on suppose qu'une mutation puisse transformer un allèle en l'autre :

$$\mathrm{A} \underset{\nu}{\overset{\mu}{\rightleftharpoons}} a$$

μ étant la probabilité qu'un allèle A mute vers a, et ν la probabilité qu'un allèle a mute vers A. La mutation est ici supposée neutre, c'est-à-dire qu'il n'y a pas de sélection naturelle agissant au locus considéré.

On notera p_t la fréquence de l'allèle A à la génération t.

7.1.1 Évolution au fil des générations

L'hypothèse de panmixie restant valide, on peut utiliser le modèle de l'urne gamétique. Quelle est la probabilité qu'un gamète tiré au hasard dans l'urne contenant les gamètes émis par la génération t porte l'allèle A ?

Soit on a tiré un allèle A de la génération parentale (probabilité p_t) qui n'a pas muté (probabilité $1-\mu$), soit on a tiré un allèle a (probabilité $q_t = 1-p_t$) qui a muté (probabilité ν), et donc, pour finir, la fréquence de l'allèle A dans la génération suivante est

$$p_{t+1} = (1-\mu)p_t + \nu q_t.$$

On trouve le point d'équilibre en résolvant $p_e = (1-\mu)p_e + \nu q_e = (1-\mu)p_e + \nu(1-p_e)$, ce qui mène à $p_e = \frac{\nu}{\mu+\nu}$. On calcule ensuite $p_{t+1} - p_e$:

$$\begin{aligned} p_{t+1} - p_e &= (1-\mu)p_t + \nu q_t - p_e \\ &= (1-\mu)p_t + \nu q_t - \big((1-\mu)p_e + \nu q_e\big) \\ &= (1-\mu)(p_t - p_e) + \nu(q_t - q_e) \\ &= (1-\mu-\nu)(p_t - p_e) \end{aligned}$$

D'où pour finir

$$\boxed{p_t = p_e + (1-\mu-\nu)^t (p_0 - p_e)}$$

On a à nouveau une « demi-période » : ma distance entre p_t et p_e est multipliée par $(1-\mu-\nu)$ à chaque génération, c'est-à-dire qu'elle est divisée par 2 toutes les $\frac{-\log(2)}{\log(1-\mu-\nu)} \simeq \frac{0{,}7}{\mu+\nu}$ générations ; en effet, on a

$$(1-\mu-\nu)^x = \frac{1}{2}$$

pour

$$x\log(1-\mu-\nu) = -\log(2),$$

et quand $\mu+\nu$ est petit (ce qui est généralement le cas), on a $\log(1-\mu-\nu) \simeq -(\mu+\nu)$. La demi-période est alors très longue.

7.2 Migration

Nous n'envisagerons que le modèle à une île : une petite population (sur l'île) reçoit, via un processus de migration, un flux de gènes d'une très grande population (sur le continent). La très grande population est supposée à l'équilibre de Hardy-Weinberg.

On considère donc un locus di-allélique A/a, et on note p_t et q_t leur fréquence à la génération t. On note p_c et q_c leur fréquence sur le continent, et m le taux de migration, c'est-à-dire la probabilité qu'un allèle à une génération donnée vienne du continent ; ou encore, la proportion d'allèles continentaux dans l'urne gamétique à chaque génération.

Un allèle de la génération $t+1$ peut venir de l'île (probabilité $1-m$), il est alors A avec probabilité p_t ; il peut venir du continent (probabilité m) il est alors A avec probabilité p_c. On a

$$p_{t+1} = (1-m)p_t + mp_c,$$

d'où

$$p_{t+1} - p_c = (1-m)(p_t - p_c),$$

et pour finir

$$\boxed{p_t = p_c + (1-m)^t(p_0 - p_c)}$$

À l'équilibre, les fréquences alléliques sont devenues égales à celles du continent ; la vitesse du processus est donnée par le terme $(1-m)^t$, d'où une « demi-période » $\mathrm{T} = \frac{\log 2}{m} = \frac{0{,}7}{m}$.

Dans ce modèle, les conséquences de la migration sont comparables à celles de la mutation ; cependant, les taux de migration peuvent être beaucoup plus importants que les taux de mutation.

7.3 Équilibre sélection-mutation

La sélection tend à faire disparaître les allèles défavorables qui peuvent être créés par des mutations récurrentes. Les deux processus allant en sens inverse l'un de l'autre, leur superposition doit amener à un point d'équilibre.

Nous nous restreindrons au cas d'un allèle a causant une maladie récessive létale : les valeurs sélectives sont $u = v$ et $w = 0$. Nous négligerons également la probabilité d'une mutation a → A qui restaure un allèle fonctionnel. Nous notons μ la probabilité d'une mutation de A à a.

$$\mathrm{A} \xrightarrow{\mu} a$$

On reprend le raisonnement sur l'urne contenant les gamètes émis par la génération t. La conversion d'allèle

A en allèle a conduit à modifier l'équation d'évolution en

$$\begin{aligned} p_{t+1} &= \frac{(1-\mu)p_t^2 + (1-\mu)p_t q_t}{p_t^2 + 2p_t q_t} \\ &= \frac{(1-\mu)p_t(p_t+q_t)}{p_t(p_t+2q_t)} \\ &= \frac{1-\mu}{1+q_t}. \end{aligned}$$

La figure 7.1 récapitule le raisonnement.

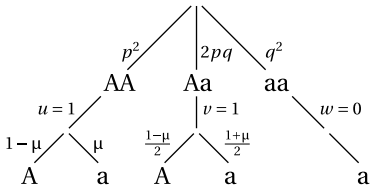


FIGURE 7.1 – Constitution de l'urne gamétique

On se contente ici de chercher le point d'équilibre : notons p_e la fréquence à l'équilibre et $q_e = 1 - p_e$. On a :

$$\begin{aligned} p_e &= \frac{1-\mu}{1+q_e} \\ 1-q_e &= \frac{1-\mu}{1+q_e} \\ 1-q_e^2 &= 1-\mu \\ q_e^2 &= \mu \end{aligned}$$

Ainsi, à l'équilibre, la fréquence de l'allèle morbide est $q_e = \sqrt{\mu}$, et l'incidence de la maladie est égale au taux de mutation : $q_e^2 = \mu$

7.3.1 Cas d'une population consanguine

Considérons à présent le cas d'une population consanguine. Les fréquences génotypiques sont

$$\begin{aligned} f_{\mathrm{AA}} &= (1-\mathrm{F})p^2 + \mathrm{F}p \\ f_{\mathrm{A}a} &= (1-\mathrm{F})2pq \\ f_{aa} &= (1-\mathrm{F})q^2 + \mathrm{F}q. \end{aligned}$$

Ceci conduit à modifier l'équation d'évolution en

$$\begin{aligned} p_{t+1} &= \frac{(1-\mu)((1-\mathrm{F})p_t^2 + \mathrm{F}p_t) + (1-\mu)(1-\mathrm{F})p_t q_t}{(1-\mathrm{F})p_t^2 + \mathrm{F}p_t + (1-\mathrm{F})2p_t q_t} \\ &= \frac{(1-\mu)\,p_t\left((1-\mathrm{F})p_t + \mathrm{F} + (1-\mathrm{F})q_t\right)}{p_t\left((1-\mathrm{F})p_t + \mathrm{F} + (1-\mathrm{F})2q_t\right)} \\ &= \frac{1-\mu}{1+(1-\mathrm{F})q_t}. \end{aligned}$$

À l'équilibre on a

$$\begin{aligned} p_e &= \frac{1-\mu}{1+(1-F)q_e} \\ 1-q_e &= \frac{1-\mu}{1+(1-F)q_e} \\ 1+(1-F)q_e-q_e-(1-F)q_e^2 &= 1-\mu \\ (1-F)q_e^2+Fq_e &= \mu \end{aligned}$$

On a montré que la fréquence des homozygotes aa est égale à μ. Là encore, l'incidence de la maladie est égale au taux de mutation : $f_{aa} = (1-F)q_e^2 + Fq_e = \mu$.

> Si on est à l'équilibre sélection-mutation, l'incidence d'une maladie récessive létale est égal au taux de mutation, indépendamment de la consanguinité de la population.

Il n'en reste pas moins vrai que dans une famille donnée, la probabilité de voir apparaître une maladie récessive est plus importante quand les parents sont apparentés que quand ils ne le sont pas.

7.4 Exercices

Exercice 1 On considère un modèle à deux îles. On note p_t la fréquence de l'allèle A dans la première île à la génération t, et m le taux de migration de la deuxième île vers la première (la proportion d'allèles provenant de l'île 2 dans l'urne gamétique). On note π_t la fréquence de l'allèle A dans la deuxième île à la génération t, et μ le taux de migration de la première île vers la deuxième.

1. Exprimer p_{t+1} en fonction de m, p_t et π_t ; exprimer π_{t+1} en fonction de μ, p_t et π_t.

2. On pose $x_t = p_t - \pi_t$. Exprimer x_{t+1} en fonction de m, μ et x_t.

3. On pose $y_t = \mu p_t + m p_t$. Exprimer y_{t+1} en fonction de m, μ et y_t.

4. Conclure.

8 Dérive génétique

Dans ce chapitre nous examinons les conséquences de l'abandon d'une des hypothèses du modèle de Hardy-Weinberg : la taille infinie de la population.

Les fréquences génotypiques observées à la création d'une nouvelle génération ne sont alors plus égales à leur espérance ; il y a une petite fluctuation, une « fluctuation d'échantillonage », au fil des générations.

Supposons qu'on prenne, dans des conditions contrôlées (en laboratoire), une centaine de populations composée chacune de 16 drosophiles hétérozygotes Aa ; des expériences de ce type ont été menées notament par Buri en 1956, et avant lui par Kerr et Wright en 1954. Les allèles A et a sont choisis codominants, de façon à ce qu'on puisse déterminer les effectifs alléliques à partir des phénotypes [1].

La fréquence de A, notée p, vaut donc $p = 0{,}5$ au début de l'expérience. On les laisse se reproduire de génération en génération, en maintenant la taille égale à 16 individus. Après une vingtaine de générations, dans certaines populations la fréquence de A vaut $p = 0$, dans d'autres $p = 1$, et enfin dans d'autres $p = \frac{1}{32}$, $\frac{5}{32}$, $\frac{21}{32}$, etc.

Ainsi, la fréquence a varié avec le temps, et elle n'a pas varié de la même façon dans toutes les populations ; ceci met en évidence le rôle joué par le hasard.

8.1 Modèle de Wright-Fisher

8.1.1 Le modèle

On suppose toujours que la taille de l'urne gamétique est infinie, et que sa composition reflète fidèlement la composition de la génération qui se reproduit (pas de sélection, pas de distortion de ségrégation) ; on suppose également qu'on a panmixie et pangamie, c'est-à-dire que les gamètes s'apparient au hasard.

La taille de la population sera supposée constante, égale à N. On se restreint au cas d'un locus autosomal di-allélique A/a ; 2N tirages sont réalisés dans l'urne parentale à chaque génération. Appelons X_t le nombre de copies de l'allèle A à la génération t ; X_t est entre 0 et 2N. On peut supposer que $X_0 = x_0$ est une constante connue ; tous les autres X_t ($t > 0$) sont des variables aléatoires. L'urne gamétique contenant les gamètes émis par la génération t contient l'allèle A en proportion $p_t = \frac{1}{2N}X_t$. On notera au besoin $q_t = 1 - p_t$.

Chacun des 2N chromosomes portés par les N individus de la génération $t+1$ est tiré au hasard dans cette urne gamétique : si X_t est fixé, X_{t+1}, le nombre de copies de l'allèle A parmi ces 2N chromosomes, suit donc une loi binomiale :

$$X_{t+1}|X_t \sim \mathcal{B}in\left(2N, p_t = \frac{1}{2N}X_t\right).$$

On parle de la loi de X_{t+1} conditionnellement à X_t. Ceci permet de simuler informatiquement l'évolution d'une telle population (X_0 est supposé connu, on tire X_1 au hasard dans la loi $\mathcal{B}in\left(2N, p_0 = \frac{1}{2N}X_0\right)$, puis X_2 dans la loi $\mathcal{B}in\left(2N, p_1 = \frac{1}{2N}X_1\right)$, etc).

La figure 8.1 montre l'évolution de la fréquence p_t au fil des générations dans des populations de taille 10, 50 et 100. On observe que plus la population est petite, plus la fréquence varie vite, et plus il est probable

1. Buri a choisi les allèles bw1 et bw75 qui déterminent la couleur des yeux et ne sont pas (a priori) sujettes à sélection

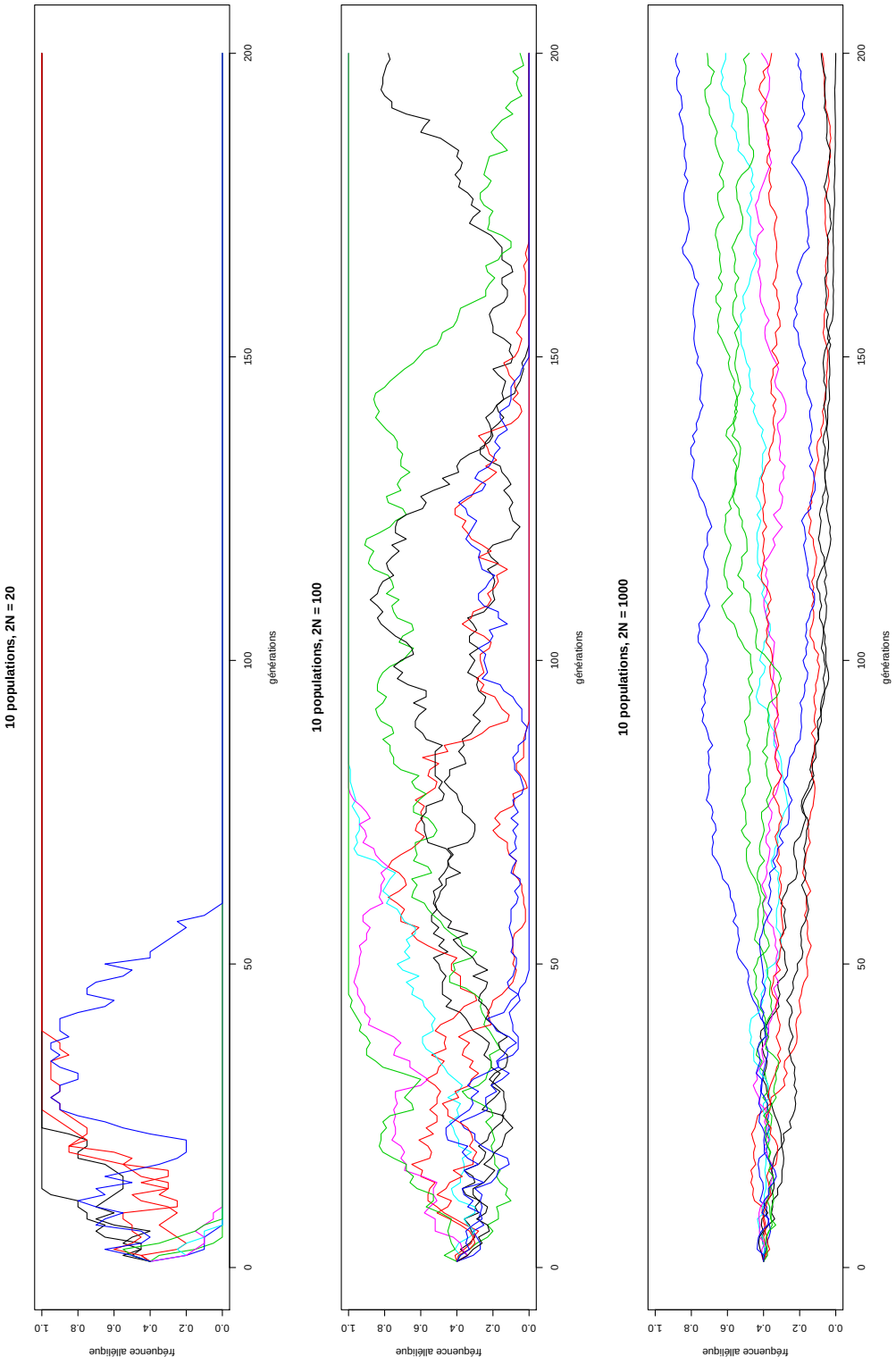


FIGURE 8.1 – 10 trajectoires typiques de la fréquence allélique au fil du temps dans des populations de taille de plus en plus grande (2N = 20, 2N = 100 et 2N = 1000). Plus la population est grande, moins les variations de fréquences sont rapides.

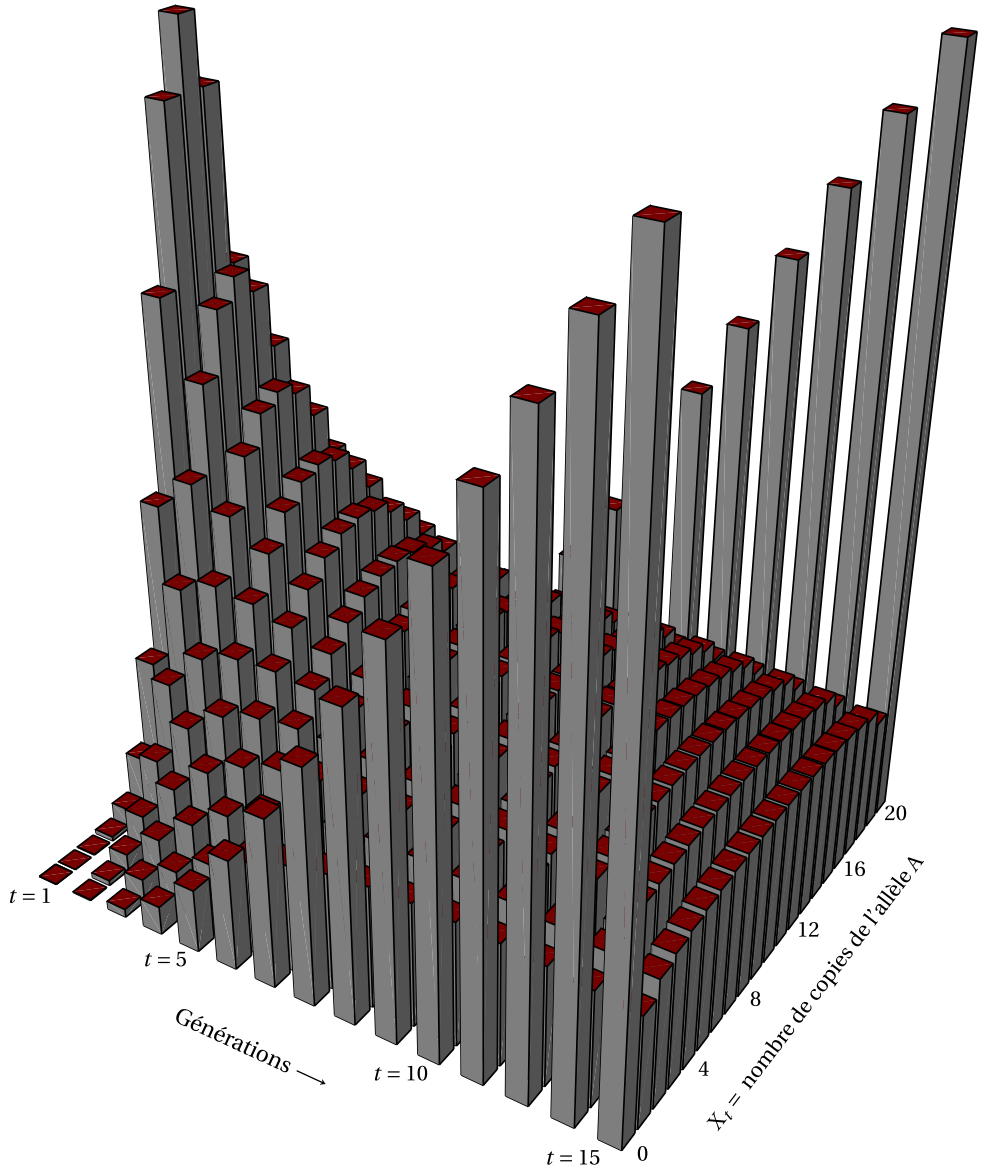


FIGURE 8.2 – Évolution de la probabilité que $X_t = k$ dans une population de 10 individus, avec k entre 0 et 20 et t variant de 1 à 15, et $X_0 = 10$. Pour $t = 1$, on a une distribution binomiale $\mathscr{B}in(20,0{,}5)$; plus t grandit, plus cette distribution s'étale, et on voit très vite croître la probabilité que $X_t = 0$ (disparition de A) et celle que $X_t = 20$ (fixation de A).

qu'après un nombre donné de générations un des allèles ait disparu.

Finissons par une serie de remarques.

Hypothèses implicites du modèle Le modèle suppose une population d'individus diploïdes et hermaphrodites. Pour une population d'individus à sexes séparés, il faudrait distinguer la taille de la population masculine et celle de la population féminine ; cela sera fait plus tard.

Remarque technique On peut expliciter les probabilités conditionnelles $\mathbb{P}(X_{t+1} = \ell | X_t = k)$:

$$\begin{aligned}\mathbb{P}(X_{t+1} = \ell | X_t = k) &= \binom{2N}{\ell} p_t^\ell (1-p_t)^{2N-\ell} \\ &= \binom{2N}{\ell} \left(\frac{k}{2N}\right)^\ell \left(1 - \frac{k}{2N}\right)^{2N-\ell}\end{aligned}$$

En utilisant de façon itérative ces relations (et avec l'aide de la formule des probabilités totales, voir la parenthèse technique ci-dessous) on peut calculer les $\mathbb{P}(X_t = k)$ pour tout t et tout k. Voir la figure 8.2 pour un représentation de ces valeurs pour $N = 10$, k allant de 0 à 20 et t de 1 à 15.

Si on note $T = [t_{k\ell}] \in \mathbb{R}^{(2N+1)\times(2N+1)}$ la matrice des probabilités de transition

$$t_{k\ell} = \mathbb{P}(X_{t+1} = \ell | X_t = k)$$

et $\pi^{(t)} = (\pi_0^{(t)}, \ldots, \pi_{2N+1}^{(t)})$ le vecteur (ligne) des $\pi_k^{(t)} = \mathbb{P}(X_t = k)$, on peut écrire

$$\begin{aligned}\pi_\ell^{(t+1)} = \mathbb{P}(X_{t+1} = \ell) &= \sum_k \mathbb{P}(X_{t+1} = \ell | X_t = k)\mathbb{P}(X_t = k) \\ &= \sum_k \pi_k^{(t)} t_{k\ell}\end{aligned}$$

et donc, en notation matricielle,

$$\pi^{(t+1)} = \pi^{(t)} T.$$

Ceci facilite le calcul (quand on utilise un ordinateur).

États absorbants Si $X_t = 0$, l'allèle A a disparu, et ne peut réapparaître : $\mathbb{P}(X_{t+1} = 0 | X_t = 0) = 1$; on dit alors que l'allèle a s'est fixé (A a disparu). De même, $\mathbb{P}(X_{t+1} = 2N | X_t = 2N) = 1$, et l'allèle A s'est fixé (a a disparu).

8.1.2 Espérance et variance des fréquences alléliques

Les fréquences alléliques p_t sont des variables aléatoires : si on recommence la même expérience, elles prendront des valeurs différentes. Seule la valeur de p_0 est fixée. On peut facilement calculer l'espérance des p_t ; on commence par l'espérance de p_{t+1} conditionnellement à p_t :

$$E(p_{t+1} | p_t) = \frac{1}{2N} E(X_{t+1} | p_t) = \frac{1}{2N} 2N p_t = p_t$$

Et l'espérance de p_{t+1} s'obtient en prenant l'espérance de ce résultat : $E(p_{t+1}) = E(p_t)$. Donc, de proche en proche on a pour tout t

$$\boxed{E(p_t) = p_0}$$

Le calcul de la variance par cette méthode est plus complexe. Il aboutit au résultat suivant :

$$\boxed{\mathrm{var}(p_t) = p_0 q_0 \left(1 - \left(1 - \frac{1}{2N}\right)^t\right) \simeq p_0 q_0 \left(1 - \exp\left(-\frac{t}{2N}\right)\right)}$$

Le calcul est présenté en encadré. La morale de l'histoire est que quand t grandit, var(p_t) s'approche de $p_0 q_0$.

8.1.3 Hétérozygotie moyenne

Il est éclairant de calculer l'hétérozygotie attendue au temps t, $\mathrm{H}_t = \mathrm{E}(2p_t q_t)$.

$$\begin{aligned}\mathrm{H}_t &= \mathrm{E}(2p_t q_t)\\ &= 2\mathrm{E}(p_t - p_t^2)\\ &= 2\mathrm{E}(p_t) - 2\mathrm{E}(p_t^2)\end{aligned}$$

On a $\mathrm{E}(p_t) = p_0$, et $\mathrm{E}(p_t^2) = p_0 - \left(1 - \frac{1}{2\mathrm{N}}\right)^t p_0 q_0$ (cf encadré sur le calcul de la variance), et donc

$$\begin{aligned}\mathrm{H}_t &= 2p_0 q_0 \left(1 - \frac{1}{2\mathrm{N}}\right)^t\\ &\simeq 2p_0 q_0 \exp\left(-\frac{t}{2\mathrm{N}}\right)\end{aligned}$$

Ainsi, l'hétérozygotie tend vers 0. On peut même définir une « demi-période », c'est-à-dire un nombre T de générations après lequel le taux d'hétérozygotie sera (en moyenne) divisé par 2 : il suffit de résoudre l'équation $\exp\left(-\frac{\mathrm{T}}{2\mathrm{N}}\right) = \frac{1}{2}$ dont la solution est $\mathrm{T} = 2\mathrm{N}\log(2) \simeq 2\mathrm{N} \times 0{,}7$.

On en déduit qu'après un temps assez long, il y a toujours fixation d'un des allèles. On a $\mathrm{E}(p_t) = p_0$, et après fixation on a $p_t = 0$ ou $p_t = 1$; donc la valeur 1 (fixation de l'allèle A) est prise avec probabilité p_0 et la valeur 0 (fixation de l'allèle a) avec probabilité $1 - p_0 = q_0$.

> Après un temps assez long, il y a toujours fixation d'un des deux allèles ; cela sera l'allèle A avec probabilité p_0 et l'allèle a avec probabilité q_0.

8.1.4 Temps de persistence

On peut montrer (nous ne le ferons pas) que le temps de persistence T, c'est-à-dire le temps pendant lequel le polymorphisme persiste avant fixation d'un allèle ou l'autre a pour espérance (en nombre de générations)

$$\mathrm{E}(\mathrm{T}) = -4\mathrm{N}\left(p_0 \log p_0 + q_0 \log q_0\right)$$

Cette expression est représentée figure 8.3. La figure 8.4 montre la distribution des temps de persistence, qui est très étalée, et très asymétrique pour les petites valeurs de p_0.

Calcul de la variance de p_t

Calculons d'abord l'espérance de p_{t+1} conditionnellement à la valeur de p_t.

$$\begin{aligned}
\mathrm{E}\left(p_{t+1}^2|p_t\right) &= \mathrm{var}(p_{t+1}|p_t) + \mathrm{E}(p_{t+1}|p_t)^2 \\
&= \left(\frac{1}{2\mathrm{N}}\right)^2 \mathrm{var}(\mathrm{X}_{t+1}|p_t) + (p_t)^2 \\
&= \frac{1}{2\mathrm{N}} p_t(1-p_t) + p_t^2 \\
&= \frac{1}{2\mathrm{N}} p_t + \left(1 - \frac{1}{2\mathrm{N}}\right) p_t^2
\end{aligned}$$

L'espérance de p_{t+1} s'obtient en prenant l'espérance du résultat ci-dessus :

$$\begin{aligned}
\mathrm{E}\left(p_{t+1}^2\right) &= \frac{1}{2\mathrm{N}}\mathrm{E}(p_t) + \left(1 - \frac{1}{2\mathrm{N}}\right)\mathrm{E}\left(p_t^2\right) \\
&= \frac{1}{2\mathrm{N}} p_0 + \left(1 - \frac{1}{2\mathrm{N}}\right)\mathrm{E}\left(p_t^2\right).
\end{aligned}$$

On a donc

$$\mathrm{E}\left(p_{t+1}^2\right) - p_0 = \left(1 - \frac{1}{2\mathrm{N}}\right)\left(\mathrm{E}\left(p_t^2\right) - p_0\right)$$

Et de proche en proche :

$$\mathrm{E}\left(p_t^2\right) - p_0 = \left(1 - \frac{1}{2\mathrm{N}}\right)^t \left(\mathrm{E}\left(p_0^2\right) - p_0\right)$$

Enfin, p_0 est une constante, donc on a $\mathrm{E}\left(p_0^2\right) = p_0^2$, d'où $\mathrm{E}\left(p_0^2\right) - p_0 = -p_0 q_0$, et

$$\mathrm{E}\left(p_t^2\right) = p_0 - \left(1 - \frac{1}{2\mathrm{N}}\right)^t p_0 q_0.$$

Calculons maintenant $\mathrm{var}(p_t)$:

$$\begin{aligned}
\mathrm{var}(p_t) &= \mathrm{E}\left(p_t^2\right) - \mathrm{E}(p_t)^2 \\
&= p_0 - \left(1 - \frac{1}{2\mathrm{N}}\right)^t p_0 q_0 - p_0^2 \\
&= p_0 q_0 - \left(1 - \frac{1}{2\mathrm{N}}\right)^t p_0 q_0
\end{aligned}$$

et pour finir

$$\mathrm{var}(p_t) = p_0 q_0 \left(1 - \left(1 - \frac{1}{2\mathrm{N}}\right)^t\right).$$

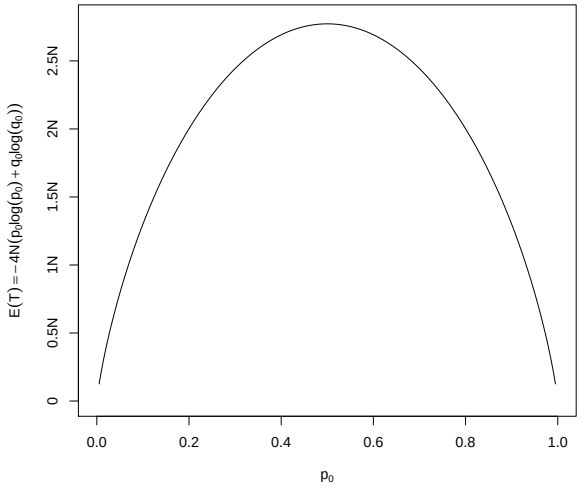


FIGURE 8.3 – L'espérance du temps de persistence en fonction de p_0

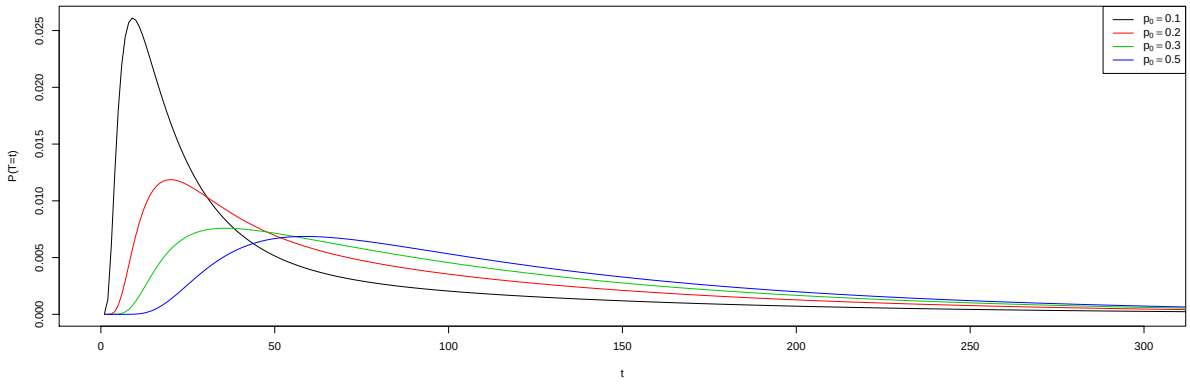


FIGURE 8.4 – Distribution du temps de persistence (N = 50, p_0 = 0,1 à 0,5)

8.2 Augmentation récurrence de la consanguinité

On considère ici une population diploïde hermaphrodite, avec possible autogamie.

Soit F_t la probabilité que les deux allèles portés par un individu de la génération t (en un locus donné) proviennent d'un ancêtre commun (soient IBD, pour *Identical By Descent*) ; on dira que cet individu est homozygote par origine, ou par « descendance », (HBD, pour *Homozygous By Descent*). Nous allons calculer F_{t+1} en fonction de F_t.

Notons que F_t est une probabilité, donc une quantité déterministe. On peut également définir A_t, la proportion d'individus HBD à la génération t : c'est une variable aléatoire dont l'espérance est F_t.

Il y a deux façons pertinentes de définir F_0 :

1. $F_0 = 0$. On considère que personne n'est HBD dans la population à $t = 0$, et on s'intéresse à la perte de diversité dans les générations qui suivent.
2. $F_0 = 1 - 2p_0q_0$. On considère que tous les allèles A ont un ancêtre commun, de même que tous les allèles a ; dans ce cas seuls les hétérozygotes ne sont pas HBD. Le statut HBD se confond avec le statut d'homozygote, et F_t est égal à $1 - H_t$.

Considérons donc les deux allèles portés par un individu de la génération $t+1$. Avec probabilité $\frac{1}{2N}$, ces deux copies sont issues d'une seule des 2N copies présentes à la génération t (auquel cas il y a eu autogamie) ; et avec probabilité $1 - \frac{1}{2N}$, ces deux copies sont issues de deux copies distinctes à la génération précédente, qui sont IBD avec probabilité F_t. Donc

$$F_{t+1} = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) F_t \tag{8.1}$$

On pourrait raisonner plus rigoureusement sur A_t : l'espérance de A_{t+1}, sachant A_t, est $\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) A_t$, et en prenant l'espérance de cette expression on retrouve l'équation ci-dessus.

On trouve la limite de F_t en résolvant l'équation $x = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) x$, qui a pour solution $x = 1$; on calcule ensuite $F_{t+1} - 1$ en fonction de $F_t - 1$:

$$F_{t+1} - 1 = \left(1 - \frac{1}{2N}\right)(F_t - 1)$$

d'où

$$F_t - 1 = \left(1 - \frac{1}{2N}\right)^t (F_0 - 1)$$

et donc si on a choisi $F_0 = 0$,

$$\boxed{\begin{aligned} F_t &= 1 - \left(1 - \frac{1}{2N}\right)^t \\ F_t &\simeq 1 - \exp\left(-\frac{t}{2N}\right) \end{aligned}}$$

et si $F_0 = 1 - 2p_0q_0$,

$$F_t = 1 - 2p_0q_0\left(1 - \frac{1}{2N}\right)^t$$

et on retrouve $H_t = 1 - F_t = 2p_0q_0\left(1 - \frac{1}{2N}\right)^t$. Notons que ceci peut permettre de retrouver la valeur de $\mathrm{var}(p_t)$.

Attention, les F_t calculés par ces formules sont des valeurs moyennes ; la proportion d'allèles IBD dans une population à une génération donnée varie autour de cette valeur, selon les fluctuations d'échantillonage. La figure 8.6 illustre ces variations.

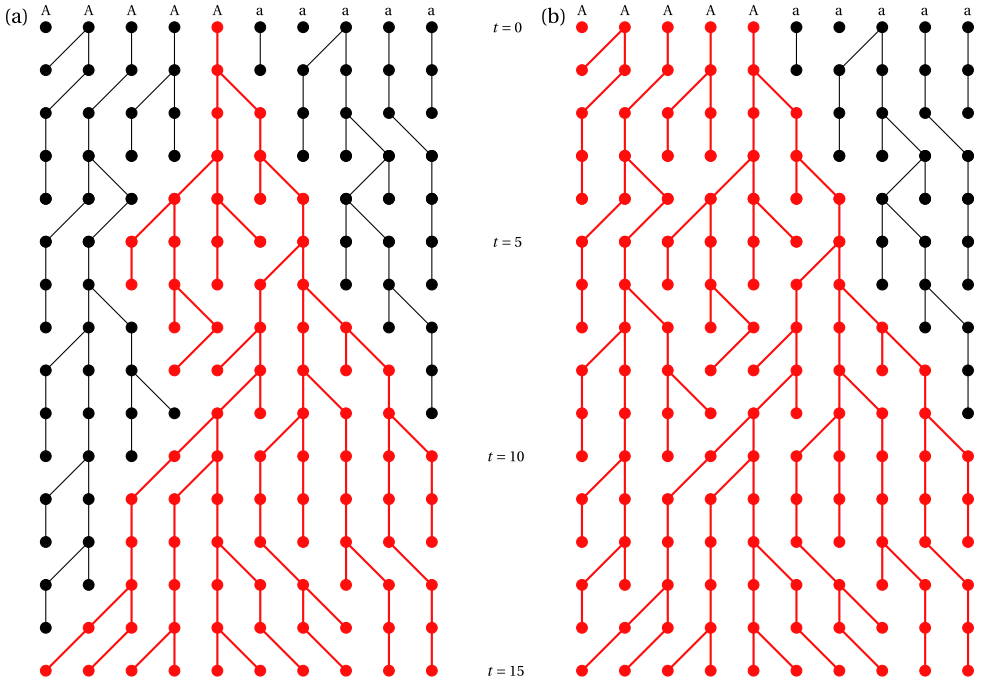


FIGURE 8.5 – On ne représente que les 2N = 10 allèles, de génération en génération. Les points représentent les allèles, et les segments la transmission d'un allèle d'une génération à l'autre ; certains allèles n'ont pas de descendant tandis que d'autres en on un ou deux. À gauche, à la génération 15, tous les allèles présents descendent du même ancêtre (ils sont IBD). À droite, si on suppose qu'à la génération 0 tous les allèles A ont un ancêtre commun, alors dès la génération 10 tous les allèles présents sont IBD.

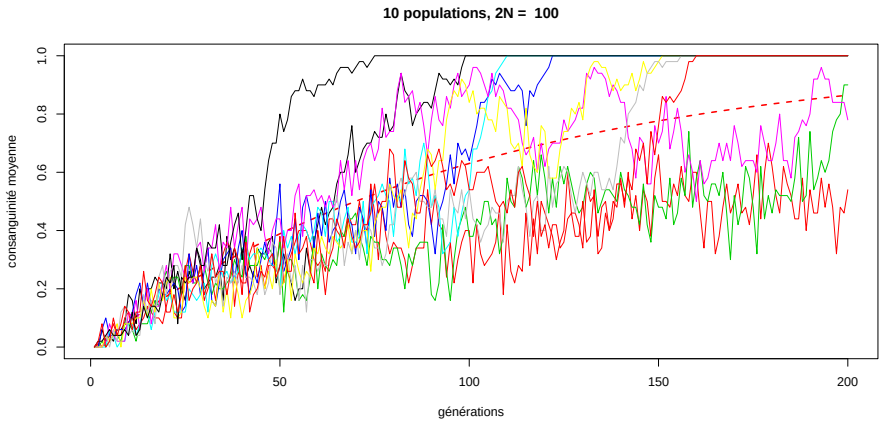


FIGURE 8.6 – Proportion d'individus portant deux allèles IBD dans 10 populations de taille N = 50

Contrairement à ce qui se passe dans une population s'écartant du régime panmictique, on ne doit pas constater d'écart important à l'équilibre de Hardy-Weinberg dans une petite population panmcitique où la valeur de F_t serait proche de 1 ; le déficit en hétérozygotes est par rapport aux fréquences dans la population d'origine, au temps $t = 0$, mais il s'accompagne d'une modification des fréquences alléliques. Quand $F_t = 1$, un des allèles est fixé, on n'observe plus d'hétérozygotes, mais il n'y a également plus qu'une sorte d'homozygotes.
Il est possible d'interpréter F_t comme un F_{ST} de structure de population : si on divise une population homogène en un grand nombre de (petites) sous-populations qu'on isole les unes des autres, et qu'on les laisse se reproduire pendant plusieurs générations, alors F_t est le F_{ST} attendu pour la population totale.

8.3 Taille efficace

Nous avons jusqu'à présent restreint la discussion à une population idéale, formée d'individus hermaphrodites avec possibilité d'autogamie. On n'excluait pas qu'un individu puisse être formé par la fusion de deux gamètes provenant d'un même individu, ce qui permet une modélisation simple du tirage dans l'urne gamétique.
Cependant le prix de cette simplicité est que la vitesse à laquelle agit la dérive génétique dans une population qui ne vérifie pas ces hypothèses idéales est différente de celle prédite par nos calculs. On peut cependant définir la *taille efficace* d'une population, qui est la taille d'une population idéale qui évoluerait à la même vitesse. Généralement, la taille efficace est plus petite que la taille réelle.
On se concentrera pour le calcule de la taille efficace sur l'évolution de F_t au cours du temps,

8.3.1 Sexes séparés

Nous allons ici considérer une population avec N_m garçons et N_f filles, et établir une équation analogue à l'équation 8.1. À la génération t, on note φ_t l'apparentement moyen entre deux individus, et F_t la probabilité qu'un individu soit HBD (sa consanguinité). On a $F_t = \varphi_{t-1}$.

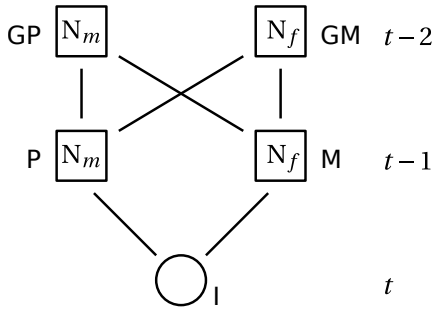


FIGURE 8.7 – Flux des gènes dans une population à sexes séparés

Les grands-parents sont considérés comme pris au hasard dans la population à $t-2$. On n'exclut donc aucune forme d'union entre apparentés ! Les deux allèles présents chez l'individu I peuvent provenir :

— de ses deux grands-pères avec probabilité $\frac{1}{4}$; ces deux grands-pères n'en sont qu'un avec probabilité $\frac{1}{N_m}$; quand c'est le cas, la probabilité que I soit HBD est $\frac{1}{2}(1+F_{t-2})$ (le grand-père peuvant être HBD avec probabilité F_{t-2}) ;
— d'une même grand-mère avec probabilité $\frac{1}{4} \times \frac{1}{N_f}$, auquel cas I est HBD avec probabilité $\frac{1}{2}(1+F_{t-2})$;

— de grands-parents différents avec probabilité $1-\frac{1}{4N_m}-\frac{1}{4N_f}$, auquel cas I est HBD avec probabilité $\varphi_{t-2} = F_{t-1}$.

On a donc

$$F_t = \left(\frac{1}{8N_m}\right)(1+F_{t-2}) + \left(\frac{1}{8N_f}\right)(1+F_{t-2}) + \left(1-\frac{1}{4N_m}-\frac{1}{4N_f}\right)F_{t-1}$$

En réorganisant les termes, on a

$$\begin{aligned} F_t &= \left(\frac{1}{8N_m}+\frac{1}{8N_f}\right)(1+F_{t-2}) + \left(1-\left(\frac{1}{4N_m}+\frac{1}{4N_f}\right)\right)F_{t-1} \\ &= \left(\frac{1}{2N_e}\right)(1+F_{t-2}) + \left(1-\frac{1}{N_e}\right)F_{t-1}, \end{aligned}$$

où on a défini la taille efficace de la population N_e comme la moyenne harmonique de $2N_m$ et $2N_f$, c'est-à-dire que N_e vérifie l'égalité suivante :

$$\boxed{\frac{1}{N_e} = \frac{1}{2}\left(\frac{1}{2N_m}+\frac{1}{2N_f}\right)}$$

On peut montrer qu'on a alors

$$F_t \simeq 1 - \exp\left(-\frac{t}{2N_e}\right)$$

Ainsi, la taille efficace (ou effectif efficace) de la population est la taille d'une population « idéale » d'individus hermaphrodites dans laquelle la dérive est de même ampleur que celle observée dans la population réelle.

La table 8.1 montre quelques valeurs de N_e pour N_m allant de 1 à 100, et N_f égal à 200. Les petites valeurs de N_m sont pertinentes pour certains animaux d'élevage.

N_m	N_f	N_e
1	100	3,96
5	100	19
50	100	133
100	100	200,00

TABLE 8.1 – Taille efficace pour différentes valeurs de N_m et N_f

8.3.2 Taille variable

On considère une population à effectifs variables, $N_0, N_1, \ldots, N_t$. On considère ici à nouveau le modèle des individus hermaphrodites. En reprenant le raisonnement, on a

$$F_t - 1 = \left(1-\frac{1}{2N_{t-1}}\right)\cdots\left(1-\frac{1}{2N_1}\right)\left(1-\frac{1}{2N_0}\right)(F_0-1)$$

Pour retrouver une équation du type $F_t - 1 = \left(1-\frac{1}{2N_e}\right)^t (F_0-1)$ il faut définir N_e tel que

$$\left(1-\frac{1}{2N_e}\right)^t = \left(1-\frac{1}{2N_{t-1}}\right)\cdots\left(1-\frac{1}{2N_1}\right)\left(1-\frac{1}{2N_0}\right)$$

En passant au logarithme et en utilisant l'approximation $\log\left(1-\frac{1}{2\mathrm{N}}\right) \simeq -\frac{1}{2\mathrm{N}}$, on obtient

$$\boxed{\frac{1}{\mathrm{N}_e} = \frac{1}{t}\left(\frac{1}{\mathrm{N}_{t-1}} + \cdots + \frac{1}{\mathrm{N}_1} + \frac{1}{\mathrm{N}_0}\right)} \tag{8.2}$$

et donc N_e est la moyenne harmonique des effectifs.

On pourra prendre comme exemple le cas d'une population dont l'effectif est réduit à un petit nombre d'individus pendant quelques générations, à la suite par exemple d'une famine (on parle de goulot d'étranglement, ou de *bottleneck*). On vérifiera que dans la somme ci-dessus, les termes $\frac{1}{\mathrm{N}_k}$ où N_k est petit ont une grande influence, et entraînent une taille efficace réduite même longtemps après l'épisode de réduction de la population.

Les effets fondateurs (populations fondées par un petit nombre d'individus) sont également un exemple de ce type.

8.3.3 Autres causes de variation de l'effectif efficace

Parmi les causes de variation de l'effectif efficace, il y a l'évitement de l'inceste ou plus généralement des unions entre apparentés ; et aussi, les écarts observés à l'hypothèse que chaque individu a une chance égale d'avoir des enfants survivants à la génération suivante, ce qui se traduit par le fait que le nombre d'enfant par individu suit une loi de Poisson. Tous ces facteurs se traduisent par un effectif efficace plus petit que la taille de la population.

Citons encore le cas des locus situés sur le chromosome X, pour lesquels l'effectif efficace est

$$\mathrm{N}_e = \frac{9\mathrm{N}_m\mathrm{N}_f}{4\mathrm{N}_m + 2\mathrm{N}_f}.$$

Enfin, si la population est divisée en sous-populations entre lesquelles subsistent des échanges génétiques, on peut également définir un effectif efficace, qui sera *plus grand* que l'effectif total de la population ; en effet, l'isolement relatif des sous-populations rend plus difficile la disparition d'un allèle, car la dérive peut aller dans des directions différentes selon les sous-populations.

8.4 Coalescence

Le modèle de Wright-Fisher est une façon de penser les petites populations « en allant de l'avant », en modélisant la façon dont une génération engendre la suivante. La théorie de la coalescence prend le parti opposé de « remonter le temps », en s'intéressant aux ancêtres des allèles pris dans une génération donnée, et en supposant incidemment la population très ancienne (et même infiniment ancienne). Dans une telle population, tous les allèles sont IBD : en remontant le temps suffisament longtemps, on finit toujours par tomber un ancêtre commun à deux allèles. Le seul moyen de conserver de la diversité est de considérer des modèles avec mutation (ce que nous ne ferons pas ici).

8.4.1 Coalescence de deux allèles

La première question posée par la théorie de la coalescence est la suivante :

> Étant donnés deux allèles pris au hasard dans une population, combien de générations faut-il remonter pour que ces deux allèles aient un ancêtre commun ?

La probabilité que deux allèles pris au hasard dans une population de taille constante égale à N (et donc comptant 2N allèles) aient le même ancêtre à la génération précédente est $\frac{1}{2N}$ (une fois l'ancêtre du premier allèle choisi, il y a une chance sur 2N de choisir le même ancêtre) ; et la probabilité qu'elles aient des ancêtres distincts à la génération précédente est donc $1-\frac{1}{2N}$. On dit alors qu'il y a eu (ou qu'il n'y pas eu) un événement de coalescence.

Calculons la probabilité que 2 allèles aient un ancêtre commun exactement $t+1$ générations plus tôt : la probabilité qu'il n'y ait pas eu d'événement de coalescence pendant t générations est $\left(1-\frac{1}{2N}\right)^t$, et qu'il y ait eu coalescence dans la génération immédiatement précédente est $\frac{1}{2N}$. La probabilité que le temps de coalescence T_2 soit égal à $t+1$ est donc

$$\boxed{\mathbb{P}(T_2 = t+1) = \frac{1}{2N}\left(1-\frac{1}{2N}\right)^t}$$

On reconnaît une loi géométrique de paramètre $p = \frac{1}{2N}$. Son espérance est $E(T_2) = 2N$ et sa variance $\text{var}(T_2) = 4N^2 - 2N \simeq 4N^2 = E(T_2)^2$. On peut également faire l'approximation par une loi exponentielle, $\mathbb{P}(T_2 = t+1) \simeq \frac{1}{2N}\exp\left(\frac{-t}{2N}\right)$.

8.4.2 Premier événement de coalescence pour k allèles

On peut généraliser la question précédente, comme suit :

> Étant donnés k allèles pris au hasard dans une population de taille 2N, combien de générations faut-il remonter pour que ces k allèles n'aient plus que $k-1$ ancêtres ?

On montrer, en raisonnant sur des bases similaires et au prix de quelques approximations que si on nomme T_k ce temps, on a

$$\boxed{\begin{aligned} E(T_k) &\simeq \frac{4N}{k(k-1)} \\ \text{var}(T_k) &\simeq E(T_k)^2. \end{aligned}}$$

8.4.3 Plus récent ancêtre commun

On peut enfin répondre à la question

> Étant donnés k allèles pris au hasard dans une population de taille 2N constante, combien de générations faut-il remonter pour que ces k allèles n'aient plus qu'un ancêtre ?

C'est $T = T_k + T_{k-1} + \cdots + T_2$, dont l'espérance vaut

$$\boxed{E(T) \simeq 4N\left(1-\frac{1}{k}\right)}$$

Il faut donc en espérance environ 4N générations avant que tous les allèles aient un ancêtre unique, ce qui est étonnamment peu ; et la moitié de ce temps est dû à T_2, le temps pour faire coalescer « les deux derniers ancêtres » (cf figure 8.8).

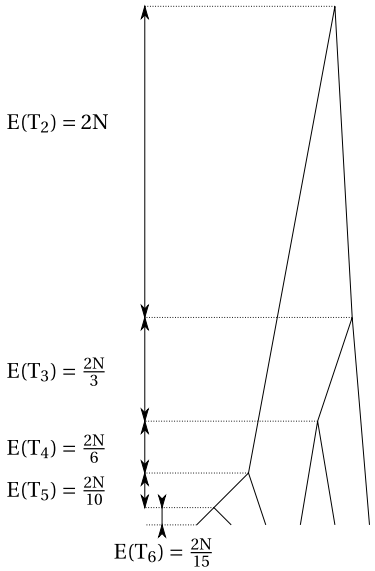


FIGURE 8.8 – Temps de coalescence pour 6 allèles

8.5 Implications en l'épidémiologie génétique

8.5.1 Bottleneck et effet fondateur

Une conséquence de la formule 8.2 est qu'il suffit de quelques générations où la population a eu un petit effectif pour que l'effectif efficace soit considérablement diminué. C'est le cas en particulier quand la population passe par un « goulot d'étranglement démographique » ou *bottleneck* (cas particuliers de population animale ayant frôlé l'extinction ou de populations humaines isolées passant par des phases de disette, épidémies, etc). Un autre cas est celui de l'effet fondateur, quand une petit population s'établit dans un nouvel habitat et y reste isolée (cas de la population québecoise, des Amish, des Huttérites, etc).

Cela peut avoir pour conséquence que la dérive favorise des allèles morbides, et que certaines maladies génétiques soient beaucoup plus fréquentes dans ces populations que dans la population humaine générale ; c'est le cas par exemple de la rétinite pigmentaire à Tristan Da Cunha, ou des mutations BRCA1 et BRCA2 chez les juifs ashkénazes.

8.5.2 Déséquilibre gamétique

De même que les fréquences alléliques changent sous l'effet de la dérive, les fréquences gamétiques changent : du déséquilibre gamétique peut être créé entre des locus en équilibre gamétique. Si ces locus sont liés, ce déséquilibre persistera au fil des générations.

En conséquence, dans les petites populations, ou dans les populations qui sont passées par un effet fondateur ou un goulot d'étranglement, les motifs du déséquilibre liaison tendent à être différents de ceux observés dans la population totale ; en particulier ils s'étendent sur une distance génétique plus importante. Ceci, ainsi que la moindre diversité allélique dans ces populations, peut augmenter la puissance statistique de la recherche des mutations (Terwilliger 1998, Drift mapping).

8.5.3 Polymorphismes conservés d'une espèce à l'autre

La morale de la théorie de la coalescence est que le temps de coalescence de tous allèles présents dans la population est très inférieur au temps de spéciation ; autrement dit, tous les polymorphismes qu'on observe dans le génome humain ont pour origine une mutation survenue chez un de nos ancêtres *H. sapiens* . On ne devrait pas retrouver les mêmes chez d'autres mammifères, même chez nos proches parents chimpanzés !

C'est cependant le cas pour certains polymorphismes (les groupes sanguins A et B, certains polymorphismes de la région HLA). Cela implique l'existence d'un mécanisme qui maintient ces polymorphismes sur un temps très long ; probablement de la sélection balancée.