SCIENCES SUP

Cours et exercices corrigés

Licence • PCEM • CAPES

GÉNÉTIQUE DES POPULATIONS

Jean-Louis Serre

DUNOD

GÉNÉTIQUE DES POPULATIONS

Consultez nos catalogues sur le Web

www.dunod.com

GÉNÉTIQUE DES POPULATIONS

Cours et exercices corrigés

Jean-Louis Serre
Professeur à l'université de Versailles-Saint-Quentin

DUNOD

Illustration : © Bios
N'Diaye Jean-Claude

ISBN 2 10 049620 4

Table des matières

Avant-propos

Une discipline scientifique se trouve toujours à l'intersection ou à la marge d'autres champs de la science et ce principe vaut particulièrement pour la génétique des populations.

D'un point de vue historique et épistémologique, la naissance, entre 1908 et 1918, de la génétique des populations apparaît comme le moment de la réconciliation entre la vision darwinienne de l'évolution et la vision mendélienne de l'hérédité, toute nouvelle à cette époque (voir Introduction).

D'un point de vue scientifique et méthodologique, la génétique des populations perpétue, à la suite de la biométrie fondée par les darwiniens, l'ouverture de la biologie à la modélisation mathématique ; elle a permis d'ouvrir de nouveaux champs théoriques et appliqués grâce à ses modèles de mesure et d'évolution de la diversité génétique :

– le renouvellement de la réflexion sur la théorie de l'évolution dont elle constitue le cœur mathématique obligatoire ;
– la biodiversité : analyse et gestion de la diversité génétique des populations et des espèces, par le recensement des espèces, la caractérisation génétique des espèces menacées, la recherche de souches sauvages des espèces domestiques et leur préservation comme source de diversité génétique, la gestion et la réintroduction d'espèces disparues ou menacées par des espèces proches, le développement de conservatoires et de banques de graines ;
– l'écologie : identification et analyse des liens interspécifique et de la base génétique de leur co-évolution, par exemple la relation hôte-parasite déterminante dans la perspective de la lutte biologique ;
– la génétique quantitative, c'est-à-dire l'analyse génétique des caractères dont la variabilité est continue, et l'analyse des interactions génome-milieu ;
– la génétique humaine et épidémiologique, où l'impossibilité éthique (et technique) de l'expérimentation exige de passer par des modèles d'analyses statistiques des données, incluant la génétique des populations.

Les domaines théoriques, auxquels la génétique des populations apporte son concours, présentent en même temps des enjeux d'application d'une grande importance, déjà évoqués pour la biodiversité et l'écologie. Le développement de la géné-

tique quantitative a permis toutes les avancées importantes de la sélection variétale végétale ou animale en agronomie et les enjeux d'aujourd'hui sont sans doute de rechercher des variétés dont la culture serait moins agressive pour l'environnement tout en demeurant économiquement rentable (agriculture raisonnée). Les enjeux de l'épidémiologie génétique humaine sont aussi innombrables et importants, qu'il s'agisse du conseil génétique et du diagnostic prénatal associé, de la possibilité de prévention des pathologies par dépistage des couples ou des individus à risque, ou de l'identification des facteurs génétique de risque qui, en interaction avec des facteurs environnementaux, sont impliqués dans les maladies multifactorielles, neurodégénératives (Alzheimer), psychiatriques (autisme, maniaco-dépression ou schizophrénie), ou physiologiques (diabètes, maladies cardio-vasculaires, maladies auto-immunes).

Il est donc logique qu'un enseignement de génétique des populations soit introduit dans tout cursus de biologie qui s'intéresse à l'évolution, la biodiversité, l'amélioration des variétés cultivées et la génétique humaine ou médicale.

La plupart des ouvrages consacrés à la génétique au sens large présentent un chapitre de génétique des populations mais n'ont pas la place de présenter autrement que de manière synthétique et résumée, les principaux modèles de la discipline. Il existe, en français, une présentation complète, détaillée et brillante de la génétique des populations[1], mais précisément en raison de ces qualités, elle est difficile et réservée à des spécialistes. Il existe aussi un ouvrage[2] qui aborde un plus grand nombre d'aspects de la génétique des populations, mais en détaillant moins les démonstrations et les conséquences, tout en proposant de nombreux exercices corrigés. L'ouvrage présent n'entend pas lui être concurrent mais complémentaire, en offrant d'autres qualités dont nous espérons qu'elles séduiront le lecteur.

Cet ouvrage est une reprise d'un ouvrage antérieur[3], actualisée par l'introduction de nouveaux exercices ou une nouvelle rédaction de certains chapitres intégrant des données nouvellement acquises. Sa démarche pédagogique reste la même, il s'agit d'abord de présenter les phénomènes et leur logique interne, de manière intuitive, afin de les faire « sentir » sans recours aux mathématiques, puis dans un deuxième temps d'introduire la formalisation mathématique des phénomènes, sans avoir peur de détailler à l'extrême certaines démonstrations ou certains raisonnements afin de constituer une aide véritable comme complément des cours magistraux. Dans le même esprit didactique, le corrigé des exercices ou des problèmes est détaillé afin de bien montrer, au-delà de la solution ponctuelle, les enjeux scientifiques et la démarche méthodologique de l'analyse en génétique des populations, et constituer ainsi un complément aux travaux dirigés et un bon entraînement aux examens.

---

1. *Génétique et évolution*. M. Solignac, G. Perriquet, D. Anxolabérère et C. Petit, Hermann, Paris, 1995.
2. *Précis de génétique des populations. Cours, exercices et problèmes résolus*. J.-P. Henry & P.-H. Gouyon, Dunod, Paris, 1999.
3. *Génétique des populations : modèles de base et applications*. J.-L. Serre, Nathan-Université, Paris, 1997 (épuisé).

Introduction

Toute théorie nouvelle est une révolution scientifique et résulte d'une « rupture dans la conception du monde ». Il en fut ainsi de la théorie darwinienne de l'évolution et de la théorie mendélienne de l'hérédité. La génétique des populations est le champ disciplinaire qui a permis la synthèse entre ces deux théories.

Aujourd'hui le concept d'évolution des espèces est si unanimement partagé et identifié à l'œuvre de Charles Darwin qu'on a perdu le fil d'une histoire si utile pour comprendre les conditions de la naissance et la finalité de la génétique des populations.

Quand Darwin publie en 1859 l'*Origine des espèces par la sélection naturelle*, de nombreux biologistes « transformistes » l'avaient précédé. Cependant ces transformistes, comme Lamarck ou Erasme Darwin (le grand-père de Charles) considéraient l'évolution des espèces comme résultant d'une transformation individuelle plus ou moins finalisée et transmissible des organismes. Chacun d'entre eux était capable de se modifier à travers un phénomène « d'adaptation » aux conditions extérieures, puis de transmettre ces modifications acquises à sa descendance.

Au contraire Charles Darwin affirme que le processus évolutif n'est pas individuel mais collectif, c'est-à-dire populationnel. Les organismes d'une population diffèrent les uns des autres, ce qui les rend plus ou moins bien « adaptés » aux conditions extérieures. L'évolution des espèces se présente alors comme un processus de tri, que Darwin appelle « sélection naturelle ». Par cette sélection, les organismes les mieux adaptés, parce qu'ils sont plus viables ou plus fertiles et donc plus féconds, laissent une descendance plus nombreuse, tout en leur transmettant leurs aptitudes.

Chez les transformistes prédarwiniens, l'évolution est un processus individuel d'adaptation assorti de l'hérédité des caractères acquis. Chez Darwin, l'évolution est un processus de tri au sein d'une population polymorphe, entre ceux qui sont, par nature, mieux ou moins bien adaptés aux conditions du moment.

Il n'est pas inutile de rappeler que cette nouvelle vision du monde biologique résulte, chez Darwin, de considérations assez étrangères à la biologie. Des études

précises sur sa biographie montrent qu'il est longtemps resté attaché à des conceptions fixistes, malgré leur incapacité à expliquer certains faits paléontologiques ou bio-géographiques qu'il avait lui-même remarqué durant son voyage sur le *Beagle* autour du monde (voir à ce sujet Camille Limoges, *La Sélection Naturelle*, Paris, PUF, 1970). Ce sont plutôt des considérations d'ordre philosophique, idéologique et social, dans cette période de libéralisme triomphant, qui l'ont amené au rejet du concept d'adaptation parfaite (et de l'utilisation qui en était faite par les tenants de la théologie naturelle) et à l'affirmation de l'existence d'organismes inadaptés, utilement écartés par la sélection. Le premier chapitre de l'*Origine des espèces par la sélection naturelle* sur les « variations dans les conditions de domestication » et la sélection artificielle des éleveurs a une visée essentiellement pédagogique, car le concept de sélection naturelle par la « survie du plus apte » trouve en fait son origine dans la théorie de Malthus. Avec lui, Darwin considère que la stabilité des effectifs dans la plupart des espèces végétales ou animales, malgré leur grande prolificité, provient d'une élimination massive et impitoyable ; Darwin ajoute seulement qu'en étant sélective, cette élimination désormais rebaptisée « survie du plus apte » est le moteur de l'évolution.

Mais la sélection appliquée à la variation interindividuelle au sein de l'espèce ne suffit pas à établir une théorie de l'évolution si ne sont pas aussi précisées les lois de l'hérédité qui permettraient aux meilleurs de transmettre à leurs descendants leurs aptitudes.

Or, de ce point de vue la théorie darwinienne de l'évolution va se trouver en échec car Darwin, et avec lui tous les biologistes de l'époque, sont dans l'incapacité d'opérer l'autre révolution conceptuelle, le rejet du concept d'hérédité mélangée, que seul Mendel entrevoit, à l'autre bout de l'Europe,

Pour tous les biologistes du XIX^{e} siècle, même les grands biologistes cellulaires comme Schleiden ou von Nageli, la fécondation est un mélange de substances parentales. Cette conception est d'ailleurs cohérente avec la ressemblance entre apparentés et le fait que les enfants présentent souvent, pour des caractères quantitatifs, des valeurs médianes à celles des parents. Galton, le fondateur de l'eugénisme, formalisera cette transmission sous le nom de « loi ancestrale de l'hérédité ». Mais accepter la conception de l'hérédité par mélange conduit obligatoirement la théorie de l'évolution par sélection naturelle à une contradiction puisque le mélange des aptitudes parentales dans la descendance conduit à leur dilution, donc à leur perte et à l'arrêt de toute sélection et évolution. Pour sauver sa théorie, Darwin évoquera la possibilité de croisements préférentiels entre les plus aptes et reviendra à la nécessité, déjà vécue par Lamarck, de l'hérédité des caractères acquis.

Certes le mendélisme recèle la clef de l'hérédité, mais les travaux de Mendel, bien que connus de nombreux naturalistes et de Darwin, n'ont pas été compris, même par Mendel, comme des travaux de portée générale sur les mécanismes de l'hérédité mais comme des expériences sur les mécanismes de l'hérédité chez les hybrides dans le cadre de l'étude de leur stabilité. De plus les travaux de Mendel étaient suspects aux yeux des darwiniens car les hybrideurs, depuis Linné, avaient une solide réputation d'anti-transformisme.

Il faut attendre le début du XX^e^ siècle pour que la portée générale du mendélisme soit reconnue et que l'alternance fécondation-méiose (observée seulement à la fin du XIX^e^ siècle, après la mort de Mendel) dans le règne animal ou végétal puisse être interprétée comme la réunion puis la séparation d'entités non miscibles, les gènes sus leurs différentes formes alléliques.

Entre-temps le mouvement darwinien, dont les préoccupations non biologiques avaient rejoint l'eugénisme, avait fondé la biométrie dont le but était de mesurer la valeur de nombreux caractères (quantitatifs) chez les individus d'une même population afin de juger de leur valeur adaptative en rapportant la variation statistique de ces caractères dans le temps (l'évolution) à la fécondité des individus. À cette fin la biométrie et l'eugénisme prirent une part prépondérante à la naissance puis au développement des statistiques (mesures, lois de distribution, définition des tests…).

En ce début du XX^e^ siècle, un courant transformiste non-darwinien (ne croyant pas à la réalité de la sélection comme moteur de l'évolution mais privilégiant le rôle de mutations brutales de l'hérédité), emmené par Bateson, trouve, dans la théorie mendélienne de l'hérédité, un cheval de bataille contre le darwinisme, accroché au concept de plus en plus intenable de « blending inheritance ». Et c'est dans le sein de l'école biométricienne et darwinienne qu'un groupe de jeunes eugénistes, dont Yule et Fisher, opéra une révolution conceptuelle majeure en réconciliant darwinisme et mendélisme. L'argument essentiel de la démarche est la démonstration mathématique, par Fisher, que la théorie mendélienne des facteurs alléliques disjoints conforte la théorie darwinienne, en lui fournissant les moyens de maintenir le polymorphisme génétique, ce que le concept de « blending inheritance » ne permettait pas. Ce faisant, ce groupe fondait la génétique des populations (et la génétique quantitative) dont le but, comme pour la biométrie, était de quantifier et de modéliser l'évolution génétique des espèces, à travers l'évolution de la fréquence des allèles des gènes, sous l'effet de sélection en faveur d'« allèles favorables ».

Le développement de la génétique des populations a conduit à de nombreuses autres hypothèses sur les mécanismes évolutifs, complémentaires à la sélection, qui seront présentés dans cet ouvrage. Cependant la synthèse entre darwinisme et mendélisme n'évacuait pas la question posée par les mutationnistes non darwiniens du début du siècle. Ce débat fut de nouveau ouvert mais non tranché par la grande lignée des généticiens évolutionnistes russes (Severzov, Timoféef-Ressovsky, Philipstschenko et Dobzhansky) et les fondateurs de la génétique des populations (Fisher, Haldane et Wright) quand ils constatent la capacité de la génétique mendélienne et du darwinisme d'expliquer la micro-évolution (à l'intérieur des vertébrés) et son incapacité à expliquer la macro-évolution (passage des arthropodes aux vertébrés).

C'est la génétique moléculaire du développement de la drosophile qui a permis de découvrir la structure et la fonction de « blocs de gènes homéotiques » dans la segmentation et le développement de l'embryon, blocs également retrouvés chez la souris et l'homme. Ce résultat a une importance scientifique considérable parce qu'il ouvre une perspective de recherche fructueuse sur la macro-évolution. En évolution et en génétique des populations, le développement moléculaire de la génétique apparaît encore comme un outil décisif dans l'avancée des connaissances.

À tous les stades de son histoire, la génétique a vu croître son importance dans la biologie en raison de sa capacité croissante à unifier des domaines auparavant disparates ; surtout depuis l'approche moléculaire, structurale et fonctionnelle, des gènes sous-jacents à l'ensemble des phénomènes biologiques. Dans tous les domaines, la physiologie, l'embryologie et le développement, la biologie cellulaire, la biologie des populations, l'écologie et l'évolution, les phénomènes en cause sont associés à l'expression de gènes spécifiques, qu'il est désormais possible de localiser et d'identifier, de cloner et de modifier à loisir par mutagenèse dirigée, et dont les effets peuvent être étudiés dans un contexte choisi, *in vitro*, *ex* ou *in vivo*. Cette biologie moléculaire du gène constitue un outil remarquable parce qu'il permet potentiellement d'ouvrir toutes les « boîtes noires » que représentaient, dans ces disciplines, des approches pertinentes mais essentiellement descriptives, compte tenu de la globalité des phénomènes ou des structures étudiés.

Par ce fait, la génétique et la génétique des populations jouent un rôle essentiel dans le projet philosophique de la science occidentale. Car si la philosophie est absente de l'activité quotidienne spontanée du scientifique, elle est pourtant, qu'il le veuille ou non et même qu'il le sache ou non, au cœur de sa démarche. Depuis la Grèce antique, par l'application des principes de la logique, du doute et de la rationalité, un corps de connaissances en astronomie, en physique, en médecine et en biologie, est en perpétuelle refonte ; son instabilité intrinsèque l'exclut définitivement du statut de cosmologie et il proclame son ambition de comprendre l'origine et la fin de l'univers et de l'homme.

D'ailleurs, l'approfondissement de l'analyse biologique à travers ses mécanismes moléculaires, a fait surgir une nouvelle problématique sur le rôle dévolu au hasard dans l'évolution des espèces, et aussi, depuis peu, dans l'expression des gènes et le développement apparemment si régulé et invariant de l'ontogenèse. Le hasard, contrairement au sens commun que suggère le titre de l'œuvre de Jacques Monod, *Le hasard et la nécessité*, n'est pas le seul hasard des mutations géniques et génomiques soumises à la sélection. Il apparaît aussi comme un acteur majeur des trajectoires évolutives des populations et des espèces (voir certains des chapitres de cet ouvrage). Mais cette question, posée en filigrane par Kimura et les partisans de la théorie neutraliste, est bien plus profonde et philosophique. Certes, sur une courte durée, les contingences physiques, chimiques, biologiques forment une nécessité qui détermine le vivant et son évolution, même si des facteurs stochastiques sont à même de perturber sa formation ou son évolution, mais après une grande durée, ces contingences semblent n'avoir eu aucun effet tant était large, à tout instant, l'éventail des possibilités évolutives, de sorte qu'il ne resterait, *a posteriori*, que le hasard comme « démiurge » principal de l'évolution des espèces. La théorie de l'évolution et la génétique des populations il y a près de 30 ans, la génétique aujourd'hui, notamment la génétique du développement et des équilibres cellulaires, se trouvent confrontées à la question du hasard, comme le furent et le sont encore les physiciens des particules et les astrophysiciens dans la recherche d'une théorie unifiée des forces et de l'origine de l'Univers.

Chapitre **1**

Définition et mesure de la diversité génétique. Constitution génétique des populations

1.1 INTRODUCTION

L'expérimentation génétique est essentiellement fondée sur des croisements entre souches pures. Au contraire, les populations naturelles sont composées d'organismes génétiquement différents pour un grand nombre de gènes. La génétique des populations se propose d'abord d'évaluer l'importance de cette diversité génétique.

Au-delà de cette estimation, la génétique des populations se propose aussi de dégager des lois permettant de rendre compte du maintien ou de l'évolution dans l'espace et dans le temps de cette diversité génétique. Partant du postulat que l'évolution des espèces est, en dernière analyse, celle de leur patrimoine génétique, la génétique des populations apparaît alors comme le noyau dur obligatoire de toute théorie de l'évolution.

Les différences génétiques entre individus, autrement dit le polymorphisme génétique des populations, sont alors conçues comme un patrimoine, un capital adaptatif, un gage de survie de l'espèce aux variations de l'environnement selon le principe darwinien de la « survie du plus apte » (voir Introduction à l'ouvrage) ; ce qui suppose évidemment l'existence préalable de telles différences génétiques entre individus. À sa naissance la génétique des populations réalise une sorte de synthèse entre la vision darwinienne de l'évolution et la vision mendélienne de l'hérédité, mais ses développements mathématiques ultérieurs ont débordé le cadre strictement darwinien.

Les modèles théoriques de la génétique des populations s'attachent à l'étude des mécanismes gouvernant l'évolution de la diversité génétique au niveau d'un gène ou deux, rarement plus ; au-delà, la démarche analytique est pratiquement impossible mais peut être remplacée par les simulations informatiques. C'est pourquoi nous

aborderons dans un premier temps la définition et la mesure du polymorphisme génétique au niveau d'un seul gène, et la mise en évidence de la structure génétique d'une population pour ce gène.

La mesure du polymorphisme pour un grand nombre de gènes permet de fournir une estimation de la diversité génétique globale existant non seulement entre individus d'une même population mais aussi entre populations, notamment chez l'homme. L'analyse de la diversité génétique entre populations humaines permet d'aborder le sujet si controversé de la définition des races. Elle permet aussi d'entreprendre une analyse phylogénétique des populations humaines dont le but est la mise en évidence de l'âge et de l'origine de l'homme à travers une recherche, dans les gènes, de l'histoire du peuplement de la planète. Les principaux résultats concernant cet aspect de la diversité génétique seront présentés dans un deuxième temps.

1.2 LES DIFFÉRENTS TYPES DE POLYMORPHISMES UTILES EN GÉNÉTIQUE DES POPULATIONS

À tout moment, une mutation peut modifier le génome contenu dans le noyau d'une cellule. Si une mutation affecte le génome d'une cellule somatique, un sous-clone muté se développera chez l'individu affecté (par exemple une tumeur). Une mutation somatique n'est pas transmissible, au contraire une mutation survenant dans la lignée germinale peut être transmise à la descendance par les quelques gamètes qui en sont porteurs. Dans ce cas le patrimoine génétique de la population se trouve « enrichi » puisque sa diversité est accrue.

1.2.1 Le polymorphisme génique

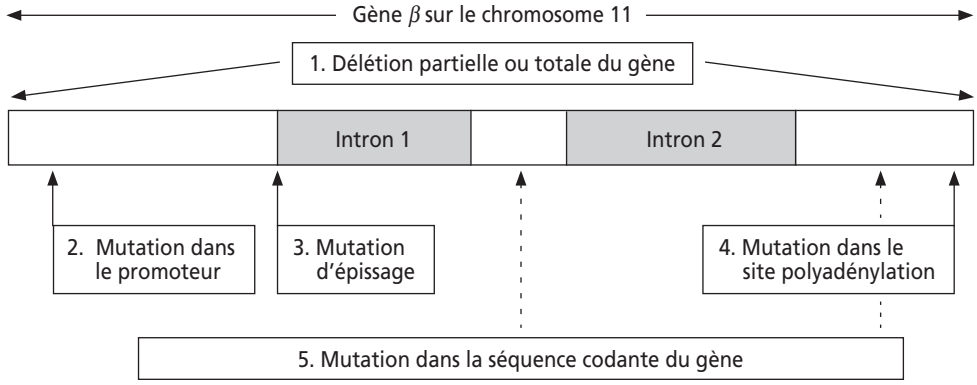
Les mutations géniques affectent un gène dans l'une de ses séquences et peuvent affecter sa fonction et retentir sur le ou les phénotypes des caractères dans lequel ce gène est impliqué. Selon sa nature (changement, perte ou insertion d'une ou plusieurs paires de bases) et son site dans le gène (figure 1.1), une mutation peut se révéler être, sur le plan fonctionnel :

– une mutation de perte de fonction, si elle entraîne la diminution ou l'absence de produit (mutation quantitative) ou la présence d'un produit moins actif ou inactif (mutations qualitatives) ;
– une mutation de gain de fonction si elle entraîne une surproduction (mutations quantitatives), ou la présence d'un produit plus actif, ou d'un produit doué d'une propriété nouvelle, absente du produit « sauvage » (mutations qualitatives), éventuellement toxique dans le cas de certaines maladies dominantes.

Sur le plan phénotypique, l'effet d'une mutation peut se révéler :

– « dominant » vis-à-vis de l'effet de l'allèle sauvage si l'hétérozygote pour ces deux allèles est de phénotype muté ;
– « récessif » vis-à-vis de celui de l'allèle sauvage si l'hétérozygote pour ces deux allèles est de phénotype sauvage et que le phénotype muté n'est observable que chez les porteurs de deux allèles mutés du gène.

Il convient de rappeler que les pertes de fonction peuvent, selon les gènes touchés, avoir un effet dominant chez certains ou récessifs chez d'autres, de la même manière que les gains de fonction, assez souvent dominants peuvent aussi avoir un effet récessif.



Nature de la mutation	Effet primaire sur l'expression du gène	Conséquence de l'effet sur le produit du gène	Pathologie associée
1	Pas de transcris ou transcris incomplet	Pas de chaîne β Perte de fonction	β-thalassémie (récessive[1])
2 (promoteur inactif)	Pas ou moins de transcription	Pas ou peu de chaîne β Perte de fonction	β-thalassémie (récessive[1])
3	Transcription mais pas de messager	Pas de chaîne β Perte de fonction	β-thalassémie (récessive[1])
4	Messager instable : peu de traduction	Pas de chaîne β Perte de fonction	β-thalassémie (récessive[1])
5 Mutation stop	Arrêt prématuré de traduction	Pas de chaîne β Perte de fonction	β-thalassémie (récessive[1])
Mutation de décalage du cadre de lecture	Chaîne aberrante et arrêt prématuré de traduction	Pas de chaîne β Perte de fonction	β-thalassémie (récessive[1])
Mutation faux-sens	Transcription et traduction	Substitution d'un acide aminé par un autre (produit modifié) Perte de fonction Gain de fonction Gain de fonction	Selon la substitution : β-thalassémie (récessive[1]) Drépanocytose (récessive[2]) Anémie hémolytique (dominante[3])

1 : récessif par compensation allélique, l'allèle sauvage étant deux fois plus transcrit chez le porteur sain.
2 : récessif par dilution de l'hémoglobine S, produit muté doué d'une propriété nouvelle, toxique pour l'hématie, sa capacité de polymériser, en pression partielle faible en oxygène, la mutation drépanocytaire étant de ce fait un gain de fonction.
3 : dominant parce que le produit muté est instable et toxique pour les globules rouges, malgré la présence de produit sauvage.

Figure 1.1 Les mutations du gène β de l'hémoglobine : effets primaires et pathologies associées.

1.2.2 Les marqueurs polymorphes de l'ADN : Indels, SNP, RFLP, STR

Le développement de la biologie moléculaire du gène et le séquençage de l'ADN, puis le séquençage des génomes, ont permis de mettre en évidence l'existence de polymorphismes moléculaires de l'ADN, une diversité génétique qui ne touche pas seulement les séquences des gènes mais se répartit sur l'ensemble du génome, aussi bien dans les séquences signifiantes (codantes ou non codantes) que dans les séquences intergéniques. De ce fait, ces polymorphismes sont très utiles dans nombre d'applications de la génétique des populations à la biodiversité, à l'étude des QTL (*Quantitative Trait Loci*) ou à l'épidémiologie génétique.

En effet, ils peuvent servir de marqueurs génétiques dans la mesure où on peut attribuer à leurs différents états le statut d'« allèles », définir alors des génotypes homozygotes ou hétérozygotes, suivre la transmission de ces marqueurs dans une généalogie, estimer leur diversité génétique dans une population ou une espèce et analyser la variation de cette diversité dans l'espace et dans le temps. Parmi les marqueurs polymorphes de l'ADN les plus utilisés, on peut distinguer :

– les Indels (Insertions-Délétions), ou polymorphismes d'insertion-délétion d'une séquence d'ADN en un site du génome, formant le plus souvent un polymorphisme di-allélique, repérable par étude de la longueur du fragment d'ADN amplifié par PCR, à partir de deux amorces flanquant le site du marqueur ;
– les SNP (*Single Nucleotide Polymorphism*), ou polymorphismes simple nucléotide, substitution d'une paire de base par une autre paire de base, en un site du génome, génique ou intergénique, bi-allélique, dont les génotypes (deux homozygotes et un hétérozygote) sont identifiables par diverses méthodes de biologie moléculaire *in vitro*[1] ;
– les RFLP (*Restriction Fragment Length Polymorphism*), ou polymorphisme de longueur de fragments de restriction, en général dus à un SNP qui fait apparaître ou disparaître, en un locus du génome, un site de reconnaissance spécifique d'une endonucléase. Les RFLP sont des marqueurs bi-alléliques (présence ou absence du site), dont le génotype est identifiable par étude de la longueur des fragments d'ADN après amplification par PCR de la séquence du génome contenant le locus et action de l'enzyme de restriction[1] ;
– les STR (*Short Tandem Repeats*), ou polymorphismes de courtes séquences de deux ou trois nucléotides répétées en tandem, encore appelées « séquences microsatellites ». Contrairement aux marqueurs précédents, les STR sont multi-alléliques, le nombre d'allèles étant égal au nombre de répétitions existant au locus considéré sur les divers chromosomes homologues de la population ou de l'espèce. Il convient de rappeler alors que le nombre d'homozygotes est alors égal à n, le nombre d'allèles, que le nombre d'hétérozygotes est égal à $n\,(n-1)/2$ et que le nombre total de génotypes possibles est égal à $n\,(n+1)/2$ (encart 1.1).

1. Voir *Les diagnostics génétiques*, J.-L. Serre et coll., Dunod, Paris, 2002.

Encart 1.1

Dénombrement des génotypes homozygotes et hétérozygotes pour un gène polymorphe présentant *n* formes alléliques.

Dans un espèce diploïde les génotypes d'un gène sont constitués des deux exemplaires portés par les deux chromosomes homologues d'origine paternelle et maternelle.

Pour un gène di-allélique, on dénombre deux homozygotes et un hétérozygote. Pour un gène présentant *n* formes alléliques distinctes, on dénombre évidement ***n* homozygotes** distincts, ***n* (*n* – 1)/2 hétérozygotes**, ce qui fait en tout ***n* (*n* + 1)/2 génotypes** distincts, et autant de phénotypes, si il y a codominance.

On peut facilement démontrer ces formules de la manière suivante. Le tableau carré ci-dessous donne tous les génotypes définis par la combinaison de deux allèles pris parmi *n*, mais il est redondant.

Allèles	*A1*	*A2*		*Ai*		*An*
A1	*A1/A1*			*A1/Ai*		
A2		*A2/A2*				
Ai	*Ai/A1*			*Ai/Ai*		
An						*An/An*

Les homozygotes correspondent aux génotypes de la première diagonale et sont évidemment aussi nombreux que les allèles eux-mêmes, soit *n*.

Les hétérozygotes correspondent donc au reste du tableau, mais comme il convient de ne pas dénombrer deux fois le même hétérozygote (voir dans le tableau l'exemple de *A1/Ai*), il faut donc diviser par deux le nombre de cases restantes après le retrait des *n* cases de la première diagonale.

On obtient donc ainsi le nombre d'hétérozygotes, soit n^2 cases totales dans le tableau moins les *n* homozygotes, le tout divisé par 2, c'est-à-dire :

$$(n^2 - n)/2 = \boldsymbol{n\,(n-1)/2}$$

1.2.3 Le polymorphisme chromosomique

Des mutations peuvent survenir à une autre échelle au sein du génome et affecter le nombre ou la structure des chromosomes. À petite échelle, celle des individus, ces mutations sont souvent associées à des pathologies et sont le plus souvent éliminées par la sélection, mais à grande échelle, celle de l'évolution, on sait désormais qu'elles

ont joué un rôle important en replaçant les uns par rapport aux autres des blocs de gènes affectant le développement. Les cytogénéticiens ont depuis longtemps précisé le nombre et l'ampleur des remaniements chromosomiques existant entre le caryotype à 46 chromosomes de l'homme et celui à 48 du chimpanzé.

Les mutations chromosomiques sont des modifications de la structure des chromosomes sans que le nombre en soit changé (figure 1.2) ; *elles correspondent à des translocations de fragments chromosomiques, des inversions, des fusions centriques, des délétions ou des duplications*. Il ne faut pas croire qu'elles sont rares (encart 1.2).

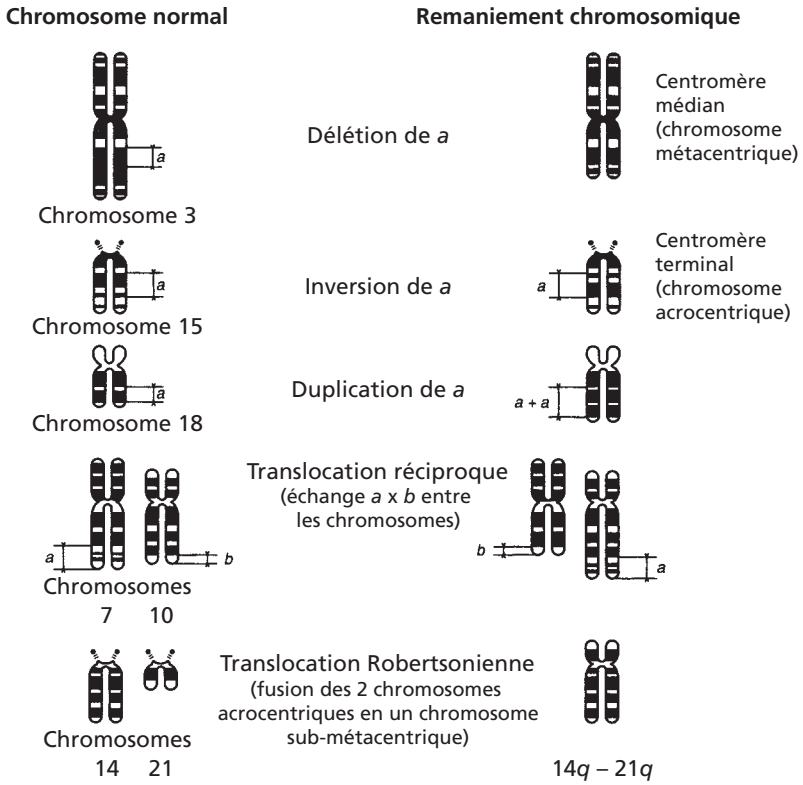


Figure 1.2 Quelques exemples de remaniements chromosomiques observés chez l'homme.

Les mutations génomiques ne concernent plus un fragment de chromosome mais un ou plusieurs chromosomes dans leur totalité ; ce sont les changements de ploïdie par perte d'un lot haploïde de chromosomes (monoploïdie) ou par gain d'un ou plusieurs lots (tri-, tétraploïdie), et les aneuploïdies par perte ou gain d'un ou plusieurs chromosomes (monosomie si un chromosome est absent ; trisomie quand il y a un chromosome surnuméraire). Ces mutations sont souvent pathologiques et létales car elles perturbent profondément le déroulement du programme génétique,

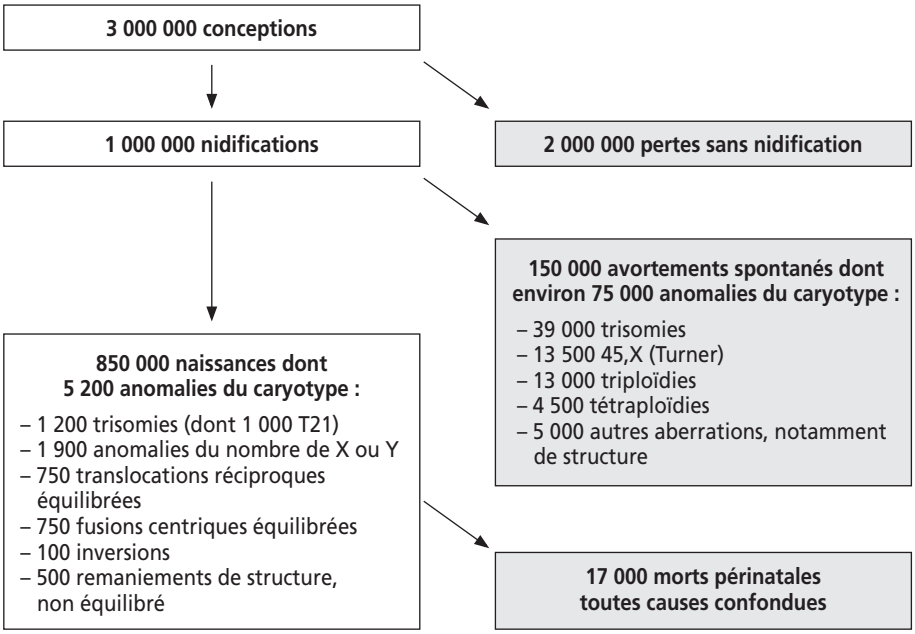
notamment lors du développement embryonnaire. Mais leur rôle a peut-être été essentiel lors de certaines étapes de l'évolution (dans le cas de la macro-évolution par exemple). Les caryotypes de l'homme et du chimpanzé, son plus proche parent, comportent respectivement 23 et 24 paires de chromosomes ; une observation fine a montré que l'une des paires humaine résultait de la fusion centrique de deux paires acrocentriques (centromère à l'une des extrémités) du chimpanzé, résultat confirmé par l'analyse cartographique des gènes portés par les chromosomes des deux espèces.

Encart 1.2

La fécondité et les anomalies du caryotype chez l'homme.

De nombreuses études démographiques (dont celles de H. Léridon à l'INED) ont permis de conclure que près des deux tiers des conceptus sont éliminés avant la nidification, puis 15 % entre la nidification et la 12e semaine.

Les études cytogénétiques des produits de fausses couches spontanées (dont celles de A. Boué, à l'INSERM) ont clairement établi qu'il s'agit d'une sélection éliminant des fœtus porteurs de mutations chromosomiques ou génomiques. En effet, ces mutations qui touchent seulement 5 nouveaux-nés sur 1 000 représentent la moitié des cas dans les avortements spontanés. Cet ensemble de données conduit au schéma suivant sur la base des 850 000 naissances annuelles en France :



1.2.4 Les différents types et niveaux de perception du polymorphisme génétique

Les premières études de génétique ont porté sur des caractères morphologiques (taille, forme, couleur des organismes ou de certaines de leurs parties, yeux, ailes, tige, fleur, graine,…). Chez l'homme, on s'intéressa assez vite aux phénotypes cliniques (atteint ou non atteint) associés aux maladies héréditaires. Tous ces phénotypes, même quand ils sont gouvernés par les allèles d'un seul gène, ne sont qu'une conséquence indirecte, physiologique ou cellulaire, en aval de l'effet primaire de ces allèles, ou de ce qu'on nomme, pour les maladies génétiques, les mutations pathogènes.

Le développement de la biochimie des protéines, de l'enzymologie et de l'immunologie, puis de la biologie moléculaire du gène, ont permis d'avoir accès soit au produit du gène, la chaîne peptidique, soit à la séquence du gène lui-même. Aussi dans bien des cas, l'analyse biochimique des protéines ou moléculaire du gène donne accès à des polymorphismes sans traduction perceptible à l'échelle morphologique ou physiologique et permet de percevoir, à cette échelle, un polymorphisme génétique non perceptible à l'échelle de l'organisme.

D'ailleurs, quand il s'agit de marqueurs polymorphes de l'ADN n'affectant aucune séquence signifiante, il n'existe aucune répercussion sur quelque phénotype de quelque caractère que ce soit et les seuls phénotypes accessibles résident dans la séquence d'ADN elle-même ou des dérivés de cette séquence, à travers les fragments de PCR et leur étude soit pour la présence ou l'absence d'un site (RFLP), soit pour la longueur de la répétition (STR). Ces phénotypes sont codominants puisque, sauf exception, leur étude permet de définir sans ambiguïté le génotype pour le marqueur étudié.

Deux exemples permettent d'illustrer comment la diversité génétique peut être perçue différemment selon le niveau d'analyse de ses conséquences dans la chaîne des effets d'une mutation, de son effet primaire à ses conséquences protéiques, cellulaires, tissulaires et physiologiques ou morphologiques.

Exemple 1.1 La drépanocytose chez l'homme

> La drépanocytose est une hémoglobinopathie très fréquente en Afrique équatoriale (environ 4 % des naissances), dans le Golfe arabique et en Asie du Sud-Est. Elle résulte d'une mutation dans le sixième codon du gène β de l'hémoglobine, mutation notée β^S. Le changement d'une base dans ce codon spécifiant un acide glutamique conduit à un triplet spécifiant la valine. La même mutation β^S est survenue cinq fois indépendamment dans les différentes populations affectées (voir chapitre 3). La question se pose évidemment, d'un point de vue évolutif, de savoir comment une telle mutation a pu survenir si « fréquemment » d'une part, et se maintenir d'autre part, alors qu'elle est fortement pathogène (voir chapitre 7).

En effet, la chaîne peptidique codée par l'allèle β^S est stable et fonctionnelle mais l'hémoglobine formée chez les homozygotes β^S/β^S, appelée hémoglobine S (HbS) diffère de l'hémoglobine A (HbA) par une propriété aux conséquences pathologiques. En faible pression partielle en oxygène, l'HbS est capable de polymériser en longues fibres déformant alors l'hématie en forme de faucille (d'où l'autre nom de la maladie : l'anémie falciforme) et altérant fortement sa plasticité. Le blocage de ces hématies dans les capillaires terminaux, où précisément la pression partielle en oxygène est très réduite, provoque des anoxies locales qui altèrent les tissus, stimule de manière autocatalytique la falciformation et provoque des chocs hémolytiques graves. Ces épisodes aigus de la maladie joints à une anémie chronique fatiguent l'organisme, dégradent irréversiblement des fonctions majeures (cardiaque, hépatique, pulmonaire, cérébrale) et finissent par entraîner la mort. On doit donc considérer que la mutation β^S affecte la viabilité en réduisant l'espérance de vie. Dans certaines populations, cette mutation n'est pas obligatoirement létale, pour des raisons génétiques, comme en Inde, où on observe une persistance assez importante (plusieurs %) d'hémoglobine fœtale après la naissance, ou, pour des raisons de milieu, comme dans les Antilles françaises, quand des soins intensifs peuvent être donnés. Les hétérozygotes ne sont pas atteints car leurs hématies contiennent un mélange d'HbS et d'HbA qui ne peut polymériser facilement.

Sur le plan clinique on distingue deux classes phénotypiques, le phénotype récessif atteint correspondant au génotype homozygote β^S/β^S et le phénotype dominant sain correspondant aux génotypes β^+/β^+ ou β^+/β^S. Du point de vue clinique, les homozygotes β^+/β^+ ne peuvent être distingués des hétérozygotes β^+/β^S : la maladie est récessive.

Il est possible néanmoins d'opérer cette distinction parce que l'hémoglobine peut être facilement obtenue et étudiée. L'électrophorèse d'un volume d'hémoglobine obtenu à partir d'individus des trois génotypes possibles donne en effet trois résultats différents, trois phénotypes électrophorétiques codominants associés à chacun des trois génotypes (figure 1.3) :

Génotype	β^+/β^+	β^+/β^S	β^S/β^S
Phénotype clinique	Sain		Atteint
Phénotype électrophorétique			

Figure 1.3 Phénotypes électrophorétiques de migration de l'hémoglobine sur gel.

Dans les conditions de l'électrophorèse, le tétramère d'hémoglobine $\alpha 2\beta 2$ se dissocie en dimères $\alpha\beta$, ce qui donne chez l'hétérozygote des dimères du type $\alpha\beta^S$ et des dimères du type $\alpha\beta^+$.

Selon qu'on perçoit l'effet des allèles et au niveau du caractère clinique ou du caractère biochimique (distance de migration électrophorétique) les génotypes sont associés à des phénotypes dominants-récessifs ou codominants.

Dans de nombreuses maladies génétiques récessives où le produit du gène impliqué est difficilement accessible ou plus simplement inconnu, une telle distinction biochimique des génotypes est impossible et constituera un des obstacles que la génétique des populations permettra de surmonter.

Exemple 1.2 Phénotype [rosy] chez la drosophile

Les drosophiles de phénotype [rosy] ont les yeux rose et diffèrent des drosophiles de phénotype sauvage noté [rosy+], aux yeux rouge brique, par la mutation d'un seul gène. Le phénotype mutant [rosy] est récessif : les hétérozygotes *ry+*/*ry* (*ry+* et *ry* étant respectivement les allèles sauvages et mutés du gène) sont de phénotype sauvage comme les homozygotes *ry+*/*ry+*.

Il a été montré que le gène impliqué dans le phénotype [rosy] codait pour la xanthine-déshydrogénase qu'il est possible de doser dans un extrait acellulaire de drosophile. Ce dosage conduit à la définition de nouveaux phénotypes : les phénotypes d'activité enzymatique. Ceux-ci sont codominants car l'hétérozygote présente un taux d'activité médian compris entre celui de l'homozygote *ry+*/*ry+* et celui de l'homozygote *ry*/*ry* (nul car déficient en enzyme). Le phénotype de l'hétérozygote pour le caractère couleur de l'œil est sauvage car une activité égale à 50 % de l'activité sauvage est largement suffisante pour assurer la chaîne de biosynthèse des pigments.

Par ailleurs des études électrophorétiques de la xanthine-déshydrogénase ont montré l'existence de deux allèles électrophorétiques, chacun codant pour une chaîne peptidique active mais différant l'une de l'autre par un seul acide aminé de charge électrique opposée. Cette différence conduit alors à une migration électrophorétique différentielle. L'un des allèles est appelé *fast* (*f*) parce qu'il code pour une chaîne à déplacement rapide (notée *F*) ; l'autre allèle est appelé *slow* (*s*), car il code pour une chaîne à déplacement plus lent (notée *S*).

Les génotypes *f*/*f*, *f*/*s* et *s*/*s* sont, en même temps, tous les trois *ry+*/*ry+* car ils présentent un même phénotype (taux élevé) pour le caractère « dosage d'activité » et un même phénotype (yeux sauvages rouge brique) pour le caractère « couleur de l'œil » ; ils ne peuvent être distingués entre eux qu'à partir de la mise en évidence, par électrophorèse, des trois phénotypes pour le caractère « migration électrophorétique ». Ceux-ci sont codominants car on peut distinguer la présence aussi bien que l'absence des chaînes rapides et lentes (figure 1.4).

Comme la xanthine-déshydrogénase est un homo-dimère, l'hétérozygote *f*/*s* présente trois bandes électrophorétiques correspondant à des dimères de type *F*/*F*, *F*/*S* ou *S*/*S*. Par ailleurs, les hétérozygotes *f*/*ry* ou *s*/*ry* ne présentent qu'un seul type de chaîne à l'électrophorèse, celle codée par l'allèle *f* ou *s*, car *ry* est une mutation amorphe entraînant l'absence de chaîne ; de ce fait, les phénotypes électrophorétiques, considérés isolément, ne peuvent permettre de distinguer sans ambiguïté les génotypes *f*/*f* et *f*/*ry* d'une part, *s*/*s* et *s*/*ry* d'autre part. Seule la connaissance conjointe du phénotype d'activité permet de faire la distinction.

Génotype au locus du gène *xdh*	*f*/*f*	*f*/*s*	*s*/*s*	*f*/*ry*	*s*/*ry*	*ry*/*ry*
Phénotypes morpho-logiques	Yeux de phénotype sauvage rouge brique [rosy+]					Yeux rose [rosy]
Phénotypes d'activité	Taux élevé +++			Taux médian +		Taux nul –
Phénotypes électro-phorétiques						

Figure 1.4 Correspondance génotypes/phénotypes pour les divers allèles du gène *xdh* (l'allèle actif *ry*+ étant noté *f* ou *s*, selon le type de chaîne codée, *F* ou *S*).

En l'absence d'informations complémentaires, il est évidemment impossible, avec les phénotypes morphologiques, de distinguer l'homozygote *ry*+/*ry*+ de l'hétérozygote *ry*+/*ry*. L'information complémentaire peut être ici le phénotype d'activité, mais on peut aussi l'obtenir par l'analyse de la descendance par test-cross avec un individu *ry*/*ry*.

1.2.5 Du génotype au phénotype : une relation souvent complexe

Le développement considérable de la génétique et des ses applications dans nombre de domaines de la biologie ou la médecine, l'a parfois transformée, notamment chez le grand public et les médias, en sciences des « certitudes ». C'est par exemple le sens implicitement donné à des formules comme le « programme génétique » ou le fait qu'un caractère soit « génétiquement déterminé ». Or, c'est oublier que l'expression des gènes ne saurait être conçue comme indépendante du milieu, de l'environnement au sein duquel ces gènes s'expriment.

Quelques exemples sont utiles pour bien rappeler et comprendre que la diversité entre individus d'une même espèce ou population, peut résulter de leur diversité génétique mais aussi de la diversité des milieux au sein desquels ils se sont construits.

Exemple 1.3 La phénylcétonurie

Maladie génétique récessive, elle touche les enfants dont les deux exemplaires du gène de la phénylalanine-hydroxylase (PHA) sont mutés et non fonctionnels. La déficience de cette enzyme bloque la transformation de la phénylalanine en tyrosine et dérive l'excès de phénylalanine vers la formation d'acide phényl-pyruvique, très toxique pour les neurones des aires cervicales de la cognition. De ce fait, ces enfants développent rapidement une arriération mentale profonde (aussi appelée idiotie phényl-pyruvique).

On peut dire que cette arriération mentale est « génétiquement programmée », mais c'est oublier que ces mêmes enfants, s'ils sont dépistés à temps par le test de Guthrie (analyse de l'excès de phénylalanine sérique, dans une goutte de sang, à la maternité, chez tous les enfants), peuvent recevoir une alimentation artificielle carencée en phénylalanine, ce qui prévient l'accumulation d'acide phényl-pyruvique et la dégénérescence des neurones frontaux. Après quelques années, quand ces neurones sont matures, ils deviennent insensibles à l'acide phényl-pyruvique et l'enfant peut reprendre une alimentation normale.

On voit donc que « génétiquement programmé » n'a ici de sens qu'en fonction du milieu au sein duquel s'exprime le « programme génétique » : dans un milieu carencé en phénylalanine, l'arriération mentale n'est pas « génétiquement programmée » !

Exemple 1.4 La nervation de l'aile de drosophile

Une souche de drosophile élevée à 20 °C présente une nervation des ailes, identique d'un individu à l'autre. Si une partie des pupes est soumise à un choc thermique de 37 °C, au début de la pupaison, les adultes auront des ailes partiellement dépourvues de nervures. La nervation des ailes est un « caractère génétiquement déterminé », mais le bloc de gènes qui gouverne ce caractère conduit à des formes différentes, des phénotypes différents, selon le milieu au sein duquel il s'est exprimé, car la température affecte le dosage d'expression d'une partie de ces gènes.

Devant cet exemple d'interaction entre gènes et environnement, on doit se souvenir que la relation entre génotypes et phénotypes définis dans un environnement donné peut être différente dans un autre environnement.

Exemple 1.5 La taille chez *Achillea* en fonction de l'altitude

L'*Achillea* (mille fleurs) est une plante qui, comme le fraisier ou le géranium, peut être multipliée par reproduction végétative (bouturage) ; tous les plants issus de la plante mère sont génétiquement identiques entre eux. On peut dire de la taille qu'elle est un « caractère génétiquement déterminé » car les boutures donnent des plantes de taille identique à celle de la plante mère.

Pourtant, à partir de plantes de grande, de moyenne ou de petite taille, des boutures replantées à basse, moyenne et haute altitude donnent des résultats très variables (figure 1.5).

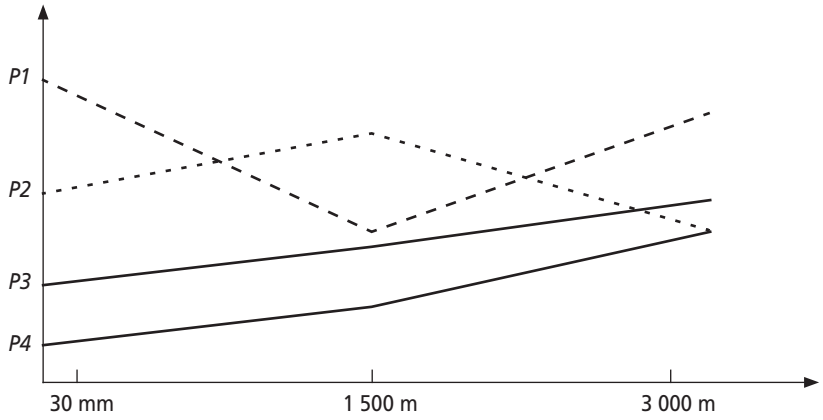


Figure 1.5

Dans tous les cas, la taille est « génétiquement programmée » puisque toutes les boutures issues d'un même plant sont génétiquement identiques et phénotypiquement de même taille… à la même altitude !

Mais les plants *P1*, les plus grands au niveau de la mer et à 3 000 m, n'ont qu'une taille moyenne à 1 500 m, alors que les plants *P2* sont plus grands que les plants *P1* à cette altitude. Peut-on dire quelle est la plante qui est « génétiquement » la plus grande, si on ne considère pas le milieu ?

Cet exemple illustre aussi la complexité de l'interaction entre les gènes et l'environnement. Alors que la moyenne altitude a un effet négatif sur la taille de *P1*, elle a un effet positif sur celle de *P2* !

Seuls *P3* et *P4* montrent un effet de même nature (positif) du milieu, permettant une interprétation additive de l'interaction gène-environnement : le programme génétique conditionne une valeur minimale de la taille et l'effet de l'environnement une valeur additionnelle. Mais ce type de conception est faux pour *P1* et *P2*.

Rappelons que beaucoup imaginent que le « programme génétique » définit pour chaque homme des capacités ou des aptitudes et que l'environnement décide du niveau qui sera atteint en pratique. Une telle conception n'est acceptable que si l'interaction gènes-milieu est additive, du type de celle affectant *P3* ou *P4*, ce qui n'est nullement démontré !

Ainsi, tous les discours sur le fondement génétique (« génétiquement déterminé ») des surdoués, de la réussite ou de l'échec scolaire, du génie musical ou de quelque autre don, sont des preuves d'une ignorance profonde de ce qu'est la relation gène-environnement dans la réalisation des caractères et la variation de leurs phénotypes, ou pire, une volonté de l'ignorer.

1.3 MESURE DE LA DIVERSITÉ GÉNÉTIQUE ET COMPOSITION GÉNÉTIQUE D'UNE POPULATION

1.3.1 La population

L'espèce est par définition un groupe génétiquement fermé au sein duquel les organismes sont susceptibles, par l'alternance méiose-fécondation de séparer ou de réunir les divers allèles de chacun des gènes et de concevoir des combinaisons génétiques nouvelles par la recombinaison génétique.

Cependant tous les individus d'une même espèce, s'ils sont potentiellement susceptibles de réaliser ce brassage peuvent en être pratiquement empêchés quand des barrières limitent les possibilités de croisements entre certains individus.

Il peut s'agir d'un isolement géographique, lié à l'existence d'une barrière naturelle comme un océan pratiquement infranchissable ou une chaîne montagneuse plus facilement franchissable, ou plus simplement par la distance qui limite la probabilité de croisements entre individus très éloignés.

Il peut s'agir d'un isolement écologique. Par exemple des plantes d'une même espèce occupant un même territoire semblent former une même population. Mais si une disparité dans la composition du sol décale la floraison entre les plantes du sol A et celles du sol B, les échanges génétiques entre les plantes des sols A et B seront limités. Il sera alors nécessaire de définir deux populations A et B. La définition et l'analyse de populations naturelles supposent donc une bonne connaissance de leur biologie et de leur biotope.

Il peut s'agir enfin, chez l'homme, de barrières culturelles, sociales ou ethniques, qui limitent plus ou moins les possibilités d'unions entre individus, même géographiquement proches.

Une espèce peut donc être subdivisée en sous-groupes au sein desquels la possibilité d'échanges génétiques entre individus est effective ; ces sous-groupes sont appelés populations et l'ensemble des allèles qu'ils partagent, pour chacun des gènes de l'espèce, en constitue le patrimoine génétique (pool allélique).

La mesure de la diversité génétique à l'intérieur des populations mais aussi entre les populations, l'origine et le devenir de ces diversités intra- et inter-populationnelles sont un enjeu important de la génétique des populations, notamment chez l'homme, en raison des polémiques qui ont accompagné la définition et l'usage du concept de race, ou les débats sur l'émergence de l'homme moderne à partir de l'*Homo erectus* (voir plus loin).

1.3.2 Variables d'état de la diversité et composition génétique d'une population

Il est possible de définir une chaîne de causalité liant la variabilité phénotypique des caractères et la diversité génétique sous jacente qui en est la cause (figure 1.6, à gauche). À chacun des niveaux hiérarchiques de la diversité, on peut associer des variables d'état qui mesurent la diversité génétique à ce niveau, les fréquences alléliques, les fréquences génotypiques, les fréquences phénotypiques (figure 1.6, à droite).

Les fréquences phénotypiques sont toujours accessibles directement par le dénombrement des phénotypes présents dans un échantillon, et la question se pose de savoir si il existe des relations mathématiques simples permettant, si on connaît la diversité à un niveau hiérarchique, d'en déduire la diversité à un autre niveau (figure 1.6, flèches doubles à droite). Si de telles relations sont disponibles alors la connaissance de la diversité en un point quelconque des niveaux hiérarchiques permettrait d'avoir une connaissance exhaustive de la diversité génétique de la population en tout autre point.

Or c'est bien le premier but de la génétique des populations que de savoir mesurer la diversité pour définir la composition génétique d'une population ou d'une espèce, au niveau des allèles et des génotypes.

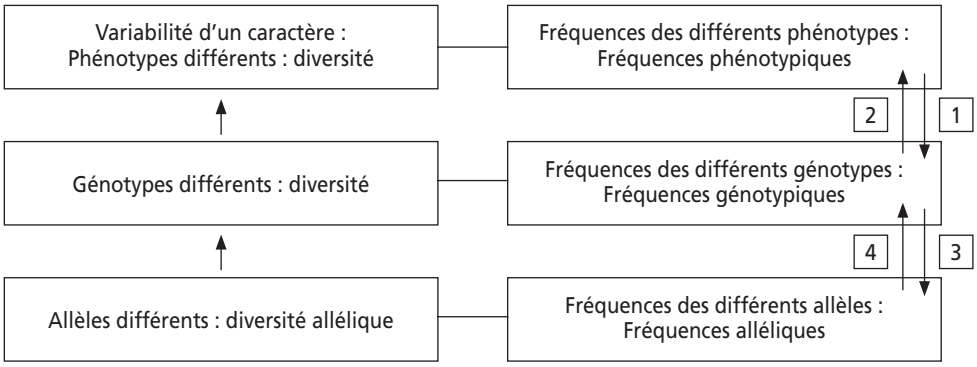


Figure 1.6 Niveaux hiérarchiques de la diversité génétique et variables d'états associés.

Dans un premier temps, on va définir la mesure de la diversité génétique relative à un gène (ou un marqueur) sachant qu'une fraction des problèmes de génétique des populations correspond à une telle situation. Ultérieurement, on généralisera nos résultats à l'étude simultanée de deux gènes ou marqueurs, situation très communément rencontrée en épidémiologie génétique, mais, comme cela a été dit, l'approche analytique devient impossible au-delà et il est nécessaire de recourir à des simulations informatiques.

1.3.3 Codominance et dominance : les limites de la mesure de la diversité génétique

Les phénotypes sont par définition directement accessibles à l'observation et les fréquences phénotypiques peuvent être calculées à partir d'un échantillon d'individus tirés aléatoirement dans la population, mais le calcul des fréquences génotypiques et alléliques n'est possible que si les phénotypes sont codominants, comme le montrent les exemples suivants.

a) Phénotypes codominants : groupe sanguin MN

Le typage est réalisé, comme pour le groupe ABO ou rhésus, par un test d'hémaglutination, les individus appartenant aux groupes [M], [N] ou [MN] selon que leurs

hématies sont respectivement reconnues par l'anticorps anti-M ou l'anticorps anti-N ou les deux (tableau 1.1), parce qu'elles sont porteuses d'une chaîne peptidique spécifiée par l'allèle L^M et/ou l'allèle L^N.

Tableau 1.1 Échantillon aléatoire d'une population européenne

Groupes sanguins	[M]	[MN]	[N]
Effectifs observés (total : 1 000)	350	500	150
Fréquences phénotypiques	350/1 000 = 0,35	500/1 000 = 0,50	150/1 000 = 0,15
Génotype	L^M/L^M	L^M/L^N	L^N/L^N

Ici, les phénotypes étudiés sont codominants, il est possible d'associer un phénotype à un seul génotype et réciproquement. Dès lors les fréquences génotypiques sont égales aux fréquences phénotypiques et les fréquences alléliques s'en déduisent aisément. La constitution génétique de la population pour le gène concerné est connue sans ambiguïté.

Deux méthodes de calcul des fréquences alléliques sont possibles, un simple comptage des allèles ou une formule probabiliste.

Méthode des comptages. Elle consiste à dire que les 1 000 individus sont porteurs de 2 000 allèles parmi lesquels les allèles *M* représentent 350×2 (pour les homozygotes qui possèdent 2) plus 500 (pour les hétérozygotes qui n'en possèdent qu'un). Les fréquences des allèles *M* et *N* sont donc respectivement :

$$f(L^M) = (350 \times 2 + 500)/2\,000 = 1\,200/2\,000 = 0{,}6$$

et $$f(L^N) = (150 \times 2 + 500)/2\,000 = 800/2\,000 = 0{,}4$$

avec $$f(L^M) + f(L^N) = 1$$

Méthode probabiliste. On peut calculer les fréquences alléliques par l'application du théorème des probabilités composées.

Considérons plus généralement les trois génotypes possibles résultant des combinaisons diploïdes des deux allèles *A1* et *A2* d'un gène et leurs fréquences respectives *D*, *H* et *R* :

– Génotypes : *A1*/*A1* *A1*/*A2* *A2*/*A2*
– Fréquences génotypiques : *D* *H* *R*

La fréquence de l'allèle *A1* peut être définie comme la probabilité de tirer cet allèle au hasard dans la population, ce qui suppose d'abord de tirer un individu, puis l'un de ses deux allèles :

– l'individu tiré peut être *A1*/*A1*, avec la probabilité *D* ; dans ce cas l'allèle tiré au hasard chez cet individu sera *A1* avec la probabilité 1,
– ou l'individu tiré peut être *A1*/*A2*, avec la probabilité *H* ; dans ce cas l'allèle tiré au hasard chez cet individu sera *A1* avec la probabilité 1/2, car l'individu est aussi porteur de *A2*.
– ou l'individu tiré peut être *A2*/*A2*, avec la probabilité *R* ; dans ce cas l'allèle tiré au hasard chez cet individu sera *A1* avec la probabilité 0, car il n'en possède pas.

On aura donc $f(A1) = D \times 1 + H \times 1/2 + R \times 0$

D'où la formule générale : $f(A1) = D + H/2$

et $f(A2) = R + H/2$

Dans l'exemple du groupe MN, on retrouve évidemment les mêmes valeurs que celles obtenues par l'autre méthode.

> ***Remarque :*** si on revient à la figure 1.6, on voit que les relations 1 et 2 sont des relations d'identité et que la relation 3 est le système d'équations qui vient d'être défini, à savoir que la fréquence d'un allèle est égale à la fréquence des homozygotes pour cet allèle plus la moitié de la fréquence des hétérozygotes. Mais la relation 4 est, pour l'instant indéterminée, car avec les fréquences alléliques, il n'est pas possible d'en déduire les fréquences génotypiques.

b) Phénotypes dominants et récessifs : groupe sanguin ABO

Il est toujours possible d'estimer les fréquences phénotypiques (tableau 1.2) mais il est ici impossible d'en déduire les fréquences génotypiques, du moins pour les phénotypes dominants présentant plusieurs possibilités génotypiques sous jacentes ; dans ce cas la fréquence du phénotype, par exemple celle de [A], est égale à la somme des fréquences des deux génotypes, I^A/I^A et I^A/I^O, mais on ne peut connaître la valeur individuelle de la fréquence de chaque génotype, indispensable pour estimer les fréquences alléliques selon les formules vues plus haut. On ne peut mesurer la diversité génétique et, en se rapportant à la figure 1.6, on reste bloqué dans l'ensemble des phénotypes car aucune des quatre relations mathématiques n'est déterminée à ce stade.

Tableau 1.2 Groupes sanguins ABO dans un échantillon aléatoire de la population française.

Groupes sanguins	[A]	[B]	[AB]	[O]
Effectifs observés (total : 500)	213	56	15	216
Fréquences phénotypiques	213/500 = **0,426**	56/500 = **0,112**	15/500 = **0,03**	216/500 = **0,432**
Génotypes	I^A/I^A ou I^A/I^O	I^B/I^B ou I^B/I^O	I^A/I^B	I^O/I^O

Dans certains cas, une information complémentaire permet de lever l'ambiguïté de la relation phénotype/génotype ; elle est apportée par la connaissance de l'ascendance ou de la descendance des individus étudiés, ou bien, comme dans la drépanocytose, par une étude biochimique donnant accès, pour la mutation étudiée, à des phénotypes codominants (voir exercice).

Cette situation de blocage est commune à tous les caractères présentant des phénotypes mutés récessifs, notamment les maladies récessives, ce qui implique qu'on ne peut pas, pour l'instant, estimer la fréquence d'un allèle pathologique ni celle des porteurs sains.

1.3.4 Degré de polymorphisme et degré d'hétérozygotie

On peut établir une mesure plus globale de la diversité génétique d'une population en intégrant sur un grand nombre de gènes, la diversité allélique établie pour chacun d'entre eux (voir paragraphe précédent). Cette diversité globale peut être appréhendée par le « degré de polymorphisme » et le « degré d'hétérozygotie ».

Connaître la constitution génétique d'une population pour un seul gène polymorphe ne permet pas de savoir si cette population est ou n'est pas très polymorphe. Elle l'est si un grand nombre de gènes sont polymorphes ; ce qui suppose, en pratique d'avoir étudié un grand nombre d'individus pour un grand nombre de gènes.

Si on étudie un très grand nombre d'individus pour un gène, on trouvera toujours un ou quelques allèles très rares correspondant notamment au bruit de fond des mutations *de novo*, dont la plupart, comme on le verra (voir chapitre 8), disparaissent en quelques générations du pool génique. Autrement dit ces allèles sont trop rares pour qu'ils fassent partie de ce pool sur un nombre significatif de générations en terme d'évolution. Le seuil qui a été arbitrairement choisi pour distinguer l'allèle rare comptabilisable dans le pool de celui qu'on délaissera est la fréquence de 1 %. Sur cette base un gène polymorphe est un gène qui présente au moins deux formes alléliques de fréquence supérieure ou égale à 1 %, et le *degré de polymorphisme* d'une population est défini par son pourcentage de gènes polymorphes (tableau 1.3).

Le degré de polymorphisme n'est cependant pas le seul, ni le meilleur indicateur de la diversité génétique globale d'une population, car il ne prend en compte ni le nombre d'allèles pour un gène, ni la valeur des fréquences alléliques mais seulement le fait que le gène est ou n'est pas polymorphe.

Or si un gène polymorphe présente un allèle *A1* très fréquent, à 99 % et un allèle rare *A2*, à 1 %, la très grande majorité des individus sera constituée d'homozygote *A1*/*A1* et une minorité sera hétérozygote *A1*/*A2* (les individus *A2*/*A2* étant sans doute exceptionnels, comme on le verra plus tard grâce à la relation de Hardy-Weinberg). Par contre si un gène polymorphe présente deux allèles *B1* et *B2* de fréquences sensiblement égales (1/2), un grand nombre d'individus sera hétérozygote.

La relation de Hardy-Weinberg (voir chapitres 2 et 3) permet de prévoir que la fréquence des hétérozygotes sera égale à :

$$H = 1 - \Sigma p_i^2 \quad \text{où } p_i \text{ est la fréquence du } i^{\text{ème}} \text{ allèle du gène}$$

Dans les deux cas précédents, on obtient pour l'hétérozygotie des valeurs respectives de 1,98 % et 50 %, ce qui montre bien que la population n'est pas génétiquement dans la même situation de diversité pour ces deux gènes di-alléliques. On notera en effet que le taux H d'hétérozygotie tend vers une valeur maximale $H_{\max}$ pour des valeurs égales des fréquences alléliques.

Pour un gène polymorphe présentant n allèles différents (avec n fréquences alléliques p_i, i variant de 1 à n) il existe n homozygotes et $n(n-1)/2$ hétérozygotes possibles (encart 1.1). Si les n allèles sont équifréquents, la valeur de cette fréquence allélique sera égale à $1/n$ et la valeur du taux d'hétérozygotie sera égale à :

$$\boldsymbol{H_{max}} = 1 - \Sigma\,(1/n)^2 = 1 - n(1/n)^2 = 1 - 1/n = \boldsymbol{(n-1)/n}$$

Pour un gène di-allélique la valeur de H_{max} est égale à 50 % (voir figure 2.3) ; mais pour un gène polymorphe avec 15 ou 20 allèles, comme il en existe dans le complexe majeur d'histocompatibilité chez l'homme, la valeur de H_{max} est égale à 93 ou 95 %. Cependant la valeur réelle de H dans la population peut être très inférieure à H_{max} si les fréquences alléliques sont différentes parce qu'un allèle est très fréquent (en effet le terme p_i^2 correspondant à cet allèle prend alors une grande importance dans la formule de *H*, comme dans le premier exemple où $H = 1{,}98$ % quand $H_{max} = 50$ %)

On définit le *taux moyen d'hétérozygotie* d'une population comme la moyenne des taux d'hétérozygotie pour un grand nombre de gènes étudiés (tableau 1.3). Il constitue une mesure de la diversité génétique globale plus précise que le degré de polymorphisme. Mais celui-ci est également utile comme indicateur simple et direct du pourcentage de gènes polymorphes dans la population ou l'espèce.

Tableau 1.3 Degrés de polymorphisme et taux d'hétérozygotie pour plusieurs gènes dans trois populations fictives.

POPULATION		I	II	III
Gène A	allèle *A1*	0,5	0,7	0,9
	allèle *A2*	0,5	0,3	0,1
Taux d'hétérozygotie pour le gène *A*		0,5	0,42	0,18
Gène *B*	allèle *B1*	0,8	0,95	1
	allèle *B2*	0,2	0,05	0
Taux d'hétérozygotie pour le gène *B*		0,32	0,095	0
Gène *C*	allèle *C1*	1	1	1
Taux d'hétérozygotie pour le gène *C*		0	0	0
Gène *D*	allèle *D1*	0,3	0,6	0,2
	allèle *D2*	0,2	0,05	0,8
	allèle *D3*	0,3	0,35	0
	allèle *D4*	0,2	0	0
Taux d'hétérozygotie pour le gène *D*		0,74	0,515	0,36
DEGRÉ DE POLYMORPHISME		75 %	75 %	50 %
TAUX MOYEN D'HÉTÉROZYGOTIE		39 %	25,75 %	13,5 %

Les populations I et II ont un même degré de polymorphisme car elles présentent trois gènes polymorphes sur quatre, contre deux seulement dans la population III. Cependant par le calcul du taux moyen d'hétérozygotie, on peut conclure que la population I présente plus de diversité que la population II, notamment en raison du fait que le gène *D* y est présent sous la forme de quatre allèles presque équifréquents alors qu'il n'est présent dans les populations II et III que sous trois et deux formes alléliques d'une part, et des fréquences très différentes entre elles d'autre part.

La population III est dépourvue de certains allèles et présente, pour les gènes polymorphes, des différences de fréquences beaucoup plus prononcées que dans les autres populations. Cette situation de moindre diversité est typiquement retrouvée dans toutes les petites populations isolées d'Amérique ou d'Océanie. Cette particularité sera expliquée par les phénomènes de dérive génétique (voir chapitre 5).

La diversité estimée sur plusieurs dizaines de gènes chez l'homme, la souris et la drosophile donne des valeurs respectives de 31, 29 et 42 % pour le degré de polymorphisme, et 10, 9 et 12 % pour le taux moyen d'hétérozygotie.

1.4 LA DIVERSITÉ GÉNÉTIQUE CHEZ L'HOMME

1.4.1 La question des races chez l'homme

Une race est une variété intraspécifique dont les individus partagent des traits qui les distinguent des individus appartenant à une autre race et dont le maintien dépend essentiellement de croisements strictement endogames, au sein d'une population strictement fermée. Le pool génique qui est à l'origine des caractères propres à la race ne peut garder sa spécificité qu'en absence de migrations susceptibles d'apporter des allèles différents qui diluerait ces caractères.

Le concept de race est parfaitement adapté à la pratique des éleveurs et à l'histoire de la domestication ; il a, dans ces conditions, une signification biologique et une utilité opératoire réelles (chien, chat, cheval, bovins, ovins, variétés ou races végétales). Mais en est-il de même chez l'homme ?

La distinction entre trois grands groupes humains caractérisés principalement par des différences de pigmentation (« races » noire, jaune et blanche) est ancienne. Puis on a assisté, surtout au XIX^{e} siècle, à de nombreuses tentatives de typologie ou de classification fondées sur une collection diverse et croissante de paramètres.

Bien évidemment les anthropologues ont défini autant de classifications raciales qu'ils avaient choisi de paramètres différents, montrant ainsi l'arbitraire de toute classification raciale dans une espèce où les populations naturelles ne sont pas longtemps strictement endogames.

Ceci n'empêche nullement de concevoir l'existence de « populations géographiques » distinctes les unes des autres à la fois par leur localisation et des caractères morphologiques d'autant plus contrastés qu'elles sont éloignées les unes des autres. Cependant aucune de ces populations n'est strictement endogame et les échanges génétiques entre populations voisines ont généré de tout temps un véritable continuum.

C'est pourquoi le terme de race n'est pas vraiment approprié à la situation génétique des populations humaines. Mais ce n'est pas tant la définition de la race chez l'homme que la connotation idéologique de l'usage qui en a été fait qui justifie la méfiance vis-à-vis de ce concept.

À la fin du XIX^{e} siècle et au début du XX^{e} siècle, de nombreux anthropologues et naturalistes pensaient, dans la lignée de Darwin, que les « races » avaient été façonnées par la sélection naturelle. Celle-ci n'y aurait laissé subsister que les variations

(comprenons la diversité génétique) adaptées à leur environnement propre. De ce fait les différences génétiques entre individus d'une même race tendraient à disparaître pour ne laisser subsister que les « types » les plus proches de la norme idéale définie par la pression de sélection naturelle. Chacune des « races » étant confrontée à des contraintes environnementales différentes se voyait caractérisée par une norme idéale différente, source naturelle des différences entre races, elles-même pérennisées par l'endogamie raciale. Dans ces conditions les seules différences génétiques importantes, au sein de l'espèce humaine, résidaient entre les races et non au sein des races.

Cette conception typologique, bien qu'erronée, ne pouvait être remise en cause à l'époque, non seulement parce qu'aucunes données sur la diversité génétique des populations n'étaient disponibles, mais aussi parce que le concept de race, dans sa définition et son usage, épousait l'idéologie dominante de cette époque du colonialisme et du ségrégationnisme triomphants. On fondait, par exemple, la condamnation des unions inter-raciales sur la base d'un discours naturaliste (donc scientifique !) arguant du fait que la descendance serait inadaptée et dégénérée car ayant hérité de variations de chacune des races, elle se serait ainsi éloignée de la norme idéale de chacune d'entre elles. Enfin, en mettant l'homme blanc au-dessus des autres dans leurs classifications hiérarchisées, les théories des races montraient leur caractère ouvertement raciste.

Les travaux de quelques généticiens des populations comme Lewontin ou Cavalli-Sforza, sur la diversité génétique des groupes sanguins et enzymatiques (puis le polymorphisme de l'ADN), ont montré que la situation génétique de l'humanité n'est en rien celle qu'imaginaient les typologistes, notamment ceux qui firent le pire usage du concept de race.

Se fondant sur l'analyse de la diversité génétique globale d'un grand nombre de gènes polymorphes, comme ABO (tableau 1.4), Lewontin aboutit à une première conclusion, déjà observée dans de nombreuses populations naturelles d'autres orga-

Tableau 1.4 Groupe sanguins ABO dans quelques populations. Effectifs rapportés à total de 1 000 individus par population pour rendre évident le calcul des fréquences phénotypiques.

Population étudiée	Effectif testé	Groupe [A]	Groupe [B]	Groupe [AB]	Groupe [O]
Anglaise	1 000	434	72	30	464
Française	1 000	426	112	30	432
Allemande	1 000	430	120	50	400
Italienne	1 000	380	110	38	472
Grecque	1 000	416	162	40	382
Turque	1 000	380	186	66	368
Indienne	1 000	190	412	85	313
Sénégalaise	1 000	226	292	50	432

nismes. Les populations humaines, quelles que soient les frontières arbitraires qu'on leur donne, sont très polymorphes et diffèrent essentiellement par les fréquences alléliques de gènes polymorphes et non, de manière systématique comme la typologie l'imaginait, par la présence d'un allèle spécifique à chaque population.

Il est assez clair (tableau 1.4) que les populations du nord de l'Europe sont plus proches entre elles que des populations de l'est, et que l'Europe forme un tout vis-à-vis de l'Inde ou du Sénégal. Mais en aucun cas on peut dire qu'un groupe sanguin est plus spécifique de l'une des populations.

Dans son analyse de la diversité génétique de l'humanité, Lewontin évalua celle-ci sous la forme d'une variance appelée V_H. Puis séparant l'humanité dans ces trois grands groupes habituels, les populations asiatiques, africaines et européennes (qu'on peut éventuellement appeler les « races » jaune, noire et blanche), Lewontin estima la diversité génétique au sein de chacune d'entre elle, sous la forme de trois variances V_J, V_N, et V_B. Aucune d'entre elle n'était inférieure à 92 % de V_H, ce qui signifie que les différences génétiques entre asiatiques, africains et européens, que personne ne peut nier, ne comptent cependant que pour 8 % des différences globales au sein de l'humanité. En d'autres termes deux hommes pris au hasard, l'un en Asie, l'autre en Afrique, différent entre eux pour 8 % de leurs gènes en raison de leur appartenance ethnique, ce qui signifie qu'ils peuvent partager un même génotype pour une bonne partie des 92 % qui restent.

Séparant alors les trois grands groupes en sous-groupes nations, ce qui a une certaine logique puisqu'ils constituent, malgré les migrations, des populations relativement endogames, Lewontin estima la diversité génétique au sein de chacune d'entre elles, sous la forme de variances V_{ni}. Aucune des variances intra-nationales n'étaient inférieures à 85 % de V_H, ce qui signifie qu'aux 8 % de la diversité génétique totale rendant compte des différences entre les trois grands « groupes raciaux » viennent s'ajouter 7 % rendant compte des différences entre populations nationales au sein de ces trois groupes (voir par exemple le tableau ci dessus pour le groupe ABO).

Ainsi 85 % de la diversité génétique de l'humanité réside au sein des populations nationales et les différences entre populations ne représentent qu'une faible partie de ce total.

Certes un tel résultat démontre l'irréalisme de toute vision typologique ou raciale de l'humanité. Mais peut-on pour autant s'en servir comme d'un argument scientifique pour nier tout fondement biologique au racisme ?

Ne suffirait-il pas au racisme de prétendre que les gènes choisis pour l'étude des différences ne sont pas pertinents ? Ne suffirait-il pas aux racistes de considérer que de tels gènes existeraient dans ces 15 % de la variabilité génétique rendant compte des différences ethniques ? Ne leur suffirait-il pas de considérer que des différences interpopulationnelles existeraient pour ces gènes cruciaux qui gouvernent des caractères le comportement, les aptitudes physiques (les 100 mètres ou le marathon) intellectuelles ou cognitives (les échecs, les mathématiques ou la musique) ?

À ce stade, il ne s'agit plus du débat sur la réalité du concept de race mais d'une autre question relative à la relation entre gènes et aptitudes. Or nous avons vu (1.2.5)

qu'il était absurde d'établir une relation stricte entre les gènes et les phénotypes en faisant abstraction du milieu. Bien évidemment, même en supposant que des différences existent entre populations, elles s'expriment dans des contextes eux-mêmes différents. De ce fait, toutes les tentatives de biologisation de problèmes sociaux ou culturels comme les problèmes d'intégration et d'échec scolaire ne sont que la manifestation d'une ignorance profonde de la relation génotype-phénotype, ou de la volonté de l'ignorer.

Mais la lutte contre le racisme ne peut s'appuyer sur la biologie ou la génétique des populations que pour arracher le masque dont tentent de se parer les racistes ; pour le reste la lutte contre le racisme est un idéal éducatif, culturel, idéologique et politique.

1.4.2 De la génétique des populations à l'origine de l'homme

Il serait prétentieux dans cet ouvrage de prétendre à un exposé complet d'une question aussi vaste et complexe que celle de l'origine de l'homme. D'ailleurs, elle met en jeu des approches parfois cohérentes, parfois contradictoires, de disciplines aussi diverses que la paléontologie, l'archéologie, l'anthropologie, la chimie-physique (pour les méthodes de datation), la génétique des populations et la linguistique.

Il est désormais clairement admis que les formes les plus anciennes d'hominidés sont apparues en Afrique avec la lignée de l'australopithèque (4 millions d'années dont la célèbre Lucy), de l'*Homo habilis* (3 millions d'années) et de l'*Homo erectus* (1,8 million d'années) dont on suit l'évolution morphologique et culturelle (objets, maîtrise du feu,...).

Il est également admis que l'absence, hors d'Afrique, de tout squelette d'hominidés antérieurs à l'*Homo erectus* n'est pas le résultat de fouilles encore infructueuses mais la conséquence d'une véritable absence. En d'autres termes l'homme a conquis l'ancien monde à partir de l'Afrique et sous sa forme erectus.

Cette conception d'une racine unique, en Afrique, il y a 1,5 million d'années, marque un véritable progrès par rapport aux conceptions polycentriques que certains anthropologues ou archéologues prônaient à la fin du XIX^{e} siècle. Pour eux, plusieurs racines humaines trouvaient leurs origines dans l'évolution parallèle des primates, les Européens descendants d'un ancêtre commun au chimpanzé, tandis que les Africains et les Asiatiques descendaient respectivement d'un ancêtre commun au gorille et à l'orang-outang !

Bien évidemment, le débat sur l'ancienneté de la divergence entre les populations humaines n'est pas « neutre » puisque les différences génétiques sont d'autant plus grandes que la séparation est ancienne et que divergence ou distance génétique sont au cœur de certaines conceptions « raciales » ou « racistes ».

Ce débat s'est prolongé et rejaillit depuis quelques années, quant à la date et surtout au lieu de d'émergence de l'homme moderne (*Homo sapiens sapiens*).

En effet, à partir du moment où de Java à l'Europe et de la Chine à l'Afrique, l'*Homo erectus* semble avoir conquis l'ancien monde, les fouilles ont permis de découvrir toutes sortes de fossiles qui peuvent aussi bien être interprétés comme des

maillons vers l'apparition des hommes modernes dans le cadre d'une théorie polycentrique (figure 1.7-a) que comme des restes de populations disparues sans descendants, dans le cadre d'une théorie unicentrique, fondée sur le repeuplement de l'ancien monde par l'homme moderne apparu, il y a 100 000 ans, entre l'Arabie et l'Éthiopie (figure 1.7-b).

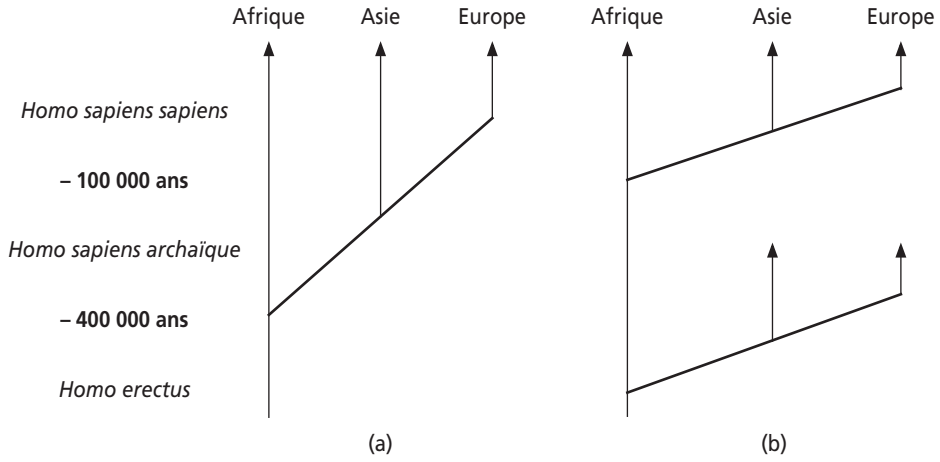


Figure 1.7 (a) : théorie polycentrique (b) : théorie unicentrique.

On peut voir une illustration d'un tel remplacement en Europe, par la disparition brutale du type néandertalien, il y a 40 000 ans au profit de l'homme moderne. Celui-ci venait probablement du Proche-Orient où il est connu à – 100 000 ans (grotte de Qafzeh en Palestine). Mais le remplacement des formes primitives, en Afrique sub-saharienne ou en Asie, par un homme moderne né entre l'Afrique de l'Est et l'Arabie n'est pas accepté par certains paléontologues qui constatent la survivance, dans ces continents de caractères propres aux formes de transition entre l'*Homo erectus* local et des fossiles de caractère moderne. Pour ces tenants du polycentrisme une évolution vers l'homme moderne aurait pu survenir parallèlement en plusieurs zones du globe.

La théorie polycentrique n'est sans doute contradictoire ni avec l'unicité d'origine de l'espèce humaine, ni avec la similitude des gènes qu'ils partagent, dans la mesure où des échanges génétiques (minimes compte tenu des distances !) auraient pu relier entre elles ces diverses zones d'évolution. Mais ses arguments, essentiellement fondés sur les collections de quelques fossiles incomplets et pas toujours parfaitement datés, rendent compte beaucoup moins bien de la très proche similitude génétique entre tous les hommes que la théorie unicentrique (figure 1.7-b). De plus, les données récentes d'analyse des marqueurs génétiques viennent confirmer assez magistralement la théorie unicentrique.

La génétique des populations a donc beaucoup apporté dans le débat sur l'homme moderne sans pour autant le résoudre définitivement et avec précision. Et ce n'est pas l'étude de l'ADN mitochondrial et la théorie de l'Ève africaine qui emporte

l'adhésion d'une origine récente et africaine de l'homme moderne ; au contraire, elle se présente plutôt comme l'exemple typique d'une étude dont la conclusion était décidée *a priori* (voir 1.4.3.a).

En fait, ce sont des travaux plus précis sur la diversité d'un grand nombre de gènes nucléaires qui permet de préciser ce qui a pu ou n'a pas pu advenir dans l'histoire de l'homme moderne.

Dès 1964 L.-L. Cavalli-Sforza et A. Edwards ont étudié la diversité des populations humaines pour cinq groupes sanguins. Ils entreprennent l'analyse de leurs données sur la base d'une idée simple : plus la ressemblance entre deux populations est grande, plus leur relation de parenté dans le temps l'est aussi. Cette étude en montrant que l'humanité se divisait en deux grands ensembles, l'Afrique et l'Europe d'une part, l'Asie, l'Amérique et l'Australie, d'autre part, montrait que l'étude des gènes permettait de retrouver des relations de filiation clairement établies comme celle qui unit l'Amérique et l'Australie à l'Asie.

L'interprétation des données ne pouvait guère aller au delà parce que le nombre de gènes étudiés, le nombre de populations et la taille des échantillons étaient insuffisants et que la validité de certaines hypothèses des modèles mathématiques était contestée.

À la fin des années 1980, Cavalli-Sforza présenta une étude sur 120 gènes ou marqueurs génétiques différents, dans 42 populations. Il y montre que la quantification de la similitude génétique entre les populations humaines, sous la forme d'un arbre phylogénétique, sépare d'abord l'Afrique du reste du monde, puis l'Asie continentale du Sud-Est asiatique. Les ramifications ultérieures conduisent pour l'Asie continentale aux populations caucasiennes (berbères, perses, européennes, indiennes), d'Asie du Nord (Chine, Mongolie, Japon), d'Amérique et d'Arctique. Elles conduisent pour l'Asie du Sud-Est aux populations de Chine du Sud, de Cochinchine, d'Australie et d'Océanie (voir 1.4.3.b).

Le degré de précision dans l'analyse de la diversité génétique (voir 1.4.3.b) fait clairement apparaître un peuplement de l'Asie par deux voies, correspondant sans doute au nord et au sud de l'Himalaya, un détail qui eut été perdu dans la nuit des temps si les populations descendaient des ancêtres locaux postulés par la théorie polycentrique.

Certes l'analyse est menée avec l'idée *a priori* d'une arborescence à partir d'une population d'origine. Mais si cette idée n'avait aucune réalité, l'analyse de la diversité aurait parfaitement pu placer les îles du pacifique avec les Bushmen, et les Berbères avec les Cheyennes. Il existe une parenté entre les populations humaines que traduit leur proximité génétique décroissante à mesure que l'on s'éloigne dans le temps. Enfin et surtout, cette reconstruction est compatible avec des données indépendantes de la génétique.

En effet, cette vision de l'histoire de l'homme moderne est confortée par les recherches en linguistique qui aboutissent *grosso modo* à une filiation des langues qui recouvre à peu près celle des gènes, ce qu'on n'attend nullement dans le cadre de la théorie polycentrique.

C'est un argument considérable en faveur d'une conception partagée aujourd'hui par le plus grand nombre de généticiens et d'anthropologues : après son apparition dans une zone large, entre l'Arabie, l'Afrique du nord ou l'Éthiopie (voir carte, figure 1.9), à une date se situant autour de –100 000 ans, l'homme moderne a très rapidement envahi l'ancien monde, puis le nouveau, en remplaçant quand ils existaient encore, les descendants des formes plus primitives, ou peut-être aussi en en absorbant quelques-uns. Cette absorption sans doute marginale dans l'histoire n'est peut-être pas perceptible à l'échelle des échantillons étudiés mais pourrait cependant rendre compte de la persistance de quelques caractères locaux.

Durant tout le paléolithique, les migrations par scissions accompagnées de l'effet fondateur et de la dérive génétique (voir chapitre 5) vont activement mais aléatoirement modifier les fréquences alléliques des gènes, de sorte que deux groupes éloignés, par le temps et la distance géographique, se retrouveront génétiquement éloignés aujourd'hui, quand on estime leur distance génétique moyenne sur un grand nombre de gènes.

Évidemment l'effet de la dérive cessa dès l'émergence de l'agriculture et de la domestication, avec l'augmentation brutale de l'effectif des populations. La sélection a agi de son côté sur quelques gènes (voir chapitre 7) en opposition ou en synergie avec la dérive (voir chapitre 8).

On pense qu'à la différence des néandertaliens, l'homme moderne disposait d'un langage dont la diversité et l'efficacité allait lui permettre un décollage culturel et les moyens de conquérir le monde en 80 000 ans.

1.4.3 Annexes

a) L'étude de l'ADN mitochondrial et la théorie de l'Ève africaine

En étudiant la diversité génétique de l'ADN mitochondrial à travers les populations humaines, Cann, en 1987, puis Vigilant, en 1991, ont prétendu faire remonter l'ensemble de l'humanité à une femme africaine, d'où la célèbre formule de l'Ève africaine, particulièrement forte pour les États-Unis.

Il faut savoir que seules les mitochondries maternelles sont transmises de générations en générations puisque le spermatozoïde ne fournit qu'un noyau dépourvu de cytoplasme. Il est intéressant de noter que tous les enfants sont identiques pour cet ADN à transmission maternelle, ce qui n'est évidemment pas le cas pour les gènes nucléaires.

Avec le temps des mutations apparaissent dans l'ADN mitochondrial, comme dans tout ADN. Partant du principe que deux individus sont d'autant plus éloignés dans le temps que leurs ADN mitochondriaux sont différents, Cann, puis Vigilant ont établi une arborescence, comme l'a fait Cavalli-Sforza pour la diversité des gènes nucléaires (avec une autre méthode statistique).

Le problème posé par les deux études successives de Cann et Vigilant est très simple : elles démontrent ce dont *a priori* les auteurs étaient persuadés !

En effet, l'arbre phylogénétique de l'humanité enfouit ses racines dans l'Afrique, mais Maddison, avec les données de Cann, ou Templeton ou Hedge, avec les données de Vigilant, ont montré qu'il y a des milliers d'autres arbres tout aussi probables, dont certains sans racine africaine : comment s'y retrouver dans une telle forêt ?

Par ailleurs la taille et la nature des échantillons étudiés laissent rêveur sur le sérieux des comités de lecture des journaux scientifiques : l'Afrique était représentée par 20 individus dans l'étude de Cann, dont seulement deux natifs d'Afrique et 18 noirs américains ; l'Australie était représentée par un seul individu dans l'étude de Vigilant.

Ceci prouve simplement que les études pour moléculaires qu'elles soient n'apportent rien en l'absence d'une méthode adéquate d'analyse des données. Les confrontations entre données paléontologiques, archéologiques, linguistiques et génétiques restent une nécessité et ne pourront guère permettre que la définition du scénario le « plus vraisemblable » avec une « fourchette large de datation ».

b) La reconstruction phylogénétique de l'homme moderne et sa traduction géographique

Il est possible de quantifier la similitude génétique entre populations et de les grouper progressivement sous la forme d'un arbre phylogénétique (sous l'hypothèse *a priori* qu'elles descendent toutes d'une population originelle). L'arbre présenté figure 1.8 est simplifié. Ce sont des arguments paléontologiques et archéologiques qui permettent de « décider » que la racine est africaine.

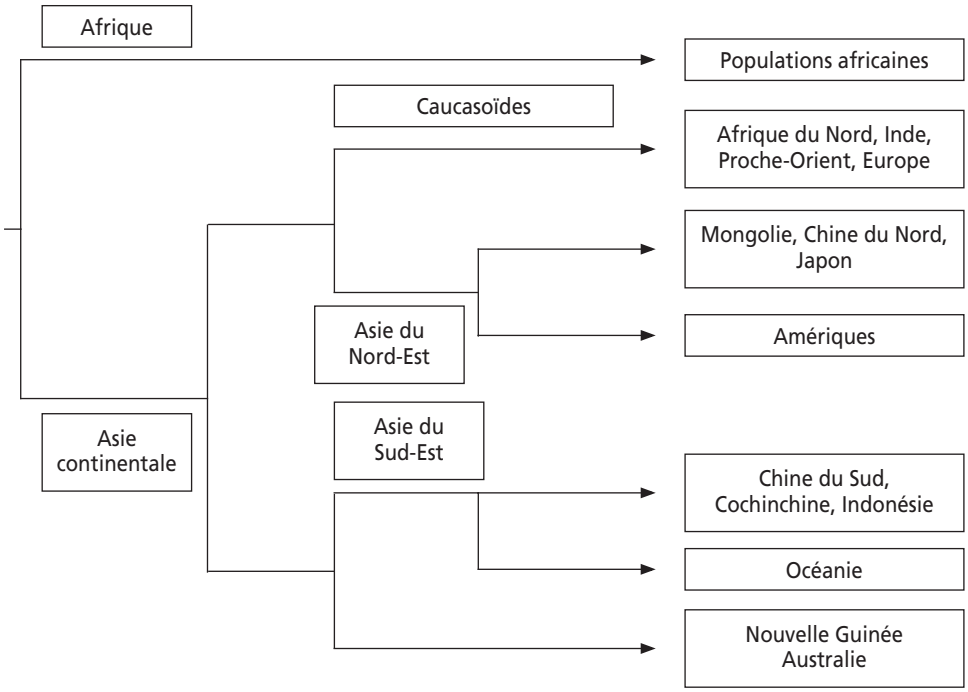


Figure 1.8

On a reporté sur la carte géographique (figure 1.9) une représentation des migrations avec les dates où elles semblent s'être produites, compte tenu de l'échelle des distances génétiques entre les populations.

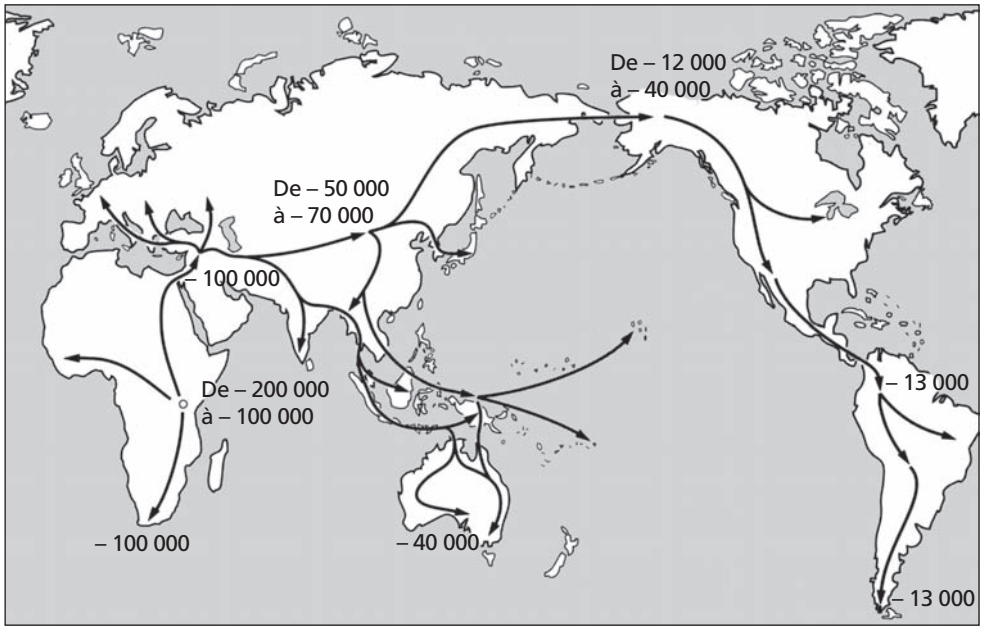


Figure 1.9

c) L'apport de la paléontologie sur les rapports entre néandertaliens et cro-magnons

La découverte, en 1856, du squelette de Néandertal, le premier homme fossile différent de l'homme actuel (*Homo sapiens sapiens*), atteste d'une présence humaine, en Europe, avec outils et rites funéraires dès 35 000 à 40 000 ans avant notre ère.

Les fossiles, découverts en Europe au XIX^e^ siècle, de type néandertal ou cro-magnon (homme moderne) ont conduit de nombreux paléontologues de l'époque à établir un lien presque linéaire entre des stades d'évolution biologique associés à la structure du squelette, notamment du crâne, et le degré de développement de l'humanité associés aux objets ou aux fresques retrouvées dans le même niveau de fouilles.

En fait la généralisation des fouilles à l'ensemble de l'ancien monde a permis de dater des hommes modernes beaucoup plus anciens que le Néandertal européen, d'abord en Palestine, dans la grotte de Qafzeh, puis en Afrique sub-saharienne, autour de 100 000 ans, et en Chine, où un fossile de caractère moderne, dont l'âge remonte à 63 000 ans fut découvert en 1988.

La découverte de Qafzeh a été importante parce qu'elle a cassé deux conceptions :

– l'européocentrisme qui voulait voir en Europe, voire en France, le berceau de l'homme moderne ;

– la relation classiquement établie entre degré d'évolution biologique et degré d'évolution culturelle : les hommes modernes de Palestine présentaient à peu près le même type de culture (moustérienne) que ceux de Néandertal, sans être des hommes de Néandertal !

RÉSUMÉ

La diversité entre individus d'une espèce ou d'une population peut provenir de leur diversité génétique mais aussi de la diversité des milieux au sein desquels s'expriment les gènes.

L'étude de la diversité génétique des individus au sein d'une population suppose qu'elle se traduise par une diversité phénotypique perceptible, au niveau morphologique ou à tout autre niveau physiologique, cellulaire ou biochimique, notamment pour les marqueurs polymorphes de l'ADN.

Pour tout gène, la diversité génétique d'une population est appréhendée par sa composition génétique. Celle-ci est, définie à trois niveaux, celui des phénotypes, avec les fréquences phénotypiques, celui des génotypes, avec les fréquences génotypiques, celui des allèles du gène, avec les fréquences alléliques.

Quand les phénotypes résultant de l'action d'un gène sont codominants, il est facile d'y associer les génotypes correspondants et d'en déduire directement les fréquences alléliques. Quand des phénotypes sont récessifs et d'autres sont dominants, le calcul direct des fréquences alléliques est impossible puisqu'on ne peut pas distinguer, pour les phénotypes dominants, les individus homozygotes et les individus hétérozygotes. Le recours au modèle de Hardy-Weinberg se révèlera alors fort utile. Dans tous les cas, le modèle de Hardy-Weinberg est le seul moyen d'accès à l'estimation des fréquences génotypiques quand on ne connaît que les fréquences alléliques.

L'étude de la diversité génétique au sein des populations conduit à l'estimation de deux paramètres : le degré de polymorphisme et le degré d'hétérozygotie.

L'estimation de la diversité dans et entre les populations humaines a montré l'inadéquation du concept de race dans l'espèce humaine :

– toutes les populations partagent, pour tous les gènes de l'espèce, le même répertoire allélique ; seules les fréquences alléliques diffèrent d'une population à l'autre ;
– 92 % de la diversité génétique totale de l'humanité est encore présente au sein de chacun des trois grands groupes asiatiques, africain et caucasien. Les différences génétiques entre ces groupes ne représentent que 8 % de la diversité génétique totale au sein de l'espèce ;
– 85 % de la diversité génétique totale de l'humanité est encore présente au sein des groupes nationaux. Les différences génétiques entre nations ne représentent que 7 % de la diversité génétique au sein d'un des trois groupes continentaux et 15 % de la diversité génétique totale au sein de l'espèce.

L'estimation de la diversité dans et entre les populations humaines a joué un rôle important dans les débats entre paléontologues, anthropologues et linguistes, sur l'âge, l'origine et l'histoire de l'homme moderne (*Homo sapiens sapiens*). Compte tenu des informations aujourd'hui disponibles dans tous ces domaines, le scénario le plus vraisemblable est le suivant :

– l'homme moderne serait apparu, il y a environ 100 000 ans, dans une population de formes plus archaïques, quelque part entre l'Afrique du nord, l'Éthiopie et la péninsule arabique ;
– il aurait rapidement envahi l'ancien monde, Afrique, Asie du sud-est et Australie autour de –50 000 ans, puis l'Europe du Sud, l'Asie continentale et la Chine du Sud autour de – 40 000 ans, la Chine du Nord et le Japon vers – 30 000, l'Europe du Nord et l'Amérique entre – 20 et – 10 000 ans.

EXERCICES

Exercice 1.1

Au Kenya, quatre enfants sur cent sont homozygotes pour la mutation drépanocytaire β^S et sont atteints de drépanocytose quelques mois après la naissance (à la naissance 50 % de l'hémoglobine est encore de type fœtal).

Question 1 : que peut-on estimer de la composition génétique de cette population aux différents niveaux hiérarchiques de la diversité ?

Une analyse sur sang du cordon a parallèlement été réalisée sur un échantillon aléatoire de nouveaux nés, permettant de montrer que 64 % des enfants ne possèdent que de l'hémoglobine A, 4 % ne possèdent que de l'hémoglobine S, les autres étant porteurs des deux types d'hémoglobine.

Question 2 : cette information complémentaire permet-elle de répondre avec plus de précision à la question 1 ?

Solution

Les données sont résumées dans le tableau 1.5.

Fréquences phénotypiques	Phénotypes électrophorétiques	Génotypes	Phénotypes cliniques	Fréquences phénotypiques
64 % (*D*)	[HbA]	β^+/β^+	sain	96 % (*P*)
32 % (*H*)	[HbA + HbS]	β^+/β^S	sain	
4 % (*R*)	[HbS]	β^S/β^S	atteint	4 % (*Q*)

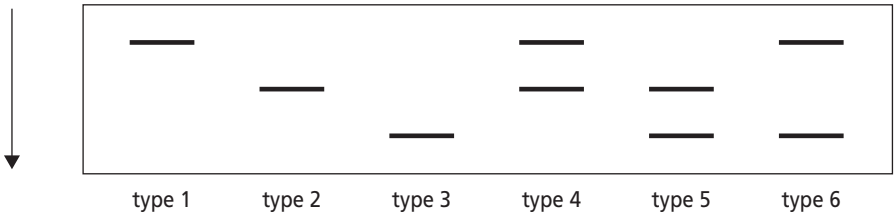
Question 1 : les fréquences phénotypiques *P* et *Q* des phénotypes cliniques pour le caractère « maladie » ne permettent pas d'en déduire les fréquences génotypiques du fait que *P* est la somme de deux fréquences génotypiques de valeur inconnue. On ne peut donc pas non plus calculer les fréquences alléliques.

Question 2 : avec le caractère « migration électrophorétiques », il est possible de distinguer trois phénotypes correspondant aux trois génotypes possibles, car les génotypes β^+/β^+ et β^+/β^S présentent des phénotypes distincts. Les fréquences génotypiques sont alors connues et le calcul de la fréquence de la mutation drépanocytaire est immédiat par l'utilisation des fréquences génotypiques égales aux fréquences des phénotypes électrophorétiques codominants *D*, *H* et *R*.

D'où : $f(\beta^S) = R + H/2 = 0{,}04 + 0{,}016 = 0{,}2$;
$f(\beta^S) = 20$ % (ce qui est considérable pour une mutation létale) ;
$f(\beta^+) = 80$ %.

Exercice 1.2

On a entrepris l'étude, chez *Drosophila melanogaster*, des variants électrophorétiques relatifs au gène codant pour une enzyme E. Cette étude a révélé l'existence des 6 phénotypes de mobilité électrophorétique suivants :



L'analyse d'un échantillon de 500 femelles et de 500 mâles extraits d'une population naturelle a fourni la distribution phénotypique suivante :

	Femelles	Mâles
Type 1	175	290
Type 2	47	155
Type 3	3	55
Type 4	185	0
Type 5	35	0
Type 6	55	0

Question 1 : l'observation des différents phénotypes de mobilité électrophorétique permet-elle de localiser le gène *E* sur un chromosome particulier ?

Question 2 : ces observations peuvent-elles être compatibles avec l'existence, dans la population, d'un allèle muté amorphe (sans expression, donc pas de produit) ?

Question 3 : peut-on estimer la composition génétique dans chacun des sexes ?

Solution

Question 1 : les phénotypes électrophorétiques sont codominants et la présence de deux bandes différentes constitue le phénotype attendu pour un hétérozygote porteur de deux allèles différents.

La présence d'une seule bande constitue le phénotype attendu pour un homozygote porteur de deux allèles identiques. Le fait que tous les mâles ne présentent qu'une seule bande et jamais deux conduit à la conclusion que le gène *E* est sur le chromosome X ; les mâles sont donc hémizygotes, ils ne possèdent qu'un seul exemplaire du gène *E* et un phénotype du type 1, 2 ou 3.

Question 2 : un génotype hétérozygote pour l'allèle amorphe conduirait à un phénotype à une bande de type du type 1, 2 ou 3 chez les femelles, simulant ainsi les phénotypes associés à des homozygotes, mais on devrait alors observer des phénotypes sans bandes chez les mâles, ce qui n'est pas le cas : il n'y a donc aucun allèle de ce type, du moins dans l'échantillon observé (ou exceptionnellement quelques exemplaires chez des femelles).

Question 3 :

Fréquences des allèles chez les mâles hémizygotes :

$f(A1) = 290/500 = 0{,}58$ $f(A2) = 155/500 = 0{,}31$ $f(A3) = 55/500 = 0{,}11$

Fréquences des allèles chez les femelles où les phénotypes sont codominants :

$f(A1) = [175 + (185+ 55)/2]/500 = 0{,}59$

$f(A2) = [47 + (185+35)/2]/500 = 0{,}314$

$f(A3) = [3+ (35+55)/2]/500 = 0{,}096$

Chapitre **2**

Le modèle général de Hardy-Weinberg

2.1 LE MODÈLE DE HARDY-WEINBERG ET LA NAISSANCE DE LA GÉNÉTIQUE DES POPULATIONS

Une anecdote est associée à l'établissement de la loi de Hardy qui allait conduire à « l'acte de naissance » de la génétique des populations. En 1908, le biologiste Punnet, faisant une conférence, à la Royal Academy de Londres, sur la transmission mendélienne de certains traits chez l'homme, se trouva interpellé par un contradicteur, G.-U. Yule, un biométricien encore partisan des lois de l'hérédité de Darwin et Galton. Celui-ci lui fit remarquer qu'il était paradoxal de ne jamais rencontrer chez l'homme les fameuses proportions 3/4 – 1/4 établies par Mendel chez le pois, et retrouvées par Bateson chez le coq ou le hamster, ou le français Cuénot chez la souris. Punnet ne sut pas bien répondre sur le moment bien qu'il entrevît que ces proportions 3/4-1/4 étaient uniquement celles de la descendance d'une F1 issue du croisement expérimental de deux souches pures, alors que, dans les populations naturelles, coexistaient les descendances de nombreux types de croisements différents. Il convenait donc de pondérer les proportions de descendants obtenues dans chaque type de croisement par la fréquence de celui-ci (voir paragraphe suivant). Il s'en ouvrit à un ami physicien, Hardy, assez à l'aise avec l'usage des probabilités. Hardy lui fournit immédiatement la solution sous la forme d'une équation donnant la fréquence des hétérozygotes dans une population naturelle, sur la base d'une transmission mendélienne des gènes et de conditions régissant la formation des couples. Au même moment le physiologiste allemand Weinberg aboutissait indépendamment au même résultat. L'observation répétée de la validité du modèle de Hardy-Weinberg au sein des populations naturelles joua alors un rôle important dans la validation du mendélisme en sortant celui-ci du laboratoire, « hors d'un milieu

contrôlé par l'homme ». Par ailleurs quelques jeunes biométriciens réalisèrent rapidement que ce modèle, en permettant d'expliquer le maintien du polymorphisme génétique, enlevait aux non-darwiniens leur argument essentiel contre l'effet de la sélection. Dès ce moment, au sein même du darwinisme, les esprits étaient prêts à l'abandon du vieux concept d'hérédité par mélange et à la conversion radicale en faveur du mendélisme. Ce fut chose faite par Fisher, dans un article célèbre de 1918, qu'il n'osa cependant publier que dans une revue secondaire d'Édimbourg, conscient de la « trahison » qu'il opérait vis-à-vis du courant darwiniste et biométricien auquel il appartenait.

2.2 LE TRANSFERT DES GÈNES D'UNE GÉNÉRATION À L'AUTRE SUIT LES ÉTAPES DU CYCLE VITAL

Le modèle de Hardy-Weinberg repose sur plusieurs conditions qu'il sera utile d'introduire en considérant le cycle vital d'un organisme à sexes séparés[1], entre les stades adultes reproducteurs de deux générations successives (figure 2.1). Les étapes critiques du cycle vital pour lesquelles chaque condition du modèle de Hardy-Weinberg sera introduite, ont été numérotées. Leur légitimité sera discutée dans un second temps.

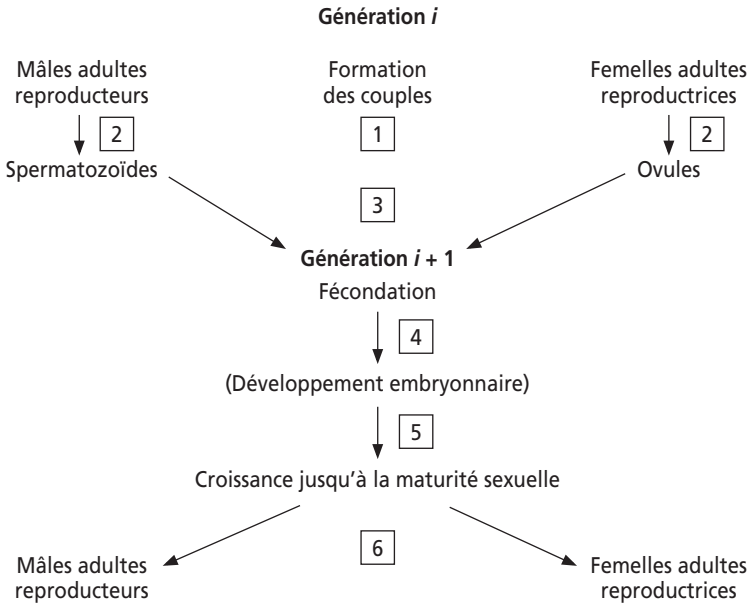


Figure 2.1 Étapes du cycle vital où sont introduites des conditions du modèle de Hardy-Weinberg.

1. Le cas des organismes à sexes non séparés, présentant une autofécondation totale ou partielle et les autres modes de reproduction particuliers seront traités ultérieurement ; ils apparaissent comme des écarts à la panmixie par rapport au modèle général de Hardy-Weinberg (voir chapitre 4).

2.3 LE MODÈLE DE HARDY-WEINBERG

Le modèle théorique général de Hardy-Weinberg sera établi dans le cas le plus simple, le plus général et le plus utile d'un gène autosomique di-allélique, pour une population d'organismes à sexes séparés, présentant des générations séparées.

La généralisation du modèle à un cas multi-allélique, le cas particulier des gènes liés au sexe, le cas des générations chevauchantes ou le cas de deux gènes étudiés simultanément seront exposés au chapitre 3.

Considérons une population constituée des adultes reproducteurs de la génération i.

Dans le cas d'un gène di-allélique (allèles *A1* et *A2*), la composition génétique de la population est constituée des :

– **trois génotypes possibles :** ***A1/A1*** ***A1/A2*** ***A2/A2***

– **de fréquences génotypiques :** ***D*** ***H*** ***R***

Ces fréquences *D*, *H* et *R* sont quelconques et seront considérées, en première analyse, égales dans les deux sexes.

On sait (voir chapitre 1) en déduire les fréquences des allèles *A1* et *A2*, respectivement nommées *p* et *q*, soit :

$$p = D + H/2$$

$$q = R + H/2$$

Quelle sera la constitution génétique (fréquences génotypiques et fréquences alléliques), à la génération suivante après un cycle vital ?

Il suffit de se reporter au cycle vital présenté dans le paragraphe précédent pour obtenir la solution en fonction des conditions que nous aurons dû définir et qui forment les « conditions de l'équilibre de Hardy-Weinberg ».

On peut établir ce modèle de deux manières différentes, soit par la formation des couples en suivant pas à pas le cycle vital, soit par le schéma de l'urne gamétique.

2.3.1 Établissement du modèle de Hardy-Weinberg par le cycle vital

a) Formation des couples : condition de panmixie

La première étape du cycle vital (**1** dans l'encadré ci-dessus) est la formation de couples reproducteurs. Des règles d'union peuvent exister : on fera l'hypothèse qu'ils se forment au hasard, les couples sont dits **panmictiques** (**condition de panmixie**).

Dans ce cas, on peut générer six types possibles de couples (tableau 2.1).

> ***Remarque :*** il convient de noter que la panmixie ne signifie pas que les six types de couples sont équiprobables ou équifréquents (1/6) mais que leurs probabilités respectives sont fonction de la fréquence, dans la population, de chacun des génotypes associés dans le couple. Pour prendre un exemple caricatural les couples noirs × noirs ne peuvent pas avoir la même probabilité ou la même fréquence dans un pays comme la Suède ou le Sénégal.

b) Probabilité et fréquences des évènements : condition d'effectif infini de la population

Dans une population naturelle concrète, ce qui nous importe, et ce qui compte, ce sont les fréquences des couples, les fréquences de leurs descendants, les fréquences des génotypes, les fréquences alléliques, et non les probabilités de ces évènements.

On sait que la fréquence d'un évènement est égale à sa probabilité si le nombre de tirages est très grand (loi des grands nombres) ; par exemple la fréquence des « piles » peut être égale à 0,7 sur dix tirages mais ne peut, sur 100 000 tirages, s'écarter notablement de sa probabilité égale à 0,5.

Afin de pouvoir considérer que les fréquences des couples ou des génotypes sont égales à leurs probabilités respectives, nous considérerons que la population est de taille infinie (concrètement suffisamment grande pour y appliquer la loi des grands nombres aux évènements étudiés). Le seuil grand/petit n'est pas définissable dans l'absolu et ne sera discuté qu'à la fin de l'ouvrage (voir chapitre 8).

Cette condition d'effectif infini s'ajoute à la condition de panmixie. Le tableau de formation des couples et de leurs descendants se présente alors ainsi :

Tableau 2.1 Fréquences des couples panmictiques et de leurs descendants pour un gène di-allélique.

Types de couples	Fréquences des couples	Fréquences des descendants *A1/A1*	Fréquences des descendants *A1/A2*	Fréquences des descendants *A2/A2*
A1/A1* x *A1/A1	D^2	1	0	0
A1/A1* x *A1/A2	$2DH$	1/2	1/2	0
A1/A1* x *A2/A2	$2DR$	0	1	0
A1/A2* x *A1/A2	H^2	1/4	1/2	1/4
A1/A2* x *A2/A2	$2RH$	0	1/2	1/2
A2/A2* x *A2/A2	R^2	0	0	1
TOTAL	1	$D^2 + DH + H^2/4$	$DH + 2DR + H^2/2 + RH$	$R^2 + RH + H^2/4$

Maintenant que les couples sont formés et leurs fréquences connues (sous les conditions de panmixie et d'effectif infini), il s'agit de réaliser les fécondations pour obtenir les adultes reproducteurs de la génération suivante de manière à avoir réalisé le cycle vital d'une génération.

c) Gamétogenèse : condition d'absence de mutations

Les individus formant les couples produisent des gamètes (étape **2** dans le cycle vital). Pour simplifier, on négligera l'effet des mutations dans le gène considéré. De ce fait on peut considérer comme nulle la fréquence des descendants *A1*/A2 chez les parents du premier type (première ligne du tableau ci-dessus), et écrire les proportions des lignes 2, 4, 7 du tableau.

d) Fécondation : condition d'absence de sélection gamétique

Les deux types de gamètes formés par les hétérozygotes sont théoriquement dans les proportions 50 :50, mais lors de la fécondation (**4** dans le cycle vital), ces proportions peuvent avoir été modifiées si un mécanisme de sélection gamétique (**3** dans le cycle vital) a joué en faveur de l'un et en défaveur de l'autre. En excluant cette éventualité (condition de l'absence de sélection gamétique), on peut écrire les proportions mendéliennes réalisées à la fécondation dans les lignes 3, 5, 6 du tableau.

e) Développement et croissance des descendants : condition d'absence de sélection zygotique

Les proportions des génotypes chez les descendants, à la fécondation, ne sont maintenues, au stade final de reproducteur adulte, que si une nouvelle condition est réalisée : l'absence de sélection zygotique (pendant l'embryogenèse ou la croissance ; **5** dans le cycle vital).

Autrement dit les trois génotypes sont supposés avoir la même espérance de vie et ne présentent, du moins pour le gène considéré, aucune mortalité différentielle.

f) Fréquences des génotypes chez les adultes reproducteurs de la génération suivante : condition d'absence de sélection et de migration

On peut calculer les fréquences génotypiques des descendants au stade d'adultes reproducteurs (sous toutes les conditions précédentes) en sommant, pour chaque génotype, le produit de la fréquence des couples dont il peut être issu par la fréquence de réalisation de ce génotype au sein de ces couples.

On obtient ainsi les trois sommes figurant en bas du tableau, mais cela suppose encore deux dernières conditions pour qu'on puisse les considérer comme les fréquences génotypiques des adultes reproducteurs de la génération $i + 1$:

– il faut d'abord considérer que les couples sont également fertiles (condition d'absence de sélection en terme de fertilité différentielle). Dans le cas contraire il faudrait pondérer chacun des termes de la somme par la fertilité respective des couples ;
– il faut ensuite considérer que ces sommes, ces fréquences génotypiques, n'ont pas été modifiées dans le temps par l'adjonction à la population d'individus extérieurs (condition d'absence de migrations). En effet un apport extérieur a peu de chance de présenter la même constitution génétique.

Sous toutes ces hypothèses, les nouvelles fréquences génotypiques deviennent :

pour les trois génotypes :		***A1/A1***	***A1/A2***	***A2/A2***
égales à :		$\boldsymbol{p^2}$	$\boldsymbol{2pq}$	$\boldsymbol{q^2}$
En effet	si $p = D + H/2$	alors	$D^2 + DH + H^2/4 = p^2$	
et	si $q = R + H/2$	alors	$R^2 + RH + R^2/4 = q^2$	

2.3.2 Établissement du modèle de Hardy-Weinberg par le schéma de l'urne gamétique

a) Panmixie et pangamie : schéma de l'urne gamétique

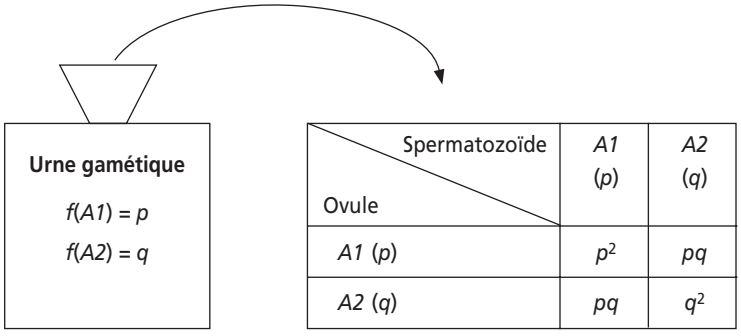
Si on considère que les couples se forment au hasard (étape **1** du cycle vital), c'est-à-dire qu'ils sont panmictiques (condition de panmixie), et que les gamètes eux-mêmes s'unissent au hasard (condition de pangamie) lors de la fécondation (**4** dans l'encadré du cycle vital), on peut alors admettre que tout se passe comme si les couples mettaient en vrac leurs gamètes dans une urne gamétique de spermatozoïdes pour les mâles et d'ovules pour les femelles, et que tout descendant est issu de l'union de deux gamètes tirés au hasard dans chacune des deux urnes. Ce schéma est nommé schéma de l'urne gamétique (figure 2.2). C'est d'ailleurs la réalité biologique pour les espèces végétales, et de nombreuses espèces animales aquatiques, qui émettent leurs gamètes dans le milieu environnant.

> ***Remarque :*** sauf exception la panmixie s'accompagne toujours de la pangamie et on considère que la condition de panmixie inclue celle de la pangamie.

Si les fréquences alléliques sont les mêmes dans les deux sexes, les urnes sont identiques (on verra plus tard le cas où elles ne le sont pas).

Compte tenu de la composition des urnes il sera alors possible de calculer la probabilité de tirage de chacun des trois génotypes possibles.

D'où le schéma suivant :



Spermatozoïde / Ovule	*A1* (*p*)	*A2* (*q*)
A1 (*p*)	p^2	*pq*
A2 (*q*)	*pq*	q^2

Figure 2.2 Schéma de l'urne gamétique et conséquences mathématiques sur la diversité génétique.

Il faut cependant considérer que les valeurs p^2, $2pq$, q^2, obtenues si facilement par ce double tirage, ne sont une réalité biologique, c'est-à-dire les fréquences de génotypes adultes reproducteurs à la génération suivante, que si des conditions additionnelles sont réalisées.

b) Conditions additionnelles

La composition génétique de l'urne selon le schéma précédent suppose que chacun des adultes reproducteurs y a placé un nombre égal de gamètes, ce qui signifie qu'il y a **absence de sélection en matière de fertilité**.

Il suppose aussi que les homozygotes *A1A1* n'y déposent que des gamètes *A1*, ce qui signifie l'**absence de mutations**.

Une fois obtenue par le dépôt des gamètes, la composition de l'urne est stable, ce qui signifie l'**absence de sélection gamétique**.

Le tirage de deux gamètes aboutit alors et seulement alors à la formation d'un génotype :

	A1/*A1*	avec	la probabilité p^2 ;
ou	*A1*/*A2*	avec	la probabilité $2pq$;
ou	*A2*/*A2*	avec	la probabilité q^2.

En fait nous nous intéresserons aux fréquences des génotypes des descendants et non aux seules probabilités de les concevoir. On sait que la fréquence d'un évènement est égale à sa probabilité si le nombre de tirages est très grand (loi des grands nombres) ; par exemple la fréquence des « piles » peut être égale à 0,7 sur dix tirages mais ne peut, sur 100 000 tirages, s'écarter notablement de sa probabilité égale à 0,5.

Afin de pouvoir considérer que les fréquences des génotypes obtenus sont égales à leurs probabilités respectives, nous considérerons que la population est de taille infinie (concrètement suffisamment grande pour y appliquer la loi des grands nombres aux évènements étudiés). Le seuil grand/petit n'est pas définissable dans l'absolu et ne sera discuté qu'à la fin de l'ouvrage (voir chapitre 8).

Cette condition de taille infinie s'ajoute aux conditions de panmixie, d'absence de mutations, de sélection gamétique et d'égale fertilité des parents.

Les proportions $p^2/2pq/q^2$ ne sont maintenues de la fécondation au stade final d'adultes reproducteurs que si deux autres conditions sont également réalisées :

– l'**absence de sélection zygotique** (pendant l'embryogenèse ou la croissance). Autrement dit les trois génotypes sont supposés avoir la même espérance de vie et ne présentent, du moins pour le gène considéré, aucune mortalité différentielle ;
– l'**absence de migrations**. En effet un apport extérieur a peu de chance de présenter la même constitution génétique ($p^2/2pq/q^2$).

Au bout du compte on aboutit évidemment au même résultat que celui obtenu par la première méthode.

2.3.3 Bilan du modèle de Hardy-Weinberg

a) La relation de Hardy-Weinberg

Les nouvelles fréquences génotypiques correspondent soit au carré des fréquences alléliques pour les homozygotes, soit au double produit des fréquences alléliques pour l'hétérozygote.

Les trois génotypes sont	***A1*/*A1***	***A1*/*A2***	***A2*/*A2***
Leurs fréquences sont égales à	p^2	$2pq$	q^2

La relation ainsi établie entre les fréquences alléliques et les fréquences génotypiques est appelée « relation de Hardy-Weinberg » ou « relation panmictique » car elle découle directement de l'hypothèse panmictique.

> ***Remarque :*** cette relation mathématique permet, en supposant que l'hypothèse panmictique soit valide, de remonter aux fréquences génotypiques (donc phénotypiques) quand on ne connaît que les fréquences alléliques (voir figure 1.6 et paragraphe 2.3.5.c).

b) L'équilibre de Hardy-Weinberg

Les fréquences alléliques sont inchangées à la génération suivante.

En effet, selon la formule de calcul des fréquences alléliques à partir des fréquences génotypiques (voir chapitre précédent), on a :

$$\boldsymbol{f(A1)} = p^2 + 2pq/2 = p^2 + pq = p\,(p + q) = \boldsymbol{p}$$

$$\boldsymbol{f(A2)} = q^2 + 2pq/2 = \mathrm{q}^2 + pq = q\,(p + q) = \boldsymbol{q}$$

Cette stabilité de la composition génétique de la population est appelée « équilibre de Hardy-Weinberg ».

c) Les conditions de l'équilibre de Hardy-Weinberg

Les conditions supposées réalisées dans le modèle de l'équilibre de Hardy-Weinberg peuvent se regrouper en trois grands groupes :

condition 1 : la population est panmictique ;
condition 2 : la population est de taille quasi infinie (loi des grands nombres applicable) ;
condition 3 : mutation, sélection, migration sont inexistantes (ou négligeables).

2.3.4 Légitimité des conditions du modèle de Hardy-Weinberg

Le modèle de Hardy-Weinberg est le modèle central de la génétique des populations. Pourtant le nombre et l'importance des conditions sous-jacentes devraient le faire apparaître comme un modèle théorique, abstrait et irréaliste. On sait bien que les mutations existent et sont la source de la variation génétique ayant permis la sélection et l'évolution. Pourtant l'étude de la plupart des gènes dans les populations naturelles donne des résultats compatibles avec ce modèle. Comment expliquer ce paradoxe ?

Tout simplement par le fait que certaines des conditions peuvent parfaitement être réalisées et que les autres, bien qu'illégitimes, n'ont d'effet perceptible que sur une longue échelle de temps. On peut donc les négliger sur l'espace de quelques générations.

Reprenons ces conditions :

a) la panmixie est la condition la plus facilement réalisée. Dans la plupart des espèces les croisements et les fécondations sont réellement aléatoires pour la

plus grande partie des gènes. Il y a évidemment quelques exceptions, comme les allèles d'incompatibilité chez certaines espèces végétales. Mais, dans ce cas, les croisements ne sont pas panmictiques pour ces seuls gènes (et les gènes qui leur sont génétiquement liés, voir chapitre 3), mais sont panmictiques pour le reste du génome.

Chez l'homme les unions ne sont pas panmictiques pour les gènes qui gouvernent la pigmentation, en raison de la discrimination raciale plus ou moins grande existant dans toutes les populations, ou pour les gènes gouvernant la taille parce que les unions tendent à associer préférentiellement un homme de taille égale ou supérieure à la femme. Mais pour tous les autres gènes, ceux qui gouvernent les groupes sanguins, les facteurs sériques ou les maladies, les unions sont panmictiques comme cela a été démontré dans la plupart des études (voir les problèmes d'application) ;

b) l'absence de mutation n'est pas une condition légitime dans l'absolu mais elle est acceptable en pratique, sur quelques générations, car, comme on le verra plus loin, l'effet des mutations sur la diversité génétique et la constitution génétique des populations, se mesure sur des centaines ou des milliers de générations ;

c) il en est de même pour l'effet de la sélection naturelle. Ce n'est évidemment pas le cas pour la sélection artificielle opérée par les agronomes afin de sélectionner races et variétés animales ou végétales ; mais dans ce cas les croisements ne sont plus panmictiques et les populations en question ne sont plus naturelles ;

d) l'absence de migration peut être acceptable pour des populations végétales ou animales réellement isolées, mais c'est une condition dont il est toujours difficile de vérifier la validité. Cette condition est évidemment légitime si on considère l'espèce dans sa globalité.

L'absence de migrations est une condition particulièrement discutable pour ce qui concerne les populations humaines (même dans les isolats). On verra de plus que les migrations, contrairement à la sélection ou aux mutations, peuvent modifier la constitution génétique des populations en peu de générations. Selon la population, le gène étudié, les conditions de l'étude ou la question posée, l'effet des migrations devra être pris en compte ou pourra être négligé ;

e) la taille infinie est par essence une condition impossible. On verra que les populations de petite taille sont le siège d'un phénomène spécifique appelé dérive génétique. On verra ensuite que les effets de la dérive et de la sélection peuvent se combiner dans l'histoire des populations. On considérera alors qu'une population est « grande » quand l'effet de la dérive est suffisamment négligeable devant celui de la sélection pour ne faire dépendre son évolution génétique que des effets sélectifs. Inversement une population sera « petite » dès lors que son parcours évolutif, malgré les effets sélectifs, peut aussi dépendre de la dérive.

On peut donc admettre en conclusion que le modèle de Hardy-Weinberg, malgré ses conditions, est acceptable sur le temps de quelques générations, avec cependant une attention particulière à accorder à l'effet des migrations.

2.3.5 L'équilibre de Hardy-Weinberg

a) Mise en évidence des situations d'équilibres allélique et génotypique

Le modèle de Hardy-Weinberg tel qu'il est défini permet de conclure qu'une population panmictique, de grande taille, sans mutations, ni sélection, ni migration, maintient son polymorphisme génétique en l'état.

En effet les fréquences alléliques, p et q à la génération i, n'ont pas varié à la génération suivante. La diversité en termes de fréquences alléliques est donc stable.

Par contre les fréquences génotypiques D, H et R, choisies quelconques à la génération i, ont varié pour prendre des valeurs particulières p^2, $2pq$, q^2 qui représentent $(p + q)^2$ le développement du carré de la somme des fréquences alléliques.

Cette relation entre les fréquences alléliques p et q d'une part et les fréquences génotypiques p^2, $2pq$, q^2 d'autre part, est une conséquence directe de la panmixie qui revient à réaliser les fécondations comme deux tirages aléatoires indépendants dans une urne (figure 2.2).

Cette relation de Hardy-Weinberg est très utile car elle va permettre d'estimer les fréquences alléliques dans l'étude d'un gène présentant, selon les allèles, des phénotypes dominants ou récessifs (voir figure 1.6 et tableau 2.2).

Dès que les fréquences génotypiques ont les valeurs p^2, $2pq$, q^2, à la génération $i + 1$, elles restent inchangées dans les générations ultérieures tant que les conditions de Hardy-Weinberg sont maintenues. En effet, l'urne gamétique $(p + q)$ restant elle-même inchangée, les fréquences génotypiques qui en sont issues, par panmixie, demeurent égales à p^2, $2pq$, q^2.

Cette situation d'équilibre des fréquences alléliques et des fréquences génotypiques appelée « équilibre de Hardy-Weinberg » est atteinte en une génération pour un gène autosomique.

b) Établissement de l'équilibre quand les fréquences alléliques diffèrent entre sexes

On avait supposé, dans l'établissement du modèle, que les fréquences alléliques étaient les mêmes dans les deux sexes. Si tel n'est pas le cas, il est facile de voir, en reprenant le schéma de l'urne gamétique, qu'il faut une première génération de panmixie pour obtenir des fréquences alléliques égales dans les deux sexes, puis une deuxième évidemment, pour obtenir les fréquences génotypiques de l'équilibre de Hardy-Weinberg.

En effet, supposons qu'à la génération g_0, la composition génétique d'une population soit différente dans chacun des sexes, ce qui peut arriver lors d'une fusion de population, on aura :

Sexe	femelle				mâle		
Génotypes	*A1A1*	*A1A2*	*A2A2*		*A1A1*	*A1A2*	*A2A2*
Fréquences	D	H	R		d	h	r
Fréquences alléliques	$p = D + H/2$			et	$u = d + h/2$		
	$q = R + H/2$				$v = r + h/2$		

À la génération suivante g_1, après un cycle de panmixie, associée aux autres conditions, on aura dans chacun des sexes la même composition génétique, donnée par le schéma du tirage dans les deux urnes gamétiques parentales de la génération g_0 :

Génotypes	*A1A1*	*A1A2*	*A2A2*	(mâle ou femelle)
Fréquences	$p.u$	$p.v + q.u$	$q.v$	
Fréquences	$P = p.u + (p.v + q.u)/2$			
alléliques	$Q = q.v + (p.v + q.u)/2$			

Les fréquences alléliques ont pris une valeur différente de celles de la génération précédente, mais égales dans les deux sexes, qui ont désormais la même urne gamétique (P et Q).

Par contre les fréquences génotypiques ne vérifient pas la relation de Hardy-Weinberg ($p.u$ n'est pas égal à P^2, ni $q.v$ à Q^2). Mais dès la génération suivante, g_2, les fréquences génotypiques vérifieront cette relation en étant égales à P^2, $2PQ$ et Q^2, si les conditions de Hardy-Weinberg ont été maintenues.

c) L'équilibre de Hardy-Weinberg n'est pas une situation quelconque

Il convient de noter que pour une même diversité allélique (mêmes fréquences alléliques) il existe une infinité de populations différentes, présentant des fréquences génotypiques différentes, mais qu'il n'en est qu'une seule à l'équilibre de Hardy-Weinberg, présentant entre fréquences alléliques et génotypiques la relation panmictique (tableau 2.2).

Tableau 2.2 Une même diversité allélique, une infinité de diversités génotypiques.

Population	*A1/A1*	*A1/A2*	*A2/A2*	*p = f(A1)*	*q = f(A2)*
1	0,7	0	0,3	0,7	0,3
2	0,6	0,2	0,2	0,7	0,3
3	0,4	0,6	0	0,7	0,3
4	0,5	0,4	0,1	0,7	0,3
5	0,49	0,42	0,09	0,7	0,3

Dans le tableau 2.2, la population 5 est strictement à l'équilibre de Hardy-Weinberg, la population 4 n'en n'est pas éloignée. En pratique on réalisera un test statistique pour statuer sur l'écart entre la diversité génétique observée dans une population et celle attendue sous le modèle théorique de Hardy-Weinberg. Si les écarts ne sont pas significatifs, on admettra que sa constitution génétique est conforme au modèle de Hardy-Weinberg.

d) Signification évolutive du modèle de Hardy-Weinberg

Le modèle de Hardy-Weinberg est fondé sur la conception mendélienne de l'hérédité (ségrégation des allèles à la méiose et réunion de deux allèles de chaque gène à

la fécondation). Il permet, soit de maintenir la diversité génétique (les allèles), notamment en l'absence de sélection, soit de prévoir une variation de la diversité, donc une évolution génétique, notamment en cas de sélection ; toutes choses que ne permettait pas une conception non mendélienne de l'hérédité. On comprend que les jeunes biométriciens darwinistes se soient laissés séduire.

Le développement des modèles de génétique des populations consistera la plupart du temps à partir de ce modèle central de Hardy-Weinberg pour voir comment la composition génétique d'une population peut être modifiée, vers quelle limite, et à quelle vitesse, si telle ou telle des conditions de cet équilibre n'est pas respectée.

2.4 APPLICATION DU MODÈLE DE HARDY-WEINBERG AU CALCUL DES FRÉQUENCES ALLÉLIQUES POUR LES CARACTÈRES PRÉSENTANT DES PHÉNOTYPES RÉCESSIFS

On a vu au chapitre 1 qu'il était facile d'estimer directement les fréquences des différents allèles d'un gène quand les différents génotypes formaient des phénotypes codominants mais que ce calcul se révélait impossible dès qu'un des phénotypes était récessif. C'est alors que le modèle de Hardy-Weinberg présente la grande utilité de fournir le moyen, indirect, d'une telle estimation. Cette utilisation sera illustrée par les quelques exemples qui suivent et dans la plupart des exercices.

2.4.1 Estimation des fréquences alléliques d'un gène responsable de l'albinisme

Il existe plusieurs formes d'albinisme dont chacune est dépendante d'un gène. La forme la plus fréquente résulte d'une mutation touchant le gène de structure de la tyrosinase, la première enzyme de la chaîne de biosynthèse des mélanines. Les individus albinos sont déficients en tyrosinase et sont porteurs de deux copies mutées du gène ; ils sont homozygotes *a/a*. Ce phénotype est récessif car les hétérozygotes *A/a*, ayant une copie fonctionnelle A du gène, sont capables d'assurer la synthèse des mélanines.

On observe, dans une population africaine, un phénotype albinos (par déficience de tyrosine) pour 10 000 habitants.

Tableau 2.3 Diversité génétique pour l'albinisme par déficit de la tyrosinase.

Phénotype	**Pigmenté**		**Albinos**
Valeurs observées des fréquences phénotypiques	$D + H = 9\,999/10\,000$		$R = 1/10\,000$
Génotypes	*A/A*	*A/a*	*a/a*
Fréquences génotypiques si Hardy-Weinberg	p^2	$2pq$	q^2
Valeurs calculées des fréquences génotypiques	9 801/10 000	198/10 000	1/10 000

Comment peut-on définir la composition génétique d'une telle population, c'est-à-dire estimer les valeurs des fréquences alléliques et génotypiques ?

Le tableau 2.3 permet de voir que l'estimation directe des fréquences alléliques est impossible, car on ne peut estimer indépendamment les trois fréquences génotypiques, *D*, *H* et *R* (on ne peut estimer que *D* + *H* d'une part, et *R* d'autre part : 2ième ligne du tableau).

Cependant, si on fait l'hypothèse que la population obéit aux conditions de l'équilibre de Hardy-Weinberg, il est possible, par la relation de Hardy-Weinberg, d'associer les fréquences des trois génotypes présents dans la population aux fréquences alléliques *p*, pour l'allèle fonctionnel *A*, et *q*, pour l'allèle muté *a* (3ième et 4ième lignes du tableau).

On voit qu'il est alors possible (dernière colonne du tableau) de poser :

$$R = q^2 \qquad \text{d'où} \qquad \boldsymbol{q = \sqrt{R}}$$

On en tire évidemment *p*, sachant que $p + q = 1$, par $\boldsymbol{p = 1 - q}$

Connaissant *p* et *q*, on peut en déduire, par le calcul, les valeurs des fréquences génotypiques (dernière ligne du tableau).

La fréquence des hétérozygotes est égale à $2pq = 2(1 - q)q$, soit $\mathbf{2}\boldsymbol{q}$, si *q* est assez faible pour être négligeable devant 1.

Dans l'exemple considéré :

- la fréquence de l'allèle responsable de l'albinisme est égale à $q = 1\ \%$;
- la fréquence des hétérozygotes, non albinos mais « conducteurs » de l'albinisme est égale à $2q = 2\ \%$.

> ***Remarque 1 :*** il est important de noter que les valeurs calculées des fréquences alléliques ou génotypiques, ne correspondent à la réalité que si la population est effectivement dans les conditions du modèle de Hardy-Weinberg, et que l'échantillon étudié est représentatif de cette population. La conformité de ces conditions peut éventuellement être vérifiée par un test statistique.
>
> ***Remarque 2 :*** dans l'exemple précédent, la fréquence des hétérozygotes, 2*pq*, est égale à 198/10 000, et proche de la valeur simplifiée *2q* soit 1 habitant sur 50. Cette fréquence peut paraître considérablement élevée en regard de celle des albinos. La même remarque vaudra dans l'exemple de la mucoviscidose, une maladie récessive touchant un enfant sur 2500, avec, dans la population française, un porteur sain pour 25 habitants.

Il faut bien comprendre, qu'en régime panmictique, la fréquence des hétérozygotes est souvent supérieure à celle des homozygotes (figure 2.3).

Quand un allèle est très rare (*q* proche de zéro), la plupart des exemplaires de cet allèle sont présents dans des individus hétérozygotes et très peu dans des homozygotes (voir aussi les dernières colonnes des tableaux 2.4 et 2.5).

Bien évidemment cette situation n'a pas la même conséquence selon que l'allèle rare est récessif ou dominant. Dans le premier cas, l'effet de l'allèle récessif n'est perceptible que chez les homozygotes, et le phénotype associé sera beaucoup plus

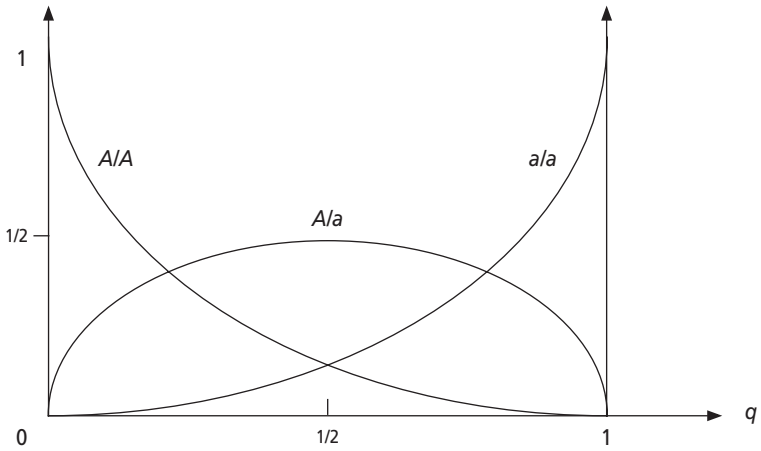


Figure 2.3 Fréquences des différents génotypes en fonction de la valeur de q [équations $(1 - q)^2$; $2q(1 - q)$ et q^2].

rare que la fréquence des hétérozygotes. Dans le deuxième cas, au contraire, l'effet de l'allèle dominant est perceptible chez les hétérozygotes et les homozygotes ; comme ceux ci sont très rares, le phénotype associé aura la fréquence des hétérozygotes.

> ***Remarque 3 :*** on remarquera par ailleurs, dans la figure 2.3, que la fréquence maximale du génotype hétérozygote, pour un gène di-allélique, est égale à 50 % et correspond à la situation où les deux fréquences alléliques sont égales (voir calcul du taux d'hétérozygotie, dans le chapitre 1).

2.4.2 Estimation des fréquences alléliques et génotypiques d'un gène responsable d'une maladie mendélienne

Dans l'exemple précédent, il était plus ou moins raisonnable de considérer la population à l'équilibre de Hardy-Weinberg, car deux de ses conditions pouvaient être contestées :

– la **panmixie** : car on peut imaginer, précisément pour ce caractère, un choix du conjoint ;
– l'**absence de sélection** : car les adultes albinos présenteront une diminution de fécondité si ils ne peuvent facilement trouver un conjoint.

Pour la plupart des maladies héréditaires qui affectent l'espérance de vie ou la fertilité, c'est-à-dire la fécondité, la condition d'absence de sélection n'est pas valide et on ne peut pas faire l'hypothèse de l'équilibre de Hardy-Weinberg. Cependant, on peut souvent admettre la panmixie pour le gène responsable de chacune de ces maladies. Ainsi, tout en sachant que les fréquences alléliques doivent être modifiées par la sélection sur plusieurs générations, il reste possible de faire une estimation ponctuelle de ces fréquences, en appliquant la relation de Hardy-Weinberg entre les fréquences des allèles, chez les parents, et les fréquences génotypiques, chez les enfants, avant que la sélection ne fasse son effet.

L'estimation des fréquences alléliques et génotypiques se révèle d'une grande utilité pour quantifier la diversité génétique en général (études de biodiversité, évaluation des ressources génétiques), et chez l'homme pour aborder les problèmes de diagnostic, de dépistage ou de santé publique.

> ***Remarque :*** bien souvent la fréquence d'un phénotype récessif rare, éventuellement une maladie, et celle de la mutation qui en est responsable, s'avèrent stables dans le temps, parce que l'effet de la sélection est contrebalancé par celui d'un autre facteur (voir chapitre 9).

a) Les maladies autosomiques récessives

Sous l'hypothèse panmictique, l'application ponctuelle de la relation de Hardy-Weinberg permet de calculer les fréquences des mutations, puis celles des génotypes, notamment des porteurs sains, et le rapport hétérozygotes/homozygotes (nombre de porteurs sains pour un patient atteint).

L'application de ces principes à quelques maladies (tableau 2.4) permet de quantifier la remarque faite plus haut sur le fait que lorsqu'un allèle récessif est rare, le phénotype récessif associé est très rare par rapport à la fréquence allélique car la plupart des exemplaires sont portés par des hétérozygotes de phénotype dominant. Ce fait est illustré par la valeur du rapport $2pq/q^2 = 2p/q = 2(1 - q)/q$ qui exprime le nombre de porteurs sains, hétérozygotes pour une mutation morbide, par rapport au nombre d'individus atteints, homozygotes pour cette mutation. Ce rapport est d'autant plus élevé que la mutation est rare.

Tableau 2.4 Incidence de maladies récessives et fréquences des allèles pathologiques.

Maladie	Population	Fréquence observée à la naissance (R)	Fréquence calculée de la mutation $q = \sqrt{R}$	Fréquence des porteurs sains $2q(1 - q)$	Nombre de porteurs sains pour un atteint $2(1 - q)/q$
Mucoviscidose	France	1/2 500	2 %	4 %	100
Phénylcétonurie	France	1/16 000	0,8 %	1,6 %	248
Galactosémie	France	1/40 000	0,5 %	1 %	400
Déficit en 21-OH	France	1/10 000	1 %	2 %	198
Déficit en 11b-OH	France	1/100 000	0,32 %	0,63 %	625
Drépanocytose	France (Antilles)	1/400	4.7 %	9,5 %	40
Drépanocytose	Afrique (voir chapitre 1)	4 %	20 %	32 %	48
β-thalassémie	Méditerranée (Italie-Grèce-Chypre)	1/100	10 %	18 %	18

> ***Remarque :*** pour la très grande majorité des maladies récessives connues (environ 1 500), la mutation est si rare que q^2 est quasi nul et que les quelques exemplaires de cette mutation sont portés par des hétérozygotes. Dans ce cas,

les quelques cas homozygotes de malades observés résultent principalement ou exclusivement de mariages entre apparentés et non de mariages panmictiques (voir chapitre 4).

b) Les maladies autosomiques dominantes

L'application de la relation de Hardy-Weinberg est également utile au calcul des fréquences alléliques et génotypiques pour les gènes dont les mutations gouvernent un phénotype dominant, et notamment pour les gènes responsables des maladies génétiques dominantes (tableau 2.5).

TABLEAU 2.5 INCIDENCE DE MALADIES DOMINANTES ET FRÉQUENCES DES ALLÈLES PATHOLOGIQUES.

Maladie	Population	Fréquence observée à la naissance (F)	Fréquence calculée de la mutation $p = 1 - \sqrt{1 - F}$	Nombre de d'hétérozygotes pour un homozygote $2(1 - p)/p$
Hypercholestérolémies familiales	Europe	1/500	10^{-3}	2 000
Neurofibromatose type I	Europe	1/3 000	$17\ 10^{-5}$	12 000
Dystrophie myotonique	Europe	1/7 000	$7\ 10^{-5}$	28 000
Maladie de Huntington	France	1/10 000	$5\ 10^{-5}$	40 000
Achondroplasie	Europe	1/20 000	$2.5\ 10^{-5}$	80 000
Syndrome de Marfan	Europe	1/25 000	$2\ 10^{-5}$	100 000
Ostéogenèse imparfaite	Europe Danemark (Fyn)	1/25 000 1/4 587	$2\ 10^{-5}$ $11\ 10^{-5}$	100 000 18 346
Rétinoblastome	Europe	1/30 000	$1{,}7\ 10^{-5}$	120 000

Dans ce dernier cas le phénotype sain est sans ambiguïté homozygote *b*/*b*, tandis que le phénotype atteint est *B*/*B* ou *B*/*b*. Sous l'hypothèse panmictique, la fréquence F des individus atteints est alors égale à $p^2 + 2pq$, ce qui revient à dire que la fréquence complémentaire $(1 - F)$, des individus sains, est égale à q^2.

Connaissant F, la fréquence des individus atteints, on peut en déduire directement :

– la fréquence de l'allèle fonctionnel, par $q = \sqrt{1 - F}$

– ou, la fréquence de la mutation responsable, par

$$p = 1 - \sqrt{1 - F}$$

– puisque $p = 1 - q$

On peut considérer, si la maladie est rare, comme c'est souvent le cas, que les homozygotes sont très rares, voire absents. Tous les individus atteints sont alors

hétérozygotes et la fréquence de la mutation morbide peut être directement estimée comme égale à la moitié de la fréquence des hétérozygotes-atteints :

$$p = F/2$$

On peut remarquer que si p est proche de zéro, l'équation de Hardy-Weinberg :

$$F = p^2 + 2p(1 - p) = 2p - p^2$$

peut être simplifiée en : $F = 2p$

d'où on tire bien que $p = F/2$.

Il est intéressant de noter que les maladies dominantes sont plutôt moins rares que les maladies récessives alors que les mutations qui en sont responsables sont, elles, environ 1 000 fois plus rares que celles responsables des maladies récessives. Ceci résulte du fait que les individus atteints d'une maladie récessive correspondent aux seuls homozygotes, alors que les patients atteints d'une pathologie dominante correspondent aux homozygotes et surtout aux hétérozygotes toujours plus nombreux (voir plus haut).

Or la sélection, qui n'a aucune prise chez les porteurs sains d'une mutation récessive, exerce son effet aussi bien chez les homozygotes que chez les hétérozygotes, pour les mutations responsables d'une maladie dominante. La pression de sélection étant moins forte sur les mutations récessives grâce à la « protection » des porteurs sains, celles-ci peuvent s'accumuler à un niveau plus élevé dans l'équilibre entre la sélection qui enlève des allèles mutés du pool génique et les mutations *de novo* qui en ramènent, puisque la sélection touche tous les porteurs pour une maladie dominante, mais seulement les homozygotes pour une maladie récessive.

> ***Remarque 1 :*** Il faut ajouter qu'un biais de recrutement accentue ce phénomène. En effet, dans le cas des maladies dominantes, de nombreux homozygotes ne sont pas recensés, soit parce que la maladie y est plus grave et provoque un décès prématuré, soit parce que les homozygotes présentent un syndrome clinique différent de celui des hétérozygotes.
> Il convient de noter que le terme de « maladie dominante » utilisé chez l'homme, ne recouvre pas vraiment le concept classique de phénotype dominant en génétique expérimentale. Le plus souvent les maladies sont dites dominantes dès que le porteur d'un seul allèle muté est atteint. En réalité, l'expérience montre, quand elle permet l'observation de quelques rares homozygotes, que ces derniers sont atteints soit plus gravement (hypercholestérolémies familiales), soit différemment, ce qui correspondrait plutôt à la définition de phénotypes codominants.
> Si cette précision est essentielle aussi bien sur le plan clinique que du point de vue fondamental de l'expression des gènes et de l'effet des mutations, cela ne change pas grand chose en ce qui concerne l'estimation des fréquences alléliques ou génotypiques, et leur utilisation du point de vue des problèmes de santé publique ou de conseil génétique.
>
> ***Remarque 2 :*** ce qui vient d'être développé pour les maladies mendéliennes chez l'homme vaut en fait pour tout phénotype rare, ce qui revient à dire que le pool génique d'une population naturelle est souvent plus riche qu'on ne l'imagine pour ce qui est des allèles récessifs.

c) Les caractères ou les maladies génétiques liés au sexe

Dans le cas d'un gène localisé sur un hétérochromosome et gouvernant un « caractère lié au sexe », la situation est sensiblement différente.

Supposons que la population obéit, pour le gène concerné, aux conditions de Hardy-Weinberg, et que les fréquences alléliques sont égales dans les deux sexes (voir chapitre 3 ce qu'il advient quand elles ne le sont pas).

Chez l'homme (ou la drosophile) les organismes de sexe mâle sont hémizygotes pour tous les gènes portés par l'hétérochromosome X. Que la mutation d'un gène ait un effet récessif ou dominant importe peu, dans ce cas, et tous les mâles d'un phénotype donné seront porteurs de l'allèle associé à ce phénotype, ce qui permet une estimation directe des fréquences alléliques par identité aux fréquences phénotypiques (tableau 2.6).

Par contre les fréquences génotypiques et phénotypiques dans le sexe femelle sont très différentes de celles du sexe mâle, même à l'équilibre de Hardy-Weinberg, comme l'illustre le tableau ci-dessous (on note p, la fréquence de l'allèle dominant, et q, celle de l'allèle récessif) :

TABLEAU 2.6

	Sexe mâle		Sexe femelle		Bilan
	atteint	non atteint	atteint	non atteint	
Maladie récessive	q	p	q^2	$p^2 + 2pq$	$q^2 < q$
Maladie dominante	p	q	$p^2 + 2pq$	q^2	$p^2 + 2pq > p$

Le bilan du tableau 2.6 permet de prévoir qu'un trait génétique, ou une maladie, lié au sexe devrait être (dans les conditions de Hardy-Weinberg) plus fréquent dans le sexe mâle s'il est récessif et plus fréquent dans le sexe femelle s'il est dominant.

C'est par exemple le cas pour le groupe sanguin humain XG, dont le phénotype récessif [XG^-] a une fréquence de 0,35 dans le sexe masculin et de 0,13 dans le sexe féminin, ce qui correspond très précisément à q et q^2 pour une fréquence de l'allèle récessif égale à 0,35 dans les deux sexes (tableau 2.7). Mais la situation des traits ou des maladies liés au sexe, chez l'homme, apparaît assez complexe. Les fréquences des génotypes ou des phénotypes observées chez les femmes sont conformes aux fréquences attendues, pour le groupe sanguin XG. C'est aussi le cas pour l'hémophilie et *rétinitis pigmentosa*, où les fréquences attendues sont si faibles qu'elles sont nulles en réalité. Une fréquence de 10^{-4} chez les garçons correspond, pour la France, à environ 78 naissances par an, mais une fréquence de 10^{-8}, pour les filles, correspond à une naissance tous les 128 ans !

La fréquence du daltonisme observée chez les filles est plus faible que la fréquence attendue parce qu'en réalité le daltonisme est gouverné par deux gènes, au moins. Les hommes daltoniens seraient mutés, pour 75 % des cas, dans le gène gouvernant

TABLEAU 2.7

Trait ou maladie [*D*] : dominant [*R*] : recessif	Fréquence observée chez les hommes *m*	Fréquence de la mutation *q* = *m* (ou *p* = *m* si [*D*])	Fréquence observée chez les femmes *f*	Fréquence attendue chez les femmes q^2 si [*R*] 2*p* si [*D*]
Groupe sanguin XG Phénotype XG^- [R]	35/100	0,35	13/100	12/100
Daltonisme [*R*]	8/100	0,08	4/1 000	6,4/1 000
Hémophilie A [*R*]	1/10 000	10^{-4}	0	10^{-8}
Hémophilie B [*R*]	1/30 000	$3{,}3\ 10^{-5}$	0	10^{-9}
Rétinitis pigmentosa [*R*]	1/300 000	$3{,}3\ 10^{-6}$	0	10^{-11}
Myopathie de Duchenne-Becker [*R*]	1/3 500	$2{,}8\ 10^{-4}$	cas exceptionnels	0 !
X-fragile [*D*]	1/1 500	$6{,}6\ 10^{-4}$	1/3 500	1/2 250 000 si [*R*] 1/750 si [*D*]

la vision du vert, et pour 25 % des cas, dans le gène gouvernant la vision du rouge. Il convient alors de répartir la fréquence observée du phénotype daltonien (0,08) en deux fréquences alléliques de 0,06 et 0,02, pour chacun des deux gènes impliqués à 75 % et 25 % dans le daltonisme. La fréquence des femmes daltoniennes ne serait donc pas le carré de 0.08, mais, en simplifiant, la somme des carrés de 0,06 et 0,02 (voir le cas de l'équilibre de Hardy-Weinberg pour deux gènes, chapitre 3).

La fréquence attendue des femmes atteintes de la myopathie de Duchenne/Becker est évidemment nulle. En effet, comme la maladie est létale dans l'enfance, aucun garçon atteint n'est susceptible de devenir père et de transmettre sa mutation à un descendant. Les quelques cas exceptionnels observés chez des filles correspondent soit à une mutation *de novo* chez le père, soit à une anomalie du caryotype comme une monosomie X (syndrome de Turner), avec un chromosome X d'origine maternelle, porteur de la mutation.

Le syndrome de l'X-fragile est une arriération mentale associée à des caractères morphologiques spécifiques, résultant d'une mutation dans le gène FMR1 du chromosome X. Si ce syndrome était récessif, la fréquence chez les filles serait égale à 1/2 250 000. Ce syndrome est donc nettement dominant. Par ailleurs des données généalogiques et moléculaires prouvent que la présence d'un allèle fonctionnel d'origine paternelle ne suffit pas à contrebalancer, chez les filles atteintes, l'effet de la mutation d'origine maternelle. Cependant la fréquence des filles cliniquement et mentalement atteintes (1/3 500) est inférieure à la fréquence attendue (1/750) dans le cas d'une maladie dominante. On explique cet écart par le phénomène d'inactivation de l'X qui conduit, dans chacune des cellules d'un organisme de sexe féminin,

à l'activité d'un seul des deux chromosomes X. Comme le chromosome X inactivé n'est pas forcément le même d'une cellule à l'autre, les tissus constituent une mosaïque cellulaire, avec une fraction de cellules ayant inactivé le X paternel et la fraction complémentaire ayant inactivé le X maternel. Chez les filles où la proportion des cellules du système nerveux central ayant inactivé l'X maternel muté est suffisamment élevée, l'effet de cette mutation sera faible ou négligeable, ce qui réduit la fréquence des femmes atteintes de 1/750 à 1/3 500.

2.5 TESTS STATISTIQUES DE VÉRIFICATION DE LA CONFORMITÉ AU MODÈLE DE HARDY-WEINBERG

Il est utile de savoir confronter les observations faites dans une population naturelle, en supposant qu'elles sont représentatives de sa composition génétique, aux valeurs théoriques attendues sous l'hypothèse que cette population est à l'équilibre de Hardy-Weinberg, pour le gène considéré. Cette démarche, qui sera détaillée dans un premier exemple, prend la forme d'un test statistique d'hypothèse utilisant la distribution d'une variable de χ^2.

2.5.1 Exemple d'un gène responsable de phénotypes codominants

Reprenons l'exemple du groupe sanguin MN dans un échantillon d'une population européenne (voir le tableau 1.1), en rappelant ou en précisant quelques points sur le raisonnement associé aux tests de χ^2.

TABLEAU **2.8**

Groupes sanguins	[M]	[MN]	[N]
Effectifs observés (total : 1 000)	350	500	150
Fréquences phénotypiques	350/1 000 = 0,35	500/1 000 = 0,50	150/1 000 = 0,15

Les phénotypes étant codominants, les fréquences génotypiques sont égales aux fréquences phénotypiques ; de ce fait les fréquences alléliques sont directement accessibles, soit :

$$f(\mathrm{M}) = p = \mathbf{0,6} \qquad \text{et} \qquad f(\mathrm{N}) = q = \mathbf{0,4}$$

Si cette population est panmictique et à l'équilibre de Hardy-Weinberg, on doit s'attendre à ce que les fréquences génotypiques ne soient pas très différentes, aux variations d'échantillonnage près, des valeurs p^2, $2pq$ et q^2, soit $0{,}6^2$, $2 \times 0{,}6 \times 0{,}4$ et $0{,}4^2$.

Sous l'hypothèse de Hardy-Weinberg, un échantillon aléatoire théorique de 1 000 individus aura une composition génétique où l'effectif de chacun des trois génotypes sera égal à sa fréquence respective (p^2, $2pq$ et q^2) multipliée par 1 000 (tableau 2.9). La question que se propose de résoudre le test statistique, sera de savoir si on peut

considérer que les différences entre effectifs observés et effectifs attendus (on dit aussi théoriques) peuvent ou ne peuvent pas être considérées comme résultant du seul hasard d'échantillonnage. Si la réponse du test à cette question est « oui », l'hypothèse théorique de Hardy-Weinberg sera acceptée ; dans le cas contraire, elle sera rejetée.

Pour cela, dans chaque classe i, on calcule le carré de l'écart entre effectifs observé (o_i) et théorique (t_i), que l'on rapporte à l'effectif théorique t_i, soit : $(o_i - t_i)^2/t_i$. En effet ce qui nous intéresse est la différence entre ce qu'on observe et ce qu'on attend, c'est-à-dire les écarts. Comme on veut s'affranchir du signe de ces écarts, car seule leurs tailles nous importent, on les élève au carré. On rapporte par ailleurs chacun de ces écarts (élevé au carré), à la valeur de l'effectif théorique car, bien évidemment un écart de 10 pour un effectif de 1 000 est beaucoup plus « petit », bien moins significatif, plus facilement dû à un hasard d'échantillonnage, qu'un écart de 10 pour un effectif de 50. On obtient donc dans notre exemple :

Tableau 2.9

Groupes sanguins	[M]	[MN]	[N]
Effectifs observés (total : 1 000)	350	500	150
Effectifs théoriques (total : 1 000)	p^2 x 1 000 = 360	$2pq$/1 000 = 480	q^2/1 000 = 160
$(o_i - t_i)^2/t_i$	0,277	0,833	0,625
$\Sigma\ (o_i - t_i)^2/t_i$	1,736		

La variable définie comme la somme de ces écarts $\Sigma\ (o_i - t_i)^2/t_i$ suit une loi de χ^2.

La fonction χ^2 est une somme de carrés et peut prendre des valeurs quelconques dans le domaine de variation [0 ; + α].

Les mathématiciens ont étudié la « fonction de densité de probabilité du χ^2 », qui permet de calculer la probabilité p avec laquelle un χ^2 peut voir sa valeur observée, atteindre ou dépasser telle ou telle valeur désignée par $\chi^2_{(p)}$ (figure 2.4).

Des tables permettent de déterminer pour un χ^2, en fonction de son nombre de degré de liberté (voir plus loin), les valeurs $\chi^2_{(p)}$ de son domaine de variation [0 ; + α], qui peuvent être atteintes ou dépassées, par le hasard d'échantillonnage, avec une probabilité p égale à 50 %, 40 %, 30 %, 20 %, 10 %, 5 %, 2.5 %, ou 1 %. Ces valeurs sont appelées seuil au risque p % (figure 2.4).

Il est aussi possible quand on a la valeur observée d'un χ^2, de calculer la probabilité p exacte qu'on a d'atteindre ou dépasser cette valeur observée.

Le test consiste à prendre une décision qui est, soit d'accepter l'hypothèse en considérant que les écarts résultent du hasard d'échantillonnage, soit de rejeter l'hypothèse en considérant que les écarts sont trop importants (la valeur observée du χ^2 est trop élevée) pour qu'ils puissent être dus à un hasard d'échantillonnage.

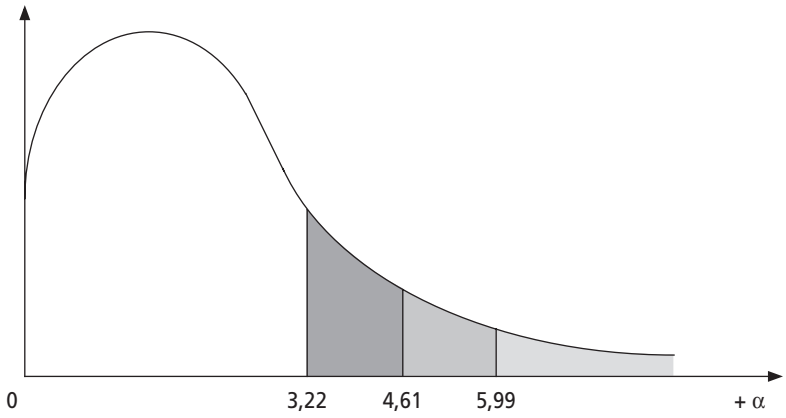


Figure 2.4 Graphe d'un χ^2 à deux degrés de liberté.

La valeur d'un χ^2 à 2 degrés de liberté a respectivement une probabilité de $p_1 = 20$ %, $p_2 = 10$ %, $p_3 = 5$ %, d'atteindre ou dépasser les valeurs 3,22, 4,61 et 5,99.
Ces valeurs désignées par $\chi^2_{(p1)}$, $\chi^2_{(p2)}$ et $\chi^2_{(p3)}$, valeurs qui ont une probabilité p_1 ou p_2 ou p_3 d'être atteintes ou dépassées, sont appelées valeurs seuil au risque p_1 ou p_2 ou p_3.
Les valeurs p_1 ou p_2 ou p_3 sont aussi appelées niveau de signification
Ces probabilités de dépassement correspondent aux surfaces en grisé (la surface totale de l'intégrale représente 100 %, ce qui signifie que la valeur observée a 100 % de chance d'être supérieure à zéro).

La décision est prise de la façon suivante :

– si la valeur observée du χ^2 est supérieure à la valeur $\chi^2_{(p)}$, on rejette l'hypothèse. Dans ce cas, il se peut que l'hypothèse soit quand même juste et qu'on la rejette par erreur, mais notre analyse mathématique a permis de quantifier ce risque, c'est précisément p, probabilité, si l'hypothèse est juste, que, par hasard, les écarts soient aussi grands et que cette valeur du χ^2 ait pu être atteinte par hasard ; *p* est donc le risque d'erreur de la décision de rejet. C'est pourquoi on ne peut prendre des valeurs trop importantes pour p, c'est-à-dire des valeurs seuils trop petites pour $\chi^2_{(p)}$. En effet si on rejette avec des risques de 10 ou 20 % ou plus, la méthode de décision n'est plus crédible. En général on n'accepte pas de prendre un risque supérieur à 5 % ;
– si la valeur observée du χ^2 est inférieure à $\chi^2_{(p)}$, en pratique à $\chi^2_{(5\,\%)}$, on accepte l'hypothèse puisque si on la rejetait, on prendrait un risque d'erreur supérieur à *p*, en pratique 5 %, ce à quoi on se refuse car ce serait admettre qu'on accepte de prendre une règle de décision où on s'autorise à se tromper très souvent.

Mais il convient bien de noter qu'un test statistique n'est que décisionnel : il ne permet pas de dire si une hypothèse est juste ou fausse, il permet seulement de prendre la décision d'accepter ou de rejeter cette hypothèse, en calculant un risque d'erreur associé à cette décision :

– si on rejette une hypothèse, cela ne signifie absolument pas qu'elle est fausse, mais qu'on a 5 chances sur 100 d'avoir tort en estimant qu'elle l'est ;
– inversement, si on l'accepte, cela ne signifie pas qu'elle est vraie, mais simplement qu'il est trop risqué de la rejeter.

Évidemment, il est plus confortable et plus significatif de rejeter l'hypothèse avec une valeur de χ^2 élevée, correspondant à un seuil décisionnel de faible risque, c'est-à-dire 1 % ou 0,1 % ou 0,01 % ! Si l'hypothèse théorique est très éloignée de la réalité, il suffira d'un petit échantillon pour obtenir un χ^2 très significatif, avec un risque d'erreur très faible ; si l'hypothèse théorique rend compte imparfaitement de la réalité, un échantillon très grand sera nécessaire pour la rejeter de manière significative, avec un risque d'erreur pas trop grand.

Enfin, il faut rappeler que la probabilité ***p*** avec laquelle une variable de χ^2 dépasse une valeur $\chi^2_{(p)}$ de son domaine de variation dépend du nombre *n* de classes (en fait le nombre de classes théoriques indépendantes appelé nombre de degrés de liberté), c'est-à-dire du nombre d'éléments de la somme constituant le χ^2. Il semble en effet naturel qu'un χ^2 constitué de 10 éléments puisse plus facilement, même par hasard échantillonnage, avoir une valeur élevée qu'un χ^2 constitué de 2 éléments. Le nombre de degré de liberté est toujours inférieur ou égal au nombre de classes moins une, car on impose au moins une relation entre les classes théoriques qui est d'avoir la même somme que les classes observées, ce qui fait que connaissant $(n - 1)$ classes on en déduit la dernière. Il convient aussi de retrancher à n le nombre de paramètres estimés à partir des observations afin de calculer les effectifs théoriques. Dans l'exemple ci-dessus il a fallu estimer *p* (q est déduit du fait que $p + q = 1$), ce qui donne un degré de liberté égal à 3 moins 1 (pour la somme des effectifs théoriques égale à 1 000) moins 1 (pour l'estimation de p) soit 1 degré de liberté.

Dans ces conditions la valeur seuil du χ^2 qui n'est dépassée, par hasard, que 5 fois sur 100, est égale à 3,84 (pour un χ^2 à deux degrés de liberté, cette valeur seuil est égale à 5,99, voir schéma ci-dessus).

Dans notre exemple la valeur observée du χ^2 est de 1,73. Comme elle est inférieure à la valeur seuil 3,84, pour laquelle le risque d'erreur est de 5 %, on accepte l'hypothèse : la population étudiée est panmictique et peut être considérée comme à l'équilibre de Hardy-Weinberg.

2.5.2 Exemple d'un gène responsable de phénotypes dominants et récessifs

Le typage du groupe sanguin ABO dans un échantillon aléatoire de 1 000 individus apporte les résultats figurés dans le tableau 2.10.

Il est évidemment impossible d'estimer directement les fréquences alléliques à partir des fréquences phénotypiques puisque la proportion des homozygotes et des hétérozygotes ne peut être mesurée dans les groupes [A] et [B].

Le calcul des fréquences alléliques peut être entrepris en recourant à l'hypothèse de l'équilibre de Hardy-Weinberg, qui permet de mettre en relation les six génotypes possibles (colonne 4) avec leurs fréquences génotypiques théoriques exprimées en fonction des fréquences *p*, de l'allèle I^A, *q*, de l'allèle I^B et *r*, de l'allèle I^O (colonne 5).

Tableau 2.10 Conformité au modèle de Hardy-Weinberg pour le groupe sanguin ABO.

Phénotype : groupe sanguin	Effectifs observés	Fréquences observées	Génotypes	Fréquences génotypiques si Hardy-Weinberg	Effectifs théoriques si Hardy-Weinberg	$(o_i - t_i)^2/t_i$
[A]	460	0,46	I^A/I^A I^A/I^O	p^2 $2pr$	90 360	0,222
[B]	140	0,14	I^B/I^B I^B/I^O	q^2 $2qr$	10 120	0,769
[AB]	50	0,05	I^A/I^B	$2pq$	60	1,666
[O]	350	0,35	I^O/I^O	r^2	360	0,277
Totaux	1 000	1		1	1 000	2,936

On pourrait, comme on l'a vu, estimer r par la racine carrée $\sqrt{f[\mathrm{O}]}$, puis reporter la valeur de *r* dans l'une des équations

$p^2 + 2pr = f[\mathrm{A}] = 0{,}46$ pour en tirer l'estimation de *p*

ou $q^2 + 2qr = f[\mathrm{B}] = 0{,}14$ pour en tirer l'estimation de *q*.

Mais il est plus judicieux de remarquer les relations suivantes :

$p^2 + 2pr + r^2 = f[\mathrm{A}] + f[\mathrm{O}] = 0{,}46 + 0{,}35 = 0{,}81$

et $q^2 + 2qr + r^2 = f[\mathrm{B}] + f[\mathrm{O}] = 0{,}14 + 0{,}35 = 0{,}49$

On remarque que :

$p^2 + 2pr + r^2 = (p + r)^2 = (1 - q)^2$

et $q^2 + 2qr + r^2 = (q + r)^2 = (1 - p)^2$

D'où $(1 - q)^2 = f[\mathrm{A}] + f[\mathrm{O}]$ et $\boldsymbol{q = 1 - \sqrt{f[\mathrm{A}] + f[\mathrm{O}]}}$

et $(1 - p)^2 = f[\mathrm{B}] + f[\mathrm{O}]$ et $\boldsymbol{p = 1 - \sqrt{f[\mathrm{B}] + f[\mathrm{O}]}}$

ce qui donne les valeurs de $q = 0{,}1$

$p = 0{,}3$

et $r = 1 - p - q = 0{,}6$

On démontre, par la théorie du maximum de vraisemblance, que les estimations obtenues par cette méthode, sont « meilleures » que celles, peu différentes, obtenues par d'autres calculs. À partir de là, il est facile de calculer les effectifs théoriques de chacun des génotypes (colonne 6) et, surtout de chacun des quatre phénotypes, afin de tester par un χ^2, les écarts (colonne 7) entre effectifs observés et effectifs attendus sous l'hypothèse de Hardy-Weinberg. On remarquera comment un écart de 10 entre effectifs n'a pas la même conséquence, selon qu'il porte sur un petit effectif (groupe [AB]) ou un grand effectif (groupe [O]).

Le nombre de degré de liberté de la variable de χ^2 est égal à :

– 4 classes ;

– moins 1 (pour la somme des effectifs théoriques égale aux effectifs observés) ;

– moins 2 (pour l'estimation, à partir des observations de p et q, r étant déduit de $1 - p - q$) ;
– soit 1 degré de liberté.

La valeur observée de la variable est égale à 2,936. Elle est inférieure à la valeur 3,84 qui n'est dépassée que 5 fois sur cent. Rejeter l'hypothèse de Hardy-Weinberg reviendrait donc à accepter un risque d'erreur supérieur à 5 % ; on accepte donc cette hypothèse.

2.5.3 Populations structurées et effet Walhund

Une population apparemment homogène peut être en réalité structurée et composée de plusieurs entités ou sous-populations différentes. Au sein de chacune il peut y avoir panmixie mais il n'y a pas de panmixie générale parce que ces sous-unités sont partiellement endogames les unes par rapport aux autres. Dans ces conditions, l'équilibre de Hardy-Weinberg est vérifié au sein des sous-populations mais n'existe pas au sein de la population générale (tableau 2.11).

Tableau 2.11

	Fréquences alléliques			**Fréquences génotypiques**		
	A	***a***		***A/A***	***A/a***	***a/a***
Sous population A (effectif égal à B)	0,9	0,1	Effectif observé	164	32	4
			Effectif attendu si *H-W*	162	36	2
	Test de H-W dans la sous-population A : $\chi^2 = 2{,}47$, non significatif					
Sous population B (effectif égal à A)	0,3	0,7	Effectif observé	20	80	100
			Effectif attendu si H-W	18	84	98
	Test de H-W dans la sous-population B : $\chi^2 = 0{,}45$, non significatif					
Population totale A + B	0,6	0,4	Effectif observé	184	112	104
			Effectif attendu si H-W	144	192	64
	Test de H-W dans la population totale : $\chi^2 = 69{,}44$, très significatif					

Chacune des sous-populations A et B est panmictique et à l'équilibre de Hardy-Weinberg pour ses fréquences alléliques et génotypiques, mais l'ensemble (A + B) qui peut être, par erreur, perçu comme homogène, présente alors des fréquences

alléliques moyennes et surtout des fréquences génotypiques avec un excès d'homozygotes et un déficit d'hétérozygotes par rapport aux fréquences attendues sous le modèle de Hardy-Weinberg.

Cet écart à la panmixie, appelé « effet Walhund », peut être vu comme un artefact quand il résulte du mélange, par l'observateur qui n'en a pas conscience, d'échantillons prélevés dans des populations en réalité séparées, mais il correspond aussi à une réalité objective quand une population est hiérarchisée ou structurée en sous-populations endogames entre lesquelles les échanges génétiques sont assez faibles, et/ou les conditions de différenciation génétique assez fortes, pour maintenir une diversité de composition génétique. C'est pourquoi l'effet Walhund sera traité en détail dans le chapitre 4, consacré aux écarts à la panmixie.

Il s'agit simplement ici de noter que cet effet Walhund est une source fréquente d'erreurs d'interprétation quand une étude ignore la structuration d'une population en deux ou plusieurs sous populations, par exemple quand un sous-ensemble de malades est considéré comme appartenant à une même population générale alors que certains sont d'origines ethniques différentes et d'intégration récente dans la population générale, ou quand on mélange, pour faire de « bons » échantillons statistiques des groupes de malades de divers pays, même européens (exemple : les études cas-témoins pour les maladies multi-factorielles). Cet exemple est donc très important en pratique car de nombreuses études peuvent se trouver biaisées si elles ignorent la structuration des populations étudiées ou des échantillons collectés et réunis et se fondent sur l'hypothèse d'une panmixie générale, notamment pour l'estimation des fréquences alléliques.

RÉSUMÉ

Les facteurs qui peuvent modifier la composition génétique d'une population sont :

– les mutations, les migrations et la sélection ;
– les variations d'échantillonnage survenant dans les populations de petite taille ;
– les modalités de choix des conjoints, c'est-à-dire les écarts à la panmixie (unions au hasard).

Le modèle théorique de « l'équilibre de Hardy-Weinberg » correspond à une population respectant les trois conditions de panmixie, de grand effectif, et d'absence de mutations, de migrations et de sélection. Sous ces trois conditions, la composition génétique de la population est invariante :

– les fréquences alléliques et génotypiques demeurent inchangées de générations en générations, tant que les conditions demeurent ;
– une relation mathématique s'établit entre les fréquences alléliques et les fréquences génotypiques : la somme des fréquences génotypiques correspond au développement du carré de la somme des fréquences alléliques, soit, pour un gène di-allélique :

$$p^2 + 2pq + q^2 = (p + q)^2$$

Cette relation entre fréquences génotypiques et alléliques, dite relation panmictique ou relation de Hardy-Weinberg, permet de déduire les fréquences génotypiques de la seule connaissance des fréquences alléliques. Elle est donc très utile pour estimer les fréquences alléliques des gènes gouvernant des caractères présentant des phénotypes récessifs, notamment les fréquences des allèles pathogènes responsables des maladies génétiques récessives et la fréquence des porteurs sains.

EXERCICES

Partie A : les tests statistiques

Deux tests statistiques sont principalement utilisés, le test de conformité et le test d'homogénéité.

Le premier consiste à tester la conformité d'une série d'observations à une loi théorique posée *a priori*. Connaissant la loi théorique sensée s'appliquer au phénomène étudié, on en déduit, sous l'hypothèse de cette loi théorique, les effectifs attendus pour chacun des types d'observations possibles. Le test consiste à mesurer « l'importance » des écarts entre les effectifs observés et les effectifs attendus (on dit aussi calculés ou théoriques) par le calcul d'un χ^2. La valeur de celui-ci permettra de décider si les écarts sont acceptables comme dus à un simple hasard d'échantillonnage, ou s'ils sont trop importants pour être dus à un tel hasard. Dans le premier cas les observations sont dites conformes à la loi théorique, et celle-ci peut être acceptée comme valide pour expliquer les observations ; dans le deuxième cas, les observations ne sont pas conformes à la loi théorique qui est alors rejetée.

Le test d'homogénéité ne se réfère à aucune loi théorique posée *a priori*. Il vise à comparer deux ou plus de deux séries d'observations afin de savoir si leurs distributions sont ou ne sont pas équivalentes.

Le test consiste à faire l'hypothèse d'homogénéité selon laquelle les deux séries d'observations présentent des distributions ne différant que par un simple hasard d'échantillonnage, ce qui revient à dire qu'elles sont identiques et conformes à une même loi (qui n'est pas posée *a priori* et qui peut même être inconnue). Sous cette hypothèse la meilleure estimation de la distribution des différents types d'observations est obtenue sur la somme de chacune des classes, ce qui permet de calculer des effectifs attendus pour chacune d'entre elles. Les écarts entre effectifs observés et attendus sont là encore testés par le calcul d'une valeur d'un χ^2. La valeur de celui-ci permettra de décider si les écarts sont acceptables comme dus à un simple hasard d'échantillonnage, ou s'ils sont trop importants pour être dus à un tel hasard. Dans le premier cas les observations sont dites homogènes et peuvent être considérées comme obéissant à une même loi (par exemple tirées d'une même population) ; dans le deuxième cas, l'homogénéité est rejetée.

On rappelle les valeurs du seuil à 5 % pour des χ^2 à *n* degrés de libertés (ddl) :

ddl	1	2	3	4	5	6	7	8	9
Valeur seuil au risque de 5 %	3,84	5,99	7,82	9,49	11,1	12,6	14,1	15,5	16,9

Exercice 2.1 Exemple de test de conformité

On dispose de deux souches pures de drosophiles, l'une de phénotype sauvage et l'autre de phénotype vestigial (ailes atrophiées).

On croise entre elles ces deux souches et on obtient une F1 de phénotype sauvage, ce qui permet de conclure qu'il est dominant et que le phénotype vestigial est récessif.

Le croisement F1 × F1 donne une F2 où on dénombre, parmi 1 000 mouches prélevées au hasard, 775 mouches de phénotype sauvage et 225 de phénotype vestigial.

Ce résultat s'interprète classiquement comme la ségrégation 2/2 typique d'un seul couple d'allèles. Justifiez cette conclusion par un test statistique.

Solution

On postule qu'il existe deux allèles *A* et *a* d'un gène et que les souches parentales (pures) sauvage et vestigiale sont respectivement *A*/*A* et *a*/*a*. Les descendants F1 sont hétérozygotes pour ce gène, leur génotype est *A*/*a* ; ils donnent tous, à la méiose, deux types de gamètes équifréquents, soit *A*(1/2), soit *a*(1/2) : manifestation de la ségrégation 2-2.

Les descendants F2 sont alors *A*/*A*, *A*/*a* et *a*/*a* dans les proportions 1/4, 1/2 et 1/4, ce qui correspond à 3/4 de phénotypes sauvages et 1/4 de phénotype vestigial... à condition de considérer que 775/1 000 et 225/1 000 sont bien « conformes » aux proportions théoriques 3/4 et 1/4.

On peut tester la *conformité* de cette hypothèse de ségrégation 2/2, c'est-à-dire la considérer comme valide, par un test de conformité qui consiste à voir si les écarts entre ce qu'on observe et ce qu'on « attend » sous l'hypothèse sont négligeables ou pas. Sous cette hypothèse de ségrégation 2/2, on attend 3/4 de phénotypes sauvages soit un effectif théorique de 750 pour un échantillon dont la taille totale est de 1 000 individus, ce qui fait par différence un effectif attendu de 250 individus de phénotype vestigial (soit 1/4). Pour chacun des deux phénotypes les écarts sont de $775 - 750 = 25$ et $225 - 250 = -25$.

Les écarts peuvent être testés par un test de χ^2. La variable de χ^2 définie ici a un degré de liberté. En effet il y a 2 classes, moins un pour l'égalité des effectifs totaux.

La valeur d'un χ^2 à 1 ddl, qui n'est dépassée que 5 fois sur 100 est égale 3,84.

La valeur observée du χ^2 est égale 3,33 ; elle peut donc être atteinte ou dépassée avec une probabilité supérieure à 5 %.

Rejeter l'hypothèse de ségrégation 2/2, associée aux proportions théoriques 3/4-1/4, reviendrait alors à accepter un seuil de décision avec un risque supérieur à 5 % : on accepte donc cette hypothèse.

> ***Remarque 1 :*** dans cet exemple les observations sont presque au seuil de signification. Si on avait observé, ne serait-ce que 777 sauvages et 223 vestigial, le test eût été significatif (valeur observée du χ^2 égale 3,88) et la conclusion génétique de ségrégation 2/2 rejetée (avec un risque d'erreur de 5 %).

C'est pourquoi il est souvent nécessaire de ne pas conclure sans avoir justifié sa conclusion par un test de conformité.

En toute rigueur, même quand le test s'avère d'avance non significatif, par exemple si les effectifs observés sont égaux à 748 et 252, sa réalisation est justifiée du point de vue de la démarche du raisonnement.

Remarque 2 : dans cet exemple la loi théorique est connue *a priori*, c'est la loi de pureté des gamètes de Mendel, autrement dit la ségrégation 2/2 d'un couple d'allèles à la méiose chez un hétérozygote.

Dans d'autres circonstances la loi connue *a priori* n'est, comme dans le cas de la relation de Hardy-Weinberg, qu'une relation mathématique entre des paramètres (les fréquences alléliques dans le cas de cette relation). Il est alors nécessaire d'estimer ces paramètres (fréquences alléliques) à partir des observations, ce qui réduira d'autant, comme on le verra, le nombre de degrés de liberté.

Exercice 2.2 Exemple de test d'homogénéité

Les individus d'une population naturelle présentent deux phénotypes alternatifs [A] et [B]. On prélève au hasard dans cette population un échantillon et on effectue de nouveau un prélèvement l'année suivante ; on observe les résultats du tableau ci-dessous.

Peut-on conclure, par un test statistique, que si la population est restée stable ou a évolué au cours de l'année ?

année	phénotype [A]	phénotype [B]
1	110	235
2	250	405

Solution

La question générale porte sur la comparaison de deux ou plus de deux séries d'observations (ici deux échantillons) classées en deux ou plusieurs catégories (ici deux phénotypes) afin de savoir si ces séries d'observations sont différentes ou non. Dans ce dernier cas, on dit qu'elles sont homogènes, et le test qui permet de rejeter ou d'accepter la conclusion d'homogénéité porte le nom de test d'homogénéité.

Le principe consiste à poser comme hypothèse nulle (hypothèse dans laquelle on se place et dont on va décider, par le test, si on la rejette ou si on l'accepte) qu'il y a homogénéité.

Dans notre cas, cela signifie qu'on considère que la population n'a pas évolué et, qu'en conséquence, les deux échantillons sont homogènes et ne peuvent présenter que de petites différences de distribution dues au hasard d'échantillonnage.

Sous cette hypothèse nulle, la population n'ayant pas évolué, les deux échantillons sont donc issus d'une même population. De ce fait les deux échantillons peuvent

être additionnés afin de fournir une estimation plus précise des fréquences des phénotypes [A] et [B], estimation ainsi réalisée sous cette hypothèse nulle.

Toujours sous cette hypothèse, en utilisant les fréquences ainsi estimées, on peut reconstituer les effectifs de deux échantillons théoriques, de même taille que celles des échantillons observés.

On montre que si l'hypothèse nulle est vraie, les écarts entre effectifs théoriques et observés ne dépendant que du hasard d'échantillonnage, la somme des carrés des écarts rapportés aux effectifs théoriques suit une loi de χ^2 à $(n-1)(m-1)$ degrés de liberté où n est le nombre de séries de résultats (nombre de lignes par exemple, ici 2) et m est le nombre de classes entre lesquelles se répartissent les différents objets de chacune des séries (ici $m = 2$).

On obtient ainsi un tableau de contingence où sont figurés, outre les effectifs observés, les sommes marginales et les effectifs théoriques, calculés sous l'hypothèse nulle (à partir de ces sommes marginales, puisqu'on somme les séries). Dans notre exemple on obtient le tableau suivant (où les effectifs théoriques sont mentionnés en italiques).

Sous l'hypothèse nulle, la meilleure estimation de la fréquence du phénotype [A] est celle qui est faite avec la somme marginale, soit 360/1 000 = 0,36, et les effectifs attendus pour deux échantillons de 345 et 655 individus sont respectivement $345 \times 0{,}36 = 124{,}2$ et $655 \times 0{,}36 = 235{,}8$.

Année	Phénotype [A]	Phénotype [B]	Totaux des échantillons
1 Effectifs attendus	110 124,2	235 220,8	345
2 Effectifs attendus	250 235,8	405 419,2	655
Totaux des phénotypes	360	640	1 000

Remarquons que la connaissance d'un des effectifs attendus permet d'en déduire l'autre par différence avec la somme marginale : il y a donc en fait $(n-1)$ effectifs théoriques à calculer par colonne, et de la même façon $(m-1)$ effectifs à calculer par ligne, ce qui conduit bien à $(n-1)(m-1)$ effectifs théoriques indépendants, soit le nombre de degrés de liberté de la variable χ^2. Pour chacune des classes, on effectue le carré de l'écart rapporté à l'effectif théorique, et on somme sur toutes les classes, ce qui donne la valeur observée de la variable χ^2, sous l'hypothèse nulle.

La variable de χ^2 définie ici a un degré de liberté.

La valeur d'un χ^2 à 1 ddl, qui n'est dépassée que 5 fois sur 100 est égale 3,84.

La valeur observée du χ^2 est égale 3,87 ! Elle est presque égale à 3,84, ce qui signifie qu'elle avait une probabilité à peu près égale à 5 % d'être observée.

Si on avait observé une valeur très supérieure à 3,84, on aurait pu rejeter l'hypothèse nulle, avec un risque très inférieur à 5 %, et conclure que les deux échantillons,

n'étant pas homogènes, ne pouvaient être considérés comme issus d'une même population, la population étudiée ayant donc évolué en l'espace d'un an.

Avec une valeur observée inférieure à 3,84, on aurait accepté l'hypothèse nulle, car la rejeter eût été prendre une décision assortie d'un risque supérieur à 5 %.

Ici, avec une valeur de 3,87, on peut se permettre de rejeter avec un risque très exact de 5 %!

On pourrait aussi se permettre d'accepter l'homogénéité, avec le calcul nécessaire de l'autre risque, dit de deuxième espèce, risque d'accepter l'hypothèse nulle alors qu'elle est fausse (et que, par hasard, on a observé, cette fois-là, des écarts exceptionnellement faibles !). Le calcul de ce risque est plus délicat et déborde le cadre de cet ouvrage.

> ***Remarque :*** rappelons cependant qu'un test ne permet pas de dire si une hypothèse est juste ou fausse mais simplement de calculer le risque d'erreur associé à la décision de rejet de cette hypothèse. Il ne faut donc pas accorder au 5 % la valeur mythique qu'on lui prête, et savoir que la décision est souvent prise en fonction des conséquences pratiques ou théoriques de celle-ci. En effet le seuil de 5 % correspond au seuil qui minimise le risque d'erreur de première espèce (rejet de l'hypothèse nulle alors qu'elle est vraie) sans trop augmenter le risque alternatif de seconde espèce.

Partie B : modèle de Hardy-Weinberg

Exercice 2.3

Dans une espèce animale, la couleur du pelage est déterminée par un gène autosomal pour lequel existent deux formes alléliques *A1* et *A2*. Les individus homozygotes *A1A1* ont un pelage noir, les individus *A2A2* ont un pelage blanc et les hétérozygotes un pelage tacheté noir et blanc. En observant les couples dans la nature, au moment du rut, on a fait les observations suivantes :

Types de couples observés	Nombres observés
[noir] x [noir]	35
[noir] x [tacheté]	250
[noir] x [blanc]	80
[tacheté] x [tacheté]	345
[tacheté] x [blanc]	260
[blanc] x [blanc]	30
Total	**1 000**

Question 1 : déterminer les fréquences phénotypiques, génotypiques, alléliques, pour le gène étudié.

Question 2 : la population est-elle panmictique, pour ce gène ?

Question 3 : la composition génétique de la population, pour le gène étudié, est-elle conforme au modèle de Hardy-Weinberg ?

Dans cette espèce, le caractère court ou long du pelage est sous la dépendance d'un autre gène autosomal di-allélique, pour lequel existe un effet de dominance : les individus de génotype *B/B* et *B/b* sont à poil court alors que les génotypes *b/b* présentent un phénotype à poil long. On a fait, en même temps que les observations sur la couleur du pelage, les observations suivantes :

Types de couples observés	nombres observés
[court] x [court]	250
[court] x [long]	520
[long] x [long]	230
Total	1 000

Question 4 : déterminer les fréquences phénotypiques, génotypiques, alléliques, pour le gène étudié.

Question 5 : la population est-elle panmictique, pour ce gène ?

Question 6 : la composition génétique de la population, pour le gène étudié, est-elle conforme au modèle de Hardy-Weinberg ? Proposez un moyen de résoudre le problème posé.

Solution

1. Étude de la couleur du pelage

Question 1 : les trois génotypes présentent un phénotype spécifique : dans ce cas de codominance le calcul des fréquences génotypiques et alléliques ne présente aucune difficulté et se fait directement, sans aucune hypothèse préalable.

En rapportant aux 2 000 individus formant les 1 000 couples observés, les nombres respectifs de noirs, de tachetés et de blancs on obtient les fréquences phénotypiques, qui, du fait de la codominance sont aussi les fréquences génotypiques (tableau suivant, ligne 4) :

Phénotype	[noir]	[tacheté]	[blanc]
Génotype	*A1/A1*	*A1/A2*	*A2/A2*
Effectifs observés	400	1 200	400
Fréquences observées	$D = 0{,}2$	$H = 0{,}6$	$R = 0{,}2$

Les fréquences alléliques s'en déduisent par : $f(A1) = D + H/2 = p = 0{,}5$

et $f(A2) = R + H/2 = q = 0{,}5$

Question 2 : test de la panmixie

On pourrait tester la panmixie en testant l'équilibre de Hardy-Weinberg dont elle est une des conditions. Mais la structure des observations nous permet de tester directement la conformité à la panmixie, sans autre hypothèse jointe : en effet, on observe des couples !

La loi postulée *a priori* est la panmixie dont il découle que les fréquences des couples sont égales aux carrés des types d'individus pour les couples d'individus identiques et aux double-produits des fréquences de chacun des individus pour les couples d'individus différents.

Sous cette hypothèse théorique, il est facile, connaissant les fréquences phénotypiques, de calculer les fréquences et les effectifs attendus des divers types de couples et de tester les écarts aux effectifs observés, par un test de χ^2. On obtient alors le tableau suivant :

Types de couples observés	Effectifs observés	Fréquence théorique	Effectifs attendus
[noir] x [noir]	35	D^2	40
[noir] x [tacheté]	250	$2DH$	240
[noir] x [blanc]	80	$2DR$	80
[tacheté] x [tacheté]	345	H^2	360
[tacheté] x [blanc]	260	$2HR$	240
[blanc] x [blanc]	30	R^2	40
Total	1 000		1 000

Les écarts peuvent être testés par un test de χ^2. La variable de χ^2 définie ici a trois degrés de liberté. En effet, il y a 6 classes, moins un pour l'égalité des effectifs totaux, moins 2 pour l'estimation de deux des trois fréquences phénotypiques à partir des observations et sans lesquelles on ne pourrait pas calculer les effectifs théoriques (la troisième fréquence est connue quand les deux premières le sont). La valeur d'un χ^2 à 3 ddl, qui n'est dépassée que 5 fois sur 100 est égale 7,8. La valeur observée du χ^2 est égale 5,83 ; elle peut donc être atteinte ou dépassée avec une probabilité supérieure à 5 %. Rejeter l'hypothèse panmictique reviendrait alors à accepter un seuil de décision avec un risque supérieur à 5 % : on accepte donc l'hypothèse de panmixie.

Question 3 : test de conformité au modèle de Hardy-Weinberg

Il consiste à calculer les fréquences génotypiques attendues sous l'hypothèse de cet équilibre, puis en multipliant par l'effectif observé d'individus (2 000) à calculer les effectifs théoriques. On obtient le tableau suivant, à partir duquel on peut tester la signification des écarts par un χ^2.

Phénotype	[noir]	[tacheté]	[blanc]
Génotype	*A1/A1*	*A1/A2*	*A2/A2*
Effectifs observés	400	1 200	400
Fréquences attendues sous l'hypothèse de *H-W*	$p^2 = 0{,}25$	$2pq = 0{,}5$	$q^2 = 0{,}25$
Effectifs attendus	500	1 000	500

La variable de χ^2 définie ici a un degré de liberté. En effet il y a 3 classes, moins un pour l'égalité des effectifs totaux, moins 1 pour l'estimation d'une des deux fréquences alléliques à partir des observations et sans lesquelles on ne pourrait pas calculer les effectifs théoriques (la deuxième fréquence est connue quand la première l'est). La valeur d'un χ^2 à 1 ddl, qui n'est dépassée que 5 fois sur 100 est égale 3,84. La valeur observée du χ^2 est égale 80 ! Elle est largement supérieure à 3,84, ce qui signifie qu'elle avait une probabilité largement inférieure à 5 % d'être observée, si l'hypothèse nulle était vraie.

Dans ces conditions, rejeter l'hypothèse de l'équilibre de Hardy-Weinberg revient à prendre un risque de se tromper (que cette hypothèse soit en réalité juste et que les écarts observés soient, par hasard, aussi exceptionnellement élevés) très inférieur à 5 % : on rejette donc l'hypothèse de l'équilibre de Hardy-Weinberg.

> ***Remarque :*** comme on a testé indépendamment la panmixie, il est possible de conclure que l'équilibre de Hardy-Weinberg n'est pas réalisé parce qu'une des autres de ses conditions ne l'est pas.

2. Étude de la structure du pelage

Question 4 : les trois génotypes ne présentent que deux phénotypes, car l'un d'eux est récessif : dans ce cas le calcul des fréquences génotypiques et alléliques ne peut se faire directement et ne peut être réalisé qu'en supposant la population, pour le gène considéré, à l'équilibre de Hardy-Weinberg.

En rapportant aux 2 000 individus formant les 1 000 couples observés, les nombres respectifs de courts et de longs, on obtient les fréquences phénotypiques :

Phénotype	[court}	[long]
Génotype	*B/B* ou *B/b*	*b/b*
Effectifs observés	1 020	980
Fréquences observées	$D + H = 0{,}51$	$R = 0{,}49$

Les fréquences alléliques s'en déduisent par la relation de Hardy-Weinberg, soit :

$$f(b) = q = \sqrt{R} = 0{,}7 \quad \text{et} \qquad f(B) = 1 - q = 0{,}3$$

Question 5 : test de la panmixie

On peut aussi, dans la situation présente, tester directement la panmixie, sans l'inclure dans le test de Hardy-Weinberg puisqu'on observe des couples. Si les couples sont panmictiques, il est facile, connaissant les fréquences phénotypiques, de calculer les

fréquences et les effectifs attendus des divers types de couples (tableau ci-dessous) et de tester les écarts aux effectifs observés, par un test de χ^2.

NB **:** si les couples sont panmictiques la fréquence des couples homogènes formés de phénotypes identiques est le carré de la fréquence du phénotype considéré, et la fréquence des couples hétérogènes est le double produit des deux fréquences phénotypiques [il suffit pour s'en assurer de faire le tableau de croisement des couples] :

Types de couples observés	Effectifs observés	Fréquence théorique	Effectifs attendus
[court] x [court]	250	f[court]^2	260,1
[court] x [long]	520	2f[court]f[long]	499,8
[long] x [long]	230	f[long]^2	240,1
Total	1 000		1 000

Les écarts peuvent être testés par un test de χ^2. La variable de χ^2 définie ici a un degré de liberté. En effet il y a 3 classes, moins un pour l'égalité des effectifs totaux, moins 1 pour l'estimation d'une des deux fréquences phénotypiques à partir des observations et sans lesquelles on ne pourrait pas calculer les effectifs théoriques (la deuxième fréquence est déduite de la première comme le complément à 1). La valeur d'un χ^2 à 1 ddl, qui n'est dépassée que 5 fois sur 100 est égale 3,84. La valeur observée du χ^2 est égale 1,63 ; elle peut donc être atteinte ou dépassée avec une probabilité supérieure à 5 %.

Rejeter l'hypothèse panmictique reviendrait alors à accepter un seuil de décision avec un risque supérieur à 5 % : on accepte donc l'hypothèse de panmixie.

Question 6 : test de l'équilibre de Hardy-Weinberg

Il consisterait à calculer les fréquences génotypiques attendues sous l'hypothèse de cet équilibre, puis en multipliant par l'effectif observé d'individus (2 000) à calculer les effectifs théoriques. Mais, dans ce cas, on retrouvera les mêmes effectifs, ce qui conduirait à un χ^2 de valeur égale à zéro, ce qui n'est pas possible.

En fait une telle variable de χ^2 aurait ici zéro degré de liberté. En effet il y a 2 classes (court et long), moins un pour l'égalité des effectifs totaux, moins 1 pour l'estimation d'une des deux fréquences alléliques à partir des observations et sans lesquelles on ne pourrait pas calculer les effectifs théoriques (la deuxième fréquence est connue quand la première l'est), ce qui fait zéro degré de liberté.

L'équilibre de Hardy-Weinberg ne peut être testé car l'information disponible n'est pas répartie sur un nombre suffisant de classes, compte tenu du nombre de paramètres à estimer. Cependant la structure des observations a permis de tester directement la panmixie.

Pour tester l'équilibre de Hardy-Weinberg plusieurs possibilités peuvent être envisagées :

– on peut changer de crible d'observation afin de disposer de phénotypes codominants (voir deux exercices suivants) ;

– on peut étudier la descendance des couples afin de déterminer les génotypes parentaux (voir un des problèmes suivants).

Exercice 2.4

On considère deux maladies de fréquence égale à 1/40 000, l'une récessive, l'autre dominante.

Question 1 : Quelles sont, dans chaque cas, les estimations des fréquences de l'allèle pathologique (justifier le calcul en quelques mots) ?

Question 2 : Calculez, dans chaque cas, le rapport [fréquences des hétérozygotes]/[fréquences des homozygotes atteints]. Quel est le sens de ce rapport ?

Question 3 : Commentez la différence entre les valeurs calculées des fréquences alléliques et des rapports, pour ces deux maladies de même fréquence dans la population.

Solution

	Maladie récessive	**Maladie dominante**
Fréquence de la maladie	R = 1/40 000	F = 1/40 000
Question 1 : fréquence de l'allèle pathologique	Sous l'hypothèse panmictique (ou *HW*) $q = \sqrt{R}$ = 0,005	Sous l'hypothèse panmictique (ou *HW*) $p = 1 - \sqrt{(1-F)}$ = 0,000 013 = 1/79 999 On peut aussi, considérant que les homozygotes sont inexistants, écrire $p = F/2$ = 1/80 000
Question 2 : valeur du rapport [fréquence des hétérozygotes]/[fréquence des homozygotes]	$2q(1-q)/q^2$ = 398	$2p(1-p)/p^2$ = 159 998

Question 3 : commentaires sur les fréquences et sur les rapports.

1. Pour deux maladies de même fréquence, on observe que la fréquence de l'allèle pathologique « dominant » est 400 fois plus faible que celle de l'allèle pathologique récessif. Ceci s'explique facilement si on considère que la sélection agit systématiquement sur tous les allèles dominants, ce qui impose une fréquence faible, alors qu'elle n'agit que sur les quelques allèles récessifs en situation homozygote et n'a aucune action sur ceux présents chez les porteurs sains, ce qui permet aux allèles récessifs, préservés de l'action de la sélection, de demeurer à un niveau de fréquence plus élevé.

2. Cette capacité d'échappement des allèles récessifs à la sélection est attestée par le rapport [fréquence des hétérozygotes]/[fréquence des homozygotes atteints] qui indique combien d'allèles pathologiques échappent à la sélection (porteurs sains) pour deux allèles qui lui sont soumis (homozygotes). Ici ce rapport est de 398 : il y a près de 200 allèles qui échappent à la sélection pour un qui lui est soumis.

3. Le rapport [fréquence des hétérozygotes]/[fréquence des homozygotes] pour une maladie dominante exprime la quantité de malades porteurs d'un seul allèle muté pour un malade porteur de deux allèles mutés (homozygotes) ; ce rapport est très

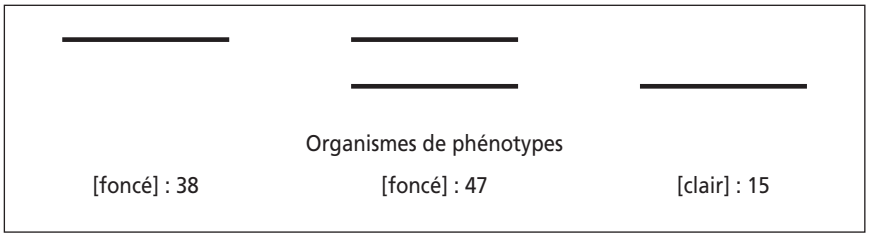
grand, ce qui signifie que les malades porteurs de deux mutations sont si rares que leur fréquence est négligeable, ce qui dispense de calculer la fréquence allélique sous l'hypothèse panmictique.

Exercice 2.5 : Test de l'équilibre de Hardy-Weinberg par changement de crible phénotypique

Le phénotype de pigmentation alaire, chez une espèce de papillon, est gouverné par un gène existant sous deux formes alléliques, notées *A* et *a*, dont les fréquences sont *p* et *q*. Une première étude a montré que le phénotype clair est récessif et correspond au génotype *aa*. Le piégeage, en milieu naturel, de 1 600 papillons, a permis de dénombrer 1 340 phénotypes foncés et 260 clairs.

Question 1 : déterminez la composition génotypique et allélique de la population et testez sa conformité au modèle de Hardy-Weinberg.

Question 2 : le gène étudié spécifie une enzyme de la chaîne de biosynthèse des pigments qu'il est possible d'étudier, en biochimie, par électrophorèse. L'étude réalisée sur 100 animaux est rapportée ci-dessous ; en quoi cette étude complémentaire permet-elle de répondre aux questions précédentes ?



Solution

Question 1 : les phénotypes n'étant pas codominants, le calcul direct des fréquences génotypiques, puis des fréquences alléliques n'est pas possible. Il faut faire l'hypothèse que la population a une composition génétique conforme à celle attendue sous l'équilibre de Hardy-Weinberg.

Dans ce cas la fréquence du phénotype récessif est égale au carré de la fréquence de l'allèle *a*, soit :

$f[\text{clair}] = 260/1\,600 = q^2$

d'où $q = \sqrt{260/1\,600} = 0{,}4$

et $p = 1 - q = 0{,}6$

Les génotypes *A/A*, *A/a* et *a/a* ont pour fréquences respectives p^2, $2pq$ et q^2.

Tester l'équilibre consisterait à calculer les fréquences génotypiques attendues sous l'hypothèse de cet équilibre, puis en multipliant par l'effectif observé d'individus (1 600) à calculer les effectifs théoriques. Mais, dans ce cas, on retrouvera les mêmes effectifs, ce qui conduirait à un χ^2 de valeur égale à zéro, ce qui n'est pas possible.

En fait une telle variable de χ^2 aurait ici zéro degré de liberté. En effet il y a 2 classes (foncé et clair), moins un pour l'égalité des effectifs totaux, moins 1 pour l'estimation d'une des deux fréquences alléliques à partir des observations et sans lesquelles on ne pourrait pas calculer les effectifs théoriques (la deuxième fréquence est connue quand la première l'est), ce qui fait zéro degré de liberté.

L'équilibre de Hardy-Weinberg ne peut être testé car l'information disponible n'est pas répartie sur un nombre suffisant de classes, compte tenu du nombre de paramètres à estimer.

Question 2 : les organismes de phénotype clair présentent tous une enzyme dont la migration est plus rapide et qui, chez ces organismes est codée par l'allèle *a*. L'enzyme codée par l'allèle A a donc une migration plus lente. Les 38 organismes présentant une bande « lente » unique correspondent à des homozygotes *A/A*, tandis que les 47 hétérozygotes présentent les deux formes enzymatiques.

Les phénotypes électrophorétiques des produits du gène étudiés sont codominants, ce qui permet le calcul direct des fréquences, soit, à partir du tableau ci-dessous :

Génotypes	*A/A*	*A/a*	*a/a*
Effectifs observés	38	47	15
Effectifs théoriques	37,82	47,36	14,82

$f(A/A) = 0{,}38$ $f(A/a) = 0{,}47$ $f(a/a) = 0{,}15$

et $f(A) = 0{,}38 + 0{,}47/2 = p =$ **0,615**

$f(a) = 0{,}15 + 0{,}47/2 = q =$ **0,385**

Ces fréquences étant estimées directement, sans aucune hypothèse *a priori*, sont, par principe, meilleures que les premières (question 1) et se substituent à elles.

En multipliant les fréquences p^2, $2pq$ et q^2 par la taille de l'échantillon, on obtient les effectifs attendus sous l'hypothèse de l'équilibre panmictique de Hardy-Weinberg (tableau ci-dessus).

La valeur du χ^2 est égale à 0,0057, ce qui très inférieur au seuil de 5 % pour un χ^2 à 1 ddl. L'équilibre de Hardy-Weinberg est donc une hypothèse largement acceptable.

> ***Remarque :*** selon qu'on observe les effets du gène à travers un crible phénotypique (pigmentation de l'aile) ou un autre (migration électrophorétique de l'enzyme codée), le phénotype de l'homozygote *a/a* sera récessif ou codominant. Cet exercice montre au passage que ce n'est pas l'allèle *a* qui est récessif ou dont l'effet est récessif, en soi ou même vis-à-vis de *A*, mais le phénotype qu'il confère, pour un caractère, à l'organisme, quand celui-ci est homozygote, ici le caractère de pigmentation alaire ; car pour un autre caractère, la migration électrophorétique, l'effet de *a* est codominant. C'est donc une facilité (une dérive) de langage que s'accordent les généticiens en parlant d'« allèles récessifs ». Cette facilité est admissible entre généticiens mais elle est trompeuse pour les élèves et les étudiants car elle obscurcit, dans l'enseignement de la génétique, la relation génotype/phénotype et l'interprétation fonctionnelle de la dominance et de la récessivité.

Exercice 2.6 : le système Rhésus chez l'homme

Un de vos amis, médecin, recueille dans une maternité les résultats de groupage sanguin rhésus pour 1 000 nouveaux-nés. 510 sont [rhésus +] et 490 sont [rhésus –].

Question 1 : se souvenant vaguement de ses cours de génétique des populations, il tente de calculer les fréquences alléliques et génotypiques mais n'arrive pas à tester l'équilibre de Hardy-Weinberg.

Pouvez-vous lui en expliquer la raison ?

Question 2 : fort de vos connaissances en génétique des populations, vous lui conseillez alors de reprendre les 1 000 dossiers d'hospitalisation, où sont également consignés les groupages de chacune des mères de ces 1 000 nouveaux-nés, afin de constituer des couples mère-enfant, ce qui lui permettra, lui affirmez-vous, de tester l'hypothèse de Hardy-Weinberg.

Votre ami vous rapporte les résultats suivants et vous demande votre aide :

Mère	Enfant	Nombre observé de couples
[+]	[+]	372
[–]	[+]	138
[+]	[–]	158
[–]	[–]	332

Expliquez-lui comment cette information permet de conclure à la panmixie et à l'équilibre de Hardy-Weinberg, pour la population étudiée, en ce qui concerne le gène gouvernant le facteur rhésus.

Solution

Question 1 : le problème ici posé est le même que celui rencontré dans l'exercice 2.5. On ne peut calculer les fréquences alléliques que sous l'hypothèse de Hardy-Weinberg, et on ne peut pas tester cette hypothèse, car on a épuisé l'information pour cela.

Sachant que les allèles du gène rhésus sont classiquement notés *D* et *d* (pour celui qui confère le groupe [–] récessif), les fréquences sont respectivement égales à $f(D) = 0{,}3$ et $f(d) = 0{,}7$.

Question 2 : bien évidemment l'information apportée par les couples mère-enfant est décisive. En effet il était impossible, précédemment, de tester l'hypothèse de Hardy-Weinberg parce que l'information était répartie en 2 classes et qu'il fallait estimer 2 paramètres (la taille de l'échantillon théorique et une des fréquences alléliques). Maintenant il est toujours nécessaire d'estimer 2 paramètres mais on dispose de quatre classes phénotypiques pour les couples mère-enfant.

En effet supposons que la population soit à l'équilibre de Hardy-Weinberg, alors les génotypes *D*/*D*, *D*/*d* et *d*/*d* sont dans les proportions p^2, $2pq$ et q^2, avec les valeurs 0,3 et 0,7 pour *p* et *q*.

On peut alors calculer la probabilité (ou la fréquence théorique) de chacun de ces quatre couples phénotypiques mère-enfant, sous cette hypothèse de Hardy-Weinberg. Attention, car il s'agit d'être exhaustif et de balayer tous les couples génotypiques possibles !

a) Le couple mère-enfant [+] × [+]

La mère peut être *D*/*D*, avec la probabilité p^2. Dans ce cas, son enfant sera [+] avec la probabilité 1.

La mère peut être *D*/*d*, avec la probabilité $2pq$. Dans ce cas son enfant sera [+] s'il reçoit D de sa mère, évènement de probabilité 1/2, ou s'il reçoit *d* de sa mère et *D* de son père, évènement de probabilité égale à $1/2 \times p$ (1/2 : probabilité que la mère donne *d* et *p* : probabilité que le père donne *D*).

Au total, la probabilité ou la fréquence attendue des couples mère-enfant [+] × [+] sera égale à :

$p^2 + 2pq\ (1/2 + 1/2 \times p) = p(1 + pq)$

b) Le couple mère-enfant [–] × [+]

La mère est *d*/*d*, avec la probabilité q^2. Dans ce cas, son enfant sera [+] si le père lui donne un allèle *D*, évènement de probabilité égale à *p*.

Au total, la probabilité ou la fréquence attendue des couples mère-enfant [–] × [+] sera égale à $pq^{2.}$

c) Le couple mère-enfant [+] × [–]

La mère peut être *D*/*D*, avec la probabilité p^2. Dans ce cas, son enfant sera [–] avec la probabilité 0.

La mère peut être *D*/*d*, avec la probabilité $2pq$. Dans ce cas son enfant sera [–] s'il reçoit *d* de sa mère, évènement de probabilité 1/2 et s'il reçoit d de son père, évènement de probabilité égale à q.

Au total, la probabilité ou la fréquence attendue des couples mère-enfant [+] × [–] sera égale à :

$2pq\ (1/2 \times q) = pq^2$

d) Le couple mère-enfant [–] × [–]

La mère est *d*/*d*, avec la probabilité q^2. Dans ce cas son enfant sera [–] si le père lui donne un allèle *d*, évènement de probabilité *q*.

Au total, la probabilité ou la fréquence attendue des couples mère-enfant [–] × [–] sera égale à q^3.

On obtient alors le tableau suivant où sont portés les effectifs attendus sous, l'hypothèse de Hardy-Weinberg, pour un échantillon de 1 000 couples mère-enfant :

Mère	Enfant	Nombre observé de couples mère-enfant	Fréquences théoriques de ces couples	Effectifs théoriques
[+]	[+]	372	$p(1 + pq)$	363
[–]	[+]	138	pq^2	147
[+]	[–]	158	pq^2	147
[–]	[–]	332	q^3	343

Les écarts peuvent être testés par un test de χ^2. La variable de χ^2 définie ici a 2 degrés de liberté. En effet, il y a 4 classes, moins un pour l'égalité des effectifs observés et théoriques, moins 1 pour l'estimation d'une des deux fréquences alléliques (la deuxième fréquence est déduite de la première comme le complément à 1). La valeur d'un χ^2 à 2 ddl, qui n'est atteinte ou dépassée que 5 fois sur 100 est égale 5,99.

La valeur observée du χ^2 est égale 1,95 ; elle peut donc être atteinte ou dépassée avec une probabilité très supérieure à 5 %. Rejeter l'hypothèse panmictique reviendrait alors à accepter un seuil de décision avec un risque supérieur à 5 % : on accepte donc l'hypothèse de l'équilibre de Hardy-Weinberg et la panmixie qui en est une des conditions.

Exercice 2.7 : le système sanguin Kidd chez l'homme

Il existe chez l'homme de nombreux systèmes sanguins, mais seuls les systèmes ABO et Rhésus sont pris en compte dans les transfusions, car ils sont très facilement la cause d'accidents transfusionnels. Par contre chez les poly-transfusés, on doit parfois tenir compte de certains des autres groupes sanguins. Par ailleurs ces groupes sont très utiles en génétique des populations pour la mesure de la diversité génétique dans ou entre les populations. Le groupe Kidd a été défini quand l'étude de sérums sanguins de poly-transfusés a permis d'isoler deux types d'anticorps, appelés anti-JKa et anti-JKb. Les érythrocytes de tout individu, soumis à un test d'hémagglutination (comme pour le typage MN, voir page 56) se répartissent en trois catégories, les deux homozygotes pour les allèles *Jka* ou *Jkb* qui ne réagissent qu'avec l'un des deux anticorps et l'hétérozygote dont les hématies sont sensibles aux deux anticorps. Un échantillon de 102 Vietnamiens a été typé et s'est trouvé réparti en 32 JKa/Jka, 32 JKb/JKb et 38 Jka/JKb.

Question 1 : calculer les fréquences alléliques et tester la conformité de la composition génétique au modèle de Hardy-Weinberg.

Question 2 : cette première étude avait délaissé trois cas particuliers de phénotype était [Jka– ; Jkb–] soit un test d'hémagglutination négatif avec les deux anticorps. Ce fait avait d'abord été interprété comme une erreur technique, mais les individus ont été reprélevés et retestés, avec un même résultat. Ce fait a conduit à formuler l'hypothèse de l'existence d'un troisième allèle, nommé Jko, qui comme l'allèle I^o du gène gouvernant le groupe ABO, ne détermine aucun antigène membranaire, allèle absent en Europe mais visiblement assez fréquent en Asie. En quoi cette nouvelle donnée modifie-t-elle les conclusions de la question précédente ?

Solution

Question 1 : les phénotypes sont codominants et le calcul direct des fréquences génotypiques permet celui des fréquences alléliques, soit $f(\text{JKA}) = p = 0{,}5$ et $f(\text{JKb}) = q = 0{,}5$.

Pour tester la conformité de la composition génétique au modèle de Hardy-Weinberg, il convient de calculer les fréquences génotypiques attendues sous ce modèle

(p^2 ; $2pq$; q^2), d'en déduire les effectifs attendus sur un échantillon de 102 individus (dernière colonne du tableau ci-dessous et de tester les écarts par un test de χ^2.

Test d'hémagglutination Phénotype	Effectifs observés	Génotypes	Effectifs attendus si Hardy-Weinberg
[Jka+ ; JKb–]	32	*JKa/JKa*	25,5
[Jka+ ; JKb+]	38	*JKa/JKb*	51
[Jka– ; JKb+]	32	*JKb/JKb*	25,5
Total	102		102

La variable de χ^2 définie ici a un degré de liberté. En effet, il y a 3 classes, moins un pour l'égalité des effectifs totaux, moins 1 pour l'estimation d'une des deux fréquences alléliques à partir des observations et sans lesquelles on ne pourrait pas calculer les effectifs théoriques (la deuxième fréquence est déduite de la première comme le complément à 1).

La valeur d'un χ^2 à 1 ddl, qui n'est dépassée que 5 fois sur 100 est égale 3,84. La valeur observée du χ^2 est égale 6,63 ; elle ne peut donc être atteinte ou dépassée qu'avec une probabilité très inférieure à 5 %. On peut donc considérer que l'importance de ces écarts est vraisemblablement due au fait que la population vietnamienne ne serait pas à l'équilibre de Hardy-Weinberg. On rejette donc cette hypothèse en sachant que le risque d'erreur (rejet alors qu'elle est vraie et que les écarts observés sont exceptionnellement grands) est largement inférieur à 5 %.

Question 2 : tout le problème doit être repris puisqu'il y a six génotypes et quatre phénotypes, dont l'un est récessif La suite de ce problème se traite comme l'exemple ABO (voir 2.5.2).

Les fréquences alléliques sont :

– pour *Jka* : $p = 1 - \sqrt{(32 + 3)/105} = 0{,}423$

– pour *Jkb* : $q = 1 - \sqrt{(32 + 3)/105} = 0{,}423$

– pour *Jko* : $r = 1 - p - q = 0{,}156$

On obtient alors pour le test de l'équilibre de Hardy-Weinberg, le tableau suivant :

Test d'hémagglutination Phénotype	Effectifs observés	Génotypes	Effectifs attendus si Hardy-Weinberg
[Jka+ ; JKb-]	32	*JKa/Jka* ou *JKa/JKo*	32,64
[Jka+ ; JKb+]	38	*JKa/JKb*	37,17
[Jka– ; JKb+]	32	*JKb/JKb* ou *JKb/JKo*	32,64
[Jka– ; JKb–]	3	*JKo/JKo*	2,55
Total	105		105

La valeur du χ^2 est égale à 0,123, ce qui très inférieur à 3,84, la valeur correspondant au seuil de 5 % pour un χ^2 à 1 ddl. L'équilibre de Hardy-Weinberg devient une hypothèse largement acceptable alors que le fait de négliger l'existence de l'allèle silencieux *Jko* avait conduit au rejet de cette hypothèse.

Exercice 2.8 Phénotype [rosy] chez la drosophile

Les drosophiles de phénotype [rosy] ont les yeux rose ; elles diffèrent des drosophiles de phénotype sauvage noté [rosy+], aux yeux rouge brique, par la mutation d'un seul gène, comme le montrent des expériences simples de croisements entre souches. Le phénotype mutant [rosy] est récessif : les hétérozygotes *ry+*/*ry* (*ry+* et *ry* étant respectivement les allèles sauvages et mutés du gène) sont de phénotype sauvage comme les homozygotes *ry+*/*ry+*.

Question 1 : on laisse se reproduire librement dans une vaste cage à population (appelée démomètre), un groupe de 2 000 mouches constitué au départ de 1 200 mouches de souche pure sauvage et 800 mouches de souche pure rosy. Après plusieurs générations on prélève 500 mouches au hasard et on observe la couleur des yeux. 90 d'entre elles sont de phénotype rosy.

Que peut-on en conclure sur les fréquences alléliques et la panmixie éventuelle dans ce démomètre ?

Question 2 : il a été montré que le gène impliqué dans le phénotype [rosy] codait pour la xanthine-déshydrogénase qu'il est possible de doser dans un extrait acellulaire de drosophile. Cela conduit à la définition d'un nouveau caractère, le « dosage d'activité » et à de nouveaux phénotypes : les phénotypes d'activité. Ceux-ci sont codominants car l'hétérozygote présente un taux d'activité médian compris entre celui de l'homozygote *ry+*/*ry+* et celui de l'homozygote *ry*/*ry* (nul car déficient en enzyme). On teste l'activité chez 100 mouches prélevées au hasard : 18 mouches de phénotype [rosy] présentent une activité nulle, tandis que 38 mouches présentent une activité très forte et que 44 mouches présentent une activité intermédiaire.

Que peut-on en conclure sur les fréquences alléliques et la panmixie éventuelle dans ce démomètre ?

Question 3 : par ailleurs, des études électrophorétiques de la xanthine-déshydrogénase ont montré l'existence de deux allèles électrophorétiques. Il s'agit de deux allèles du gène codant chacun pour une chaîne peptidique active mais différant l'une de l'autre par un seul acide aminé de charge électrique opposée. Cette différence conduit alors à une migration électrophorétique différentielle. L'un des allèles est appelé *fast* (*f*) parce qu'il code pour une chaîne à déplacement rapide (notée F) ; l'autre allèle est appelé *slow* (*s*), car il code pour une chaîne à déplacement plus lent (notée S).

Les génotypes *f*/*f*, *f*/*s* et *s*/*s* sont tous les trois *ry+*/*ry+* et présentent un même phénotype d'activité (taux élevé) et un même phénotype morphologique (yeux sauvages rouge brique) ; ils ne peuvent être distingués qu'à partir de la mise en évidence, par électrophorèse, des trois phénotypes électrophorétiques.

Ceux-ci sont codominants car on peut distinguer la présence aussi bien que l'absence des chaînes rapides et lentes (voir tableau ci-dessus où la migration est du haut vers le bas). Le tableau ci-dessus donne la correspondance génotype/phénotype pour les divers allèles du gène rosy (l'allèle actif *ry+* étant noté *f* ou *s*, selon le type de chaîne codée, *F* ou *S*).

Génotype au locus du gène *xdh*	*f/f*	*f/s*	*s/s*	*f/ry*	*s/ry*	*ry/ry*
Phénotypes morphologiques	Yeux de phénotype sauvage rouge brique [rosy+]					Yeux rose [rosy]
Phénotypes d'activité	Taux élevé +++			Taux médian +		Taux nul –
Phénotypes électrophorétiques						
Effectifs observés	0	1	37	1	43	18

Comment interpréter le phénotype à trois bandes ? Et les phénotypes à une bande pour des individus de génotypes différents, par exemple *f/f* et *f/ry* ?

Peut-on préciser le génotype des deux individus *f/s* et *f/ry* et l'origine de ces génotypes, sachant que :

– la souche pure sauvage dont étaient tirées les 1 200 fondatrices est homozygote pour l'allèle *s*,

– la souche pure Rosy, qui a fourni les 800 fondatrices Rosy, a été obtenue par mutagenèse d'une souche sauvage homozygote pour l'allèle *f* ? (cette question suppose d'avoir consulté le chapitre 3).

Solution

Question 1 : dans la mesure où les deux phénotypes ne sont pas codominants, il faut faire l'hypothèse de l'équilibre de Hardy-Weinberg pour calculer les fréquences alléliques.

Dans ces conditions, si *q* est la fréquence de l'allèle *ry* et *p* celle de *ry+*, on a :

$q = \sqrt{90/500} = 0{,}424$ et $p = 1 - q = 0{,}576$

On ne peut évidemment pas tester l'hypothèse de l'équilibre de Hardy-Weinberg (voir problèmes précédents).

Question 2 : il est alors possible avec ces phénotypes codominants de calculer directement les fréquences génotypiques et alléliques puis de tester l'équilibre de Hardy-Weinberg.

Phénotype	**Dosage élevé**	**Dosage intermédiaire**	**Dosage nul**
Génotypes	*ry+/ry+*	*ry+/ry*	*ry/ry*
Effectifs observés	38	44	18
Effectifs théoriques	36	48	16

Fréquences génotypiques : $f(ry+/ry+) = 0{,}38$ $f(ry+/ry) = 0{,}44$ $f(ry/ry) = 0{,}18$

Fréquences alléliques : $f(ry+) = 0{,}38 + 0{,}44/2 = p = 0{,}6$

et $f(ry) = 0{,}15 + 0{,}44/2 = q = 0{,}4$

> ***Remarque :*** ces fréquences étant estimées directement, sans aucune hypothèse *a priori*, sont, par principe, meilleures que les premières et se substituent à elles.

En multipliant les fréquences p^2, $2pq$ et q^2 par la taille de l'échantillon, on obtient les effectifs attendus sous l'hypothèse de l'équilibre panmictique de Hardy-Weinberg. La valeur du χ^2 est égale à 0,694, ce qui très inférieur au seuil de 5 % pour un χ^2 à 1 ddl. L'équilibre de Hardy-Weinberg est donc une hypothèse largement acceptable.

Question 3.a : on remarquera que la xanthine-déshydrogénase étant un homodimère, l'hétérozygote *f/s* présente trois bandes électrophorétiques correspondant à des dimères de type *F/F*, *F/S* ou *S/S*.

Par ailleurs les hétérozygotes *f/ry* ou *s/ry* ne présentent qu'un seul type de chaîne à l'électrophorèse, celle codée par l'allèle *f* ou *s*, car *ry* est une mutation amorphe entraînant l'absence de chaîne ; de ce fait les phénotypes électrophorétiques, considérés isolément, ne peuvent permettre de distinguer sans ambiguïté les génotypes *f/f* et *f/ry* d'une part, *s/s* et *s/ry* d'autre part. Seule la connaissance conjointe du phénotype d'activité permet de faire la distinction.

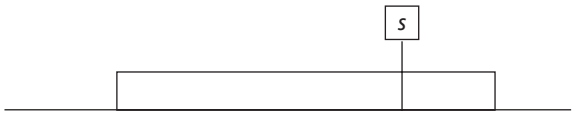
En l'absence d'informations complémentaires il est évidemment impossible, avec les phénotypes morphologiques, de distinguer l'homozygote *ry+/ry+* de l'hétérozygote *ry+/ry*. L'information complémentaire peut être ici le phénotype d'activité, ou en règle générale l'analyse de la descendance par test-cross avec un individu *ry/ry*.

Question 3.b : parmi les 44 hétérozygotes *ry+/ry*, 43 présentent un allèle fonctionnel de type *s*, ce qui est conforme à l'origine de cet allèle (les fondatrices sauvages) ; par contre un hétérozygote présente un allèle fonctionnel de type *f*.

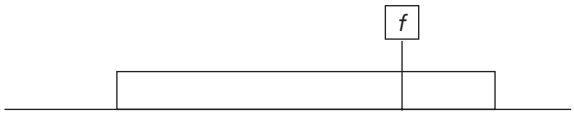
De la même manière, parmi les homozygotes *ry+/ry+*, 37 présentent deux gènes sauvages identiques de type *s*, tandis qu'un organisme présente un gène fonctionnel *ry+* de type *s* et un gène fonctionnel *ry+* de type *f*.

On pourrait imaginer que ces deux allèles *ry+* de type *f* puissent être apparus par mutation, mais il est plus vraisemblable de considérer que ces allèles sont apparus par recombinaison intra-génique entre le site de mutation électrophorétique et le site de mutation fonctionnelle.

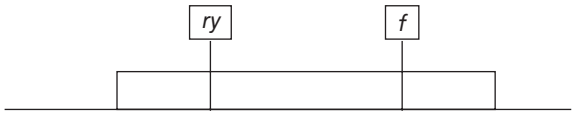
En effet, l'allèle de la souche sauvage d'origine peut être visualisé ainsi :



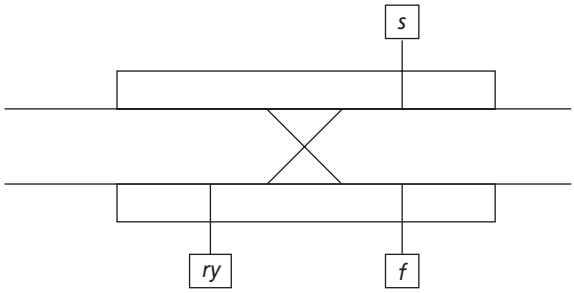
L'allèle fonctionnel de la souche sauvage, dont a été tirée la souche mutante rosy, ainsi



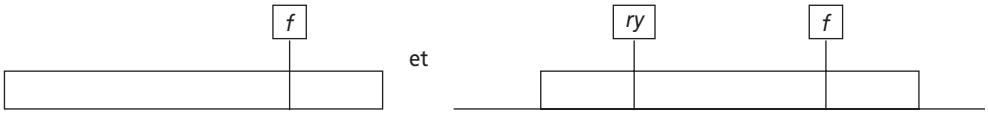
Et l'allèle muté, non fonctionnel, rosy, ainsi :



Aussi, dans les générations qui suivirent la fondation, le génotype des hétérozygotes *ry+/ry* avait la structure suivante :



Alors, un éventuel mais rare crossing-over intragénique, entre le site de la mutation *ry* affectant la fonctionnalité du gène (sans pour autant abolir sa séquence : par exemple une mutation non-sens) et le site de mutation électrophorétique (qui est presque toujours une mutation ponctuelle de substitution d'un acide aminé), aura construit deux nouveaux allèles :



Évidemment seul le produit du premier peut être observé sur les gels puisqu'il est fonctionnel alors que l'allèle recombiné réciproque code pour un produit non fonctionnel qui ne peut être visualisé sur le gel.

Ceci est un exemple typique de déséquilibre gamétique et de déséquilibre de liaison, non pas ici entre deux gènes mais entre deux sites ou marqueurs intra-géniques.

En effet la combinaison en cis, aussi appelée haplotype, (*ry+*,*f*) a une fréquence de 2 sur 200. Autrement dit la fréquence des gamètes (*ry+*,*f*) est égale à 1 %.

Par contre, les fréquences « alléliques » des sites *ry+*, *ry*, *f* et *s*, n'ont pas changé, puisqu'il y a équilibre de Hardy-Weinberg (cela se vérifie facilement pour la fréquence des sites ry+ qui est toujours égale à 0,6).

Si la population était à l'équilibre gamétique, alors la fréquence des gamètes ou haplotypes (*ry+*, *f*) devrait être égale à $0{,}6 \times 0{,}4$, soit 0,24.

Il demeure un déséquilibre gamétique $\Delta = 0{,}01 - 0{,}24 = -0{,}23$.

Ce déséquilibre durera très longtemps car les deux sites sont très proches, très liés, et on peut alors parler de déséquilibre de liaison (voir 3.5.4).

Exercice 2.9 : modèle de Hardy-Weinberg et conseil génétique

La mucoviscidose est une maladie létale, à mode de transmission autosomique récessif (monogénique).

Question 1 : un enfant sur 2 500 est atteint. Quelle est la fréquence de l'allèle pathogène et celle des porteurs sains ?

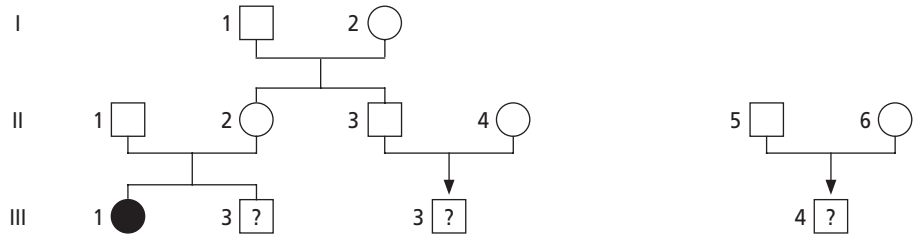
Question 2 : quelle est la fréquence des couples à risques ?

Sachant celle-ci, quelle est la fréquence de naissance d'un enfant atteint ?

Question 3 : il est désormais possible de faire un diagnostic prénatal de mucoviscidose, chez les couples qui, ayant déjà eu un premier enfant atteint, sont identifiés comme couple à risque. Le fait d'interrompre la grossesse pour les embryons « atteints », porteurs des deux copies parentales mutées, va-t-il avoir un effet sur la fréquence allélique de la mutation dans la population ? Est-ce de l'Eugénisme ?

Question 4 : cette possibilité de diagnostic anténatal amène à la consultation d'autres couples que les couples à risque 1/4, comme l'illustre la généalogie ci-dessous.

Pouvez-vous calculer le risque pour chacun des couples II-1 × II-2, II-3 × II-4 et II-5 × II-6 d'avoir un enfant atteint.



Solution

Question 1 : on peut utiliser le modèle de Hardy-Weinberg, bien que la maladie soit létale et qu'il y ait sélection si on considère qu'il y a panmixie pour ce gène et que la sélection a peu d'effets sur les fréquences alléliques sur l'espace de quelques générations, surtout si cette fréquence est faible (voir chapitre 7). Dans ce cas, si q est la fréquence de l'allèle pathogène chez les parents, la fréquence des enfants atteints sera égale à q^2, dont on nous donne la valeur 1/2 500.

D'où $q = \sqrt{1/2\,500} = 1/50$, soit $q = 2$ %.

La fréquence des porteurs sains, c'est-à-dire des hétérozygotes est égale à $2pq = 2q(1 - q) = 2q$, le double de la fréquence de la mutation, si q est petit devant p. La fréquence des porteurs sains est égale à 4 %, soit une personne sur 25.

Question 2 : les couples à risque sont ceux qui peuvent donner naissance à un enfant atteint, comme la maladie est létale avant l'age de la reproduction (de plus les garçons atteints sont stériles), les seuls couples à risque sont les couples de porteurs sains, c'est-à-dire d'hétérozygotes. Sous l'hypothèse panmictique, la fréquence de ces couples est égale à $2pq \times 2pq$ soit 16/10 000.

Comme, pour ces couples, le risque (mendélien) de donner naissance à un enfant atteint est de 1/4, la fréquence des enfants atteints dans la population sera égale à la fréquence de ces couples multipliée par le risque 1/4, soit 1/2 500 !

Question 3 : le diagnostic anténatal suivi de l'IVG en cas de diagnostic génétique défavorable (le fœtus est porteur de deux copies mutées du gène : l'enfant sera atteint de mucoviscidose) ne modifiera nullement les fréquences alléliques. Comme, de toute façon, les individus atteints meurent sans laisser de descendance, le fait que l'homme anticipe, par son diagnostic et son IVG ce que la nature aurait réalisé plus tard, ne change rien aux phénomènes qui régissent l'évolution ou le maintien des fréquences alléliques.

En d'autres termes, l'action de l'homme si elle peut et doit prêter à un jugement en terme d'éthique, ne saurait être ici qualifiée d'eugénisme.

L'eugénisme est une pratique qui consiste à abaisser la fréquence de certains gènes jugés délétères ; ce n'est pas le cas ici, malgré les discours de certains scientifiques ou politiques.

Tenter de disqualifier le diagnostic génétique et l'IVG, pour des maladies graves et incurables, en lui apposant le sceau d'infamie de l'eugénisme, montre que ni le terme d'eugénisme, ni la génétique des populations, ne sont correctement appréhendés par ces « décideurs », ce qui peut brouiller leur jugement pour discerner à quel moment ou en quel endroit un véritable eugénisme est éventuellement mis en œuvre.

Question 4 : risque pour chacun des couples II-1 × II-2, II-3 × II-4 et II-5 × II-6 d'avoir un enfant atteint.

a) Pour le couple II-1x II-2, le risque d'avoir un enfant atteint est évidemment 1/4.

b) Pour le couple II-3 × II-4 le risque d'avoir un enfant atteint dépend de deux évènements :

– II-3 est-il porteur sain ? Évènement dont le risque *R1* est élevé puisque sa sœur l'est.

– II-4 est-il aussi porteur sain ? Évènement dont le risque *R2* est connu : il est égal à la fréquence des porteurs sains dans la population générale, soit 1/25.

Le risque d'avoir un enfant atteint sera égal à $R1 \times R2 \times 1/4$, soit ***R1*/100**.

Quelle est alors la valeur de *R1* ?

Si II-2 est porteur sain, c'est qu'elle a reçu un allèle muté de l'un de ses parents.

On sait donc que l'un des parents de II-2 est *N/m*, le deuxième peut être aussi *N/m* avec la probabilité 1/25 (fréquence des hétérozygotes dans la population générale), ou bien *N/N*, avec la probabilité complémentaire 24/25.

Dans le premier cas, la probabilité que II-3 soit hétérozygote est égale à 1/2, si on ne connaît pas son phénotype ; mais sachant qu'il n'est pas atteint, cette probabilité est de 2/3.

Dans le deuxième cas, la probabilité que II-3 soit hétérozygote est égale à 1/2.

Globalement le risque *R1* que II-3 soit hétérozygote est égal à $(1/25 \times 2/3) + (24/25 \times 1/2)$, soit 38/75.

Le risque de naissance d'un enfant atteint est donc $R1 \times R2 \times 1/4$, soit $38/75 \times 1/25 \times 1/4$, soit environ 1/197. C'est 12,5 fois plus que le risque dans la population générale.

> ***Remarque :*** très souvent la mutation chez II-2 a été identifiée au cours de l'étude diagnostique réalisée pour le dépistage anténatal (ou indirectement identifiée par un marqueur polymorphe de l'ADN). Il est alors possible de diagnostiquer la présence ou l'absence de cette mutation chez II-3. Si la réponse est l'absence, II-3 est alors *N/N* et le risque pour l'enfant à naître devient nul ; il est même inutile d'analyser la mère. Si la réponse est la présence de la mutation, le risque *R1* devient égal à 1, et le risque global pour l'enfant à naître monte à 1 % (25 fois plus que le risque général).
>
> La question qui se pose est de pouvoir identifier une éventuelle mutation chez la mère, quand on sait que plus de 1 000 mutations différentes du gène CFTR sont connues. Certes on peut étudier les plus fréquentes, mais c'est très long, très cher, et sans garantie de ne pas passer à côté d'une des mutations qui n'aurait pas été recherchée. Il faut alors faire une étude complète du gène.

c) Pour le couple II-5 × II-6, le risque d'avoir un enfant atteint est évidemment celui de la population générale, soit 1/2 500.

Dépister les couples à risque, en dépistant les hétérozygotes, renvoie, mais pour les deux individus de tout couple, au problème soulevé, à la remarque précédente pour la seule mère II-4.

Il est impossible d'imaginer tester, par la biologie moléculaire, en temps utile, à moindre frais et avec fiabilité tous les couples, mais des stratégies de dépistage en deux temps sont à l'étude qui confineraient l'analyse moléculaire de l'ADN dans un petit groupe à risque, défini par un premier tri, rapide, fiable et peu coûteux.

Exercice 2.9

Au Danemark, la mucoviscidose affecte un nouveau-né sur 4 700, alors qu'en France elle affecte un nouveau né sur 2 500.

Au Danemark, 87 % des chromosomes porteurs d'un allèle pathologique sont porteurs de la mutation ΔF508 tandis qu'en France la fréquence relative des chromosomes ΔF508 parmi les chromosomes porteurs d'un gène muté n'est que de 70 % en moyenne.

Dans ce problème :

– la fréquence de l'allèle « normal » (fonctionnel), noté *N*, sera notée *p*, la fréquence de l'allèle Δ*F* sera notée *q1* et celle de l'ensemble des allèles non-Δ*F* sera notée *q2* ;

– dans le cas où il n'est pas utile de distinguer entre les allèles Δ*F* et non Δ*F*, la fréquence globale des allèles mutés sera notée $q = q1 + q2$;

– les valeurs des fréquences seront données avec une précision de cinq chiffres après la virgule ;

– les fréquences des homozygotes *N/N*, des porteurs sains *N*/Δ*F* et des porteurs sains *N*/non-Δ*F* seront notées respectivement *D*, *H1* et *H2*.

Question 1 : quelles sont, dans chacun des deux pays :

a) Les valeurs de *q*, *q1* et *q2* ? Justifiez vos réponses.

b) Les fréquences respectives *D*, *H1* et *H2* des homozygotes *N/N*, des porteurs sains *N*/Δ*F* et des porteurs sains *N*/non*ΔF*.

Vous établirez d'abord les fréquences *D*, *H1* et *H2* en fonction de *p*, *q1* et *q2*, en justifiant les étapes du raisonnement et vous ferez ensuite l'application numérique pour chacune des populations (les fréquences des individus atteints seront négligées face à *D*, *H1* et *H2*).

Question 2 : quels sont les différents types de couples dans la population si on distingue les deux types de porteurs sains *N*/ Δ*F* et *N*/non*ΔF* ?

Vous donnerez leurs fréquences respectives en fonction de *D*, *H1* et *H2*, puis de *p*, *q1* et *q2* (pas d'application numérique)

Question 3 : en supposant qu'un dépistage systématique des porteurs de Δ*F* soit entrepris dans chacune des populations, quelle serait la proportion des couples à risque qui pourraient être dépistés avant la première naissance ? Vous établirez cette proportion de façon littérale (en fonction de *D*, *H1* et *H2*, ou en fonction de *p*, *q1* et *q2* puis en fonction de *C*), puis vous ferez l'application numérique.

Question 4 : en supposant une efficacité parfaite du dépistage, puis un diagnostic prénatal avec interruption systématique en cas d'atteinte pathologique, quelle serait alors la nouvelle fréquence de la maladie dans chacune des populations ?

Dans quelle population, le dépistage aurait-il la plus grande efficacité et pourquoi ?

Solution

Question 1 : il convient de faire l'ensemble des calculs de ce problème sous le modèle de Hardy-Weinberg. Bien qu'avec la mucoviscidose la condition d'absence de sélection ne soit pas valide, on sait que l'effet de la sélection sur les fréquences alléliques et génotypiques est assez faible sur quelques générations pour pouvoir être négligé.

a) Sous le modèle de *H-W*, la fréquence *R* des enfants atteints d'une maladie récessive est égale à q^2, où *q* est la fréquence de l'allèle pathologique.

En prenant la racine de *R*, on en déduit la valeur de *q*, soit :

– au Danemark : $q = 0{,}01458$; avec $q1 = 0{,}87q = 0{,}01268$ et $q2 = 0{,}00190$

– en France : $q = 0{,}02000$; avec $q1 = 0{,}70q = 0{,}01400$ et $q2 = 0{,}00600$

b) La fréquence des porteurs sains est égale à $H = 2q(1 - q)$, et $H = 2q$ si *q* est petit.

En fonction de *q1*, on aura, au Danemark, $H1 = 2q1(1 - q) = 0{,}87 \times 2q(1 - q) = 0{,}87H$

Le rapport *H1*/*H* est égal au rapport *q1*/*q*, d'où on tire les valeurs :

– au Danemark : $H = 0{,}02875$ avec $H1 = 0{,}87H = 0{,}0250$ et $H2 = 0{,}00374$

Si on néglige $(1 - q)$, $H = 0{,}02917$ avec $H1 = 0{,}02538$ et $H2 = 0{,}00379$

– en France : $H = 0{,}03920$ avec $H1 = 0{,}70H = 0{,}02744$ et $H2 = 0{,}01176$

Si on néglige $(1 - q)$, $H = 0{,}04000$ avec $H1 = 0{,}02800$ et $H2 = 0{,}01200$

La fréquence D des homozygotes N/N est exactement égale à p^2, soit 0,97104 au Danemark et 0,96040 en France, mais elle est égale à $D = 1 - H$, si on néglige la fréquence très faible des malades, soit 0,97125 au Danemark et 0,96080 en France.

Question 2 : les six types de couples et leurs fréquences respectives sont donnés par le tableau ci-dessous :

Types de couples	N/N x N/N	N/N x $N/\Delta F$	N/N x $N/\text{non}\Delta F$	$N/\Delta F$ x $N/\Delta F$	$N/\Delta F$ x $N/\text{non}\Delta F$	$N/\text{non}\Delta F$ x $N/\text{non}\Delta F$
Fréquences en fonction de D, $H1$ et $H2$	D^2	$2D.H1$	$2D.H2$	$H1^2$	$2H1.H2$	$H2^2$
Fréquences en fonction de p, $q1$ et $q2$	p^4	$4p^3.q1$	$4p^3.q2$	$4p^2.q1^2$	$8p^2.q1.q2$	$4p^2.q2^2$

Question 3 : les couples à risques dépistables sont ceux où les deux conjoints sont $N/\Delta F$, leur proportion parmi les couples à risque est donc égale à

$H1^2/(H1^2 + H2^2 + 2H1.H2)$ ou $4q1^2/(4q1^2 + 4q2^2 + 8q1.q2)$

On remarquera que c étant la proportion d'allèles étant ΔF (0,87 ou 0,70 selon les deux cas), on peut écrire que $q1 = cq$ ou que $H1 = cH$, et la relation devient simplement égale à c^2, ce qui est le résultat donné par le simple tableau des couples possibles de porteurs sains, chacun des conjoints pouvant être $N/\Delta F$ avec la probabilité c et $N/\text{non}\Delta F$ avec la probabilité $(1 - c)$.

Application numérique : la proportion c^2 des couples dépistables est égale à 0,7569 au Danemark et 0,49000 en France.

Question 4 : si on dépiste X % des couples à risque et qu'on pratique un DPN, on évite X % des naissances et il en demeure $(1 - X)$ %, la fréquence de la maladie est donc multipliée par $(1 - X)$.

Au Danemark, elle passerait de 1/4 700 à 1/19 333, et en France de 1/2 500 à 1/4 902.

Le dépistage serait beaucoup plus efficace au Danemark (fréquence de la maladie divisée par plus de quatre) qu'en France (fréquence simplement divisée par deux) parce que la fréquence relative de $\Delta F508$ est beaucoup plus élevée au Danemark qu'en France.

En effet, $X = c^2$ et l'efficacité est donnée par le rapport $1/(1 - X)$ soit $1/(1 - c^2)$; si $c = 0,7$, ce rapport est égal 2 et si $c = 0,9$, ce rapport devient égal à 5.

Chapitre **3**

Généralisation du modèle de Hardy-Weinberg

3.1 INTRODUCTION

Sous le modèle de Hardy-Weinberg, établi dans le chapitre précédent, la composition génétique d'une population demeure stable, inchangée d'une génération à l'autre, si trois conditions sont réunies :

1. absence ou effet négligeable sur quelques générations, des mutations, de la sélection et des migrations ;
2. taille infinie ou assez grande pour que les fréquences des évènements soient égales à leurs probabilités (loi des grands nombres). Sous ces deux conditions, les fréquences alléliques demeurent inchangées ;
3. panmixie (incluant la condition de pangamie), c'est-à-dire croisements au hasard.

Sous cette condition, une relation mathématique, désignée par relation panmictique ou relation de Hardy-Weinberg, permet de calculer les fréquences génotypiques à partir des seules fréquences alléliques. En effet, cette relation, conséquence directe du schéma de l'urne gamétique, stipule que la somme des fréquences génotypiques est égale au développement du carré de la somme des fréquences alléliques, soit :

$$[p^2 + 2pq + q^2] = (p + q)^2$$

Les fréquences alléliques demeurant inchangées, sous l'effet des deux premières conditions, les fréquences génotypiques, résultant de la troisième, demeurent également inchangées. La situation d'équilibre ainsi réalisée définit l'équilibre de Hardy-Weinberg.

Mais ce modèle de Hardy-Weinberg a été établi dans une situation biologique et génétique simple d'un seul gène, autosomique et di-allélique, dans une population d'organismes à sexes et à générations séparés. Il est maintenant nécessaire et utile de montrer comment le modèle de Hardy-Weinberg peut être généralisé au cas des gènes portés par les hétérochromosomes, à celui de gènes pluri-alléliques, au cas des populations à générations chevauchantes, ainsi qu'à l'analyse de la composition génétique d'une population, et de son évolution, quand deux gènes sont pris en compte simultanément.

3.2 GÉNÉRALISATION DU MODÈLE DE HARDY-WEINBERG À UN GÈNE PLURI-ALLÉLIQUE

Un gène pluri-allélique présente n formes alléliques notées A_1 , A_2 , ..., A_n.

Leurs fréquences alléliques sont respectivement notées p_1 , p_2 , ..., p_n.

Sous les trois conditions du modèle de Hardy-Weinberg rappelées dans l'introduction, ces fréquences alléliques sont stables et définissent la composition d'urne gamétique, invariante d'une génération à l'autre. Sous la condition de panmixie, le schéma de tirage aléatoire de deux gamètes dans cette urne permet de constituer les descendants de la génération suivante. Il est facile de voir que le tableau à quatre cases obtenu dans le cas d'un gène di-allélique, est ici remplacé par un tableau à n cases (tableau 3.1) :

TABLEAU 3.1

Gamètes (fréquence)	***A1*** ***p_1***	***A_2*** ***p_2***		***A_i*** ***p_i***		***A_n*** ***p_n***
A1 ***p_1*** ***A_2*** ***p_2***	*A/A1* p_1^2	A_2/A_2 p_2^2		$A1/A_i$ $p_1\,p_i$		
A_i ***p_i***	$A_i/A1$ $p_i\,p_1$			$A1_i/A_i$ p_i^2		
A_n ***p_n***						A_n/A_n p_n^2

Les fréquences génotypiques des n homozygotes sont égales au carré de la fréquence de l'allèle dont ils sont porteurs ; et les fréquences génotypiques des $n(n-1)/2$ hétérozygotes (pour le dénombrement des génotypes, voir encart 1.1, page 9) sont égales au double produit des fréquences des deux allèles dont ils sont porteurs.

La relation de Hardy-Weinberg liant fréquences alléliques et génotypiques d'un gène di-allélique :

$$(p + q)^2 = p^2 + 2pq + q^2$$

se généralise donc, dans le cas d'un gène pluri-allélique, à :

$$(p_1 + p_2, +,...., + p_n)^2 = \Sigma p_i^{\,2} + \Sigma\, 2\, p_i\, p_j \quad \text{avec } (i > j)$$

NB : si on prend les cases du tableau, la fréquence d'un hétérozygote quelconque A_iA_j s'écrit $p_i\, p_{j\,+}\, p_i\, p_j$, mais il faut bien supposer que *i* est différent de *j*, car alors ce serait un homozygote et la formule serait inexacte puisque qu'on aurait deux fois le carré p_i^2. Si, maintenant, on souhaite remplacer la somme $p_i\, p_{j\,+}\, p_i\, p_j$ par $2p_i\, p_j$, il faut bien supposer que non seulement *i* est différent de *j*, mais aussi que l'un des indices est plus grand que l'autre car, si *i* et *j*, tout en étant différents, pouvaient prendre toutes les valeurs possibles, on compterait par exemple $2p_3p_4$ et aussi $2\, p_4p_3$, c'est-à-dire qu'on compterait deux fois les hétérozygotes A_3A_4.

Le modèle de Hardy-Weinberg se généralise ainsi très facilement au cas d'un gène pluri-allélique.

> ***Remarque 1 :*** le taux d'hétérozygotie H (voir chapitre 1) est défini comme la fréquence des hétérozygotes dans la population, est donc égal à $1 - \Sigma p_i^{\,2}$.
>
> ***Remarque 2 :*** comme dans le cas d'un gène di-allélique, les fréquences entre les sexes sont égalisées en une génération de panmixie, si elles différaient à la génération précédente.

3.3 GÉNÉRALISATION DU MODÈLE DE HARDY-WEINBERG À UN GÈNE PORTÉ PAR UN HÉTÉROCHROMOSOME

3.3.1 Fréquences alléliques dans chacun des sexes et dans la population

Dans une espèce où le sexe femelle est homogamétique (XX) et le sexe mâle hétérogamétique (XY ou X0), les organismes homogamétiques possèdent deux exemplaires des gènes portés par le chromosome X, et les organismes hétérogamétiques n'en possèdent qu'un seul.

Si les deux sexes sont en nombre égal dans la population, un tiers des chromosomes X sont portés par des mâles (si c'est le sexe hétérogamétique) et deux tiers des chromosomes X le sont par des femelles. Les fréquences *p* et *q* des allèles *A1* et *A2* d'un gène di-allélique du chromosome X, s'écrivent alors :

$p = p_m/3 + 2\, p_f/3$, où p_m et p_f sont les fréquences de *A1* dans les sexes mâle et femelle.

$q = q_m/3 + 2\, q_f/3$, où q_m et q_f sont les fréquences de *A2* dans les sexes mâle et femelle.

Si les fréquences sont égales dans les deux sexes, on a alors

$$p_m = p_f = p \qquad \text{et} \qquad q_m = q_f = q$$

3.3.2 Équilibre de Hardy-Weinberg pour un gène hétérosomique avec des fréquences alléliques égales dans chacun des sexes

Sous les conditions du modèle de Hardy-Weinberg, la construction de la génération suivante par l'application du schéma des urnes gamétiques montre une situation d'équilibre (tableau 3.2), sachant que la moitié des spermatozoïdes sont porteurs d'un chromosome X, lui-même porteur de *A1* (probabilité *p*) ou *A2* (probabilité *q*), et l'autre moitié sont porteurs d'un chromosome Y :

TABLEAU 3.2

	Spermatozoïdes X porteur de *A1* 1/2(p)	**Spermatozoïdes X porteur de *A2* 1/2(q)**	**Spermatozoïdes Y 1/2**
Ovules avec X porteur de *A1* (p)	$1/2 \cdot p^2$	$1/2 \cdot pq$	$1/2 \cdot p$
Ovules avec X porteur de *A1* (q)	$1/2 \cdot pq$	$1/2 \cdot q^2$	$1/2 \cdot q$
Fréquences génotypiques dans la population	*pour le sexe homogamétique* $(p^2 + 2pq + q^2)/2$		*pour le sexe hétérogamétique* $(p + q)/2$

La somme, sur les deux sexes, des fréquences génotypiques, est bien égale à 1, et les fréquences génotypiques, au sein de chacun des sexes, sont évidemment ($p^2 + 2pq + q^2$) pour le sexe homogamétique et ($p + q$) pour le sexe hétérogamétique, la pondération par 1/2 dans le tableau résultant du fait qu'on y raisonne sur la population totale avec un sexe ratio équilibré.

On observe que les fréquences alléliques n'ont pas changé et sont restées égales à *p* et *q* dans chacun des deux sexes. On retrouve évidemment, dans le sexe homogamétique, la relation de Hardy-Weinberg.

Conclusion : pour un gène porté par un hétérochromosome, avec des fréquences alléliques égales dans chacun des deux sexes, l'équilibre de Hardy-Weinberg est caractérisé par une relation de Hardy-Weinberg restreinte aux seules fréquences génotypiques du sexe homogamétique.

> ***Remarque :*** l'absence d'hétérozygotes chez les organismes mâles hémizygotes entraîne des différences importantes entre les fréquences des phénotypes chez les hommes et celles des phénotypes correspondant aux homozygotes chez les femmes. Ainsi le génotype *A2*/Y a une fréquence *q* chez les hommes alors que l'homozygote *A2A2* a une fréquence q^2, beaucoup plus petite, chez les femmes. Ceci explique la différence considérable entre les nombres d'hommes ou de femmes atteints d'une maladie récessive liée au sexe (voir les exemples chapitre 2).

3.3.3 Évolution de la composition génétique d'une population vers l'équilibre de Hardy-Weinberg quand les fréquences alléliques sont inégales entre les sexes

Cette situation peut naturellement toucher une population dès que survient un phénomène de mélange ou d'extinction (par exemple une guerre chez l'homme) affectant un sexe plus que l'autre.

On a vu, dans le cas d'un gène autosomique, que des fréquences alléliques différentes dans chacun des sexes sont égalisées en une génération de panmixie. Au contraire, dans le cas d'un gène hétérosomique, les fréquences alléliques ne s'égalisent qu'asymptotiquement, après plusieurs générations.

Prenons une génération initiale, notée g_0, où les fréquences alléliques diffèrent entre les sexes, pour un gène du chromosome X, et suivons l'évolution de l'une d'entre elle, par exemple celle de *A1* (l'évolution de la fréquence de *A2* est évidemment le complément à 1 de celle de *A1*).

À la génération g_0, on a : $f(A1) = p_{m,0}$ dans le sexe mâle

et : $f(A1) = p_{f,0}$ dans le sexe femelle

Avec, pour l'ensemble de la population, tous sexes confondus, en supposant un sexe ratio équilibré, une fréquence moyenne de l'allèle *A1* égale à :

$$\boldsymbol{p_0 = p_{m,0}/3 + 2\,p_{f,0}/3}$$

NB : si le sexe ratio est équilibré, il y a cependant seulement un tiers de chromosomes X présent dans le sexe hétérogamétique, contre deux tiers des X présent dans le sexe homogamétique, et la fréquence moyenne de l'allèle *A1*, sur l'ensemble des sexes, doit tenir compte de ce rapport.

Après une génération de panmixie, quelles seront les fréquences alléliques dans chacun des sexes ?

Comme tous les mâles de la génération g_1 ont reçu un chromosome X de leur mère, celui-ci porte l'allèle *A1* avec la fréquence de cet allèle, chez les mères de la génération g_0, d'où, à la génération g_1 :

$$\boldsymbol{p_{m,1} = p_{f,0}}$$

Chez les femelles de la génération g_1, la moitié des chromosomes X sont d'origine maternelle et la moitié d'origine paternelle. La moitié des chromosomes X, d'origine maternelle, porteront l'allèle *A1* avec la fréquence $p_{f,0}$, et la moitié des chromosomes X, d'origine paternelle, porteront *A1* avec la fréquence $p_{m,0}$, d'où, à la génération g_1 :

$$\boldsymbol{p_{f,1} = (p_{f,0} + p_{m,0})/2}$$

Ces deux relations constituent des relations de récurrence liant les fréquences alléliques d'une génération à celles de la génération précédente. Ce qui est valable entre les générations g_0 et g_1 , l'est aussi entre les générations g_{n-1} et g_n, ce qui permet d'écrire les deux relations de récurrence générales :

$$\boldsymbol{p_{m,n} = p_{f,n-1}}$$

$$\boldsymbol{p_{f,n} = (p_{f,n-1} + p_{m,n-1})/2}$$

Les fréquences alléliques vont donc varier d'une génération à l'autre. Dans une telle circonstance le généticien des populations se posera toujours deux questions :

– vers quel état évoluera la composition génétique de la population ? Notamment existe-t-il un état d'équilibre polymorphe ?
– avec quelle vitesse cette évolution se réalisera-t-elle ?

Dans le cas présent, il est facile de répondre à la première question, en considérant la différence des fréquences alléliques entre les sexes. À la génération g_n, elle est égale, par définition, à :

$$\boldsymbol{p_{m,n} - p_{f,n}}$$

Mais, en fonction des relations de récurrence définies précédemment, la différence des fréquences alléliques à la génération g_n peut s'exprimer en fonction des fréquences alléliques à la génération g_{n-1}, d'où :

$$p_{m,n} - p_{f,n} = p_{f,n-1} - (p_{f,n-1} + p_{m,n-1})/2$$

soit

$$\boldsymbol{p_{m,n} - p_{f,n} = (-1/2)\,(p_{m,n-1} - p_{f,n-1})}$$

D'une génération à l'autre, la différence de fréquences alléliques entre les sexes est divisée par deux et change de signe. Si cette différence est divisée par deux à chaque génération, elle va tendre vers zéro, situation où les fréquences alléliques seront alors égales dans les deux sexes.

On peut exprimer mathématiquement cette évolution, en repartant de la différence ci-dessus qui peut s'écrire, en fonction des fréquences à la génération g_{n-2}, comme :

$$\boldsymbol{p_{m,n} - p_{f,n} = (-1/2)^2\,(p_{m,n-2} - p_{f,n-2})}$$

ce qui donne, en remontant à la génération g_0 :

$$\boldsymbol{p_{m,n} - p_{f,n} = (-1/2)^n\,(p_{m,0} - p_{f,0})}$$

Il est clair que le terme $(-1/2)^n$ tend vers zéro quand n augmente, ce qui signifie bien que la différence des fréquences alléliques entre les sexes tend vers zéro, et que les fréquences alléliques tendent vers une valeur d'équilibre égale entre les sexes.

Quelle est cette valeur d'équilibre ?

Rappelons que la fréquence globale dans la population, à la génération g_0 a été définie comme :

$$\boldsymbol{p_0 = p_{m,0}/3 + 2\,p_{f,0}/3}$$

À la génération g_1, on aura : $p_1 = p_{m,1}/3 + 2\,p_{f,1}/3$

En appliquant les relations de récurrence donnant $p_{m,1}$ et $p_{f,1}$ en fonction de $p_{m,0}$ et $p_{f,0}$, on peut écrire que

$$p_1 = p_{f,0}/3 + 2\,[(p_{m,0} + p_{f,0})/2]/3$$

d'où

$$p_1 = p_{m,0}/3 + 2p_{f,0}/3$$

et

$$\boldsymbol{p_1 = p_0}$$

Autrement dit, la valeur de la fréquence globale ou moyenne entre les sexes ne varie pas d'une génération à l'autre. Ce qui est valable entre les générations g_0 et g_1, l'est aussi entre les générations g_{n-1} et g_n, puis à l'équilibre, quand les fréquences alléliques sont égales dans les deux sexes.

Si on note p_e, la valeur de la fréquence de l'allèle *A1*, quand elle est devenue égale dans chacun des deux sexes, cette valeur p_e vérifie aussi la relation :

$$p_0 = p_e/3 + 2\,p_e/3$$

On en tire que

$$\boldsymbol{p_0 = p_e}$$

La valeur d'équilibre p_e vers laquelle tendent les fréquences alléliques de chacun des deux sexes est donc la fréquence moyenne initiale p_0 qui reste inchangée d'une génération à l'autre.

Mais cette égalité des fréquences alléliques dans chacun des sexes, n'est obtenue qu'asymptotiquement, quand le terme $(-1/2)^n$, affectant la différence entre ces fréquences, peut être considéré comme nul. En pratique, l'équilibre est atteint au bout de 8 à 10 générations, ce qui représente 4 à 5 mois chez la drosophile mais environ un siècle chez l'homme.

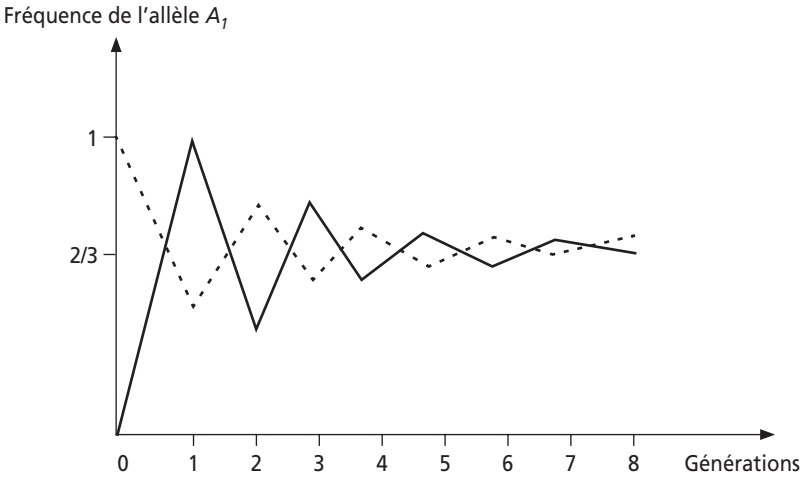


Figure 3.1 Évolution de la valeur de la fréquence de l'allèle *A1*.
Trait plein : sexe masculin
Pointillés : sexe féminin

L'ensemble de ces résultats est illustré par la figure 3.1. On y a représenté l'évolution de la fréquence de l'allèle *A1*, dans le cas extrême d'une population où, à la génération g_0, toutes les femelles sont *A1A1* ($p_{f,0} = 1$), et les mâles *A2* ($q_{m,0} = 1$, donc $p_{m,0} = 0$). Dans une telle population, la fréquence globale entre les sexes est égale à $p_0 = 0/3 + 2/3 = 2/3$.

On retrouve sur cette figure l'illustration des relations de récurrence, avec :

– une fréquence de *A1* chez les mâles égale à la fréquence de *A1* chez les mères ;
– une fréquence de *A1*, chez les femelles, égales à la moyenne arithmétique des fréquences dans les deux sexes, à la génération précédente ;
– l'inversion du signe de la différence de fréquences entre les sexes.

3.4 GÉNÉRALISATION DU MODÈLE DE HARDY-WEINBERG AU CAS DES GÉNÉRATIONS CHEVAUCHANTES

Quand les générations ne sont pas séparées, il y a un renouvellement continu des individus. Pendant une fraction de temps dt, une fraction de la population, proportionnelle à la fraction de temps *dt* considérée, décède et est remplacée par de nouvelles naissances.

Si la population est grande et panmictique, ces nouvelles naissances résultent d'un tirage de gamètes dans une urne dont la composition en allèles (*A1* : *p* et *A2* : *q*) est invariante, en l'absence de mutations, de sélection et de migrations.

Les nouvelles naissances d'hétérozygotes, pendant un temps *dt*, sont alors proportionnelles au produit $2pq.dt$.

Si la fréquence des hétérozygotes au temps *t* est notée H_t, la fréquence, au temps $t + dt$, sera H_{t+dt}.

Pendant la fraction *dt* de génération, une fraction $H_t.dt$ va décéder pour être remplacée par les nouvelles naissances représentant $2pq.dt$. D'où la relation :

$$H_{t+dt} = H_t - H_t.dt + 2pq.dt$$

soit
$$(H_{t+dt} - H_t)/dt = -H_t + 2pq$$

ou
$$dH_t/dt = -H_t + 2pq$$

En intégrant cette équation du temps 0 au temps *t*, on obtient :

$$H_t = 2pq - (2pq - H_0)\, e^{-t}$$

On voit qu'avec l'augmentation du temps *t*, la fréquence des hétérozygotes tend vers la valeur permanente $2pq$, caractéristique de l'équilibre de Hardy-Weinberg.

3.5 MODÈLE DE HARDY-WEINBERG APPLIQUÉ À L'ANALYSE DE LA COMPOSITION GÉNÉTIQUE D'UNE POPULATION POUR DEUX GÈNES ÉTUDIÉS SIMULTANÉMENT

Le modèle de Hardy-Weinberg, tel qu'il a été présenté jusqu'ici s'applique aisément aux caractères dont les variations phénotypiques peuvent ne dépendre que d'un seul gène (caractères monogéniques ou monofactoriels). Mais il est utile de généraliser ce modèle à l'étude simultanée de plusieurs gènes ou polymorphismes de l'ADN. En effet, cette situation est fréquemment rencontrée dans de nombreux problèmes d'analyse génétique. Mais dès qu'on s'intéresse à plus de deux gènes ou polymorphismes, la situation devient si compliquée qu'elle ne permet pas une mise sous équation. On se contentera, dans ce paragraphe, de considérer la composition génétique d'une population simultanément pour deux marqueurs di-alléliques, qu'il s'agisse de gènes ou de simples polymorphismes de l'ADN (voir page 8).

L'étude de cette situation est d'un grand intérêt car elle permet d'introduire un nouveau concept, le déséquilibre gamétique, qui se révèle d'une grande utilité, d'une part en recherche fondamentale, (cartographie des gènes, analyse de la diversité génétique des populations et de l'origine de certaines mutations, modèle

d'analyse de la composante génétique des pathologies en génétique épidémiologique humaine), d'autre part en recherche appliquée (diagnostic et conseil génétique, dépistage des risques génétiques et santé publique).

3.5.1 Fréquences alléliques et fréquences gamétiques

Quand on étudie la diversité génétique relative à un gène, on confond toujours les fréquences alléliques et gamétiques, car si allèles et gamètes ne sont pas des objets de même nature, ils sont, dans ce cas particulier, superposables : un gamète contient un allèle du gène, et la fréquence du gamète *A1* est égale à la fréquence de l'allèle *A1*.

Dès que l'étude génétique d'une population porte sur plus d'un gène, gamètes et allèles ne sont plus superposables, fréquences alléliques et gamétiques non plus.

Dans le cas de deux gènes (ou SNP) di-alléliques A et B, on définit :

– pour le premier gène : les allèles ***A1*** **de fréquence *p***
et ***A2*** **de fréquence *q***

– pour le deuxième gène : les allèles ***B1*** **de fréquence *u***
et ***B2*** **de fréquence *v***

Les fréquences gamétiques correspondent à quatre objets différents, contenant un allèle de chacun des deux gènes, soit les gamètes :

(*A1, B1*) **de fréquence g_{11}**
(*A1, B2*) **de fréquence g_{12}**
(*A2, B1*) **de fréquence g_{21}**
(*A2, B2*) **de fréquence g_{22}**

où la place de l'indice se réfère au gène (gène *A* en première position et *B* en seconde), et la valeur de l'indice se réfère à la forme allélique du gène (valeur 1 pour *A1* ou *B1*, valeur 2 pour *A2* ou *B2*).

3.5.2 Équilibre et déséquilibre gamétique

Il existe bien évidemment une relation entre les valeurs des fréquences gamétiques et celles des fréquences alléliques, *p* et *q* du gène *A*, et *u* et *v*, du gène *B*, mais cette relation n'est pas évidente.

On peut imaginer que les allèles des deux gènes sont réunis indépendamment les uns des autres dans les gamètes, qu'ils y sont associés aléatoirement. Dans ce cas, on dit qu'il y a équilibre gamétique, et les quatre fréquences gamétiques s'écrivent :

pour **(*A1, B1*)** $g_{11} = p.u$
(*A1, B2*) $g_{12} = p.v$
(*A2, B1*) $g_{21} = q.u$
(*A2, B2*) $g_{22} = q.v$

Mais cette situation d'équilibre gamétique n'est ni obligatoire, ni même courante, dans les populations naturelles, même quand elles sont à l'équilibre de Hardy-Weinberg pour chacun des deux gènes (voir plus loin).

Quand la situation d'équilibre gamétique, pour les deux gènes considérés, n'est pas réalisée, on dit qu'il y a déséquilibre gamétique. Celui-ci est défini, pour chacun des quatre gamètes, comme la différence entre la fréquence réelle du gamète et la fréquence théorique qui serait la sienne, si l'équilibre gamétique était réalisé, soit

pour
$$(A1, B1) \quad \Delta_{11} = g_{11} - p.u$$
$$(A1, B2) \quad \Delta_{12} = g_{12} - p.v$$
$$(A2, B1) \quad \Delta_{21} = g_{21} - q.u$$
$$(A2, B2) \quad \Delta_{22} = g_{22} - q.v$$

3.5.3 Genèse d'un déséquilibre gamétique

De nombreuses circonstances dans l'histoire génétique des populations peuvent générer de tels déséquilibres, en fait tous les mécanismes que le modèle de Hardy-Weinberg a supposé inexistants, ou négligeables à l'échelle de quelques générations, les mutations, la sélection, les migrations, la dérive (qui survient quand l'effectif est « petit »), et certains types d'écarts à la panmixie.

a) Genèse d'un déséquilibre gamétique à la suite de migrations

Deux populations, même si elles sont génétiquement proches, ont rarement des compositions génétiques identiques. Dans ces conditions, tout phénomène de migration ou de fusion entre elles, va générer des déséquilibres gamétiques entre deux gènes polymorphes. Parmi les populations humaines, on peut citer le Brésil ou l'île Maurice, dont les populations sont issues d'un mélange de trois continents.

La genèse d'un déséquilibre gamétique est illustrée par l'exemple de la figure 3.2 (pour une meilleure clarté pédagogique des valeurs extrêmes ont été choisies) :

– les populations A et B sont à l'équilibre gamétique puisque les fréquences gamétiques représentent bien les produits des fréquences alléliques ;
– dans la population C, issue d'une fusion à parts égales de A et B, un déséquilibre gamétique survient tout simplement parce que les allèles *A1*, *A2*, *B1* et *B2*, même s'ils ont des fréquences égales dans la population, ne sont pas répartis au hasard chez les individus. En effet, les individus étant *A1B1*/*A1B1* ou *A2B2*/*A2B2*, ils ne peuvent produire que des gamètes (*A1*,*B1*) ou (*A2*,*B2*) et jamais des gamètes (*A1*,*B2*) ou (*A2*,*B1*).

Évidemment, dès la génération suivante, des gamètes recombinés (*A1*,*B2*) et (*A2*,*B1*) pourront être formés par les individus *A1B1*/*A2B2* issus des premiers couples panmictiques *A1B1*/*A1B1* × *A2B2*/*A2B2*.

On comprend donc que les déséquilibres gamétiques ainsi générés vont progressivement disparaître sous l'effet de la recombinaison génétique. Mais il est évident que cette disparition sera d'autant plus lente que la recombinaison génétique est rare (voir plus loin). En effet si les gènes *A* et *B* sont génétiquement indépendants (indépendance physique ou liaison physique avec une distance assez grande pour que de nombreux crossing-over induise une ségrégation indépendante à la méiose), les individus *A1B1*/*A2B2* fourniront autant de gamètes parentaux (*A1*,*B1*) et (*A2*,*B2*) que de

gamètes recombinés (*A1*,*B2*) et (*A2*,*B1*). Par contre, si les gènes *A* et *B* sont très liés, alors les individus *A1B1*/*A2B2* fourniront essentiellement des gamètes parentaux (*A1*,*B1*) et (*A2*,*B2*) et très rarement des gamètes recombinés (*A1*,*B2*) et (*A2*,*B1*). Dans ce cas, le déséquilibre gamétique se maintiendra longtemps.

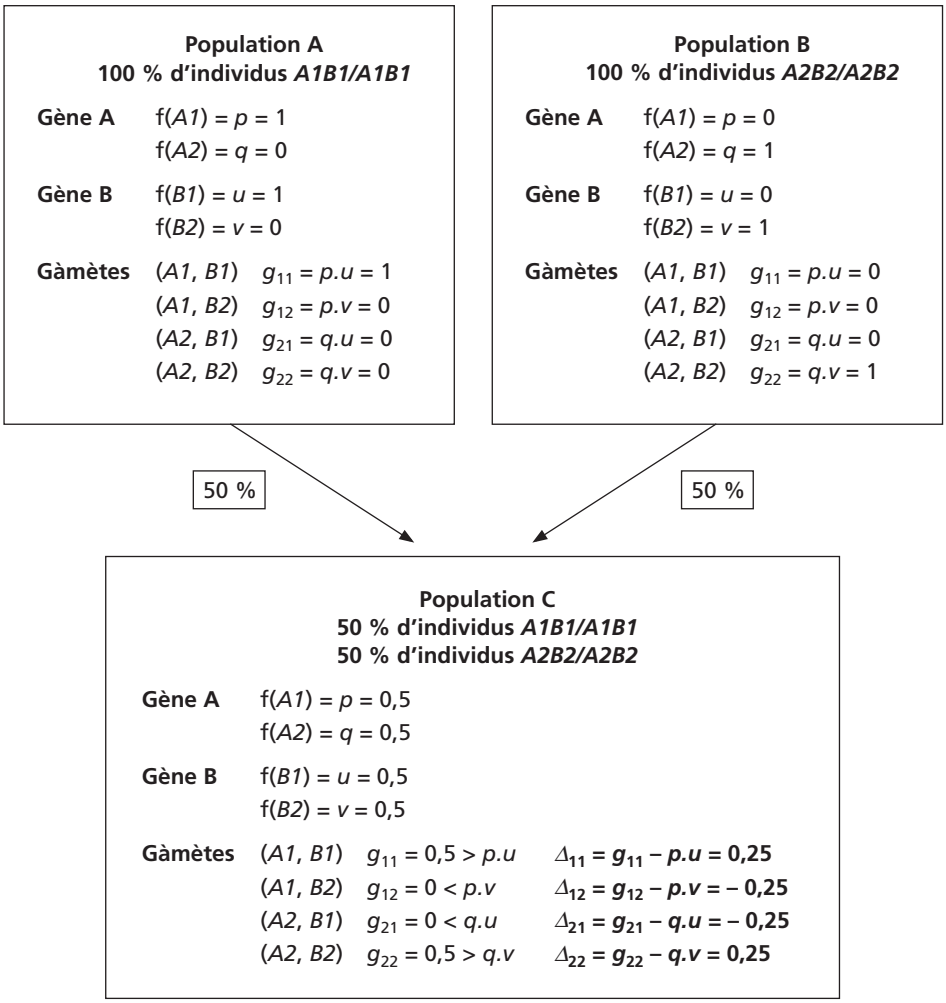


Figure 3.2

b) Genèse d'un déséquilibre gamétique à la suite d'une mutation

Le gène humain *CFTR* (*Cystic Fibrosis Trans-membrane conductance Regulator*), sur le chromosome 7, code pour une chaîne peptidique de 1 480 acides aminés constituant un canal chlorure. Ce canal joue un rôle important dans la fluidification des sécrétions du pancréas et de l'épithélium pulmonaire et une pathologie assez grave, souvent létale, survient par absence de ce canal chez les individus dont les deux

exemplaires du gène sont affectés d'une mutation de perte de fonction. Il s'agit de la mucoviscidose, caractérisée par une dégénérescence fibreuse du pancréas et des poumons.

Plus de 1 000 mutations pathologiques différentes du gène *CFTR* ont été identifiées. Parmi ces mutations, la plus courante, retrouvée en France chez 70 % des porteurs sains, est une délétion d'un triplet de bases qui se traduit par la perte d'une phénylalanine, en position 508 de la chaîne peptidique. C'est pourquoi elle a été nommée ΔF508.

En amont du gène *CFTR*, à 50 kilobases environ, existent deux SNP générant un marqueur di-allélique de type RFLP (*Restriction Fragment Length Polymorphism*) pour les enzymes de restriction TaqI et PstI. Le nombre et la longueur des fragments de restriction, générés en cet endroit du génome, sera fonction de l'absence ou de la présence de ces sites facultatifs. La visualisation de ces fragments de restriction, par une étude *in vitro* de l'ADN, permet de définir le statut, à cet endroit, des chromosomes 7 portés par un individu.

Tableau 3.3

Sites de restriction taqI	Sites de restriction PstI	Haplotypes	Fréquence Générale de l'haplotype	Fréquence relative de l'haplotype sur un chromosome non muté	Fréquence relative de l'haplotype sur un chromosome porteur de ΔF508
–	–	A	32,6 %	33,8 %	1,3 %
–	+	B	22,4 %	19,4 %	91,4 %
+	–	C	33,6 %	34,8 %	0,5 %
+	+	D	11,4 %	12,1 %	6,8 %

Selon qu'à cet endroit du chromosome 7, les sites sont présents (+) ou absents (–) on peut définir quatre associations en cis appelées « haplotypes », ainsi qu'ils figurent dans la première colonne du tableau 3.3. Celui-ci présente aussi la fréquence de chacun de ces haplotypes sur les chromosomes non mutés de la population française et sur les chromosomes porteurs de la mutation ΔF508.

Un chromosome 7 porteur de la mutation ΔF508 est ainsi porteur dans 91,4 % des cas de l'haplotype B, en amont du gène *CFTR*. Par contre un chromosome 7 non porteur d'une mutation du gène *CFTR*, ne porte l'haplotype B que dans 19,4 % des cas.

Il existe donc une association (non fonctionnelle mais seulement physique ou cartographique) entre l'haplotype B et la mutation ΔF508. Cette association vient du fait que la mutation ΔF508, qui est très ancienne (on évalue son « âge » à près de 40 000 ans, ce qui explique sa dispersion dans toutes les populations européennes), est survenue sur un chromosome 7 de type B, mais que les deux RFLP sont si près du gène *CFTR* que les recombinaisons par crossing over ont été trop rares, même en 40 000 ans, pour réassocier cette mutation ΔF508 avec les autres haplotypes et faire disparaître l'association entre B et ΔF508. Cette association constitue un déséqui-

libre gamétique, aussi appelé dans ce cas où il se maintient en raison de la liaison génétique, « déséquilibre de liaison » (traduction du linkage disequilibrium de la littérature anglophone).

Les gamètes porteurs de ΔF508 ont une fréquence de 1,4 %, puisqu'ils représentent 70 % de gamètes porteurs d'un exemplaire muté du gène *CFTR*, dont la fréquence est égale à 2 % dans la population (voir tableau 2.4).

La fréquence, dans ce sous ensemble des gamètes (des chromosomes 7) porteurs de la mutation ΔF508, de ceux qui sont en même temps porteurs de l'haplotype B est égale à 91,4 % (tableau 3.3), ce qui signifie que les gamètes porteurs à la fois de ΔF508 et de l'haplotype B ont une fréquence de $1{,}4 \times 91{,}4 =$ **1,27 %.**

Si il y avait équilibre gamétique, la fréquence de l'haplotype B parmi les chromosomes porteurs de la mutation ΔF508 serait égale à celle observée parmi les chromosomes 7, soit 22,4 % et les gamètes porteurs à la fois de ΔF508 et de l'haplotype B auraient une fréquence de $1{,}4 \times 22{,}4 =$ **0,31 %.** Le déséquilibre gamétique est donc égal à la différence, soit 0,96 %, la mutation ΔF508 est quatre fois plus souvent associée à l'haplotype B que s'il y avait équilibre gamétique.

3.5.4 Évolution d'un déséquilibre gamétique et définition du déséquilibre de liaison

a) Évolution d'un déséquilibre gamétique

Comme nous l'avons entrevu dans les deux exemples précédents, un déséquilibre gamétique tendra à diminuer sous l'effet de la recombinaison génétique. Cette évolution sera d'autant plus lente que les gènes étudiés *A* et *B* sont liés et très proches. Il est possible d'estimer le temps nécessaire pour atteindre l'équilibre en fonction du taux de recombinaison ***r*** entre les locus des deux gènes *A* et *B*.

Considérons $\boldsymbol{g_{11,i-1}}$, la fréquence du gamète (*A1*,*B1*), à la génération $i - 1$, et exprimons la valeur de cette fréquence à la génération suivante.

À la génération (*i*), les gamètes (*A1B1*) seront constitués de deux fractions :

– d'une part les gamètes (*A1B1*) de la génération précédente qui n'ont pas recombiné, évènement de probabilité $(1 - r)$, où r est le taux de recombinaison entre les locus des deux gènes ;
– d'autre part, les nouveaux gamètes (*A1*,*B1*) formés à la suite d'un évènement de recombinaison (de probabilité *r*) associant l'allèle *A1* (de fréquence *p*) et l'allèle *B1* (de fréquence *u*), ce qui s'écrit :

$$\boldsymbol{g_{11,i} = (1 - r)\, g_{11,i-1} + r.p.u}$$

Si on retranche *p.u* aux deux membres de cette équation de récurrence, on fait apparaître l'équation de récurrence du déséquilibre gamétique, soit :

$$\boldsymbol{\Delta_{11,i} = (1 - r)\, \Delta_{11,i-1}}$$

En remontant à la génération 0, où le déséquilibre est apparu, cette équation donne :

$$\boldsymbol{\Delta_{11,i} = (1 - r)^i\, \Delta_{11,0}}$$

Il apparaît donc clairement, ce que les exemples laissaient intuitivement entrevoir, que le déséquilibre tend vers zéro, au cours des générations, et que cette évolution dépend fortement de la liaison ou de l'absence de liaison génétique entre les locus des deux gènes.

Si les gènes sont indépendants, r est égal à 1/2 et $(1 - r)^i$ tend très vite vers zéro, en pratique en 8 à 10 générations. Mais si les gènes sont très liés, le déséquilibre peut perdurer pendant des centaines ou des milliers de générations, comme l'illustre la figure 3.3.

Un déséquilibre gamétique généré pour deux gènes *A* et *B*, génétiquement indépendants (physiquement liés ou non, cela est indifférent) aura donc disparu en 8 à 10 générations (un siècle chez l'homme).

Mais l'équilibre gamétique demandera 66 générations si les gènes sont liés avec un taux de 10 %, ce qui représente déjà 15 siècles chez l'homme (la distance au baptême de Clovis !), et signifie que les déséquilibres générés par les fusions entre Gaulois, romains, celtes et autres germains peuvent encore perdurer pour des gènes distants de moins de 10 centi-Morgans

Quant aux déséquilibres existant entre gènes liés à moins de 1 cM ($r < 0,01$), comme les gènes du complexe majeur d'histocompatibilité, ils perdurent depuis des millénaires (voir exercice 3.5) et peuvent aisément remonter à l'origine de l'homme moderne, voire avant.

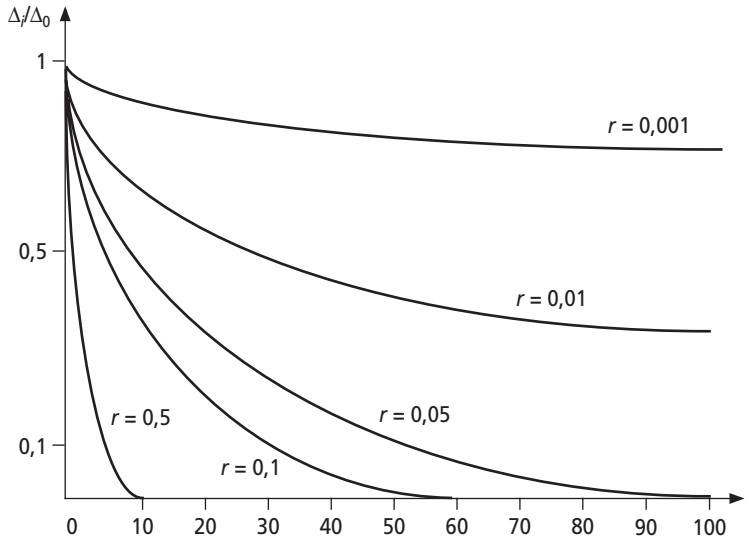


Figure 3.3 Évolution du déséquilibre gamétique en fonction du taux de recombinaison *r*, exprimé par la variation du rapport Δ_i/Δ_0, entre le déséquilibre à la génération *i* et le déséquilibre initial.

b) Déséquilibre gamétique et déséquilibre de liaison

Les Anglo-Saxons ont donné à un déséquilibre gamétique qui perdure en raison d'une liaison génétique intense entre les locus de deux gènes le nom de *linkage disequilibrium*, aussitôt traduit en français par déséquilibre de liaison.

La formule est très utilisée, et souvent à tort, parce qu'elle est trompeuse. En effet le déséquilibre de liaison suppose qu'il existe, entre les deux gènes considérés, à la fois une liaison génétique et un déséquilibre gamétique.

Or une liaison génétique peut exister entre deux gènes sans qu'il y ait déséquilibre gamétique, et, inversement, un déséquilibre gamétique peut exister pour deux gènes sans qu'il y ait liaison génétique.

Il suffit, pour s'en convaincre de reprendre l'exemple 1 de la fusion des populations A et B en une population C, en considérant que les gènes sont physiquement indépendants. Cette indépendance physique ne change rien, ni à l'existence, ni à l'intensité du déséquilibre gamétique ; elle entraîne simplement la disparition rapide de celui-ci en 8 à 10 générations.

Ainsi l'observation d'un fort déséquilibre gamétique n'est absolument pas l'indication d'une liaison génétique, comme certains utilisateurs mal informés ou mal formés ont pu l'imaginer.

Cette remarque est importante car de nombreuses études d'épidémiologie génétique mettent en évidence des associations entre des marqueurs génétiques (gènes ou sites polymorphes de type RFLP) et des maladies, et il serait erroné d'interpréter ce type de résultat comme l'indication d'une liaison génétique entre ce ou ces marqueurs et le ou les gènes impliqués dans les maladies étudiées.

3.5 UTILITÉ DU DÉSÉQUILIBRE DE LIAISON DANS LES ANALYSES GÉNÉTIQUES

La mesure du déséquilibre gamétique, quand il s'agit d'un déséquilibre de liaison, se révèle d'une grande utilité dans au moins plusieurs domaines de la recherche fondamentale à la recherche appliquée.

3.5.1 Analyse de la diversité génétique des populations et de leurs parentés

La mesure d'une association gamétique constitue une voie d'accès à l'étude de l'origine ou de l'ancienneté de certaines mutations. Les nombreuses études sur la mutation ΔF508 du gène *CFTR* (voir ci-dessus) en sont un exemple. Ce sont également les analyses moléculaires des séquences polymorphes adjacentes au gène β de l'hémoglobine, et associées à la mutation drépanocytaire, qui ont permis de démontrer que cette mutation drépanocytaire était survenue cinq fois, indépendamment, dans l'histoire de l'humanité, en cinq lieux différents (l'Inde, le Cameroun, le Sénégal, le Bénin et le pays Bantou). Il y a un nombre croissant d'études de la parenté et de l'éloignement de populations et/ou de mutations, fondées sur l'analyse informatique des données relatives aux polymorphismes moléculaires, de type microsatellite ou SNP, adjacents à tel ou tel gène, à partir du principe que l'hétérogénéité des marqueurs moléculaires autour d'une mutation est croissante avec l'ancienneté de celle-ci aussi bien qu'avec l'ancienneté de la séparation entre deux populations.

3.5.2 Épidémiologie génétique

Les analyses de marqueurs polymorphes de l'ADN, localisés avec précision dans le génome humain, en rapport avec l'étude de familles où ségrèguent les allèles du gène impliqué dans une maladie mono-factorielle (monogénique) ont permis la localisation, puis l'identification, d'un nombre croissant des gènes responsables des maladies mendéliennes (voir *homozygosity mapping*, page 146).

On arrive depuis peu, grâce à de nouvelles méthodes statistiques, dont la puissance doit être améliorée, à localiser des facteurs de susceptibilité génétique impliqués dans les maladies pluri-factorielles. Plusieurs facteurs génétiques de risque ont ainsi été localisés pour des maladies auto-immunes comme le diabète non insulino-dépendant, la sclérose en plaque ou la maladie cœliaque, et l'un des buts ultimes de ce type de recherches est d'identifier des facteurs de risques de maladies communes, c'est-à-dire répandues et dont la part génétique reste mystérieuse, telles les maladies neurologiques (épilepsie) ou psychiatriques (schizophrénie, maniaco-dépression, autisme).

3.5.3 Dépistage et diagnostic génétique

La mise en évidence d'une association aussi forte que celle qui existe entre l'haplotype B et la mutation ΔF508 du gène *CFTR* permet, pour la mucoviscidose (mais la méthode est la même pour de nombreuses autres maladies), d'appliquer le théorème de Bayes au calcul du risque qu'un individu soit porteur d'une mutation du gène *CFTR*, sachant qu'il est ou n'est pas porteur de tel ou tel haplotype RFLP. On conçoit facilement que ce risque est maximal si ses deux chromosomes 7 sont de type B alors qu'il est minimal s'ils sont tous deux de type C (voir le tableau 3.3).

Par ailleurs ces associations sont très utiles dans le diagnostic génétique anténatal. L'analyse moléculaire de l'ADN d'un premier enfant atteint permet d'identifier les chromosomes parentaux porteurs de la mutation responsable de la maladie (ou le chromosome parental incriminé, quand il s'agit d'une maladie dominante) ; par la suite, l'analyse moléculaire de l'ADN fœtal permet d'en déduire son statut génotypique et de savoir s'il sera atteint ou non, laissant alors aux parents la possibilité d'interrompre la grossesse.

RÉSUMÉ

Le modèle de l'équilibre de Hardy-Weinberg établi pour un gène autosomique diallélique peut être généralisé à l'étude de situations génétiques particulières.

– Dans le cas d'un gène multi-allélique, le modèle se généralise aisément. La relation panmictique demeure entre la somme des fréquences génotypiques et le carré de la somme des fréquences alléliques. Elle est établie en une génération de panmixie.

– Dans le cas de générations non séparées (chevauchantes), on démontre que l'équilibre de Hardy-Weinberg, pour un gène, est obtenu asymptotiquement, concrètement en quelques générations.

– Dans le cas d'un gène situé sur un hétérosome et gouvernant un caractère « lié au sexe », on démontre que les fréquences alléliques, si elles ne sont pas égales dans chacun des sexes, convergent vers une valeur limite en quelques générations.

Lorsque cet équilibre des fréquences alléliques est atteint, les fréquences génotypiques vérifient la relation de Hardy-Weinberg, dans le sexe femelle, qui est dizygote.

Mais les fréquences génotypiques restent égales aux fréquences alléliques, dans le sexe mâle qui est hémizygote.

L'étude de la composition génétique d'une population pour deux gènes considérés simultanément amène à dissocier les fréquences alléliques des fréquences gamétiques qui ne leur sont plus superposables.

Les fréquences gamétiques sont les produits des fréquences alléliques, pour deux gènes, si et seulement si il n'y a pas de déséquilibre gamétique, que les allèles de chacun des deux gènes sont associés aléatoirement dans les gamètes.

Dans le cas contraire on définit un déséquilibre gamétique. On montre que celui-ci diminuera de générations en générations, d'autant plus vite que la liaison génétique entre les locus des deux gènes est faible.

Si cette liaison génétique est très forte, le déséquilibre gamétique peut perdurer pendant très longtemps ; il est souvent appelé, dans ce cas, « déséquilibre de liaison ».

EXERCICES

Exercice 3.1

Reprendre l'exercice 1.2 et ajouter les questions suivantes :

Question 4 : quelle hypothèse doit-on préalablement tester avant de tester la conformité de la composition génétique au modèle de Hardy-Weinberg ? Faites-le.

Question 5 : tester cette conformité au modèle de Hardy-Weinberg.

Solution

Question 4 : les fréquences alléliques changent à chaque génération si elles ne sont pas égales entre les sexes et on ne peut tester la conformité au modèle de Hardy-Weinberg que si on est à l'équilibre avec des fréquences égales entre les sexes. Dans le cas présent, il faut construire un tableau de contingence présentant pour chaque classe d'allèles et pour chacun des sexes, les effectifs observés d'allèles et attendus (*en italiques*) sous l'hypothèse d'homogénéité (égalité des fréquences alléliques).

	A1	**A2**	**A3**	**Somme marginale des allèles**
Femelles	590 *586,6*	314 *312,6*	96 *100,6*	1 000
Mâles	290 *293,3*	155 *156,3*	55 *50,3*	500
Sommes marginales des sexes	880	469	151	1 500

La valeur du χ^2 est alors calculée ainsi :

$$\chi^2 = \frac{(590-586{,}6)^2}{586{,}6} + \frac{(314-312{,}6)^2}{312{,}6} + \frac{(96-100{,}6)^2}{100{,}6} + \frac{(290-293{,}3)^2}{293{,}3} + \frac{(155-156{,}3)^2}{156{,}3} + \frac{(55-50{,}3)^2}{50{,}3}$$

Le nombre de degrés de liberté (ddl) est ici égal à (nombre de lignes – 1) (nombre de colonnes –1) = 2

La valeur observée du $\chi 2$ est égale à 0,72, soit très inférieure à 5,99, la valeur qui est atteinte ou dépassée dans 5 % des cas pour un $\chi 2$ à 2ddl. Rejeter l'hypothèse d'homogénéité serait prendre un risque très supérieur à 5 %, on l'accepte donc.

On peut donc recalculer les fréquences alléliques sur l'ensemble de la population, sexes confondus, afin de mieux estimer celle-ci (l'échantillonnage en nombre d'allèles sera supérieur).

$f(A1) = 0{,}586$ $f(A2) = 0{,}313$ $f(A3) = 0{,}101$

Question 5 : on ne peut tester le modèle de Hardy-Weinberg que chez les femelles en testant la validité de la relation panmictique liant les fréquences alléliques aux fréquences génotypiques.

Effectifs	A1A1	A2A2	A3A3	A1A2	A2A3	A1A3
Observés	175	47	3	185	55	35
Théoriques H-W	*172,3*	*49*	*5,1*	*183,7*	*59,3*	*31,6*

$$\chi^2 = \frac{(175-172{,}3)^2}{172{,}3} + \frac{(47-49)^2}{49} + \frac{(3-5{,}1)^2}{5{,}1} + \frac{(185-183{,}7)^2}{183{,}7} + \frac{(55-59{,}3)^2}{59{,}3} + \frac{(35-31{,}6)^2}{31{,}6}$$

La valeur observée est égale à 1,67, largement inférieure à la valeur seuil de 7,8, pour un risque de 5 % et un ddl de 3 (6 classes, moins la somme à imposer, moins deux fréquences à estimer).

Rejeter l'hypothèse de Hardy-Weinberg serait prendre une décision avec un risque très supérieur à 5 %, on accepte donc cette hypothèse et la population peut être considérée comme à l'équilibre de H-W.

Exercice 3.2 Calcul des fréquences alléliques pour plusieurs gènes avec effet de dominance et d'épistasie

Les principaux types de pelages du chat sont gouvernés par trois gènes.

Le premier gène gouverne l'absence de coloration (phénotype dominant) ou la coloration (phénotype récessif), par le biais d'un couple d'allèles *W* et *w*+.

Le deuxième gène, lié au sexe (dont le déterminisme chromosomique est semblable à celui de l'homme), gouverne la dominante de couleur (jaune ou noire), par le biais de deux allèles *y* (pour yellow) et *y+* (l'allèle « sauvage »). Chez les femelles, on distingue trois phénotypes codominants, jaune, noir et mosaïque en « écailles de tortue » (pour les hétérozygotes *y/y+*).

Le troisième gène conditionne la tigration du pelage, par le biais de trois allèles :

– *Ta* (abyssin : couleur unie) dont l'effet domine celui de *Ts* et *Tl* ;

– *Ts* (sauvage à bandes étroites), dont l'effet domine celui de *Tl* ;

– *Tl* (sauvage à bandes larges).

Un échantillon aléatoire de 1 000 chats est prélevé dans une population et se classe en 60 chats blancs, 61 chats jaunes, dont 55 mâles et 6 femelles, 804 chats noirs, dont 445 mâles et 359 femelles, 75 chats « écailles de tortue ». La tigration des chats jaunes ou noirs se répartit de la façon suivante :

Tigration	Type abyssin	Sauvage à bandes étroites	Sauvages à bandes larges
Chats jaunes	13	22	26
Chats noirs	118	290	396

Question 1 : calculer les fréquences alléliques et tester la panmixie pour le premier gène.

Question 2 : calculer les fréquences alléliques et tester la panmixie pour le deuxième gène, après avoir testé préalablement une autre hypothèse, laquelle ?

Question 3 : calculer les fréquences alléliques et tester la panmixie pour le troisième gène.

Question 4 : y a-t-il équilibre gamétique pour les deuxième et troisième gènes dans la population ?

Solution

Question 1 : fréquences alléliques et test de la panmixie pour le gène de coloration

60 chats sont blancs et de génotype *W/W* ou *W/w+*, tandis que 940 ont le génotype homozygote *w+/w+* correspondant au phénotype récessif coloré. Sous l'hypothèse de Hardy-Weinberg, la fréquence de l'allèle w+ est égale à

$$f(w+) = q = \sqrt{940 / 1\,000} = 0{,}969 \text{ et } f(W) = p = 0{,}031$$

Comme on n'observe pas directement des couples (voir exercice 2.3), on ne pourrait tester la panmixie qu'à travers le test plus global de la conformité de la composition génétique au modèle de Hardy-Weinberg, or ne peut tester le tester car on a épuisé l'information disponible en estimant les fréquences alléliques (voir exercices chapitre 2).

Question 2 : fréquences alléliques pour le deuxième gène et test de la panmixie.

Les mâles sont hémizygotes et les femelles dizygotes présentent des phénotypes co-dominants. Dans chacun des sexes les fréquences alléliques peuvent être calculées directement.

	Allèle *y+*	Allèle *y*
Mâles	0,89	0,11
Femelles	0,901	0,099

On peut contrairement à la question précédente, tester le modèle de Hardy-Weinberg, puisqu'on dispose de trois classes de phénotypes co-dominants chez les femelles, mais il convient, au préalable de savoir si les fréquences alléliques sont égales entre les sexes afin de pouvoir appliquer la formule mathématique (r^2, $2rs$ et s^2) permettant le calcul des fréquences génotypiques à partir des fréquences alléliques r et s.

Cela revient à construire un test homogénéité (voir exercice 2.2), d'où le tableau de contingence suivant (les effectifs attendus sont italiques) :

	Effectifs d'allèles *y+*	Effectifs d'allèles *y*	Sommes marginales
Mâles	445 *448,55*	55 *51,45*	500
Femelles	793 *789,45*	87 *90,55*	880
Sommes marginales	1 238	142	1 380

La variable de χ^2 définie ici a $(n - 1)(m - 1) = 1$ degré de liberté. La valeur d'un χ^2 à 1 ddl, qui n'est dépassée que 5 fois sur 100 est égale 3,84. La valeur observée du χ^2 est égale 0,0159 ! Elle est très inférieure à 3,84, ce qui signifie qu'elle avait une probabilité très supérieure à 5 % d'être observée par le simple hasard d'échantillonnage. On peut donc accepter l'hypothèse nulle et conclure que les fréquences alléliques sont les mêmes dans les deux sexes.

Dans ce cas les meilleures estimations sont celles obtenues sur la somme des deux échantillons, soit respectivement $r = 0{,}897$ et $s = 0{,}103$ pour *y+* et *y*.

On peut alors tester l'équilibre de Hardy-Weinberg :

– soit en comparant les effectifs observés et attendus pour le seul sexe féminin (test avec un χ^2 à $3 - 2 = 1$ ddl) ;

– soit en comparant les effectifs observés et attendus pour les deux sexes (test à $5 - 3 = 2$ ddl, car il faut estimer une des deux fréquences alléliques et imposer les deux effectifs totaux d'hommes et de femmes).

Dans tous les cas l'hypothèse de l'équilibre de Hardy-Weinberg est acceptable.

Question 3 : étude du gène de tigration.

Compte tenu des effets de dominance on peut associer les phénotypes à un ou plusieurs génotypes, avec leurs fréquences respectives sous l'hypothèse de Hardy-Weinberg (les allèles *Ta*, *Ts* et *Tl* ayant respectivement les fréquences *u*, *v* et t) :

Tigration	type abyssin *Ta/Ta* ou *Ta/Ts* ou *Ta/Tl*	sauvage à bandes étroites *Ts/Ts* ou *Ts/Tl*	sauvages à bandes larges *Tl/Tl*
Fréquences si H-W	$u^2 + 2uv + 2ut$	$v^2 + 2vt$	t^2
Chats jaunes	13	22	26
Chats noirs	118	290	396
Somme	131	312	422

En utilisant une méthode semblable à celle utilisée pour le système ABO (voir chapitre 2 et exercice 2.7), on peut calculer les fréquences alléliques sous l'hypothèse de Hardy-Weinberg :

On voit que $f(Tl/Tl) = 422/865 = t^2$

d'où $t = \sqrt{422/865}$ soit $t = 0{,}698$

Par ailleurs, on voit que $(v^2 + 2vt + t^2) = (312 + 422)/865$

comme $(v^2 + 2vt + t^2) = (v + t)^2 = (1 - u)^2$

on en déduit que $u = 1 - \sqrt{(312 + 422)/865}$ soit $u = 0{,}079$

Par complément à un, on a $v = 0{,}223$

En estimant ainsi les fréquences alléliques sur l'ensemble des chats jaunes ou noirs, on a supposé qu'elles étaient les mêmes chez ces deux types. Il aurait fallu, au préalable s'en assurer par un test homogénéité (voir chapitre 2 et exercice 2.2), à partir du tableau de contingence suivant (les effectifs attendus sous l'hypothèse nulle homogénéité étant en italiques).

Tigration	type abyssin	sauvage à bandes étroites	sauvages à bandes larges	Sommes marginales
Chats jaunes	13 *9,24*	22 *22*	26 *29,7*	61
Chats noirs	118 *121,76*	290 *290*	396 *392,3*	804
Sommes marginales	131	312	422	865
Effectifs attendus sous Hardy-Weinberg	*131,27*	*312,29*	*421 ,44*	*865*

La variable de χ^2 définie ici a $(n - 1)(m - 1) = 1$ degré de liberté. La valeur d'un χ^2 à 1 ddl, qui n'est dépassée que 5 fois sur 100 est égale 3,84. La valeur observée du χ^2

est égale 2,14 ! Elle est inférieure à 3,84, ce qui signifie qu'elle avait une probabilité un peu supérieure à 5 % d'être observée. On peut donc accepter l'hypothèse nulle, car la rejeter serait prendre une décision assortie d'un risque supérieur à 5 %.

À partir de là, les fréquences alléliques étant égales entre les deux types de chats, il est possible de tester le modèle de Hardy-Weinberg sur l'ensemble (dernière ligne du tableau avec des effectifs attendus presque égaux aux effectifs observés) et conclure à l'équilibre de Hardy-Weinberg et donc à la panmixie.

Question 4 : équilibre ou déséquilibre gamétique pour les gènes de couleur et de tigration :

On sait que l'équilibre de Hardy-Weinberg peut être réalisé pour deux gènes, le cas ici, sans que, pour autant il y ait équilibre gamétique (voir chapitre 3).

Si il y a équilibre gamétique, la fréquence du gamète (y, Ta) doit être égale à $s.u$ et la fréquence du double homozygote égale à $(s.u)^2$, chez les femelles, car les mâles sont hémizygotes pour le couple d'allèles *y+* et *y.* Dans ces conditions on attend, chez les femelles, les fréquences suivantes pour les différents phénotypes, en fonction de leurs fréquences génotypiques, sous l'hypothèse de l'équilibre gamétique :

Tigration	type abyssin	sauvage à bandes étroites	sauvages à bandes larges
Chattes jaunes	$(su)^2 + 2susv + 2sust$	$(sv)^2 + 2svst$	$(st)^2$
Chattes noires	$(ru)^2 + 2rurv + 2rurt$	$(rv)^2 + 2rvrt$	$(rt)^2$

Ce tableau permet de calculer les effectifs attendus de chacun des doubles phénotypes couleur et tigration sous l'hypothèse de l'équilibre gamétique (en italiques dans le tableau ci-dessous) et de comparer leurs écarts aux effectifs observés, par un test de conformité. Le tableau ci-dessous est construit :

NB : dans ce tableau qui ressemble à un tableau de contingence pour un test homogénéité (voir plus haut) les effectifs théoriques ne sont pas ceux qui seraient calculés pour un test homogénéité, parce qu'ils ne correspondent pas à la même hypothèse nulle.

Dans le test homogénéité, on supposerait que les deux types de chattes, jaunes ou noires présentent la même fréquence de type abyssin ; on estimerait donc celle-ci comme égale à la somme des abyssins (54) rapportée au total jaune + noir (365), puis on calculerait l'effectif attendu sur un échantillon de 6 chattes jaunes (d'où 0,888 et non 0,708). Les sommes marginales seraient ici très utiles.

Ci-dessous l'effectif théorique est calculé par le produit de la fréquence attendue sous l'hypothèse d'équilibre gamétique, avec un total de 440 chats femelles, dont 75 présentent un phénotype « écaille de tortue, et 365 une coloration non écaille de tortue. Les sommes marginales sont ici inutiles.

La variable de χ^2 définie ici a un nombre de degrés de liberté difficile à calculer car les effectifs observés, s'ils ont participé, avec ceux des mâles, à l'estimation des trois fréquences alléliques nécessaires au calcul des effectifs théoriques, n'ont pas

été utilisés dans la réalisation du test ; les fréquences alléliques sont préalablement connues à la réalisation du test sur le seul sous échantillon des femelles réparties en doubles phénotypes.

Tigration	type abyssin	sauvage à bandes étroites	sauvages à bandes larges
Chats jaunes (sexe femelle)	1 0,708	2 1,68	3 2,27
Chats noirs (sexe femelle)	53 53,7	130 127,45	176 172,16

On peut se mettre dans les conditions les plus sévères pour le jugement des écarts et considérer qu'il y a 2 ddl (6 classes –1 pour l'effectif total, –3, pour l'estimation des fréquences). Dans ce cas, la valeur d'un χ^2 à 1 ddl, qui n'est dépassée que 5 fois sur 100 est égale 5,99.

De toute façon, la valeur observée du χ^2 est égale 0,56 ! Donc, quelque soit le nombre de ddl, l'hypothèse d'équilibre gamétique est vérifiée.

L'importance de la variance échantillonnage pour les classes d'évènements rares (effectifs théoriques ou attendus inférieurs à 5) ne permet pas de faire un test de χ^2 sans regrouper certaines classes afin de dépasser ce seuil de 5, pour chaque effectif théorique.

Cependant le seul biais que peut introduire la prise en compte de classes rares est d'induire un écart excessif conduisant à un rejet non justifié par des écarts réellement significatifs.

Au contraire, dans notre cas, où l'hypothèse nulle est acceptable, la prise en compte de classes rares ne nuit pas à la conclusion.

Exercice 3.3

Voir question 3.b) de l'exercice 2.8

Exercice 3.4 Évolution du déséquilibre gamétique

Le groupe Rhésus est gouverné, pour simplifier, par un couple d'allèles *D* et *d* (les homozygotes d/d étant de phénotype récessif rhésus négatif) d'un gène localisé sur le bras court du chromosome 1.

À 5 cM environ, vers le centromère, se trouve le locus d'un gène gouvernant la synthèse de 6Phospho-Gluconate déshydrogénase (6PGD) dont on connaît deux formes de mobilité électrophorétique différente, qui seront nommées *6l* (pour lente) et *6r* (pour rapide).

Sur le bras long du même chromosome, mais à une distance largement supérieure à 50 cM, réside le gène gouvernant le groupe sanguin Duffy, pour lequel existe, comme pour ABO, trois allèles *Fya*, *Fyb* et *Fyo* (on supposera, dans le problème que ce dernier allèle est suffisamment rare, dans les populations étudiées, pour être négligé).

Deux populations d'origine différente se mêlent, dans des proportions égales, en une population unique. La première était constituée d'individus rhésus négatif et de phénotype [6l] et [Fya] ; la deuxième était constituée d'individus rhésus positif (homozygotes) et de phénotype [6r] et [Fyb].

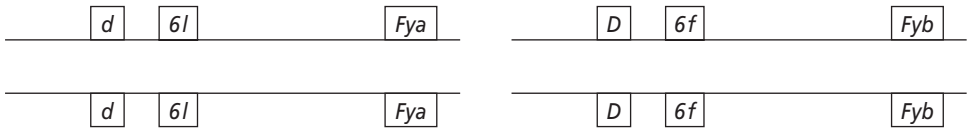
Question 1 : quelle sera la composition génétique de la nouvelle population à la génération de fusion, notée g_0 ?

Question 2 : Y a-t-il déséquilibre gamétique ?

Question 3 : quelle sera la composition de la population après 10 générations de panmixie ?

Solution

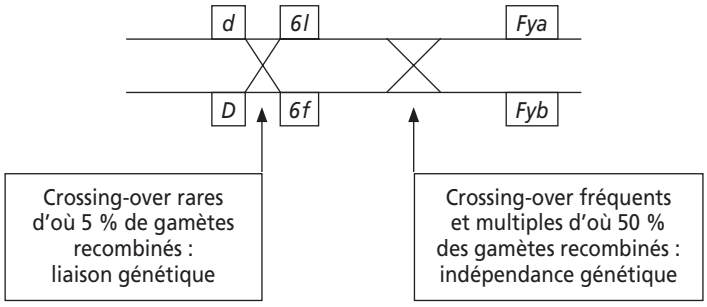
Question 1 : la composition génétique de la nouvelle population sera constituée de deux génotypes de même fréquence.



Question 2 : il y a déséquilibre gamétique car les fréquences allèliques sont toutes égales à 1/2, mais seuls quatre gamètes sont présents sur huit possibles.

Les valeurs des déséquilibres sont égales à +/– 0,25 selon les cas, quand on considère les gènes deux à deux (voir page 99), et sont égales à + 0,375/– 0,125 quand on les considère tous les trois simultanément (tableau ci-dessous).

Question 3 : après 10 générations de panmixie, les huit gamètes seront présents du fait des recombinaisons par crossing-over qui seront survenues, pendant les 10 générations précédentes, entre les locus des gènes chez les individus doubles hétérozygotes pour chaque paire de gènes. Dès la première génération, chez les individus triple hétérozygotes, on a pu obtenir des recombinaisons entre les locus, mais avec des probabilités différentes :



L'évolution du déséquilibre gamétique est donnée par l'équation $\Delta_i = \mathbf{(1 - r)^i}\ \Delta_0$ (voir figure 3.3).

Du fait de la liaison génétique relativement forte entre les locus rhésus et 6PGD, la valeur du Δ (calculée par l'équation précédente) sera encore égale à +/– 0,15 pour les divers types de gamètes possibles entre ces deux gènes, ce qui donnera les fréquences gamétiques suivantes :

$f(D, 6f) = 0{,}25 + 0{,}15 = 0{,}40$

$f(D, 6l) = 0{,}25 - 0{,}15 = 0{,}10$

$f(d, 6f) = 0{,}25 - 0{,}15 = 0{,}10$

$f(d, 6l) = 0{,}25 + 0{,}15 = 0{,}40$

À la génération $i = 10$, la valeur de Δ sera pratiquement nulle pour ce qui concerne les locus rhésus et Duffy, ou 6PGD et Duffy. De ce fait, les quatre gamètes précédents seront indifféremment *Fya* ou *Fyb* puisque les deux allèles *Fya* et *Fyb* sont équifréquents dans la population. Il est alors possible d'en déduire les fréquences de l'ensemble des gamètes après 10 générations et d'estimer le déséquilibre gamétique global pour les trois gènes concernés.

Gamètes	Fréquence initiale	Fréquence théorique	Déséquilibre initial	Fréquence après 10 générations	Déséquilibre après 10 générations
D, 6f, Fya	0	1/8	–0,125	0,20	0,075
D, 6f, Fyb	0,5	1/8	+0,375	0,20	0,075
D, 6l, Fya	0	1/8	–0,125	0,05	–0,075
D, 6l, Fyb	0	1/8	–0,125	0,05	–0,075
d, 6f, Fya	0	1/8	–0,125	0,05	–0,075
d, 6f, Fyb	0	1/8	–0,125	0,05	–0,075
d, 6l, Fya	0,5	1/8	+0,375	0,20	0,075
d, 6l, Fyb	0	1/8	–0,125	0,20	0,075

On remarquera que les fréquences des gamètes pour les gènes rhésus et Duffy, ou 6PGD et Duffy sont bien égales à 0,25, ce qui correspond à l'absence de déséquilibre entre ces gènes. On remarquera surtout que le déséquilibre a été réduit pour tous les gamètes mais a changé de signe pour deux d'entre eux. En effet la redistribution rapide des allèles *Fya* et *Fyb* vis-à-vis des allèles des deux autres gènes s'accompagne d'une redistribution lente des allèles de ces deux autres gènes (rhésus et 6PGD) entre eux qui maintiennent un déséquilibre de +/– 0,15.

Exercice 3.5

Calculer le nombre *i* de générations pour annuler un déséquilibre gamétique, ou un déséquilibre de liaison, en fonction du taux *r* de recombinaison. On considère que le Δ est nul dès que $(1 - r)^i$ est inférieur à 1/1 024.

Quelle échelle de temps cela représente-t-il si on considère que la durée d'une génération est comprise entre 20 et 25 ans ?

Solution

On applique l'équation $\Delta_i = (1 - r)^i \Delta_0$
Dont on tire l'inégalité $\Delta_i/\Delta_0 = (1 - r)^i < 1/1\,024$
Ce qui donne :

Taux *r* de recombinaison	**0,5 (indépendance)**	**0,1 (10 centi-Morgans)**	**0,01 (1 centi-Morgan)**	**0,005 (0,5 centi-Morgan)**
Nombre *i* de générations	10 générations	66 générations	690 générations	1 380 générations
Échelle de temps	200 à 250 ans	1 320 à 1 650 ans	13 800 à 17 250 ans	27 600 à 34 500 ans
Périodes de l'histoire	Autour de la Révolution Française	Du dernier siècle de l'Empire Romain à l'émergence de l'Islam	Des cultures paléolithiques à l'émergence de l'agriculture et de l'élevage (néolithique)	Première expansion démographique de l'humanité* : transition du paléolithique ancien au paléolithique récent

* Voir figure 5.5 (page 184).

Chapitre **4**

Les écarts à la panmixie : consanguinité, autogamie, homogamie

4.1 INTRODUCTION

Non seulement la plupart des croisements réalisés dans les populations expérimentales ne sont pas panmictiques, mais il existe aussi, dans les populations naturelles, de nombreuses exceptions à la panmixie, ce qui amène aux questions suivantes.

– Ces écarts à la panmixie peuvent-ils modifier la composition génétique de ces populations ?
– Si oui, une nouvelle situation d'équilibre polymorphe peut-elle être générée ?
– Quel sera le temps nécessaire pour l'atteindre ?

On conviendra, dans tout ce chapitre, que seule parmi les conditions de l'équilibre de Hardy-Weinberg, celle de panmixie est invalidée ; les populations naturelles seront donc supposées d'effectif infini (ou assez grand pour pouvoir y appliquer la loi des grands nombres), sans mutations, ni migrations, ni sélection (ou du moins avec un effet négligeable sur quelques générations).

On peut distinguer deux types d'écarts à la panmixie, c'est-à-dire de « choix » du conjoint (le choix n'étant conscient que pour l'espèce humaine).

Le choix du conjoint fondé sur la relation de parenté existant entre eux. Ce type d'écart à la panmixie est bien connu chez l'homme, où certaines unions sont prohibées (tabou de l'inceste) et d'autres, selon les cultures, plus ou moins favorisées (cousins germains et autres apparentements). Mais il existe aussi des unions entre apparentés, involontaires, dans certaines espèces animales à sexes séparés, et bien évidemment dans les espèces végétales où l'autofécondation (autogamie) totale ou partielle est possible.

Le choix du conjoint fondé sur la similitude ou la dissemblance phénotypique ou génotypique. Si le choix est conditionné par la ressemblance phénotypique, on parle d'homogamie ; si le choix est conditionné par la dissemblance phénotypique, ou l'existence de différences génétiques (allèles d'incompatibilité chez certains végétaux), on parle d'hétérogamie. La similitude ou la dissemblance pour les caractères considérés, peuvent dépendre, pour les gènes qui les gouvernent, soit du phénotype, soit du génotype.

4.2 CHOIX DU CONJOINT EN FONCTION DE LA PARENTÉ ET CONSANGUINITÉ

4.2.1 Trois définitions et une propriété

La rédaction de ce chapitre peut paraître « anthropocentrée » car la consanguinité est un élément important des écarts à la panmixie chez l'homme, mais les définitions qui suivent ont une portée générale pour tous les croisements où intervient un critère de parenté et s'appliquent donc aussi bien aux populations végétales, animales qu'aux populations humaines.

Qu'elles soient volontaires ou non, les unions entre apparentés correspondent au schéma de la figure 4.1, et donnent lieu à trois définitions (introduites par le généticien français Gustave Malécot en 1948) :

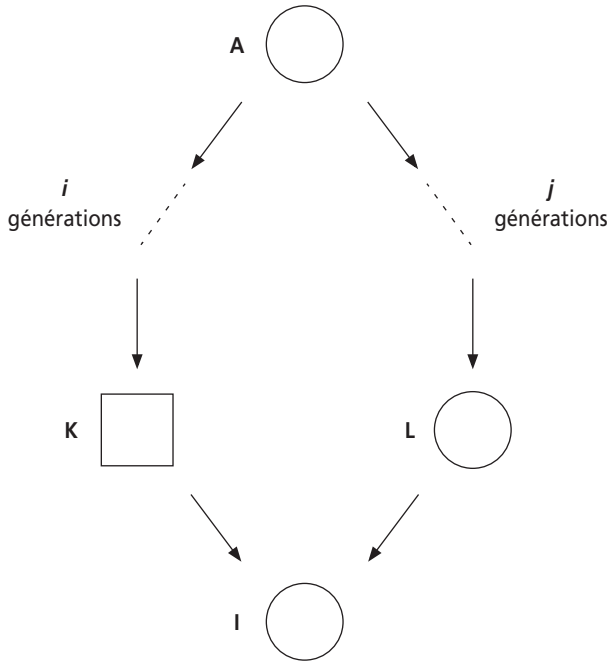


Figure 4.1

Les deux conjoints *K* et *L* sont apparentés puisqu'ils partagent un ancêtre commun *A*.

De ce fait, ils peuvent aussi partager des exemplaires de certains gènes, copies identiques d'un exemplaire ancestral (présent chez *A*).

Définition 1 : on dit que deux exemplaires d'un gène sont **identiques par ascendance** (ou descendance) s'ils sont copies d'un même exemplaire ancestral de ce gène.

> ***Remarque :*** deux exemplaires d'un gène, identiques par ascendance, sont porteurs de la même information allélique (on néglige l'éventualité de mutations dans ce chapitre), mais à l'inverse deux exemplaires alléliques identiques d'un gène ne sont pas forcément identiques par ascendance. L'identité par ascendance est une information généalogique dont on tire les conséquences génétiques (identité allélique) alors que l'identité allélique est une information génétique sans contenu informatif du point de vue généalogique.

Le partage par *K* et *L*, d'allèles identiques par ascendance est la conséquence génétique de leur parenté généalogique. On conçoit bien que leur similitude génétique sera d'autant plus grande que le nombre d'ancêtres est lui-même élevé et que les nombres de générations, *i* et *j*, séparant ces ancêtres des deux conjoints, sont faibles. Il convient donc de définir puis de quantifier cette proximité génétique résultant de la parenté.

La proximité génétique de deux apparentés *K* et *L* est définie et estimée par leur coefficient de parenté, noté ϕ_{KL}.

Définition 2 : le coefficient de parenté ϕ_{KL}, de deux individus *K* et *L*, est défini comme la probabilité qu'un exemplaire d'un gène, tiré au hasard chez *K*, soit identique par ascendance à un exemplaire, du même gène, tiré au hasard chez *L*.

Chacun des deux conjoints *K* et *L* peut alors transmettre au descendant *I* une copie, identique par ascendance, du gène ancestral. La situation d'homozygotie particulière réalisée chez ce descendant *I*, pour le gène considéré, dont les deux exemplaires sont identiques par ascendance, définit la consanguinité.

La consanguinité d'un individu est la conséquence génétique de l'apparentement de ses parents. Plus la parenté des conjoints sera élevée, plus la consanguinité de leurs enfants le sera elle-même. Il convient donc de définir la consanguinité, de la quantifier, et d'établir la relation existant entre la parenté des conjoints et la consanguinité de leurs enfants.

La consanguinité d'un individu *I* est définie et mesurée par son coefficient de consanguinité noté f_I.

Définition 3 : le coefficient de consanguinité f_I d'un individu *I* est défini comme la probabilité que les deux exemplaires d'un gène quelconque soient identiques par ascendance.

Propriété : relation entre parenté et consanguinité : il résulte des définitions adoptées que le coefficient de consanguinité f_I d'un individu I est égal au coefficient de parenté ϕ_{KL} de ses parents K et L.

En effet, procréer consiste à tirer au hasard un exemplaire de chacun des gènes chez chacun des parents. Aussi estimer l'identité par ascendance des deux exemplaires d'un gène chez un individu I (f_I) revient à estimer l'identité par ascendance de deux exemplaires de ce gène, l'un tiré au hasard chez le père K et l'autre chez la mère L, ce qui revient donc à estimer le coefficient de parenté des parents (ϕ_{KL}).

Cette relation entre parenté des conjoints et consanguinité des enfants sera très utile lors de l'établissement de formules de récurrence.

> **Remarque 1 :** les définitions et les formules de ces coefficients (voir plus loin) s'appliquent aux gènes autosomaux pour lesquels tous les individus sont dizygotes ; elles doivent être adaptées dans le cas particulier des gènes du chromosome X.
>
> **Remarque 2 :** la définition de la parenté et sa mesure s'appliquent à toute paire d'individus, qu'ils soient de sexe différent ou de même sexe (par exemple deux frères). Le fait qu'ils n'aient ou ne puissent avoir aucun descendant ne change rien à leur parenté. Seule la consanguinité résulte de la parenté d'un couple fécond.
>
> **Remarque 3 :** la mesure de la parenté, ou de la consanguinité, n'est pas absolue mais dépend de l'information généalogique disponible sur les individus étudiés. En réalité les individus d'une population sont tous apparentés, dès lors qu'on remonte suffisamment dans leur ascendance, mais nous verrons que la parenté résultant d'ancêtres communs très éloignés devient très vite négligeable, sauf dans les petites populations (ce qui n'est pas le cas dans ce chapitre).

4.2.2 Mesure de la parenté et de la consanguinité

a) Formule générale relative à un ancêtre

Il suffit d'appliquer les définitions qui viennent d'être adoptées au schéma général de la figure 4.1. Mesurer la consanguinité de I revient à mesurer la parenté de K et L, et consiste donc à mesurer la probabilité de l'évènement « deux exemplaires d'un gène donné tirés au hasard, l'un chez K, l'autre chez L, sont identiques par ascendance à un exemplaire ancestral présent chez A ».

Pour réaliser un tel évènement, il faut réaliser conjointement deux évènements indépendants :

– l'**identité de provenance** : à savoir que les deux exemplaires du gène, tirés l'un chez K, l'autre chez L, proviennent tous deux de l'ancêtre A, car, si ce n'était pas le cas, ils ne risqueraient pas d'être identiques par ascendance.
– l'**identité de copie chez A, sachant l'identité de provenance** : à savoir que les deux exemplaires tirés, l'un chez K, l'autre chez L, sachant qu'ils viennent de A (identité de provenance), soient bien la copie d'un même exemplaire ancestral de A.

Pour estimer le coefficient de ϕ_{KL}, il faut donc estimer la probabilité de ces deux évènements, puis en faire le produit, puisqu'ils sont indépendants.

La probabilité de l'évènement « identité de provenance » est égale à $(1/2)^{i+j}$. En effet la probabilité que l'exemplaire du gène tiré chez *K* vienne, à chaque génération de l'ascendant de la chaîne de parenté qui le relie à *A* est égale à 1/2 ; comme il y a *i* générations entre *K* et *A*, la probabilité que l'exemplaire tiré chez *K* vienne de *A* est égale à $(1/2)^i$. De la même façon la probabilité que l'exemplaire du gène tiré chez *L* vienne de *A* est égale à $(1/2)^j$. La probabilité conjointe de ces deux évènements indépendants est donc le produit $(1/2)^i \times (1/2)^j$ soit $(1/2)^{i+j}$.

Le calcul de la probabilité de l'évènement «identité de copie chez *A*, sachant l'identité de provenance » suppose d'établir toutes les situations possibles de transmission de *A* vers *K* et *L*. On est ici dans la situation où les deux gènes tirés, l'un chez *K*, l'autre chez *L*, viennent effectivement de A. Le tableau 4.1 détaille les quatre situations possibles, au regard des deux allèles homologues du gène présents chez A, symbolisés par ° et par *.

Deux fois sur quatre, **soit avec la probabilité 1/2**, les copies transmises par *A* vers *K* et *L*, sont des copies du même exemplaire ancestral, ce qui réalise la situation d'identité par ascendance.

Deux fois sur quatre, **soit avec la probabilité 1/2**, les copies transmises par *A* vers *K* et *L*, ne sont pas des copies du même exemplaire. Cependant les deux exemplaires du gène présents chez *A* peuvent être déjà identiques par ascendance, si *A* est lui-même consanguin et la probabilité qu'elles le soient est par définition f_A.

TABLEAU 4.1

L'ancêtre A transmet à K → et transmet à L	une copie de son allèle ° avec la probabilité 1/2	ou une copie de son allèle * avec la probabilité 1/2
une copie de son allèle ° avec la probabilité 1/2	les allèles tirés chez *K* et *L* sont ° et ° avec la probabilité 1/4	les allèles tirés chez *K* et *L* sont * et ° avec la probabilité 1/4
ou une copie de son allèle * avec la probabilité 1/2	les allèles tirés chez *K* et *L* sont ° et * avec la probabilité 1/4	les allèles tirés chez *K* et *L* sont * et * avec la probabilité 1/4

L'évènement « identité de copie chez *A*, sachant l'identité de provenance » peut donc être réalisé selon ces deux modalités exclusives dont les probabilités s'additionnent, soit : $1/2 + 1/2 \times f_A$

Le coefficient de parenté de *K* et *L* est le produit des probabilités respectives des deux évènements indépendants « identité de provenance » et « identité de copie chez *A*, sachant l'identité de provenance », soit :

$$\phi_{KL} = (1/2)^{i+j}\,(1/2 + 1/2 \times f_A)$$

Très souvent 1/2 est mis en facteur, ce qui fait apparaître une puissance $(1/2)^{i+j+1}$, mais il est pédagogiquement plus utile de laisser la formule inchangée, de manière à bien séparer les probabilités des deux évènements nécessaires à la réalisation de la parenté des conjoints ou de la consanguinité de l'enfant.

b) Coefficients de parenté et de consanguinité en cas d'ancêtres multiples

Si les deux individus *K* et *L* ont plusieurs ancêtres communs, les exemplaires du gène, tirés l'un chez *K* et l'autre chez *L*, peuvent être identiques par ascendance, soit à un exemplaire ancestral d'un premier ancêtre A_1, soit à un exemplaire ancestral du deuxième A_2, soit à un exemplaire ancestral d'un autre A_t, évènements mutuellement exclusifs.

De ce fait les coefficients de parenté relatifs à chacun des ancêtres s'additionnent pour donner la parenté totale du couple considéré, soit :

$$\phi_{KL} = \Sigma_t \, (1/2)^{it+jt} \, (1/2 + 1/2 \times f_{A_t})$$

où *t* est l'indice relatif à chacun des ancêtres A_t, *it* et *jt* étant les nombres de générations séparant respectivement *K* et *L* de l'ancêtre A_t.

Cette formule n'est cependant applicable qu'au cas où les différents ancêtres A_t ne sont pas eux-mêmes apparentés entre eux. En effet, si des ancêtres communs sont eux mêmes apparentés, *K* et *L* peuvent partager des exemplaires identiques par ascendance ne venant pas d'un même ancêtre mais de deux ancêtres différents apparentés, ce qui introduit des termes supplémentaires (voir calcul de la parenté dans les croisements systématiques frères sœurs ou la généalogie de la reine Hatshepsout).

c) Réseaux généalogiques complexes

Il arrive parfois que le réseau généalogique entre un couple d'individu *K* et *L* et un ancêtre commun *A* soit très complexe ; notamment quand l'ancêtre commun est éloigné, il peut exister plusieurs chemins généalogiques (aussi appelés chaînes de parenté) liant l'individu *K* et son ancêtre *A* ; de même entre *L* et *A*, ainsi que l'illustre la figure 4.2 (voir aussi généalogie de la reine Hatshepsout).

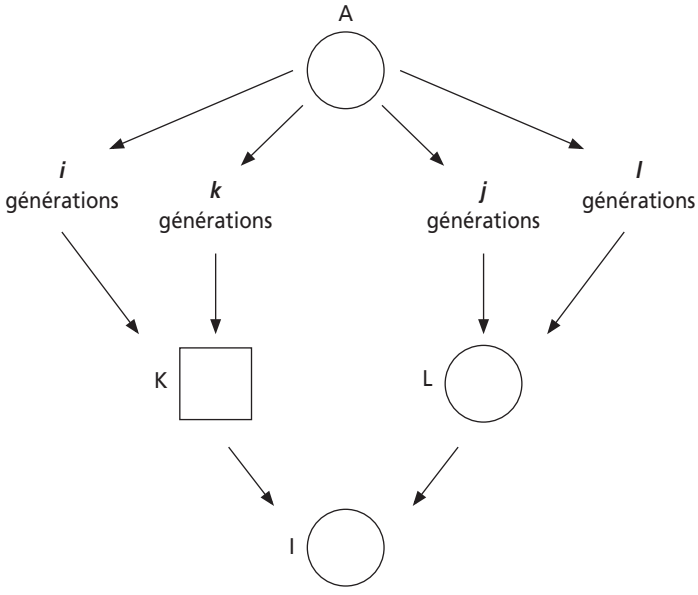


Figure 4.2

La parenté résultant de la transmission de gènes de *A* vers *K* et *L*, à travers les chaînes de parenté *i* et *j* est exclusive de celle résultant de la transmission à travers les chaînes *k* et *j*, ou *i* et *l*, ou *k* et *l*.

Aussi la parenté totale entre *K* et *L* est la somme des coefficients de parenté calculés pour tous les chemins généalogiques mutuellement exclusifs, soit :

$$\phi_{KL} = [(1/2)^{i+j} + (1/2)^{i+l} + (1/2)^{k+j} + (1/2)^{k+l}](1/2 + 1/2 \times f_A)$$

d) Coefficients des parentés les plus courantes

La figure 4.3 présente les parentés les plus courantes, avec, dans chacun des cas les ancêtres communs figurés en grisé. L'application de la formule dans ces différentes situations aboutit aux valeurs présentées dans le tableau 4.2.

Les ancêtres communs n'ayant pas d'ascendance connue sont considérés comme non consanguins, et la valeur de f_A est prise égale à zéro pour chacun d'eux.

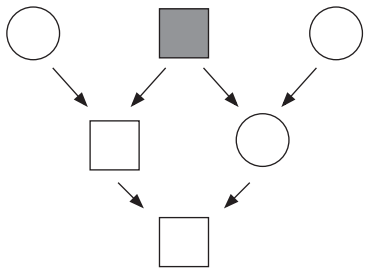
Comme dans tous ces cas les valeurs *i* et *j* sont les mêmes pour les différents ancêtres communs, il suffit de calculer le coefficient de parenté pour un ancêtre et de le multiplier par le nombre d'ancêtres pour avoir la parenté totale.

TABLEAU 4.2

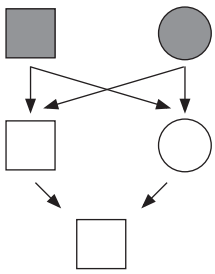
Relation de parenté	Demi-germains	Frère-sœur	Cousins germains	Oncle-nièce	Double cousins germains
Valeur de *i*	1	1	2	2	2
Valeur de *j*	1	1	2	1	2
Coefficient de parenté pour un ancêtre	1/8	1/8	1/32	1/16	1/32
Nombre d'ancêtres	1	2	2	2	4
coefficient Total de parenté	1/8	1/4	1/16	1/8	1/8

Remarque 1 : la parenté entre cousins germains est égale à 1/16 et la consanguinité d'un enfant issu d'une telle union est aussi de 1/16, ce qui signifie que pour 1 gène sur 16, soit 6,25 % du génome, les deux exemplaires du gène sont identiques par ascendance, réalisant ainsi une situation d'homozygotie.
On conçoit bien que si l'un des ancêtres est porteur sain d'une mutation pathogène (récessive), la consanguinité accroît le risque de voir apparaître cette pathologie récessive. Cet accroissement du risque sera calculé ultérieurement à l'échelle individuelle et à celle de la population.

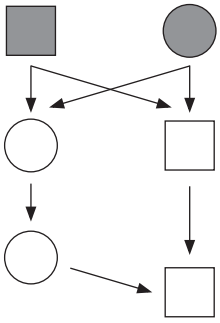
Remarque 2 : les unions demi-germains, doubles cousins germains et oncle-nièce ont la même conséquence génétique en termes de consanguinité mais n'ont évidemment pas le même statut chez l'homme. La première est toujours prohibée comme incestueuse, la seconde ne présente pas d'entraves sur le plan civil ni même sur le plan religieux (une dispense est toujours nécessaire mais accordée dans la religion catholique), la troisième suppose, en France, l'autorisation du président de la république.



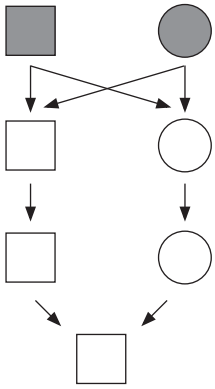
Demi-germains
(demi-frère/sœur)



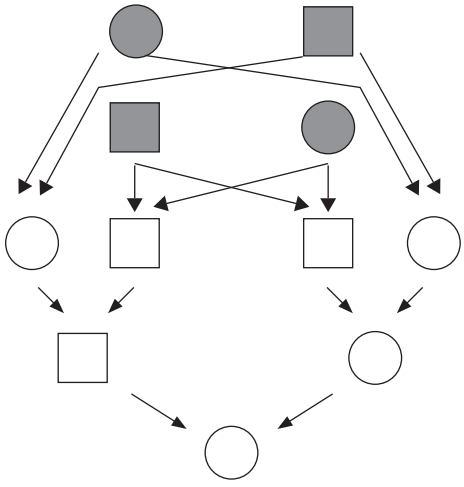
Germains
frère/sœur



Oncle-nièce
(tante-neveu)



Cousins germains



Doubles cousins germains

Figure 4.3

e) Le coefficient de parenté et la réalité biologique : définition des IBD

Le coefficient de parenté, dans sa simplicité algébrique, résume la totalité des situations biologiques possibles, nombreuses et complexes dès qu'on tente de les détailler. On peut le montrer pour le cas des frères-sœurs.

Pour ce faire, on va distinguer chacun des deux exemplaires paternels et maternels d'un gène, ainsi que cela est indiqué dans la figure 4.4.

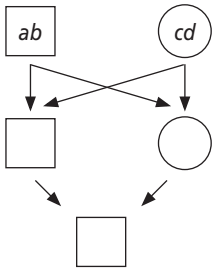


Figure 4.4

Les parents peuvent ainsi générer quatre types d'enfants, *a/c*, *a/d*, *b/c* et *b/d*, ce qui donne seize types possibles de couples de germains (tableau 4.3).

TABLEAU 4.3

Frère → / Sœur ↓	*a/c*	*a/d*	*b/c*	*b/d*
a/c	IBD = 2	IBD = 1	IBD = 1	IBD = 0
a/d	IBD = 1	IBD = 2	IBD = 0	IBD = 1
b/c	IBD = 1	IBD = 0	IBD = 2	IBD = 1
b/d	IBD = 0	IBD = 1	IBD = 1	IBD = 2

Le nombre d'exemplaires d'un gène pour lequel existe une relation d'identité par descendance entre les germains est un paramètre appelé (dans les articles scientifiques qui l'utilisent) IBD pour « *Identity By Descent* ». Il peut prendre les valeurs 2, 1 ou 0.

Les seize situations toutes différentes les unes des autres se fondent alors en trois classes distinctes selon que ces germains sont doublement identiques (IBD = 2 ; première diagonale : 4/16), totalement différents (IBD = 0 ; deuxième diagonale : 4/16) ou semi-identique (IBD = 1 ; le reste du tableau : 8/16).

Si on tire au hasard un exemplaire d'un gène chez le frère et un exemplaire du même gène chez la sœur, quelle est la probabilité qu'ils soient identiques par ascendance ?

Si les germains sont dans la situation IBD = 0, cette probabilité est nulle !

S'ils sont dans la situation IBD = 1, cette probabilité est égale à 1/4. En effet on tire chez chacun d'eux, avec une probabilité 1/2, l'exemplaire identique qu'ils partagent en commun.

Si les germains sont dans la situation IBD = 2, cette probabilité est égale à 1/2 (1/4 pour chacun des deux exemplaires identiques qu'ils partagent en commun).

Ces trois situations IBD ayant elles-mêmes des probabilités respectives de 1/4, 1/2 et 1/4, la probabilité de tirer au hasard, chez l'un et chez l'autre des germains, deux exemplaires d'un gène identiques par ascendance, c'est-à-dire leur coefficient de parenté, est bien égal à :

$$(1/4 \times 0) + (1/2 \times 1/4) + (1/4 \times 1/2) = 1/4 \text{ !}$$

Le coefficient de parenté 1/4 entre frères-sœurs résume bien la totalité des seize situations génétiques possibles détaillées dans le tableau 4.3.

> ***Remarque 1 :*** deux germains qui sont dans la situation IBD = 2 pour un gène peuvent être dans la situation IBD = 1 pour un autre et IBD = 0 pour un troisième. En fait deux germains sont dans la situation IBD = 2 pour un quart de leurs gènes, IBD = 1 pour la moitié et IBD = 0 pour le dernier quart.
>
> ***Remarque 2 :*** ce paramètre IBD est aujourd'hui très utilisé dans la cartographie des gènes impliqués dans des caractères polygéniques, comme les caractères quantitatifs (QTL : *Quantitative Trait Loci*) ou les maladies complexes (maladies auto-immunes, neurologiques, cardio-vasculaires, obésité, diabète non insulino-dépendant). La stratégie consiste à étudier des paires de germains présentant des valeurs semblables du caractère ou des phénotypes cliniques identiques (paires de germains atteints) pour un ensemble de marqueurs polymorphes de l'ADN, répartis sur l'ensemble du génome. Les paires de germains atteints sont sans doute identiques pour les gènes impliqués dans le phénotype étudié. On s'attend dès lors à ce que, pour tout marqueur localisé au voisinage d'un de ces gènes, la distribution des valeurs IBD (2, 1 et 0) de ce marqueur diverge de la distribution aléatoire 1/4-1/2-1/4 vers une distribution $(1/4 + x) - (1/2 + y) - (1/4 - x - y)$.

f) Un exemple de généalogie complexe : la généalogie de la reine-pharaon Hatshepsout

La reine Hatshepsout (XVIII^{e} dynastie du moyen empire, à Thèbes-Louqsor, 1504 av. JC) est issue de deux unions frères sœurs, et d'une union demi frère sœur, autorisées chez les pharaons, qui n'étaient pas des hommes mais des dieux, et.... bien utiles pour garder le pouvoir dans la famille.

Elle a exercé le pouvoir pharaonique et s'est fait construire un temple-tombeau tout à fait exceptionnel : Deir El Bahari, sur la rive des morts à Louqsor.

Afin de garder le pouvoir dans la famille, elle a épousé son demi-neveu et a donné naissance au pharaon Aménophis II.

Quelle était la valeur de la consanguinité de la reine Hatshepsout, celle de sa mère et celle du pharaon Aménophis II ?

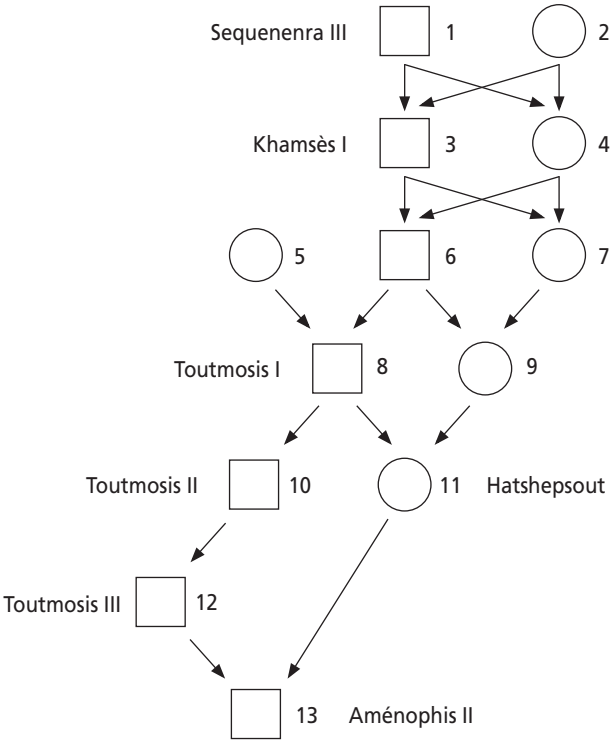


Figure 4.3

➤ Consanguinité de la mère de la reine Hatshepsout

Calculer f_9 revient à calculer $\phi_{6\text{-}7}$: parenté de deux germains issus d'une union entre germains.

L'application de la formule de récurrence établie plus loin pour les croisements systématiques entre germains donne la valeur de 3/8.

Mais on peut aussi la calculer en faisant la somme de toutes les probabilités relatives à tous les chemins généalogiques possibles, relatifs à tous les ancêtres communs de 6 et 7, avec pour chaque ancêtre commun la probabilité 1/2 que les deux gènes tirés chez 6 et 7 soient copies du même exemplaire (et donc identiques par ascendance).

Les ancêtres communs à 6 et à 7 sont 1, 2, 3 et 4. Aucun des quatre n'est *a priori* consanguin.

Les chemins généalogiques joignant 6 et 7 à chacun de leurs ancêtres communs et les probabilités d'identité par ascendance résultant de chacun de ces chemins, pour chacun de ces ancêtres sont donnés par le tableau 4.4.

TABLEAU 4.4

Ancêtre commun	Chemins généalogiques	Probabilité du chemin généalogique	Probabilité d'identité par ascendance par chemin généalogique
1	6-3-1-4-7 6-4-1-3-7	$(1/2)^4$ $(1/2)^4$	$(1/2)^5 = 1/32$ $(1/2)^5 = 1/32$
2	6-3-2-4-7 6-4-2-3-7	$(1/2)^4$ $(1/2)^4$	$(1/2)^5 = 1/32$ $(1/2)^5 = 1/32$
3	6-3-7	$(1/2)^2$	$(1/2)^3 = 1/8$
4	6-4-7	$(1/2)^2$	$(1/2)^3 = 1/8$
TOTAL			3/8

➤ Consanguinité de la reine Hatshepsout

Calculer f_{11} revient à calculer $\phi_{8\text{-}9}$.

Les ancêtres communs à 8 et à 9 sont 1, 2, 3, 4 et 6.

Aucun des quatre premiers n'est *a priori* consanguin, mais 6, issu d'un croisement frère-sœur à une consanguinité égale à 1/4.

Nous allons montrer deux modes de calcul de la consanguinité, donnant évidemment le même résultat :

– ou bien on détaille, comme ci-dessus, tous les chemins généalogiques possibles, et dans ce cas on ne tient jamais compte de la consanguinité des ancêtres communs intermédiaires, car elle est déjà prise en compte dans les chemins généalogiques qui remontent au-dessus de lui ;
– ou bien on tient compte de la consanguinité des ancêtres intermédiaires, mais dans ce cas il ne faut plus compter les chemins généalogiques qui remontent au-dessus de cet ancêtre intermédiaire.

a) calcul détaillé : ce calcul se fait par la somme de toutes les probabilités relatives à tous les chemins généalogiques possibles, relatifs à tous les ancêtres communs de 8 et 9, avec pour chaque ancêtre commun l'unique probabilité 1/2 que les deux gènes tirés chez 8 et 9 soient copies du même exemplaire (et donc identiques par ascendance). Les chemins généalogiques joignant 8 et 9 à chacun de leurs ancêtres et les probabilités d'identité par ascendance résultant de chacun de ces chemins, pour chacun de ces ancêtres sont donnés par le tableau 4.5.

b) calcul emboîté : ce calcul se fait par la seule prise en compte de la parenté et/ou de la consanguinité des seuls grands-parents, ce qui permet de distinguer quatre cas mutuellement exclusifs dans lesquels on prend en compte la parenté ou la consanguinité accumulées dans les générations en amont.

– *1er cas :* les deux gènes de 11 peuvent venir tous les deux de 6, avec la probabilité 1/4, et sont dans ce cas identiques par ascendance avec la probabilité $(1/2 + f_6/2)$. Comme 6 est issu d'une union entre germains, f_6 est égale à 1/4. Donc $(1/2 + f_6/2) = 5/8$. La probabilité d'identité par ascendance est donc, pour ce cas, égale à $1/4 \times 5/8 = 5/32$.

TABLEAU 4.5

ancêtre commun	chemins généalogiques mutuellement exclusifs	probabilité du chemin généalogique	probabilité d'identité par ascendance par chemin généalogique
1	8-6-3-1-4-7-9 8-6-4-1-3-7-9 8-6-3-1-4-6-9* 8-6-4-1-3-6-9*	$(1/2)^6$ $(1/2)^6$ $(1/2)^6$ $(1/2)^6$	$(1/2)^7 = 1/128$ $(1/2)^7 = 1/128$ $(1/2)^7 = 1/128$ $(1/2)^7 = 1/128$
2	8-6-3-2-4-7-9 8-6-4-2-3-7-9 8-6-3-2-4-6-9* 8-6-4-2-3-6-9*	$(1/2)^6$ $(1/2)^6$ $(1/2)^6$ $(1/2)^6$	$(1/2)^7 = 1/128$ $(1/2)^7 = 1/128$ $(1/2)^7 = 1/128$ $(1/2)^7 = 1/128$
3	8-6-3-7-9 8-6-3-6-9*	$(1/2)^4$ $(1/2)^4$	$(1/2)^5 = 1/64$ $(1/2)^5 = 1/64$
4	8-6-4-7-9 8-6-4-6-9*	$(1/2)^4$ $(1/2)^4$	$(1/2)^5 = 1/64$ $(1/2)^5 = 1/64$
6	8-6-9	$(1/2)^2$	$(1/2)^3 = 1/8$
TOTAL			1/4

* Tous ces chemins sont responsables de la consanguinité chez 6 et sont globalement pris en compte dans l'autre mode de calcul.

– *2e cas :* les deux gènes de 11 peuvent venir l'un de 5, l'autre de 6, avec la probabilité 1/4, et ne sont jamais, dans ce cas, identiques par ascendance.
– *3e cas :* les deux gènes de 11 peuvent venir l'un de 5, l'autre de 7, avec la probabilité 1/4, et ne sont jamais, dans ce cas, identiques par ascendance.
– *4e cas :* les deux gènes de 11 peuvent venir l'un de 6, l'autre de 7, avec la probabilité 1/4, et sont, dans ce cas, identiques par ascendance avec une probabilité égale au coefficient de parenté de 6 et 7, égal, comme on l'a vu, à 3/8. La probabilité d'identité par ascendance est donc, pour ce cas, égale à $1/4 \times 3/8 = 3/32$.

La somme des probabilités trouvées pour chacun des quatre cas est égale à la consanguinité de la reine Hatshepsout, sa consanguinité, soit $5/35 + 3/32 = 8/32$ ou 1/4.

Cet exemple montre bien que les formules de Malécot, appliquées à des ancêtres intermédiaires, peuvent dispenser d'une définition exhaustive des chemins généalogiques jusqu'aux ancêtres les plus éloignés, à condition de ne rien oublier. Mais la logique des programmes informatiques est fondée sur la définition de tous ces chemins généalogiques, car les ordinateurs sont bêtes mais rapides.

On remarquera au passage que la reine Hatshepsout, malgré une ascendance exceptionnellement « incestueuse » n'a pas plus de consanguinité qu'un simple enfant de frère-sœur. Sa mère avait une consanguinité bien supérieure et il a suffit d'une « pièce rapportée », 5, pour casser la progression de la consanguinité.

C'est un résultat qui sera retrouvé, à l'échelle de la population, lors de l'étude de l'effet combiné de la dérive et des migrations (voir chapitre 5).

➤ Consanguinité du pharaon Aménophis II

Calculer f_{13} revient à calculer $\phi_{11\text{-}12}$.

Il convient, comme dans le raisonnement ci-dessus de tenir compte du fait que les deux exemplaires d'un gène tirés chez Aménophis II :

– peuvent venir de 8, avec une probabilité de 1/8, et être identiques, dans ce cas, avec une probabilité de 1/2, car 8 n'est pas consanguin,
– ou venir, l'un de 8 et l'autre de 9, avec une probabilité de 1/8, et être, dans ce cas identiques avec une probabilité égale à la parenté entre 8 et 9, soit 1/4 (consanguinité de la reine Hatshepsout).

D'où un total de $1/8 \times 1/2 + 1/8 \times 1/4 = 3/32$.

4.2.3 Croisements consanguins systématiques

Le terme de « croisement consanguin » est trop utilisé pour le remettre en question mais il convient de remarquer que ce n'est pas le croisement qui est consanguin mais son produit, le croisement étant en fait entre apparentés.

Dans de nombreux cas, notamment pour l'autofécondation, les populations sont divisées en « lignées » et ne peuvent donc pas être considérées comme des populations mendéliennes de grande taille. La consanguinité relevant des croisements consanguins systématiques sera cependant présentée dans ce chapitre et sera revue sous une autre forme dans le chapitre sur la dérive génétique.

a) Autofécondation totale

C'est, dans la nature, un mode de reproduction possible chez les végétaux monoïques ou certains animaux hermaphrodites ; elle peut aussi résulter de l'action de l'expérimentateur.

Si une population est constituée d'organismes strictement autoféconds, les individus homozygotes *A1*/*A1* ou *A2*/*A2* constitueront des souches pures autogames (chaque individu se fournit ses gamètes) et les éventuels hétérozygotes *A1*/*A2* donneront 1/4 *A1*/*A1*, 1/2 *A1*/*A2* et 1/4 *A2*/*A2*.

Si la fréquence des hétérozygotes, à une génération quelconque $(n - 1)$ est égale à H_{n-1}, leur fréquence H_n, à la génération suivant, est divisée par 2, et ainsi de suite à chaque génération, comme Mendel l'avait d'ailleurs remarqué dans son étude sur le pois.

On déduit facilement (voir 3.3) de la récurrence à un seul terme $H_n = H_{n-1}/2$, que si H_0 est la fréquence initiale, on a, après n générations :

$$\boldsymbol{H_n = (1/2)^n H_0}$$

La fréquence des hétérozygotes tend donc très vite vers zéro (en 8 à 10 générations, voir 3.3).

La population n'est plus constituée que de lignées pures autogames dans des proportions qui peuvent être prédites.

Si les homozygotes *A1*/*A1* sont dans une proportion D_0 au départ, leur fréquence, à la génération suivante, est égale à :

$$D_1 = D_0 + 1/4\ H_0$$

(les homozygotes *A1*/*A1* de la génération précédente donnent des homozygotes *A1*/*A1*, et les hétérozygotes donnent un quart de descendants homozygotes *A1*/*A1*)

puis on aura $D_2 = D_1 + 1/4\ H_1$

etc.,

et $D_n = D_{n-1} + 1/4\ H_{n-1}$

La somme de ces égalités, suivie des simplifications algébriques, conduit à :

$$D_n = D_0 + 1/4\ [H_0 + H_1 + ... + H_{n-1}]$$

soit $D_n = D_0 + 1/4\ [H_0 + 1/2\ H_0 + ... + (1/2)^{n-1}\ H_0]$

ou $D_n = D_0 + H_0/4\ [1 + 1/2 + ... + (1/2)^{n-1}]$

où $[1 + 1/2 + ... + (1/2)^{n-1}]$ représente la somme des n premiers termes d'une progression géométrique[1] dont l'élément initial est 1 et la raison 1/2. Cette somme S_n est égale à :

$$S_n = 2\ (1 - (1/2)^n)$$

On peut alors écrire que la fréquence des homozygotes *A1*/*A1* évolue selon la relation :

$$D_n = D_0 + H_0/4 \times 2\ (1 - (1/2)^n)$$

Qui peut aussi s'écrire

$$\boldsymbol{D_n = D_0 + H_0/2 \times (2^n - 1)/2^n}$$

On démontre, la même façon, que la fréquence R des homozygotes évolue selon la relation :

$$\boldsymbol{R_n = R_0 + H_0/2 \times (2^n - 1)/2^n}$$

On en déduit alors les limites de D_n et R_n quand n tend vers l'infini :

$$D = D_0 + H_0/2 \qquad \text{et} \qquad R = R_0 + H_0/2$$

Elles sont égales aux fréquences alléliques initiales qui étaient :

$$f(A1) = p = D_0 + H_0/2$$

et $$f(A2) = q = R_0 + H_0/2$$

On peut donc conclure, de cet ensemble de calculs, que l'autofécondation totale aboutit aux résultats suivants :

1. la totalité des hétérozygotes disparaissent de la population, il ne subsiste que des lignées pures homozygotes ;

2. le nombre de lignées pures fixées à l'état homozygote est proportionnel aux fréquences alléliques initiales ;

1. On calcule la somme des termes d'une progression géométrique S_n en retranchant, à celle-ci, cette même somme multipliée par la raison géométrique, soit $S_n \times 1/2$. Tous les termes s'annulent sauf le premier et le dernier, ce qui donne, dans notre cas : $S_n - S_n \times 1/2 = 1 - (1/2)^n$
D'où on tire $S_n = 2\ (1 - (1/2)^n)$

3. ces fréquences alléliques initiales demeurent inchangées, seule évolue la composition génétique en termes de fréquences génotypiques ;

4. ces conclusions démontrées pour un gène sont valables pour l'ensemble du génome, l'autofécondation aboutit donc à des lignées pures pour tous les gènes. Il faut cependant souligner que la plupart si ce n'est toutes les lignées pures seront quand même génétiquement différentes entre elles, selon qu'elles ont fixé tel ou tel allèle de tel ou tel gène ; si, par exemple, on considère trois gènes di-alléliques, il y a huit lignées pures possibles, et 2^n si on considère n gènes di-alléliques. En pratique, comme les effectifs, même grands, ne sont jamais infinis, de nombreuses lignées potentiellement possibles au départ ne seront pas réalisées à l'arrivée, mais l'ensemble de ces lignées pures seront génétiquement différentes.

b) Autofécondation ou autogamie partielle

Dans les populations végétales naturelles, l'autofécondation n'est jamais totale et un pourcentage d'allogamie (union de gamètes provenant d'individus différents) maintient une fraction d'hétérozygotie. Ce pourcentage d'allogamie varie de 3 à 4 % chez le riz ou la tomate à 9 à 10 % chez le blé et le haricot, et 30 à 35 % chez le tabac et le colza.

Dans ce cas, il convient de reprendre les relations de récurrence du paragraphe précédent en tenant compte du fait qu'une fraction λ des fécondations sont autogames et qu'une fraction $(1 - \lambda)$ sont allogames, ce qui donne :

– pour les hétérozygotes : $H_n = \lambda(H_{n-1}/2) + (1 - \lambda)\, 2p_{n-1}\, q_{n-1}$
En effet la fraction λ des croisements autogames voit la fréquence des hétérozygotes divisée par deux, comme on l'a vu, alors que la fraction des croisements allogames, sous l'hypothèse panmictique, voit la formation d'hétérozygotes dans la proportion $2p_{n-1}\, q_{n-1}$, le produit des fréquences alléliques de *A1* et *A2*, à la génération $(n - 1)$;

– pour les homozygotes *A1*/*A1* : $D_n = \lambda[\, D_{n-1} + H_{n-1}/4] + (1 - \lambda)\, p_{n-1}^2$
et *A2*/*A2* : $R_n = \lambda[\, R_{n-1} + H_{n-1}/4] + (1 - \lambda)\, q_{n-1}^2$
En effet, la fraction des croisements autogames voit, comme on l'a vu, la fréquence de chaque homozygote augmentée du quart de celle des hétérozygotes, alors que la fraction des croisements allogames, sous l'hypothèse panmictique, voit la formation d'homozygotes dans la proportion p_{n-1}^2, pour *A1*/*A1*, et q_{n-1}^2, pour *A2*/*A2*.

> ***Remarque 1 :*** On peut tout de suite noter que les fréquences alléliques demeurent inchangées d'une génération à l'autre, en effet, à la génération n, on peut écrire que :
>
> $$f(A1) = p_n = D_n + H_n/2 = \lambda[\, D_{n-1} + H_{n-1}/4] + (1 - \lambda)\, p_{n-1}^{\,2} + \lambda\, H_{n-1}/4 + (1 - \lambda)\, p_{n-1}\, q_{n-1}$$
>
> soit $p_n = \lambda[\, D_{n-1} + H_{n-1}/2] + (1 - \lambda)\, [p_{n-1}^{\,2} + p_{n-1}\, q_{n-1}]$
>
> et $p_n = \lambda[\, p_{n-1}] + (1 - \lambda)\, [p_{n-1}]$
>
> et donc $p_n = p_{n-1}$

L'autofécondation, qu'elle soit totale ou partielle, ne modifie nullement les fréquences alléliques, et donc la diversité génétique des populations à ce niveau, mais modifie la composition génétique en termes de fréquences génotypiques. Dans un régime d'autofécondation totale les hétérozygotes disparaissaient, dans un régime d'autofécondation partielle, ils seront seulement diminués jusqu'à un niveau d'équilibre qui dépend du taux λ d'autogamie.

Remarque 2 : le niveau d'équilibre dépendant du taux λ d'autogamie est la valeur H, solution limite de la relation de récurrence

$$H_n = \lambda(H_{n-1}/2) + (1 - \lambda)\, 2p_{n-1}\, q_{n-1}$$

ou, puisque p et q sont invariants : $H_n = \lambda(H_{n-1}/2) + (1 - \lambda)\, 2pq$

soit, H tel que : $H = \lambda(1/2)\, H + (1 - \lambda)\, 2pq$

d'où on peut tirer : $\boldsymbol{H = 4pq(1 - \lambda)/(2 - \lambda)}$

On peut vérifier, au passage, que pour la valeur $\lambda = 1$, correspondant à l'autofécondation totale, la fréquence des hétérozygotes tend bien vers zéro.

On peut utiliser cette relation en sens inverse et calculer le taux d'autogamie d'une espèce, en supposant que la fréquence observée des hétérozygotes, dans la nature, est à l'équilibre, ce qui donne :

$$\boldsymbol{\lambda = (2H - 4pq)/(H - 4pq)}$$

Remarque 3 : La valeur d'équilibre de la fréquence des hétérozygotes est inférieure à la valeur attendue $2pq$ dans une population panmictique.

En effet, $H = 4pq(1 - \lambda)/(2 - \lambda)$

s'écrit $H = 2pq(2 - 2\lambda)/(2 - \lambda) = 2pq\,[(2 - \lambda)/(2 - \lambda) - \lambda/(2 - \lambda)\,]$

Soit $\boldsymbol{H = 2pq - 2pq\,\lambda/(2 - \lambda)}$

Dans une population panmictique, le paramètre λ est nul, et on retrouve bien une valeur $H = 2pq$. Si, au contraire, λ n'est pas nul, la valeur de H est inférieure à $2pq$, et l'abaissement de H est conditionné par la valeur du paramètre λ d'écart à la panmixie.

Quand il y a autofécondation, λ est égal à 1 et on retrouve $H = 0$, absence d'hétérozygotes.

Remarque 4 : Il est utile de savoir « résoudre » une récurrence, c'est-à-dire d'obtenir le terme X_n en fonction d'une équation algébrique où n n'est plus un indice mais une variable, ce qui permet par exemple de calculer directement le millième terme de la récurrence $X_{1\,000}$, sans avoir à calculer les 999 précédents.

Les récurrences vues jusqu'ici (3.3 ou 4.2.3.a) étaient simples car le terme X_n ne dépendait que d'un seul terme en X_{n-1}. Il suffisait alors de faire le produit des récurrences de l'ordre 0 à l'ordre n, puis de simplifier, pour avoir X_n en fonction de X_0 et d'un paramètre k à la puissance n, soit $X_n = k^n\, X_0$.

Ici la récurrence $H_n = \lambda(1/2)\, H_{n-1} + (1 - \lambda)2pq$ est plus complexe puisqu'on voit apparaître un deuxième terme, constant, $(1 - \lambda)\, 2pq$.

La méthode générale de résolution consiste à écrire l'équation de l'écart entre le terme de la récurrence et sa limite (voir 3.5, déséquilibre gamétique). Dans le cas présent cela donne :

$$H_n - H = [\lambda(H_{n-1}/2) + (1 - \lambda)2pq] - H$$

d'où, en remplaçant H par sa valeur :

$$H_n - H = [\lambda(H_{n-1}/2) + (1 - \lambda)2pq] - [4(1 - \lambda)pq/(2 - \lambda)]$$

soit, après développement et factorisation :

$$H_n - H = (\lambda/2)[H_{n-1} - H]$$

Sous cette forme $X_n = k\ X_{n-1}$, il est facile d'obtenir la solution simple déjà évoquée $X_n = k^n\ X_0$, soit, dans le cas présent :

$$H_n - H = (\lambda/2)^n[H_0 - H]$$

ou $$\boldsymbol{H_n = (\lambda/2)^n[H_0 - H] + H}$$

Le paramètre λ va donc déterminer à la fois le niveau d'équilibre des hétérozygotes $H = 4(1 - \lambda)\ pq/(2 - \lambda)$, mais aussi la vitesse avec laquelle cet équilibre est atteint.

En effet $(\lambda/2)^n$ tendra d'autant plus vite vers zéro que λ est petit (figure 4.6), et la fréquence d'équilibre des hétérozygotes est d'autant moins inférieure à la fréquence panmictique $2pq$ que λ est faible, et d'autant plus vite atteinte.

Quant aux fréquences des homozygotes à l'équilibre, il est facile de les calculer, sachant que les fréquences alléliques sont restées inchangées et que la fréquence des hétérozygotes est égale à H.

À partir de l'équation $D_n = \lambda[\ D_{n-1} + (1/4)\ H_{n-1}\ ] + (1 - \lambda)\ p^2$

Il suffit de remplacer D_n et D_{n-1} par la valeur D à l'équilibre et H_{n-1} par $H = 2pq - 2pq\ \lambda/(2 - \lambda)$,

d'où : $D = \lambda[\ D + (1/4)[2pq - 2pq\ \lambda/(2 - \lambda)] + (1 - \lambda)\ p^2$

soit, pour *A1*/*A1* $\boldsymbol{D = p^2 + pq\ \lambda/(2 - \lambda)}$

et, pour *A2*/*A2* $\boldsymbol{R = q^2 + pq\ \lambda/(2 - \lambda)}$

On peut vérifier que les valeurs de D et R sont bien p^2 et q^2 dans une population panmictique ($\lambda = 0$) et p et q dans une population strictement autogame ($\lambda = 1$).

> ***Remarque 5 :*** l'autogamie induit une diminution de la fréquence des hétérozygotes d'un facteur $2pq\ \lambda/(2 - \lambda)$, compensée par une augmentation équivalente de la fréquence des homozygotes, répartie équitablement entre eux, soit $pq\ \lambda/(2 - \lambda)$, sans que les fréquences alléliques soient modifiées.

C'est un résultat général qui sera retrouvé dans toutes les équations prenant en compte un type quelconque d'écart à la panmixie (autogamie, homogamie ou unions consanguines), les fréquences génotypiques s'écriront comme celles de Hardy-Weinberg, modifiées par un facteur Δ, fonction du paramètre d'écart à la panmixie, avec, pour les gènes di-alléliques, un excès de chacun des homozygotes égal à la moitié du déficit en hétérozygotes, soit :

pour *A1*/*A1* $\boldsymbol{D = p^2 + \Delta pq}$

pour *A1*/*A2* $\boldsymbol{H = 2pq - 2pq\ \Delta}$

pour *A2*/*A2* $\boldsymbol{R = q^2 + \Delta pq}$ avec, ici, $\boldsymbol{\Delta = \lambda/(2 - \lambda)}$

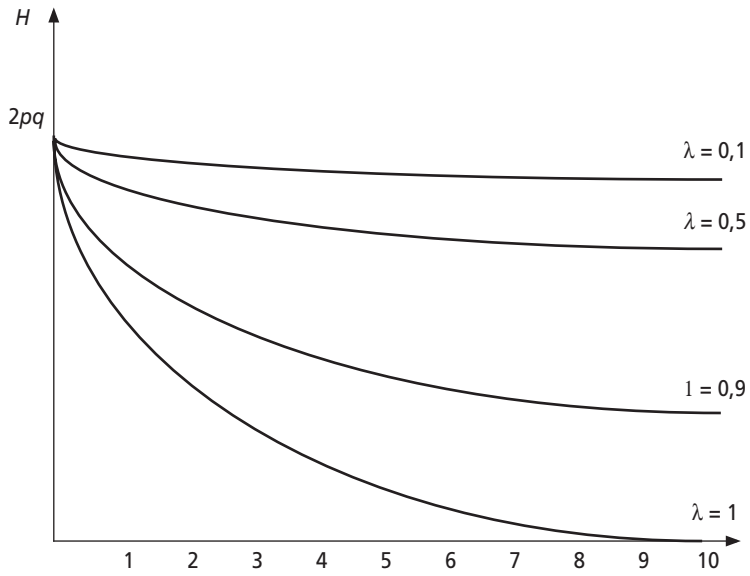


Figure 4.6 Limite et vitesse de l'évolution de la fréquence des hétérozygotes, dans une population autogame, à partir de leur fréquence panmictique $2pq$, en fonction du taux λ d'autogamie.

c) Croisements frère x *sœur systématiques*

Les croisements frères sœurs systématiques ou récurrents (figure 4.7) sont les croisements les plus « autogames » possibles dans les espèces à sexes séparés. Il ne sera pas surprenant de voir qu'ils conduisent, comme l'autofécondation, mais plus lentement qu'elle, à des lignées pures. C'est par ce type de croisements qu'on construit des lignées pures, en génétique expérimentale, chez la drosophile ou la souris.

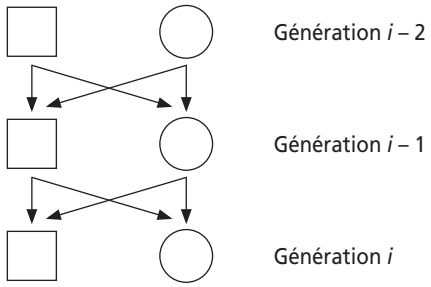


Figure 4.7

Dans un schéma comme celui de la figure 4.7, il y a un très grand nombre de chemins généalogiques entre un très grand nombre d'ancêtres communs, les parents frère-sœur de la génération $i - 1$, leurs grands-parents frère-sœur de la génération $i - 2$, etc. jusqu'à la génération initiale notée 0.

Dans un cas général, seul un calcul informatique, fondé sur un algorithme simple mais exigeant une mémoire et une vitesse de calcul suffisantes permet le recensement de tous ces éléments et le calcul de la consanguinité, ou de la parenté. Mais dans le cas particulier des croisements systématiques frères sœurs, il est possible d'établir une relation de récurrence simple, à partir du schéma général de la figure 4.7, et de calculer le coefficient de parenté entre deux germains de la génération i en ne tenant compte que des parents frère-soeur de la génération $i - 1$.

Comme on l'a vu, le coefficient de parenté est le produit de deux probabilités, la première est relative à l'origine des deux exemplaires du gène tirés chez les membres du couple, la deuxième est relative à l'identité par ascendance, sachant la provenance des deux exemplaires tirés.

Trois situations différentes, de probabilité connue, sont possibles quant à l'origine des exemplaires d'un gène, l'un tiré chez le frère, l'autre chez la sœur, à la génération i :

– une fois sur quatre, les deux exemplaires tirés viennent de leur père à la génération $i - 1$ (situation S1) ;
– une fois sur quatre, les deux exemplaires tirés viennent de leur mère à la génération $i - 1$ (situation S2) ;
– une fois sur deux, un exemplaire est d'origine paternelle et l'autre exemplaire d'origine maternelle (situation *S3*).

Dans chacune de ces situations *S1*, *S2* et *S3*, il est facile d'estimer la probabilité avec laquelle les deux exemplaires du gène, tirés chez le frère et la sœur de la génération *i*, sont identiques par ascendance :

– si les deux exemplaires tirés sont d'origine paternelle (situation *S1*), ils sont identiques par ascendance avec la probabilité $[1/2 + f_{i-1}/2]$, où f_{i-1} est le coefficient de consanguinité du père (voir 4.2.2.a) ;
– si les deux exemplaires tirés sont d'origine maternelle (situation *S2*), ils sont identiques par ascendance avec la probabilité $[1/2 + f_{i-1}/2]$, où f_{i-1} est le coefficient de consanguinité de la mère ;
– si les deux exemplaires tirés sont l'un d'origine paternelle, et l'autre d'origine maternelle (situation *S3*), ils pourront quand même être identiques par ascendance, car les parents sont, dans notre cas, apparentés. Par définition, la probabilité que deux exemplaires d'un gène venant de deux individus de la génération $i - 1$, soient identiques par ascendance est égale à leur coefficient de parenté, soit ϕ_{i-1}.

> ***Remarque 1 :*** dans les croisements frère-sœur récurrents, les individus des deux sexes jouent un rôle symétrique. De ce fait, les individus d'une même génération, quel que soit leur sexe, auront une même consanguinité, et une même parenté entre eux.
>
> ***Remarque 2 :*** comme l'origine des deux exemplaires tirés chez le couple de la génération *i*, a été définie de manière exhaustive, et que dans chacun des cas, la probabilité d'identité par ascendance a pu être définie, la mesure de la parenté à la génération est totalement estimée. En effet toute la parenté résultant des ancêtres d'ascendance supérieure a été comptabilisée (voir généalogie de la reine

Hatshepsout), soit par la prise en compte de la consanguinité f_{i-1} des frère sœurs à la génération $i-1$, soit par la prise en compte de leur parenté ϕ_{i-1}.

Le coefficient de parenté entre deux frère-sœur de la génération i s'écrit alors :

$$\phi_i = [1/2 + f_{i-1}/2] \times 1/4 + [1/2 + f_{i-1}/2] \times 1/4 + \phi_{i-1} \times 1/2$$

soit
$$\phi_i = 1/4 + f_{i-1}/4 + \phi_{i-1}/2$$

Sachant la propriété liant la consanguinité d'un individu à la parenté de ses parents, on peut dire que le coefficient de consanguinité d'un individu à la génération i, noté f_i, est égal au coefficient de parenté de ses parents, à la génération $i-1$, noté ϕ_{i-1}. De même $f_{i-1} = \phi_{i-2}$.

La formule de récurrence obtenue au dessus peut alors être écrite de manière homogène, soit sur la parenté, soit sur la consanguinité, ce qui donne :

– en parenté : $\phi_i = 1/4 + \phi_{i-2}/4 + \phi_{i-1}/2$
– en consanguinité : $f_{i+1} = 1/4 + f_{i-1}/4 + f_i/2$
– ou, indexée sur i : $f_i = 1/4 + f_{i-2}/4 + f_{i-1}/2$

Cette formule de récurrence montre que les croisements frères-sœurs récurrents aboutissent à une lignée pure puisque la valeur limite f ou ϕ vérifiant la formule ne peut être égale qu'à 1.

Les premières valeurs de la récurrence des croisements frères-sœurs sont données dans la figure 4.8. Après quelques générations, cette équation de récurrence admet une solution algébrique dont la forme simplifiée et applicable est :

$$f_i = 1 - e^{-i/4}$$

Remarque 1 : on peut considérer qu'après une quinzaine de générations, on dispose d'une souche presque pure (c'est-à-dire homozygote pour f = 98 % du génome).

Remarque 2 : on notera la similitude de cette équation avec celle de la consanguinité résultant de la dérive (voir chapitre 5). En fait, les croisements systématiques entre frères-sœurs reviennent à considérer une population constituée à chaque génération d'un individu de sexe masculin et d'un individu de sexe féminin, ce qui correspond bien, dans l'équation de la dérive, à un effectif efficace de 2. On peut aussi appliquer à cette équation le calcul de la période qui sera développée dans ce chapitre.

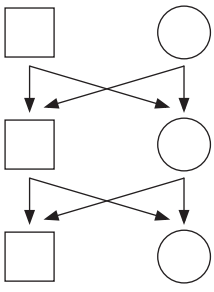


Figure 4.8 Valeurs des premiers termes de la récurrence dans les croisements frères-sœurs.

4.3 COMPOSITION GÉNÉTIQUE DES POPULATIONS CONSANGUINES

Une partie de ce sous-chapitre est plus spécialement consacrée aux populations humaines, dans la mesure où la consanguinité liée au choix du conjoint est une réalité culturelle et sociologique spécifique que doivent prendre en compte de nombreux chercheurs ou praticiens, mais elle peut aussi s'appliquer à des populations animales ; les conséquences de l'autofécondation ou de l'autogamie partielle sur la composition génétique des populations végétales ont été vues plus haut, 4.2.3.a et 4.2.3.b). L'effet combiné de la consanguinité et de l'effet Walhund s'applique à tout type de populations.

4.3.1 Choix du conjoint en fonction de la parenté et composition génétique de la population

a) Coefficient moyen de parenté et de consanguinité dans une population non panmictique

Dans une population humaine la réalité matrimoniale est hétérogène quant aux modalités de rencontre entre conjoints. On peut considérer qu'une fraction des unions est panmictique et qu'une fraction ne l'est pas, obéissant à des règles sociales, culturelles ou religieuses.

Dans ces circonstances, on peut définir un paramètre F_i, égal à la moyenne des coefficients de parenté ϕ des différents couples de la génération i. Ainsi défini, le coefficient F_i est le coefficient de parenté d'un couple tiré au hasard, à la génération i, et aussi le coefficient de consanguinité d'un enfant tiré au hasard à la génération suivante $i + 1$.

> ***Exemple 4.1***
>
> Une grande population où 16 % des mariages sont réalisés entre cousins germains alors que les autres (84 %) sont panmictiques présentera un coefficient moyen de parenté égal à :
>
> $$F = (16/100) \times (1/16) + (84/100) \times 0 = 1/100$$
>
> et le coefficient moyen de consanguinité, à la génération suivante, sera aussi égal à 1 %.

b) Composition génétique des populations consanguines

Le choix du conjoint en fonction de la parenté aura-t-il un effet sur la composition génétique (fréquences alléliques et génotypiques) ? Si oui, lequel ?

Considérons le cas simple d'un gène di-allélique *A1* et *A2*, de fréquences p et q, chez les parents, et calculons les fréquences génotypiques et alléliques chez les enfants de la génération suivante, sachant que ces enfants présentent une consanguinité moyenne égale à F, le coefficient moyen de parenté résultant de la fraction des unions entre parents apparentés.

Tirons un enfant au hasard, avec quelle probabilité sera-t-il *A1A1* ? (ce serait p^2, si la population était panmictique).

Les deux exemplaires du gène *A*, présents chez cet enfant, peuvent :

– soit « être identique par ascendance », évènement de probabilité F ;

– soit ne pas « être identique par ascendance », évènement contraire de probabilité $(1 - F)$.

Dans le premier cas, la probabilité que le premier exemplaire du gène *A* soit *A1* est égale à *p*, la fréquence de *A1* chez les parents, mais la probabilité que le deuxième exemplaire soit aussi *A1* est égale à 1, puisqu'on est dans le cas où les deux exemplaires du gène sont « identiques par ascendance » ; si le premier est *A1*, le deuxième l'est forcément aussi.

Dans le deuxième cas, la probabilité que le premier exemplaire du gène A soit *A1* est égale à *p*, la fréquence de *A1* chez les parents, et la probabilité que le deuxième exemplaire soit aussi *A1* est encore égale à *p*, puisqu'on est dans le cas où les deux exemplaires du gène ne sont pas « identiques par ascendance » ; si le premier est *A1*, le deuxième ne l'est pas forcément : il ne l'est que si on le tire à nouveau, avec la probabilité *p*.

D'où la fréquence des génotypes *A1A1* chez les enfants :

$$f(A1A1) = Fp + (1 - F)p^2$$

Un raisonnement identique, où *A2* et *q* remplacent *A1* et *p*, conduit à la fréquence attendue des génotypes *A2A2*, chez les enfants :

$$f(A2A2) = Fq + (1 - F)q^2$$

Avec quelle fréquence un enfant tiré au hasard sera-t-il hétérozygote *A1A2* (ce serait $2pq$, si la population était panmictique) ?

A priori les deux exemplaires du gène *A*, présents chez cet enfant, peuvent :

– soit « être identique par ascendance », évènement de probabilité *F* ;

– soit ne pas « être identique par ascendance », évènement contraire de probabilité $(1 - F)$.

Dans le premier cas, la probabilité que le premier exemplaire du gène *A* soit *A1* est égale à *p*, la fréquence de *A1* chez les parents, mais la probabilité que le deuxième exemplaire soit *A2* est alors égale à 0, puisqu'on est dans le cas où les deux exemplaires du gène sont « identiques par ascendance » ; si le premier est *A1*, le deuxième l'est forcément aussi.

Dans le deuxième cas, la probabilité que le premier exemplaire du gène *A* soit *A1* est égale à *p*, la fréquence de *A1* chez les parents, et la probabilité que le deuxième exemplaire soit *A2* est égale à *q*, puisqu'on est dans le cas où les deux exemplaires du gène ne sont pas « identiques par ascendance » ; mais on peut aussi avoir le premier *A2* et le deuxième *A1*, avec la probabilité *qp*.

D'où la fréquence des génotypes *A1A2* chez les enfants :

$$f(A1A2) = Fp_0 + (1 - F)2pq$$

Ces fréquences génotypiques peuvent, après développement et mise en facteur, être écrites de la façon suivante :

$$f(A1A1) = p^2 + Fp(1 - p)$$
$$f(A1A2) = 2pq - 2Fpq$$
$$f(A2A2) = q^2 + Fq(1 - q)$$

ce qui conduit à plusieurs remarques et conclusions.

Conclusion 1 : dans une population consanguine résultant d'unions entre apparentés, on observe un accroissement de la fréquence des homozygotes et une diminution de celle des hétérozygotes.

Ce résultat est logique puisque l'union entre apparentés conduit à une plus grande homozygotie des enfants comme conséquence de leur consanguinité.

Conclusion 2 : dans une population consanguine résultant d'unions entre apparentés, les fréquences alléliques ne sont nullement modifiées.

En effet la fréquence de l'allèle *A1*, calculée comme la somme de la fréquence du génotype *A1A1* et de la demi-fréquence du génotype hétérozygote, donne :

$$f(A1) = p^2 + Fp(1 - p) + pq - Fpq$$

soit $f(A1) = p^2 + Fpq + pq - Fpq$

d'où $f(A1) = p^2 + pq = p(p+q) = p$

Remarque 1 : la pratique des « mariages consanguins » ne modifie nullement la diversité génétique des populations, en terme de fréquences alléliques. En d'autres termes, et contrairement aux idées qui ont longtemps couru dans les milieux de la médecine ou de l'anthropologie, les mariages consanguins ne conduisent pas, dans une grande population, à une « détérioration du stock génique de la population ». Et si la « consanguinité » s'accompagne d'une augmentation de la morbidité, pour les maladies récessives, la raison n'est pas dans l'augmentation de la fréquence des allèles morbides mais dans l'augmentation de la fréquence des homozygotes, notamment les individus atteints. En effet, la consanguinité ne modifie pas la diversité génique mais peut modifier profondément la diversité génotypique, faisant passer la fréquence d'un homozygote de la fréquence panmictique q^2 à la fréquence $(q^2 + Fpq)$ qui lui est d'autant plus supérieure que q est proche de zéro (dans ce cas, p est proche de 1, Fpq est égal à Fq et est bien supérieur à q^2). Mais la fréquence de l'allèle pathogène reste inchangée et égale à q.

Remarque 2 : pour les maladies dominantes, la morbidité des populations consanguines est diminuée par rapport à la situation panmictique, puisque la fréquence des individus atteints est égale à $(p^2 + 2pq - Fpq)$ au lieu de $(p^2 + 2pq)$, dans la situation panmictique. Cependant cette diminution n'est pas aussi perceptible que l'augmentation de la fréquence des maladies récessives. En effet, p étant proche de zéro (fréquence d'un allèle morbide dominant, voir chapitre 2), la fréquence d'une maladie dominante est proche de $2p$ en situation panmictique et de $2p - Fp$ dans une population consanguine, or Fp est négligeable devant $2p$, alors que Fq ne l'était pas devant q^2.

Conclusion 3 : l'accroissement de la fréquence des homozygotes, comme la diminution de celle des hétérozygotes, prend la forme d'un écart à la panmixie.

En effet, on retrouve pour chacun des génotypes les fréquences p^2, $2pq$ et q^2, caractéristique de l'équilibre de Hardy-Weinberg, chacune augmentée ou diminuée d'un écart fonction des fréquences alléliques d'une part, de l'intensité de la consanguinité mesurée par F d'autre part ($\Delta = F$, dans remarque 5, page 132).

On remarquera d'ailleurs que si, à une génération, aucune union entre apparentés n'est réalisée, alors $F = 0$, et les équations sont celles de l'équilibre de Hardy-Weinberg.

c) Cas d'un gène pluri-allélique

Dans ce cas les écarts à la panmixie induits par les unions entre apparentés donnent, par application des formules du chapitre 3 :

$$f(AiAi) = p_i^2 + F\,p_i(1 - p_i)$$
$$f(AiAj) = 2\,p_i\,p_j - 2F\,p_i\,p_j \qquad (\text{avec } i > j)$$

Il faut remarquer ici que les écarts n'ont plus la symétrie qu'ils avaient dans le cas d'un gène di-allélique. Ils peuvent être très élevés pour certains génotypes et nuls pour d'autres, tout en restant positifs pour les homozygotes et négatifs pour les hétérozygotes.

d) Calcul des fréquences alléliques dans une population consanguine

Le calcul des fréquences alléliques dans une population consanguine ne peut plus se fonder sur la relation de Hardy-Weinberg, définie et appliquée au chapitre 2. En effet la fréquence R des enfants atteints d'une maladie récessive ne sera plus égale à q^2 mais sera égale à $q^2 + Fq(1 - q)$.

L'estimation de la valeur de q est la solution de l'équation $R = q^2 + Fq(1 - q)$, comprise entre 0 et 1.

Résoudre cette équation suppose connues les valeurs de F et de R. La valeur de R est assez bien estimée pour les maladies assez fréquentes mais reste très difficile à estimer pour les maladies rares. Par ailleurs la valeur de F suppose des recensements assez exhaustifs.

On peut contourner ces difficultés par l'utilisation de la formule de Dalhberg[1], fondée sur le seul rapport k, nombre d'enfants atteints, dont les parents sont cousins germains, rapporté au nombre total de cas observés.

---

1. **La formule de Dalhberg**
La formule de Dalhberg (1948) permet de calculer la fréquence q de l'allèle pathogène responsable d'une maladie génétique récessive, dans une population consanguine où une fraction c des unions concerne des cousins germains, dont la parenté est égale à 1/16.
Considérons un enfant issu de cet ensemble de couples panmictiques $(1 - c)$ ou cousins germains (c) :
– ou bien il est issu d'un couple panmictique, évènement de probabilité $\mathbf{(1 - c)}$, et, dans ce cas, il sera atteint avec la probabilité $\boldsymbol{q^2}$;
– ou bien il est issu d'un couple de cousins germains, évènement de probabilité $\boldsymbol{c}$, et, dans ce cas, il sera atteint avec la probabilité $\phi q + (1 - \phi)q^2$ (voir plus haut la démonstration avec F, remplacé ici par ϕ), soit $\mathbf{(q + 15\,q^2)/16}$, dans le cas présent ($\phi = 1/16$).
Si on connaît (par les statistiques hospitalières) la fraction k des enfants atteints dont les parents sont cousins germains, on peut écrire ce rapport k des enfants atteints et issus de cousins germains à la totalité des enfants atteints comme :

$$k = [c\,(q + 15\,q^2)/16]/[c\,(q + 15\,q^2)/16) + (1 - c)\,q^2]$$

soit après développement, simplification et mise en facteur

$$k = c\,(1 + 15q)/[16q + c(1 - q)]$$

d'où $\boldsymbol{q = c(1 - k)/[k(16 - c) - 15c]}$

En effet, la plupart du temps la majorité des unions entre apparentés concernent des cousins germains (ϕ = 1/16). Le généticien Dalhberg a défini un estimateur adapté à cette situation ; la valeur de la fréquence q de l'allèle pathogène d'une maladie récessive est donnée par :

$$q = c(1 - k)/[k(16 - c) - 15c]$$

où c est la fréquence de mariages entre cousins germains, dans la population, et k, la fraction des enfants atteints dont les parents sont cousins germains.

4.3.2 Consanguinité, effet Walhund et « statistiques *F* » de Wright

L'effet Walhund (voir 2.5.3) est généré par la structuration d'une population, ou d'une espèce, en sous entités génétiquement différenciées présentant donc des compositions génétiques différentes pour un ensemble de gènes et conduisant au fait que la population générale n'est pas une réelle entité panmictique. De nombreuses causes peuvent concourir, ensemble ou non, à cette différenciation parmi lesquelles :

– la taille de l'ère de répartition, qui peut générer un isolement spatial si les individus de l'espèce étudiée ont une mobilité assez réduite pour empêcher le croisement d'individus trop éloignés ;
– une endogamie de sous populations, c'est-à-dire des obstacles aux échanges génétiques, résultant d'obstacles naturels (reliefs, océans, ..) ou culturels (religion, culture, ethnie, ...) ;
– une variation des conditions de milieu (sélection) conduisant à des différenciations locales

a) Écart à la panmixie associé à l'effet Walhund

Soit une population générale G, structurée en n sous populations, au sein desquelles l'allèle A d'un gène présente les fréquences $p_1,\ldots, p_i,\ldots., p_n$, et l'allèle a les fréquences $q_1,\ldots, q_i,\ldots., q_n$.

Considérons, par ailleurs, que chaque sous population présente une composition génétique conforme au modèle de Hardy-Weinberg, avec des fréquences génotypiques telles qu'elles sont rapportées dans le tableau 4.6 (lignes 1 à n).

La composition génétique de la population générale G peut s'écrire comme la somme des fréquences génotypiques de chaque sous population, pondérées par leurs poids respectifs, c'est-à-dire leur espérance (tableau 4.6, avant dernière ligne supérieure, où t_i est la taille de la sous population i, et T est la taille de la population générale G, c'est-à-dire Σt_i).

On peut aussi écrire que les fréquences moyennes P et Q des allèles A et a, dans G, sont égales à leurs espérances $E(p)$ et $E(q)$, soit :

$$P = E(p) = \Sigma\,(t_i/T)\,p_i$$
$$Q = E(q) = \Sigma\,(t_i/T)\,q_i$$

On peut remarquer que le fait que $q_i = (1 - p_i)$ conduit à

$$Q = E(1 - p_i) = 1 - \Sigma\,(t_i/T)\,p_i = 1 - P$$

Tableau 4.6

Sous-populations	Génotypes		
	A/A	***A/a***	***A/a***
1	p_1^2	$2\,p_1\,q_1$	q_1^2
2	P_2^2	$2\,p_2\,q_2$	q_2^2
…			
i	P_i^2	$2\,p_i\,q_i$	q_i^2
…			
n	P_n^2	$2\,p_n\,q_n$	q_n^2
Population générale *G*	$\Sigma\,(t_i/T)\,p_i^2$ $\boldsymbol{P^2 + V(p)}$	$\Sigma\,(t_i/T)\,2p_i.q_i$ $2PQ - 2V(p)$	$\Sigma\,(t_i/T)\,q_i^2$ $\boldsymbol{Q^2 + V(p)}$
Population théorique panmictique	P^2	$2PQ$	Q^2

Si la population générale *G* était une unité panmictique, sa composition génétique serait celle donnée par la dernière ligne du tableau 1.6. qui diffère de la composition générale du fait de la structuration de cette population en sous populations de composition différente.

En effet, du fait de la variabilité des fréquences entre les sous populations, il est possible de définir la variance des fréquences des allèles *A* et *a* au sein de la population générale *G*, soit :

$$V(p) = \Sigma\,(t_i/T)\,(p_i - P)^2$$

$$V(q) = \Sigma\,(t_i/T)\,(q_i - Q)^2$$

et $V(q) = V(p)$ puisque $qi = (1 - p)$ et que $Q = (1 - P)$

Or, sachant que la variance d'une variable est égale à l'espérance de son carré moins le carré de son espérance, il est possible d'écrire que :

$$V(p) = \Sigma\,(t_i/T)p_i^2 - P^2$$

$$V(q) = V(p) = \Sigma\,(t_i/T)q_i^2 - Q^2$$

D'où on peut tirer que :

$$\Sigma\,(t_i/T)p_i^2 = P^2 + V(p)$$

$$\Sigma\,(t_i/T)q_i^2 = Q^2 + V(p)$$

$$\Sigma\,(t_i/N)\,2p_i.q_i = \Sigma\,(n_i/N)\,2p_i.(1 - p_i) = 2PQ - 2V(p)$$

Ces équations permettent de mettre en évidence l'effet Walhund (tableau 1.6, avant dernière ligne inférieure, en gras) comme un écart à la panmixie où la fréquence réelle des homozygotes est égale à leur fréquence théorique sous Hardy-Weinberg (P^2 ou Q^2) augmentée de la variance de cette fréquence, alors que la fréquence des hétérozygotes est égale à la fréquence théorique $2PQ$ diminuée de la covariance des fréquences (égale ici au double de la variance).

La composition réelle de la population n'est donc celle attendue sous le modèle de Hardy-Weinberg que si la variance $V(p)$ est égale à zéro, c'est-à-dire en absence de structuration en sous populations génétiquement différenciées.

Cet écart à la panmixie apparaît comme formellement équivalent à celui qui serait généré par une fraction d'unions consanguines, un phénomène qui conduit effectivement à la mise à l'écart de la panmixie une fraction des gamètes et définit deux sous populations dans la population générale, celle où les unions sont panmictiques et celle où elles ne le sont pas (voir 4.3.1, et remarque 5, page 132).

On peut même, formellement définir une « population consanguine équivalente », celle dont le coefficient moyen de consanguinité F conduirait au même écart. En pratique, on peut estimer, dans une population, à la fois la fréquence réelle Ho des hétérozygotes et les fréquences alléliques P et Q, et écrire :

$$Ho = 2PQ - 2FPQ = 2PQ - 2V(p)$$

D'où on peut tirer une estimation de F :

$$\boldsymbol{F = 1 - Ho/2PQ = V(p)/PQ}$$

On peut si l'écart à la panmixie est induit par un taux de consanguinité, estimer par F, ou s'il résulte d'un effet Walhund, estimer l'équivalent de F en terme de variance des fréquences allélique par la relation $\boldsymbol{V(p) = FPQ}$.

Le paramètre F, associé au déficit d'hétérozygote dans les populations consanguines, joue un rôle équivalent à l'*indice de fixation* $\lambda/(2-\lambda)$ associé à la perte d'hétérozygotie dans les espèces partiellement autogames (voir 4.2.3.b), et de la même façon pour le rapport $V(p)/PQ$ dans une population structurée présentant un effet Walhund. Dans ce dernier cas, on voit que la valeur de l'équivalent $F = V(p)/PQ$ est d'autant plus élevée que la différenciation génétique entre sous populations est importante.

b) Statistiques « F » de Wright

Pour tenir compte du fait que les écarts à la panmixie peuvent être induits simultanément par plusieurs causes, le généticien américain Sewall Wright a introduit un groupe de paramètres désigné par statistiques F. Il a ainsi défini :

– F_{IS}, mesure de l'écart à la structure théorique de Hardy-Weinberg au sein d'une sous-population résultant d'un écart à la panmixie dus à des comportements individuels (*IS* signifiant « Individual within Sub-population »), comme les unions entre apparentés ou l'autogamie partielle. Ce paramètre est associé à la perte d'hétérozygotie en raison de l'écart à la panmixie interne à la sous-population i selon l'équation :

$$\boldsymbol{H_i = 2p_iq_i\,(1 - F_{ISi})}$$

– F_{ST}, mesure du déficit d'hétérozygotes dans la population générale G, en raison du seul effet Walhund résultant d'une différenciation génétique entre sous populations autre que l'autogamie interne (*ST* pour « Sub-population within Total »). Comme cela a été montré plus haut :

$$\boldsymbol{F_{ST} = V(p)/PQ}$$

– F_{IT}, mesure du déficit global d'hétérozygotes dans la population générale *G*, en raison de l'effet Walhund et de l'autogamie interne aux sous populations (*IT* pour « Individual within Total »), qui peut s'écrire :

$$\boldsymbol{H_o = 2PQ\,(1 - F_{IT})}$$

Si toutes les sous populations *i* présentent des valeurs presque égales du paramètre F_{ISi}, soit F_{IS}, il est alors possible de mesurer la moyenne des écarts à la panmixie internes aux sous populations, soit :

$$Ho = \Sigma\,[2p_iq_i\,(1 - F_{IS}\,)].(t_i/T)$$

où t_i est la taille de la sous population *i*, et *T* est la taille de la population générale *G*.

Soit : $Ho = (1 - F_{IS}\,)\,\Sigma\,(2p_iq_i\,).(ti/T)$

Comme on a montré plus haut que $\Sigma\,(2p_iq_i\,).(ti/T) = 2PQ - 2V(p)$

et que $F_{ST} = V(p)/PQ$

on peut écrire que $\Sigma\,(2p_iq_i\,).(ti/T) = 2PQ - 2PQ.F_{ST} = 2PQ(1 - F_{ST})$

et *Ho* peut s'écrire $Ho = 2PQ.(1 - F_{IS}\,).(1 - F_{ST})$

D'où on tire de $Ho = 2PQ.(1 - F_{IT})$

que $\boldsymbol{(1 - F_{IT}) = (1 - F_{IS}\,).(1 - F_{ST})}$

Cette formule de Wright permet donc de préciser comment, dans la population générale *G*, s'articulent les effets propres aux sous populations et la structuration spatiale en sous populations.

Cette formule peut aussi être considérée sous un angle probabiliste quand on considère qu'une statistique *F* est une probabilité de tirer deux exemplaires identiques par ascendance d'un allèle, soit en raison du régime de croisement au sein de la sous population (F_{IS}), soit en raison de la restriction de tirage résultant de la structuration de la population générale *G* en sous populations (F_{ST}). Dans ces conditions, on peut définir F_{IT} comme la probabilité de tirer, dans la population générale, deux exemplaires d'un gène identiques par ascendance. De ces définitions, il est possible, par le théorème des probabilités composées, d'énoncer que :

$$F_{IT} = F_{IS} + (1 - F_{IS}).\,F_{ST}$$

ou $$F_{IT} = F_{ST} + (1 - F_{ST}).\,F_{IS}$$

qui conduisent toutes les deux à l'équation $\boldsymbol{(1 - F_{IT}) = (1 - F_{IS}\,).(1 - F_{ST})}$

Ces statistiques *F* permettent d'analyser la structuration de la variabilité au sein des populations naturelles et sont d'ailleurs applicables à tous facteurs de différenciation intra ou inter sous populations.

4.3.3 Consanguinité, conseil génétique et santé publique

Dans la mesure où la consanguinité accroît la probabilité d'homozygotie, il est facile de comprendre qu'elle accroît le risque de pathologie récessive résultant de la présence d'une éventuelle mutation morbide chez l'un des ancêtres communs du couple apparenté. Mais les problèmes, bien que de même nature, n'ont pas la même ampleur selon qu'on les considère à l'échelle individuelle, qui relève du conseil génétique, ou à l'échelle collective de la population, qui relève de la santé publique.

a) Consanguinité, risque familial et conseil génétique

À l'échelle d'un couple, le risque de naissance d'un enfant atteint d'une maladie génétique récessive est égal au risque panmictique q^2, augmenté du supplément de risque $\phi q(1 - q)$ associé à la parenté ϕ du couple, ce qui donne, pour l'enfant consanguin, un risque global égal à $q^2 + \phi q(1 - q)$. Le tableau 4.6 donne une idée des choses pour un couple de cousins germains ($\phi = 1/16$).

TABLEAU 4.7

Maladie	Fréquence de la mutation pathogène (q)	Risque panmictique q^2	Supplément de risque $\phi q(1 - q)$ avec $\phi = 1/16$	Accroissement relatif du risque
Mucoviscidose	2 %	1/2 500	3/2 500	x 4
Phénylcétonurie	0,8 %	1/16 000	8/16 000	x 9
Galactosémie	0,5 %	1/40 000	13/40 000	x 14

Le supplément de risque absolu $\phi q(1 - q)$ peut s'écrire ϕq dès que la fréquence q est petite (dès que l'allèle muté responsable de la pathologie est rare). Ce supplément de risque est évidemment d'autant plus faible que cet allèle est rare, mais l'accroissement relatif du risque égal à $[q^2 + \phi q]/q^2$ est d'autant plus élevé que l'allèle pathogène, et partant la maladie sont rares (dernière colonne du tableau 4.7). À la limite les enfants atteints d'une maladie très rare seront tous consanguins comme cela est illustré par le tableau 4.8 qui présente, en fonction de la fréquence q de l'allèle pathogène responsable d'une maladie récessive, le rapport des enfants atteints, consanguins issus de cousins germains ($f = 1/16$) aux enfants atteints non consanguins, c'est-à-dire le rapport $[q^2 + fq]/q^2 = (q + f)/q$.

TABLEAU 4.8

Valeur de q	0,5	0,4	0,3	0,1	0,01	0,001	0,0001
Valeur de $(q + f)/q$	1,06	1,09	1,15	1,56	7,19	63,4	625

Pour une maladie dont l'allèle responsable n'est pas trop rare ($q = 0,01$), ce rapport est égal à 7,19 : pour un enfant atteint issu d'un couple panmictique, sept enfants atteints sont issus de cousins germains.

Pour une maladie dont l'allèle responsable est dix fois plus rare ($q = 0,001$), ce rapport monte à 63,4 : pour un enfant atteint issu d'un couple panmictique, 63 enfants atteints sont issus de cousins germains !

Cependant, à l'échelle d'un couple de cousins germains ou de tout autre parenté, il est peu probable que les quelques ancêtres communs aient été porteurs sains de toutes les mutations pour toutes les maladies, le risque réel n'est pas, pour chaque couple, la somme des risques sur toutes les maladies. De plus il s'agit, dans le cas présent, de la conception d'un ou de quelques enfants.

Ce ne sera plus le cas, à l'échelle collective de la population, où les risques familiaux s'additionnent et peuvent, sur la masse des naissances, se manifester concrètement par un supplément de morbidité important en termes de santé publique.

b) Consanguinité, risque collectif et santé publique

À l'échelle de la population, le risque de naissance d'un enfant atteint d'une maladie génétique récessive est égal au risque panmictique, q^2, augmenté du supplément moyen de risque $Fq(1-q)$ associé à la parenté moyenne F des couples, ce qui donne, pour les enfants, un risque global égal à $q^2 + Fq(1-q)$. Ce risque, pour chacune des maladies, doit être, à l'échelle de la population multiplié par le nombre de maladies génétiques qui y sont présentes, ce qui donne un accroissement considérable de la morbidité, comme le révèle l'exemple suivant.

Dans cet exemple (tableau 4.9), on compare le taux de morbidité d'une population panmictique, où sont recensées 1 103 maladies plus ou moins fréquentes, au taux de morbidité d'une population où 16 % des mariages sont réalisés entre cousins germains, ce qui donne une valeur de F égale à 1 %.

Afin de juger concrètement de l'incidence de la consanguinité en matière de santé publique, les risques calculés ont été appliqués à une population qui compterait un million de naissances annuelles (la France plus le Benelux).

TABLEAU 4.9

Maladies récessives recensées	Très rares	Rares	Assez fréquentes (phénylcétonurie)	Fréquente (mucoviscidose)	Total
Nombre de maladies recensées	1 000	100	2	1	1 103
Fréquence q de la mutation pathogène	10^{-4}	10^{-3}	10^{-2}	2.10^{-2}	
Risque panmictique q^2	10^{-8}	10^{-6}	10^{-4}	4.10^{-4}	
Nombre de cas annuels (sur 10^6 naissances)					
– par maladie	0,1	1	100	400	
– pour toutes les maladies	10	100	200	400	710
Accroissement Fq du risque (pour F=0,01)	10^{-6}	10^{-5}	10^{-4}	2.10^{-4}	
Cas annuels supplémentaires					
– par maladie	1	10	100	200	
– pour toutes les maladies	1 000	1 000	200	200	2 400

On constate donc que 1 % de consanguinité multiplie par plus de quatre le nombre de naissances d'enfants atteints d'une pathologie génétique récessive.

Ce résultat, à l'échelle de la population, n'est pas contradictoire avec la conclusion tirée précédemment à l'échelle individuelle, considérant qu'assez souvent les ancêtres communs avaient peu de chances d'être porteurs sains. En reprenant les données du tableau 4.8, on peut conclure qu'il y a sur le million de naissances, 160 000 (16 %) issues de couples cousins germains, ce qui signifie que sur les 160 000 couples apparentés, 2 400 ont joué de « malchance » et que 157 600 ont été « chanceux ». Ce type de résultat statistique est le même que celui des victimes de la route pour lequel chaque automobiliste a un risque individuel très faible d'être accidenté, ce qui n'empêche pas une centaine d'entre eux d'y laisser la vie lors du week-end de Pâques.

Les problèmes de santé publique qui résultent de la pratique des unions entre apparentés ne sont pas seulement quantitatifs (multiplication du nombre des cas) mais aussi qualitatifs, car la société et le corps médical se trouvent confrontés à une grande diversité de pathologies différentes exigeant une variété de spécialistes et de thérapies (quand elles existent). En effet les trois maladies principales (tableau 4.9) qui représentent 85 % de la pathologie en régime panmictique (600 cas sur 710), ne représentent plus que 32 % de celle-ci quand F= 0,01. On remarque que des pathologies très rares (q = 0,0001) pour lesquelles on observait une naissance par siècle présentent alors une naissance par an !

> ***Remarque 1 :*** l'accroissement de morbidité en santé publique, illustré par le tableau 4.9, est d'autant plus réaliste que le nombre de maladies génétiques existant dans une grande population excède largement le nombre pris en exemple (on connaît plusieurs milliers de maladies génétiques) et que certaines populations, comme l'Irak, présentent des taux moyen de consanguinité avoisinant 3 %.
>
> ***Remarque 2 :*** une morbidité élevée peut chuter très rapidement si les unions deviennent panmictiques. Le régime panmictique est un phénomène associé au développement économique, à l'urbanisation, à la transition démographique et à l'acculturation. En France la valeur moyenne F qui était déjà très faible au début du XX^{e} siècle ($8{,}6.10^{-4}$) a encore chuté à $2{,}3.10^{-4}$ au début des années 1960, en plein milieu du dépeuplement rural. De fortes valeurs de F se rencontrent encore dans les populations où de fortes traditions culturelles président au choix du conjoint (par l'homme évidement), en Afrique ou en Orient plus ou moins proche.

4.3.4 Consanguinité et cartographie des gènes : « *homozygosity mapping* »

Le génome humain a été balisé par la localisation d'un grand nombre de marqueurs génétiques constitués par des polymorphismes moléculaires de l'ADN (chapitre 1). Avec les premiers d'entre eux, les RFLP, Botstein a défini en 1980 une stratégie d'assignation chromosomique et de cartographie d'un gène impliqué dans une maladie monogénique, consistant à identifier le ou les marqueurs présentant une

liaison génétique avec ce gène. Le principe consiste à tester, dans un échantillon de familles avec plusieurs enfants dont au moins un est atteint, la ségrégation de la maladie avec celle de marqueurs répartis sur le génome. Dans le cas d'une maladie dominante, le problème est assez simple, dans le cas d'une maladie récessive où les deux parents porteurs sont porteurs sains d'une mutation pathogène, le problème est moins facile et suppose à la fois un échantillon assez fourni de familles et une informativité du marqueur testé (parents hétérozygotes). Si la ségrégation du marqueur testé est indépendante de celle de la maladie (en fait de l'allèle pathogène dominant ou des allèles morbides récessifs), cela signifie que le gène impliqué dans la maladie n'est pas localisé au voisinage du marqueur ; dans le cas contraire, l'établissement d'une liaison génétique entre le locus du marqueur et celui du gène locus permet soit d'identifier le chromosome porteur du locus du gène « maladie », soit de préciser la localisation de ce gène sur le chromosome par des calculs de distance. C'est ainsi qu'ont été localisés les gènes impliqués dans la maladie de Huntington, la mucoviscidose, la forme héréditaire de cancer du sein, la forme précoce de la maladie d'Alzheimer et plusieurs dizaines d'autres pathologies.

Mais cette stratégie de cartographie génétique par recherche systématique d'une liaison génétique entre le locus du gène impliqué dans la maladie et plusieurs dizaines de marqueurs du génome, est non seulement très lourde sur le plan pratique, mais elle suppose aussi d'avoir un grand nombre de familles informatives (au moins deux enfants dont au moins un atteint) pour permettre au test statistique de la liaison génétique d'avoir une puissance suffisante. Aujourd'hui, les marqueurs de type microsatellite sont largement utilisés car ils sont presque toujours informatifs, très nombreux, bien répartis sur le génome, faciles à tester dans des routines totalement ou partiellement automatisées (PCR multiplex + séquenceur) ce qui permet d'inclure toutes les familles recensées et d'étudier plus facilement des maladies assez rares. Toutefois, la puissance de cette stratégie reste trop faible quand une maladie est si rare que quelques cas seulement sont recensés, maladies appelées « orphelines » car la recherche ne peut s'y intéresser, non par manque d'intérêt mais plutôt par impuissance.

Il existe cependant une exception pour les maladies récessives très rares présentes dans les populations consanguines. Comme on l'a vu, la consanguinité, malgré une fréquence très faible de la mutation pathologique, peut induire un écart à la panmixie si élevé que la maladie sera observable dans un petit groupe de familles très consanguines. Botstein a alors proposé en 1987, une stratégie adaptée à ces circonstances exceptionnelles, appelée « *homozygosity mapping* ». Comme, dans chacune de ces familles consanguines, les enfants atteints sont porteurs de deux exemplaires identiques par ascendance pour le gène impliqué dans la pathologie, cette identité par ascendance doit aussi s'appliquer aux marqueurs polymorphes dans le voisinage du gène, qui auront eu d'autant moins de chances d'être recombinés par crossing-over qu'ils sont plus proches de celui-ci. La cartographie par homozygotie se propose de faire la recherche systématique des marqueurs polymorphes du génome qui, chez les enfants atteints, sont systématiquement ou significativement en situation d'homo-

zygotie, car cela indique que le gène impliqué dans la pathologie a toutes les chances de se trouver dans le voisinage d'un de ces marqueurs. Cette stratégie a été utilisée avec succès dans la localisation de nombreux gènes impliqués dans des maladies assez rares pour n'être observées que dans quelques familles de populations consanguines, voire dans une seule famille avec plusieurs branches et un nombre assez grand d'enfants atteints (en pratique, il suffit de quatre à sept enfants atteints et autant de non atteints pour pouvoir localiser un gène impliqué dans une maladie récessive).

4.4 L'HOMOGAMIE

Le choix du conjoint peut être fondé sur la similitude ou la dissemblance, génotypique ou phénotypique. Si le choix est conditionné par la ressemblance génotypique ou phénotypique on parle d'homogamie ; si le choix est conditionné par la dissemblance phénotypique ou génotypique (allèles d'incompatibilité chez certains végétaux), on parle d'hétérogamie. Seule l'homogamie sera envisagée dans ce chapitre.

L'homogamie génotypique et l'homogamie phénotypique se recouvrent quand il s'agit de phénotypes codominants, mais ne se recouvrent pas quand le gène étudié gouverne des phénotypes dominants et récessifs. Par ailleurs l'homogamie peut être stricte ou totale, ou bien seulement partielle.

4.4.1 L'homogamie génotypique totale

Il s'agit du cas simple où, pour un gène donné, chaque génotype ne se croise qu'avec un génotype identique. On peut, dans le cas simple d'un gène di-allélique, établir les relations de récurrence sur les fréquences génotypiques. À la génération *i*, la composition génotypique de la population est la suivante :

Génotypes	*A1/A1*	*A1/A2*	*A2/A2*
Fréquences	D_i	H_i	R_i

Chacun des génotypes ne se croisant qu'avec un génotype identique, les génotypes *A1/A1* se croisent entre eux et ne donnent que des descendants *A1A1* ; de même les *A2A2* ne donnent que des descendants *A2A2*. Quant aux *A1A2*, croisés entre eux, ils donnent 1/4 de *A1A1*, 1/4 de *A2A2* et 1/2 de *A1A2*. D'où les relations suivantes, entre les générations *i* et *i* + 1 :

$$\begin{aligned} D_{i+1} &= D_i + H_i/4 \\ H_{i+1} &= H_i/2 \\ R_{i+1} &= R_i + H_i/4 \end{aligned}$$

Ainsi, de générations en générations, la fréquence des homozygotes va croître à mesure que celle des hétérozygotes diminue (de moitié à chaque génération).

Formellement les relations de récurrence sont les mêmes que pour l'autofécondation et, à la limite, on obtient des lignées pures, mais pour le gène *A* uniquement, puisque le critère de choix ne porte que sur les génotypes pour ce gène ; dans le cas de l'autofécondation, l'homozygotie porte sur tout le génome.

On remarquera que les fréquences alléliques, comme dans le cas de l'autofécondation, restent inchangées, en effet :

$$f(A1)_i = D_i + H_i/2$$

et $$f(A1)_{i+1} = D_{i+1} + H_{i+1}/2$$

soit $$f(A1)_i = [D_i + H_i/4] + H_i/4 = D_i + H_i/2 = f(A1)i$$

Ce modèle simple n'a qu'un intérêt théorique car dans la réalité l'homogamie est le plus souvent partielle et phénotypique, sauf peut-être pour des gènes d'incompatibilité de croisement.

4.4.2 L'homogamie génotypique partielle

Il existe, dans ce cas, une fraction λ des individus se croisant de manière homogame et une fraction $(1 - \lambda)$ se croisant de manière panmictique.

Les génotypes *A1A1* de la génération $i + 1$ seront, pour une part λ, issus de génotypes *A1A1* de la génération i, croisés par homogamie, et résulteront, pour l'autre part $(1 - \lambda)$, d'unions panmictiques, avec une probabilité égale à la probabilité de tirages, dans l'urne gamétique de deux gamètes *A1*, soit p^2. Un même raisonnement, pour les deux autres génotypes conduira aux équations de récurrence suivantes :

$$\boldsymbol{D_{i+1} = \lambda\,[D_i + H_i/4] + (1 - \lambda)\,p^2}$$
$$\boldsymbol{H_{i+1} = \lambda\,[H_i/2] + (1 - \lambda)\,2pq}$$
$$\boldsymbol{R_{i+1} = \lambda\,[R_i + H_i/4] + (1 - \lambda)\,q^2}$$

Vers quelle situation d'équilibre va évoluer la composition génétique de cette population ?

On vérifie aisément, comme dans le cas précédent, que les fréquences alléliques restent inchangées et on se doute que l'évolution ne peut conduire à des lignées pures pour le gène *A* puisqu'il y a toujours un peu d'hétérogamie pour générer des hétérozygotes.

Si H_e est la fréquence des hétérozygotes à l'équilibre, on doit avoir la relation suivante :

$$H_e = \lambda\,[H_e/2] + (1 - \lambda)\,2pq$$

d'où on tire que $$\boldsymbol{H_e = 4pq(1 - \lambda)/(2 - \lambda)}$$

Ici encore le résultat est semblable à celui qu'on obtient pour une autofécondation partielle, à la différence qu'il concerne un seul gène et non tout le génome.

> ***Remarque :*** la fréquence H des hétérozygotes devient égale à $2pq$ si la population est panmictique, c'est-à-dire si $\lambda = 0$. Dès qu'il y a un peu d'homogamie, la fréquence des hétérozygotes devient inférieure à ce qu'elle serait en régime panmictique, puisque $2(1 - \lambda)/(2 - \lambda) < 1$

On a donc, en régime homogame, **pour le gène concerné**, ce qu'on observe, en régime consanguin, pour tout le génome, un excès d'homozygote et un déficit d'hétérozygote. Sachant qu'à l'équilibre $H_e = 4pq(1 - \lambda)/(2 - \lambda)$, on peut aussi l'écrire sous la forme (voir 4.2.3.b) :

$$H_e = 2pq - [\lambda/(2 - \lambda)]\,2pq$$

Ce qui permet d'écrire, à l'équilibre :

$$D_e = p^2 + [\lambda/(2 - \lambda)]\, pq$$
$$H_e = 2pq - [\lambda/(2 - \lambda)]\, 2pq$$
$$R_e = q^2 + [\lambda/(2 - \lambda)]\, pq$$

On remarque d'ailleurs que pour $\lambda = 1$, on obtient bien $H = 0$, $D = p$ et $R = q$.

4.4.3 L'homogamie phénotypique

On peut rappeler que, pour des phénotypes codominants, l'homogamie phénotypique est aussi une homogamie génotypique.

Quand l'homogamie porte sur des phénotypes récessifs et dominants, elle est encore une homogamie génotypique pour les couples de phénotypes récessifs, mais ne l'est plus pour les couples de phénotypes dominants qui peuvent être de génotypes homozygotes et/ou hétérozygotes. Le traitement mathématique est un peu plus compliqué, mais il est logique de s'attendre à un résultat de même nature que pour l'homogamie génotypique : la fréquence des hétérozygotes diminuera.

Dans le cas d'une homogamie phénotypique totale, la fréquence des hétérozygotes tendra vers zéro, comme pour l'homogamie génotypique totale, mais plus lentement. Dans le cas d'une homogamie phénotypique partielle, la fréquence des hétérozygotes tendra, comme pour l'homogamie génotypique partielle, vers une limite H_e. Cette limite sera non seulement atteinte plus lentement, mais elle aura aussi une valeur différente.

4.4.4 Homogamie et maintien du polymorphisme

Il convient de rappeler que, dans les divers cas d'homogamie, le choix du conjoint ne porte que sur un ou quelques gènes. Par conséquent, *l'effet de l'homogamie ne concerne que le (ou les) gène(s) qui gouverne(nt) le caractère sur lequel porte le choix du conjoint*.

Dans le cas d'homogamie partielle, l'excès d'homozygotes ne portera que sur ce (ou ces) seul(s) gène(s). Dans le cas limite de l'homogamie totale, les individus deviendront homozygotes pour ce (ou ces) seul(s) gène(s). Pour les autres gènes, les croisements sont toujours panmictiques, la diversité génétique, au niveau génotypique, est toujours maintenue dans les proportions de Hardy-Weinberg.

> ***Remarque :*** pour les gènes dans le voisinage du gène impliqué dans l'homogamie, un déséquilibre gamétique peut être instauré, que le temps fera disparaître, plus ou moins vite, au rythme des crossing over entre leur locus respectifs (voir chapitre 3).

Par contre, dans les unions entre apparentés, l'effet de la consanguinité porte sur l'ensemble du génome. La consanguinité induira un écart positif à la panmixie chez les homozygotes et un écart négatif chez les hétérozygotes pour tous les gènes. Dans le cas limite des croisements consanguins systématiques, la consanguinité conduira

à des lignées pures homozygotes pour tous les gènes, mais à des lignées pures différentes entre elles selon les allèles fixés pour les différents gènes.

À grande échelle, c'est-à-dire au sein de très grandes populations, la consanguinité, malgré la réduction de l'hétérozygotie, peut demeurer compatible avec un certain maintien de la diversité allélique, mais en pratique, c'est un facteur souvent important de perte allélique (voir chapitre 5).

L'homogamie est beaucoup moins un obstacle au maintien de la diversité allélique. Même quand elle est totale, elle ne concerne qu'un seul gène ou quelques-uns, et quand elle est partielle, elle ne conduit qu'à un excès d'homozygotes et un déficit d'hétérozygotes que pour cette petite fraction du génome, la population gardant pour le reste du génome, à la fois sa diversité allélique et sa diversité génotypique (hétérozygotes).

RÉSUMÉ

Le choix du conjoint en fonction de la parenté conduit à une descendance consanguine.

La parenté de deux individus *K* et *L*, relativement à un ancêtre commun *A*, s'écrit :

$$\phi_{k,l} = (1/2)^{i+j}\,(1/2 + f_A/2)$$

où *i* et *j* sont les nombres de générations entre *K* et *A* et entre *L* et *A*, et f_A, le coefficient éventuel de consanguinité de *A*. La consanguinité d'un individu est définie comme égale à la parenté de ces parents.

Dans une grande population où une fraction d'unions, non panmictiques, survient entre apparentés, on peut définir un coefficient moyen de parenté *F*, égal au coefficient moyen de consanguinité, à la génération suivante.

Dans ces conditions, la composition génétique de la population est modifiée de manière telle que, pour un gène di-allélique, les fréquences génotypiques sont égales à :

$$f(A1A1) = p^2 + Fp(1-p)$$
$$f(A1A2) = 2pq - 2Fpq$$
$$f(A2A2) = q^2 + Fq(1-q)$$

Le choix du conjoint en fonction de la parenté ne modifie pas les fréquences alléliques mais génère seulement un excès d'homozygotes et un déficit d'hétérozygotes par rapport aux fréquences caractéristiques de la relation de Hardy-Weinberg, fondée sur la panmixie.

Cet écart à la panmixie est responsable d'un accroissement de risque pathologique, pour les maladies génétiques récessives, dans la descendance des couples apparentés. Cet accroissement relativement limité à l'échelle d'un couple est très prononcé parce qu'additif, à l'échelle d'une population.

Les croisements systématiques (autofécondation totale ou partielle) et les croisements homogames (choix du conjoint en fonction de la similitude génotypique ou

phénotypique) conduisent également à des écarts à la panmixie, formalisés à l'équilibre, pour un gène di-allélique, par des équations de même nature :

$$D_e = p^2 + [\lambda/(2 - \lambda)]\, p(1 - p)$$
$$H_e = 2pq - [\lambda/(2 - \lambda)]\, 2pq$$
$$R_e = q^2 + [\lambda/(2 - \lambda)]\, q(1 - q)$$

où λ est un paramètre exprimant le degré d'autofécondation ou d'homogamie. Quand λ est égal à 0 (absence d'autofécondation ou d'homogamie), on retrouve les équations de Hardy-Weinberg. Quand λ est égal à 1 (autofécondation ou homogamie totale), la fréquence des hétérozygotes est nulle à l'équilibre, au profit de souches pures, sans changement des fréquences alléliques.

Le phénomène d'écart à la panmixie, et l'évolution éventuelle vers des lignées pures, concerne tout le génome quand il s'agit d'autofécondation, mais ne concerne que les seuls gènes gouvernant les caractères sur lesquels porte le choix du conjoint, quand il s'agit d'homogamie.

EXERCICES

Exercice 4.1

Question 1 : quel est le taux de parenté d'un couple de doubles cousins germains ?

Question 2 : quel sera le taux de consanguinité de l'enfant d'un tel couple ?

Question 3 : quel sera le risque, pour cet enfant, d'être atteint de β-thalassémie, sachant que la fréquence de la mutation du gène β, responsable de cette hémoglobinopathie, est égale à 2 %, dans cette population.

Question 4 : Vous discuterez de la différence d'appréciation entre le risque individuel pour un tel couple et le risque collectif de santé publique dans cette population, sachant que 8 % des mariages concernent des doubles cousins germains et que 16 % des mariages concernent de simples cousins germains.

Solution

Question 1 : 1/8

Question 2 : 1/8

Question 3 : $q^2 + fq(1 - q)$, soit si q est petit, $q^2 + fq$

Avec $q = 2/100$ et $f = 1/8$, $q^2 + fq = 4/10\,000 + 1/400$ qui peut s'écrire $q^2 + fq = 4/10\,000 + 25/10\,000$

Le risque panmictique de 4/10 000 est donc accru de 25/10 000, il est multiplié par 7,25 !

Question 4 : un risque de 29/10 000 reste encore faible à l'échelle individuelle, mais entraîne un surplus de pathologie considérable à l'échelle collective.

En effet, avec 8 % de mariages entre doubles cousins germains et 16 % de mariage entre simples cousins germains, le taux moyen F de parenté, c'est-à-dire aussi le taux moyen de consanguinité, est égal à $F = 0{,}08 \times 1/8 + 0{,}16 \times 1/16 = 0{,}02$, soit 2 %.

À l'échelle collective, on attend une fréquence d'enfants atteints égale à $q^2 + Fq$ = 4/10 000 + 4/10 000.

Soit le doublement du nombre de naissances d'enfants atteints.

Par ailleurs, ce qui vaut pour cette maladie vaut pour les autres, et le surplus de pathologie peut devenir considérable, comme l'illustre le tableau 4.8.

Exercice 4.2 : Parenté, consanguinité et calcul des fréquences alléliques

On étudie, dans une grande population, la fréquence des mariages entre apparentés. Les résultats suivants sont observés.

Mariages entre					
Cousins germains	**Cousins inégaux**	**Cousins issus de cousins germains**	**Doubles cousins germains**	**Oncle et nièce**	**Tante et neveu**
16 %	32 %	6,4 %	0,8 %	1,6 %	0,8 %

Question 1 : quel est le taux de parenté d'un couple pris au hasard dans une telle population ?

Question 2 : quel sera le taux de consanguinité de l'enfant d'un tel couple ?

Question 3 : quel sera le risque, pour cet enfant, d'être atteint de β-thalassémie, sachant que la fréquence q de la mutation du gène β, responsable de cette hémoglobinopathie, est égale à 2 %, dans cette population ?

Question 4 : quel sera le risque, pour cet enfant, d'être atteint d'une maladie neurodégénérative dominante, dont l'allèle pathogène responsable a une fréquence p = 1/1 000 ?

Question 5 : une étude portant sur une myopathie (maladie musculaire) génétique récessive non liée au sexe, a montré une fréquence d'un enfant atteint sur 4 103 naissances. Quelle est la fréquence de la mutation responsable de cette myopathie ?

Solution

Question 1 : le paramètre F est la moyenne des ϕ ; soit :

$F = 0{,}16 \times 1/16 + 0{,}032 \times 1/32 + 0{,}064 \times 1/64 + 0{,}008 \times 1/8 + 0{,}016 \times 1/8 + 0{,}08 \times 1/8$

F = 0,025, soit 2,5 %.

Question 2 : le taux de consanguinité d'un enfant pris au hasard sera égal au taux de parenté d'un couple pris au hasard, soit le taux moyen F = 2,5 %.

Question 3 : $q^2 + Fq(1 - q)$, soit si q est petit, $q^2 + Fq$

Avec q = 2/100 et f = 1/8, $q^2 + fq$ = 4/10 000 + 1/2 000

qui peut s'écrire $q^2 + fq$ = 4/10 000 + 5/10 000

Le risque panmictique de 4/10 000 est donc accru de 5/10 000, il est plus que doublé !

Question 4 : la maladie étant dominante, la fréquence de celle-ci est la somme des fréquences des génotypes homozygotes et hétérozygotes, soit :

$p^2 + Fp(1 - p) + 2pq - 2Fpq$, soit si p est petit,

$p^2 + Fp + 2p - 2Fp = p^2 + 2p - Fp$

La consanguinité, en réduisant la fréquence des hétérozygotes, réduit la fréquence des maladies génétiques dominantes. Avec q = 1/1 000 et F = 0,025, on a :

$p^2 + 2p - Fp$ = 1/1 000 000 + 2/1 000 – 1/40 000;

qui peut s'écrire $p^2 + 2p - Fp$ = 1/1 000 000 + 2 000/1 000 000 – 25/1 000 000

Le risque panmictique est diminué de 25/1 000 000, mais il faut bien réaliser que ce risque panmictique est essentiellement dû à la présence des hétérozygotes (2000 fois plus nombreux que les homozygotes) où se manifeste l'effet dominant de la mutation pathogène.

De ce fait, la diminution du risque en raison de la consanguinité est minime en pratique.

Question 5 : il est impossible d'estimer la fréquence de la mutation responsable de la myopathie en prenant la racine carrée de 1/4103, puisque la population n'est pas panmictique.

Par contre, on connaît F, et on peut écrire que :

f(enfants atteints) = $q^2 + Fq(1 - q)$,

si q est la fréquence recherchée de l'allèle pathogène.

D'où 1/4 103 = $q^2 + Fq(1 - q)$ dont on tire aisément que : ***q* = 0,0075**

NB : ne sachant pas que la population était consanguine et calculant la fréquence sous l'hypothèse panmictique, on l'aurait estimée par la racine carrée de la fréquence des enfants atteints, soit avec une valeur de 0,016, deux fois plus élevée qu'elle n'est en réalité, puisque de nombreuses naissances constituent le surplus dû à la consanguinité.

Exercice 4.3

Question 1 : on considère une population dans laquelle une maladie génétique autosomique et récessive, létale dans l'enfance, touche un nouveau-né sur 4 900.

a) Quelle est la fréquence f des porteurs sains ? On demande une réponse argumentée et justifiée.

b) Quelle est la probabilité qu'un couple ait un enfant atteint, sachant que l'un des membres de ce couple a une sœur ayant un enfant atteint ?

c) Que devient cette probabilité quand une analyse moléculaire atteste que cette personne est, comme sa sœur, porteuse saine ?

d) Quel serait le risque, pour un couple de cousins germains, dans la population générale, d'avoir un enfant atteint ?

Question 2 : apprenant que la fréquence des mariages entre cousins germains, dans la population précédente, est en fait assez élevée, de l'ordre de 32 %, en quoi cela change-t-il vos estimations précédentes, pour les questions **a**) et **b**) ?

Question 3 : voir chapitre 7

Question 4 : lors d'une consultation de conseil génétique, un couple vous apprend plusieurs choses :

– qu'ils sont l'un pour l'autre oncle et nièce ;

– que leur ancêtre commun féminin (mère de l'un et grand-mère de l'autre) est-elle même issue d'un mariage entre cousins germains ;

– que leur ancêtre commun masculin (père de l'un et grand père de l'autre) est lui même issu d'un mariage oncle-nièce (les deux ancêtres communs n'étant pas apparentés).

Établir la généalogie et calculer la consanguinité de l'enfant que ce couple envisage de concevoir ? Commentez votre résultat.

Solution

Question 1 :

a) Sous l'hypothèse du modèle de Hardy-Weinberg, sachant que l'effet de la sélection est négligeable sur quelques générations, on peut écrire que la fréquence de la maladie est égale à q^2 où q est la fréquence de l'allèle pathologique.

De l'équation $q^2 = 1/4900$, on peut tirer que $q = 1/70$ (1,49 %)

La fréquence f des porteurs sains est égale à $2q(1 - q) = 2 \times 1/70 \times 69/70 = 2{,}82$ %,

Valeur proche de $2q = 1/35$.

b La probabilité que le membre du couple, dont la sœur a eu un enfant atteint, soit porteur sain, est égale à 1/2 ; la probabilité que l'autre membre du couple, venant de la population générale, soit porteur sain est égale à 1/35 et la probabilité, dans ces conditions, qu'ils aient un enfant atteint est égale à 1/4. Le risque final, pour un tel couple, d'avoir un enfant atteint est égal à $1/2 \times 1/35 \times 1/4$, soit 1/280 (0,35 %).

c) Sachant que le membre du couple, dont la sœur a eu un enfant atteint, est porteur sain, la probabilité qu'un tel couple ait un enfant atteint devient égale à $1 \times 1/35 \times 1/4$, soit 1/140 (0,70 %).

d) Cette probabilité est égale au risque panmictique, soit q^2, augmentée du supplément de risque dû à la parenté des parents, soit $\phi.q\,(1 - q)$, où ϕ est le taux de parenté entre cousins germains, soit 1/16.

On obtient donc la valeur du risque $q^2 + 1/16 \times q\,(1 - q) = 1/4\,900 + 4{,}3/4\,900 = 0{,}108$ %

Le risque est multiplié par 5.

Question 2 :

Dans ce cas, on ne peut plus considérer la population comme panmictique et la fréquence des enfants atteints n'est pas égale à q^2 mais à $q^2 + F.q\,(1 - q)$ où F est le coefficient moyen de consanguinité.

Dans le cas présent, $F = 0{,}32 \times 1/16 = 2$ %.

Ce qui conduit à une valeur de $q = 0{,}00746$, soit une valeur deux fois plus faible que la valeur estimée sous l'hypothèse inexacte du modèle de Hardy-Weinberg (1/70 = 0,0142).

Les porteurs sains ont alors pour fréquence $f = 2q(1 - q) - 2Fq(1 - q) = 0{,}0145 =$ 1,45 %, soit là encore une réalité à peu près la moitié de l'estimation faite sous l'hypothèse inexacte de H-W.

De ce fait, les autres risques doivent aussi être divisés par deux, si on suppose que le conjoint du couple n'est pas apparenté à celui dont la sœur a eu un enfant atteint.

Question 3 : voir chapitre 7

Question 4 :

La consanguinité de l'enfant à naître est égale à la parenté du couple qui dépend de ses deux ancêtres communs, eux mêmes consanguins, avec un taux $\phi = 1/16$ pour l'ancêtre féminin et $\phi = 1/8$ pour l'ancêtre masculin.

Par rapport au premier ancêtre, la parenté est égale à $(1/2)^3 \times (1/2 + 1/2 \times 1/16)$

Par rapport au deuxième ancêtre, la parenté est égale à $(1/2)^3 \times (1/2 + 1/2 \times 1/8)$

Si on effectue la somme, on obtient la parenté du couple venu en consultation,

Soit $\phi = 1/8 + 1/16 \times 1/16 + 1/16 \times 1/8$, soit en réduisant au même dénominateur

$\phi = 32/256 + 1/256 + 2/256 = 35/256 = 0{,}1367 = 13{,}7$ %

Commentaires :

Si les deux ancêtres n'étaient pas consanguins, la valeur de f serait simplement égale à 1/8 soit 0,125 = 12,5 %.

Le fait qu'ils soient consanguins, avec une valeur relativement élevée, ajoute 3/256 = 0,012 = 1,2 %.

Si on avait négligé ce supplément, on aurait sous-estimé la parenté en faisant une erreur relative de 8,7 % (1,2/13,7).

Exercice 4.4

On étudie le groupe MN dans une grande population dans laquelle on sait qu'une fraction importante des couples est apparentée ; les résultats sont donnés dans le tableau ci-dessous :

Phénotypes	[M]	[MN]	[N]
Génotypes	*M/M*	*M/N*	*N/N*
Effectifs observés	4 950	4 100	950

Question 1 : estimez les fréquences alléliques et testez le modèle de Hardy-Weinberg en calculant les effectifs théoriques sous cette hypothèse et en testant la signification statistique des écarts. Commentez votre résultat.

Question 2 : utilisez l'effectif des hétérozygotes pour estimer une valeur du coefficient F. Commentez brièvement sachant qu'une population avec 16 % de mariages entre cousins germains admet une consanguinité moyenne de 1 %.

Question 3 : peut-on tester la conformité de la composition génétique avec la valeur estimée de *F* ?

Question 4 : une maladie autosomique récessive présente, dans cette population, une fréquence de 3 enfants atteints pour 10 000 naissances, quelle est la fréquence de l'allèle pathologique et celle des porteurs sains ? Comparez vos résultats à ceux que vous auriez obtenus si vous aviez négligé le contexte de parenté attaché à de nombreux couples de cette population et que vous aviez supposé, à tort, la panmixie.

Solution

Question 1.a) : Les fréquences alléliques peuvent être directement estimées par la formule $D + H/2$ puisque les phénotypes sont codominants, soit : $f(M) = p = 0{,}7$ et $f(N) = q = 0{,}3$.

Question 1.b) : les effectifs, sous l'hypothèse de H-W sont respectivement égaux à 4 900, 4 200 et 900.

Le test statistique de la signification des écarts donne une valeur observée du χ^2 égale à 5,67, ce qui est significatif pour un test avec 1 ddl (3 classes moins 1 pour la taille de l'échantillon théorique moins 1 pour l'estimation d'une fréquence allélique). On peut donc se permettre, avec un risque d'erreur inférieur à 5 %, de rejeter l'hypothèse du modèle de HW ; ce qui n'est pas surprenant puisqu'on observe excès d'homozygote et déficit d'hétérozygotes conformes à ce qu'on peut attendre dans une population où il y a une fraction importante d'unions entre apparentés.

Question 2 : on pose que la fréquence des hétérozygotes est égale à la fréquence attendue sous H-W moins l'écart à la panmixie, soit :

$f(\mathrm{MN}) = 2pq - 2\mathrm{F}pq$

$f(\mathrm{MN}) = 2p(1-p) - 2Fp(1-p)$

D'où on tire que $F = [f(MN) - 2p(1-p)]/2p(1-p)$

Avec $p = 0{,}7$ et $f(\mathrm{MN}) = 0{,}41$, on estime $\boldsymbol{F = 0{,}0238}$.

Ce qui correspond à peu près à 40 % de mariages entre cousins germains.

Question 3 : on a épuisé l'information en estimant une fréquence allélique et F ; en effet on a trois classes moins 1 pour la taille de l'échantillon, moins 2 pour les paramètres estimés, il ne reste aucun degré de liberté.

Question 4 : comme on sait qu'il y a une consanguinité moyenne égale à $F = 0{,}0238$, il est nécessaire de considérer que la fréquence des enfants atteints n'est pas égale seulement à q^2, comme si la population était panmictique, mais à q^2 augmentée du supplément résultant de la consanguinité, soit :

$\boldsymbol{f(m/m) = q^2 + Fq(1-q)}$ avec $f(m/m) = 0{,}0003$ et $F = 0{,}0238$

D'où on tire que **q = 0,91 %**

La fréquence des porteurs sains, hétérozygotes est égale à

F(N/m) $= 2q(1-q) - 2Fq(1-q) =$ **1,76 %**

NB : si on avait négligé la consanguinité, on aurait estimé q en prenant la racine carrée de la fréquence des atteints, soit 1,7 %, et celle des porteurs sains de 3,4 %, un biais de près de 100 % !!!

Exercice 4.5

Question 1 : vous voyez en consultation un couple de cousins germains, nommés *K* et *L*, issus d'une grande population. Quelle est la proportion de gènes pour laquelle leur futur enfant sera porteur de deux exemplaires identiques par ascendance ? Faire un seul schéma généalogique utile pour tout l'exercice.

Question 2 : quelle est la proportion de gènes pour laquelle leur futur enfant sera porteur de deux exemplaires identiques par ascendance **à un gène du grand père** ?

Question 3 : sachant que le grand-père ancêtre des deux cousins germains *K* et *L* était lui même enfant de cousins germains, quel sera le coefficient de consanguinité du futur enfant du couple *K-L* ?

Question 4 : commentez ce résultat, sachant qu'on est dans une grande population ; votre commentaire serait-il identique si le couple était issu d'une petite population ? Justifiez votre réponse.

Question 5 : enfin le couple vous apprend que leur grand-père et leur grand-mère communs étaient également eux-mêmes cousins germains ; quel sera le coefficient de consanguinité d'un futur enfant du couple *K-L* ? Commentez votre réponse concernant les contributions respectives à ce coefficient de consanguinité de la parenté des ancêtres ou de leur consanguinité.

NB : pour cette question 5, il convient de tenir compte de la possibilité que *K* et *L* puissent partager des exemplaires identiques par ascendance même si ils ne viennent pas d'un même ancêtre puisque ceux ci sont apparentés.

Solution

Question 1 : le calcul de la parenté entre *K* et *L* donne la valeur de 1/16.

En effet, $i = 2$ et $j = 2$, on prend, en l'absence de toute information sur *GP* et *GM*, les deux grands-parents communs, une valeur de f égale à 0 pour tous deux, d'où une parenté de 1/32 entre *K* et *L*, vis-à-vis de *GP,* et symétriquement de 1/32 vis-à-vis de *GM*.

Question 2 : la réponse est évidemment 1/32.

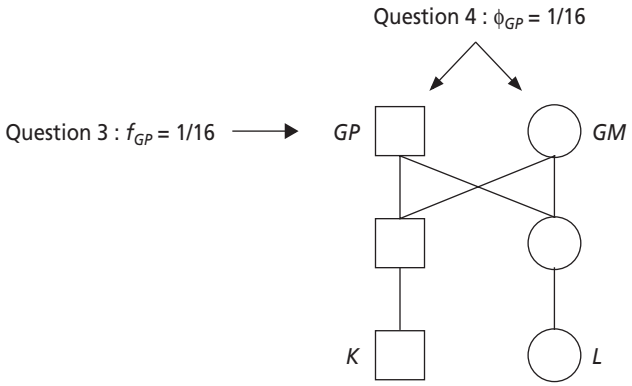
Question 3 : dans ce cas, il faut tenir compte de la valeur de $f_{GP} = 1/16$

La valeur devient $\phi_{KL} = 1/16 + 1/32 \times 16$

Soit, si on réduit au même dénominateur : $\phi_{KL} = (32 + 1)/512$

Question 4 : la prise en compte de la consanguinité du grand-père, qui est l'un des plus fort taux possibles dans une grande population, ne modifie pas tellement le coefficient de parenté des cousins (1/512 à ajouter à 32/512) ; le fait de négliger la consanguinité antérieurement accumulée n'est donc pas trop préjudiciable… dans une grande population où il n'y a pas de consanguinité cumulative par dérive génétique !

Au contraire, dans une petite population, les ancêtres communs peuvent avoir un fort taux de consanguinité cumulée et, de plus, être eux mêmes apparentés, ce qui va encore augmenter la parenté entre les cousins *K* et *L* (voir question suivante).



Question 5 : il faut alors ajouter à la valeur précédemment trouvée, 33/512, les valeurs de deux probabilités :

– probabilité qu'un allèle tiré chez *K* et venant de *GP* soit identique par ascendance à un allèle tiré chez *L* et venant de *GM*, soit $1/4 \times 1/4 \times 1/16$;

– probabilité qu'un allèle tiré chez *K* et venant de *GM* soit identique par ascendance à un allèle tiré chez *L* et venant de *GP*, soit $1/4 \times 1/4 \times 1/16$.

La somme de ces deux probabilités est égale à 1/128 soit 4/512, d'où :

$$\phi_{KL} = (32 + 1 + 4)/512$$

On voit alors que la contribution de la parenté entre ancêtres est plus importante que leur consanguinité respective.

Exercice 4.6

L'étude de la composition génétique pour un gène bi-allélique permet de définir et d'estimer les fréquences génotypiques, selon la structure suivante :

Génotypes	*A*/*A*	*A*/*a*	*a*/*a*
Fréquences observées	*D*	*H*	*R*

Avec les fréquences alléliques $f(A) = p = D + H/2$

$f(a) = q = R + H/2$

Question 1 : on teste la conformité au modèle de Hardy-Weinberg par un test de χ^2.

Montrez que le χ^2 peut s'écrire sous la forme $\chi^2 = N.F^2$, où *N* est la taille de l'échantillon et *F* est l'indice de fixation (voir 4.3.2), soit $F = 1 - H/2pq$.

Il convient de se rappeler qu'une fréquence observée peut s'écrire sous la forme de la somme de la fréquence panmictique théorique et d'un écart à la panmixie.

Question 2 : utilisez cette relation pour déduire la taille de l'échantillon nécessaire pour mettre en évidence par ce test statistique, au risque de 5 %, un écart à la panmixie induit par un taux de consanguinité égal à $F = 1$ %, ou $F = 10$ %.

Solution

Question 1 : Le test consiste à calculer la valeur observée du χ^2 en calculant la somme des carrés des écarts entre effectifs observés et attendus, rapportés à ces derniers, soit

$\chi^2 = [N\,(D - p^2)]^2/Np^2 + [N(H - 2pq)]^2/N2pq + [N\,(R - q^2)]^2/Nq^2$

d'où, en simplifiant par N :

$\chi^2 = N\,[(D - p^2)^2/p^2 + (H - 2pq)^2/2pq + (R - q^2)^2/q^2]$

ce qui montre, au passage, que la valeur du χ^2 est égale à celle obtenue sur les écarts de fréquence multipliée par la taille de l'échantillon, raison pour laquelle on précise qu'en pratique on réalise toujours un test sur des effectifs et non sur des fréquences.

Or, on peut écrire les fréquences observées D, H et R sous la forme :

$D = p^2 + Fpq$ d'où $D - p^2 = Fpq$

$H = 2pq - 2Fpq$ $H - 2pq = -2Fpq$

$R = q^2 + Fpq$ $R - q^2 = Fpq$

De sorte que tester les écarts respectifs entre D, H et R et p^2, $2pq$ et q^2 revient à tester les écarts entre Fpq ou $-2Fpq$ et zéro, en remplaçant les écarts $(D - p^2)$, $(H - 2pq)$, $(R - q^2)$ dans l'équation du χ^2 par Fpq ou $-2Fpq$, soit :

$\chi^2 = N\,[(Fpq)^2/p^2 + (-2Fpq)^2/2pq + (Fpq)^2/q^2]$

d'où $\chi^2 = N.F^2\,[(pq)^2/p^2 + (-2pq)^2/2pq + (pq)^2/q^2]$

soit $\chi^2 = N.F^2\,[q^2 + 2pq + p^2]$

donc $\boldsymbol{\chi^2 = N.F^2}$

Question 2 : au risque de 5 %, le rejet de l'hypothèse panmictique est acquis pour les valeurs observées du χ^2 supérieures à 3,84.

Il faut donc que $\chi^2 = NF^2 > 3,84$

soit $N > 3,84/F^2$

Si $F = 0,01$, il faut un échantillon minimal de 38 400 individus pour pouvoir conclure à des écarts significatifs, et rejeter la panmixie, un échantillon de 384 suffit si $F = 0,1$.

Exercice 4.7

On étudie la diversité génétique dans une espèce végétale pour un gène dont les deux allèles *A* et *a* gouvernent l'aspect lisse ou velouté du fruit, le phénotype lisse étant récessif. Plusieurs études ont permis de montrer que cette espèce présente 60 % d'autogamie.

Question 1 : sur un échantillon de 1 000 plants, 180 présentent des fruits lisses ; quelles sont les valeurs des fréquences alléliques ?

Question 2 : on étudie la diversité génétique pour un gène dont les deux allèles *R* et *B* gouvernent la couleur de la fleur, les homozygotes *R/R* étant rouges, les hétérozygotes *R/B* étant violets et les homozygotes *B/B* étant bleus. On dénombre respectivement 575 plants à fleurs rouges, 250 à fleurs violettes et 175 à fleurs bleues. Calculez les fréquences alléliques et voyez si les effectifs observés sont conformes aux effectifs attendus dans cette espèce.

Solution

Question 1 :

Phénotypes	**Phénotype velouté, dominant**		**Phénotype lisse, récessif**
Effectifs observés	**820**		**180**
Génotypes	A//A	A//a	a//a
Fréquences sous l'hypothèse d'un écart à la panmixie par autogamie partielle (λ = 0,6)	$p^2 + pq\ \lambda/(2-\lambda)$	$2pq - 2pq\ \lambda/(2-\lambda)$	$q^2 + q(1-q)\ \lambda/(2-\lambda)$
Calcul de la valeur de q	sous l'hypothèse précédente, avec la valeur de λ = 0,6, on peut écrire $q^2 + q(1-q)\ \lambda/(2-\lambda) = 180/1\,000$, d'où on tire que $q = 0,3$ et $p = 0,7$ on ne peut vérifier la validité du modèle puisqu'on a épuisé l'information en estimant une fréquence allélique		

Question 2 :

Phénotypes	**Phénotype [fleur rouge]**	**Phénotype [fleur violette]**	**Phénotype [fleur bleue]**
Effectifs observés	575	250	175
Génotypes	R//R	R//B	B//B
Fréquences alléliques	f(R) = p = f(R//R) + f(R//B)/2 = 0,7 f(B) = q = f(B//B) + f(R//B)/2 = 0,3		
Fréquences sous l'hypothèse d'un écart à la panmixie par autogamie partielle et valeurs avec λ = 0,6	$p^2 + pq\ \lambda/(2-\lambda)$ 0,575	$2pq - 2pq\ \lambda/(2-\lambda)$ 0,250	$q^2 + q(1-q)\ \lambda/(2-\lambda)$ 0,175
Effectifs attendus sur un échantillon de 1 000	580	240	180
Test statistique de validation du modèle	On peut vérifier la validité du modèle car on a estimé une fréquence allélique, le paramètre λ nous est donné, il fait partie du modèle et n'est pas estimé à partir de l'échantillon ; il reste un degré de liberté. La valeur observée du χ^2 est égale à 0,598 et est très inférieur au seuil du risque à 5 % (3,87) ce qui ne permet pas de rejeter l'hypothèse et conduit à accepter la validité du modèle.		

Exercice 4.8

La fièvre méditerranéenne familiale (FMF), ou fièvre périodique, est une maladie autosomique récessive très fréquente dans les populations du Proche-Orient, arabes ou non arabes (turque, perse, arménienne, juive sépharade) ; elle se caractérise par des épisodes fébriles récurrents accompagnés de douleurs abdominales et, dans chez une proportion variable de patients, de complications sévères aboutissant à la perte de la fonction rénale par amylose des tissus. Les effets de cette maladie, surtout les risques d'amylose, peuvent être prévenus par la prise régulière de colchicine.

Question 1 : une étude épidémiologique d'une population du Proche-Orient, a montré que la FMF y était la maladie récessive la plus fréquente avec un cas atteint sur 900 individus.

Quels sont les génotypes possibles dans la population, si un allèle normal fonctionnel et un allèle pathologique sont respectivement désignés par *N* et *m* ?

Quelles sont les fréquences de ces génotypes si on désigne par *p* et *q* les fréquences respectives de *N* et *m* ? Précisez quelles sont les conditions que vous supposez valides pour établir ces fréquences.

Application numérique : déduisez la valeur de *q* et la fréquence des porteurs sains.

Question 2 : le gène impliqué dans la FMF a été localisé sur le chromosome 16 puis identifié et séquencé ; les mutations pathologiques les plus courantes ont été aujourd'hui identifiées mais certaines, plus rares, ne le sont pas encore. Nous désignerons dans la suite du problème par *m1*, l'allèle pathologique le plus fréquemment rencontré, et par *m2*, l'ensemble des autres allèles pathologiques, identifiés ou inconnus, et par *q1* et *q2*, leurs fréquences respectives dans la population (ce qui signifie que $q = q1 + q2$).

Quels sont les différents génotypes possibles dans la population et, **sous les conditions définies plus haut**, leurs fréquences en fonction de *p*, *q1* et *q2* ?

Que remarquez-vous concernant les proportions entre les différents génotypes des malades, quand on considère que $q = q1 + q2$? Quelle est la condition, parmi celles définies plus haut, qui entraîne cette observation ?

Question 3 : en s'intéressant à présent au seul échantillon de 250 malades dont l'ADN a été étudié, on observe les résultats suivants :

Génotypes	*m1//m1*	*m1//m2*	*m2//m2*	Total	Valeur du χ^2 testant les écarts
Effectifs observés	55	80	115	250	
Effectifs attendus				250	

Calculez, à partir de ces observations, les fréquences *r1* et *r2* des allèles *m1* et *m2*, **dans l'échantillon des malades** (à ne pas confondre avec les fréquences *q1* et *q2* dans la population) et déduisez les effectifs attendus des trois génotypes, si la population est conforme à la condition définie précédemment et en fonction de la remarque faite sur les proportions entre les différents génotypes des malades.

Par un test statistique de χ^2, testez les écarts entre effectifs observés et attendus, et concluez sur la validité de cette condition.

Question 4 : on souhaite trouver la cause de ces écarts entre effectifs observés et attendus parmi les génotypes des malades.

Peut-on, dans le principe, faire l'hypothèse que la consanguinité pourrait être à l'origine de ces écarts, sachant que la population étudiée présente un taux élevé de mariages entre apparentés ? (quelques lignes suffisent).

Quelle serait, connaissant les valeurs de *r1* et *r2*, la valeur du paramètre *F* qui pourrait expliquer de tels écarts ? Vous ferez le calcul à partir de la fréquence des hétérozygotes.

Pourquoi la valeur de *F* exclue que les écarts puissent être expliqués par la seule consanguinité ?

NB : cette réponse est confirmée par l'étude de l'échantillon de malades desquels ont été enlevés tous ceux dont les parents étaient apparentés et où les écarts restent significatifs.

Question 5 (indépendante de la question 4) : on remarque alors que l'échantillon étudié n'est pas homogène car issu de deux communautés religieuses « génétiquement différentes » et « isolées » sur le plan sociologique et matrimonial puisque sans échange de conjoints entre elles. Les données ont donc été scindées (tableau ci dessous) ; vous testerez l'hypothèse panmictique dans chaque communauté (même si des mariages entre apparentés surviennent plus ou moins fréquemment en leur sein), et vous montrerez ainsi que les effectifs observés sur l'ensemble ne sont pas significativement différents des effectifs attendus **calculés par la somme des effectifs attendus dans chaque communauté.** Quel est l'effet ainsi illustré ?

	Génotypes	*m1//m1*	*m1//m2*	*m2//m2*	Total	Valeur du χ^2 testant les écarts
Communauté A	**Effectifs observés**	9	46	105	160	
	Effectifs attendus				160	
Communauté B	**Effectifs observés**	46	34	10	90	
	Effectifs attendus				90	
Total des effectifs observés		55	80	115	250	
Total des effectifs attendus					250	

Solution

Question 1 : les génotypes et leurs fréquences, en supposant valides les conditions de Hardy-Weinberg, notamment la panmixie, sont :

Génotypes	*N//N*	*N//m*	*m//m*
Fréquences théoriques	p^2	$2pq$	q^2
Fréquences observées			R = 1/900

De ces équations, on peut tirer que $R = q^2$

d'où la fréquence q des allèles pathologiques : $q = 1/30$

La fréquence des porteurs sains, c'est-à-dire des hétérozygotes est égale à $2pq$, soit peu différente de

$2q = 1/15$

Question 2 : si on suppose le modèle de Hardy-Weinberg, on aura les génotypes suivants avec leurs fréquences respectives :

Génotypes	*N//N*	*N//m1*	*N//m2*	*m1//m1*	*m1//m2*	*m2//m2*
Fréquences théoriques	p^2	$2p.q1$	$2p.q2$	$q1^2$	2q1.q2	$q_2{}^2$

On remarque qu'au sein des malades, les fréquences génotypiques suivent, du fait de la panmixie, la même loi mathématique que pour l'ensemble des génotypes au sein de la population et que leurs fréquences respectives résultent du développement de $q^2 = (q1 + q2)^2 = q1^2 + 2q1.q2 + q_2{}^2$

Question 3 :

Génotypes	*m1//m1*	*m1//m2*	*m2//m2*	total	valeur du χ^2 testant les écarts
Effectifs observés	55	80	115	250	25,74 >> 3,84 (1 ddl)
Effectifs attendus	$r1^2 \times 250 = 36{,}1$	$2r1.r2 \times 250 = 117{,}8$	$r2^2 \times 250 = 96{,}1$	250	

On montre que $r1 = 55/250 + (80/250)/2 = 0{,}38$

$r2 = 1 - r1 = 0{,}62$

Et on tire, sous l'hypothèse de la panmixie, les fréquences théoriques, soit $r1^2$, $2r1.r2$ et $r2^2$.

D'où les effectifs théoriques ou attendus et un χ^2 très significatif (pour 1 ddl).

Question 4 : en principe, sachant qu'on observe un fort excès d'homozygotes et un déficit d'hétérozygotes par rapport aux effectifs attendus sous l'hypothèse panmictique et qu'il y a des mariages entre apparentés dans la population ce qui favorise,

par la consanguinité, de tels écarts, on peut supposer que ces écarts puissent résulter de la consanguinité. Dans ce cas, la fréquence des hétérozygotes s'écrit :

$$H = 2p.q - 2F.p.q \quad \text{d'où} \quad \boldsymbol{F = 1 - H/2p.q}$$

L'application numérique donne une valeur de F **= 0,321** qui est bien trop élevée pour rendre compte de tels écarts puisqu'en supposant que 100 % des couples soient cousins germains, F n'aurait alors qu'une valeur de 0,0625… et 100 % des couples frères-sœurs donneraient une valeur de 0,25 !!!

Question 5 :

	Génotypes	*m1//m1*	*m1//m2*	*M2//m2*	Total	Valeur du χ^2 testant les écarts
Communauté A	**Effectifs observés**	9	46	105	160	1,65
	Effectifs attendus	6,4	51,2	102,4	160	
Communauté B	**Effectifs observés**	46	34	10	90	0,90
	Effectifs attendus	44,1	37,8	8,1	90	
Total des effectifs observés		55	80	115	250	1,49
Total des effectifs attendus		50,5	89	110,5	250	

Les fréquences *r1* et *r2* ont pour valeurs 0,2 et 0,8 dans la population *A*, mais 0,7 et 0,3 dans la population *B*.

D'où les effectifs attendus et les valeurs de χ^2 correspondantes, dont on voit qu'aucune n'est plus significative (inférieures à 3,84 ; ddl = 1, dans tous les cas).

On a ici une illustration de l'effet « Walhund » caractérisé par des écarts significatifs sur le plan statistique mais sans aucun sens sur le plan biologique si on considère que la population « unique » étudiée est fictive puisque formée de deux sous populations qui sont les réelles entités panmictiques.

Conclusion : quand on étudie, sans le savoir, un échantillon d'individus issus de populations différentes, notamment par les fréquences des allèles étudiés, on génère, par une sorte d'artefact, des écarts qui peuvent être interprétés comme résultant d'une consanguinité importante, au sein d'une population virtuelle, alors que ces écarts ne sont dus qu'à ce mélange, au sein duquel on calcule des fréquences qui ne correspondent pas à celles d'une population réelle.

Exercice 4.9

L'étude génétique de la descendance, chez une espèce dioïque (appareils floraux mâles et femelles séparés), a montré constamment, sur toute son aire de répartition, un taux d'allogamie de 80 %. Mais on se doute que la taille de l'aire de répartition

de l'espèce est telle qu'on ne peut la considérer comme une entité unique ; par ailleurs, la variabilité du milieu en températures, en précipitations et en insectes pollinisateurs, peut laisser supposer une certaine différenciation en sous populations présentant des différences génétiques.

On entreprend une collecte d'échantillons et leur étude en groupant alternativement les individus en lots différents selon la taille de la surface au sein de laquelle ils ont été recueillis (voir tableau ci-dessous). On notera qu'une même plante peut avoir été comptée plusieurs fois car les surfaces de taille inférieures à la surface totale sont chevauchantes.

On étudie dans chacun des lots plusieurs gènes ou marqueurs di-alléliques. Le tableau ci-dessous donne le résultat obtenu pour un gène dont les fréquences alléliques sont égales à 0,6 et 0,4.

Question 1 : la progression du taux *H* d'hétérozygotie en fonction de la décroissance de la taille de la parcelle étudiée est-elle logique avec ce que l'on sait de l'effet Walhund ? Calculer F_{IT} à partir de *H* (voir index de fixation).

Question 2 : compléter le tableau en calculant les statistiques *F*, après avoir rappelé la formule de Wright. La progression de F_{ST} est-elle logique ?

Question 3 : quelle valeur, pour quel paramètre, devrait-on observer, dans une population dont la taille (la surface) permettrait une même probabilité de fécondation allogame entre tous les individus ?

Question 4 : à partir de quelle surface se trouve-t-on dans la condition définie à la question précédente, si on considère que la valeur observée du paramètre étudié ne diffère pas de plus de 5 % de la valeur attendue ?

Surface de répartition de la population étudiée	Taux d'hétérozygotie	Valeur de F_{IT}	Valeur de F_{IS}	Valeur de F_{ST}
Totale : 10 000 km²	0,307			
Partielles : 2 500 km²	0,345			
Partielles : 1 000 km²	0,365			
Partielles : 400 km²	0,380			
Partielles : 100 km²	0,385			

Solution

Question 1 : cette progression est logique car, du fait de l'effet Walhund, on mesure un déficit d'autant plus grand qu'est grande la différenciation qui s'instaure entre les sous populations au sein d'une grande population subdivisée.

Si on prend la population totale, les différences entre sous populations sont telles qu'il y a un fort déficit d'hétérozygotes ; si on étudie un lot d'individus plus proches dans l'espace, il y a moins de différenciation entre sous-groupes et le taux d'hétérozygotie est moins diminué.

Question 2 : la formule de Wright est la suivante :

$$(1 - F_{IT}) = (1 - F_{IS}).(1 - F_{ST})$$

ou

$$(F_{IT}) = F_{IS} + (1 - F_{IS}).(1 - F_{ST})$$

sachant que $(1 - F_{IT})$ est la chute d'hétérozygotie dans la population totale, $(1 - F_{IS})$ est la chute d'hétérozygotie résultant de l'autogamie partielle et $(1 - F_{IS})$ est la chute d'hétérozygotie résultant de la structuration de la population totale en sous populations différenciées.

On sait que le taux d'hétérozygotie de la population étudiée est égal à

$$H = 2pq.\ (1 - F_{IT})$$

D'où on tire les valeurs de F_{IT} (tableau, troisième colonne) valeur de la corrélation (ressemblance) gamétique totale qui augmente bien avec la taille de la population du fait de l'effet Walhund.

Surface de répartition de la population étudiée	Taux d'hétérozygotie	Valeur de F_{IT}	Valeur de F_{IS}	Valeur de F_{ST}
Totale : 10 000 km²	0,307	0,3604	0,2	0,2005
Partielles : 2 500 km²	0,345	0,2812	0,2	0,1015
Partielles : 1 000 km²	0,365	0,2396	0,2	0,0495
Partielles : 400 km²	0,380	0,2083	0,2	0,0104
Partielles : 100 km²	0,385	0,1979	0,2	-0,0026

La valeur de F_{IS} est le taux d'autogamie, c'est-à-dire le complément à 1 du taux d'allogamie, soit 20 % dans tous les cas, ce qui permet d'estimer F_{ST} (dernière colonne) et de remarquer que F_{ST} décroît avec la taille du lot, indiquant qu'à partir d'un certain seuil, il y a plus de sous populations différenciées, que F_{ST} est alors quasi nul et qu'il n'y a plus d'effet Walhund. La valeur négative dans les populations sur 100 km² n'est pas significativement différente de zéro.

Question 3 : sur une surface sans effet Walhund, sans différenciation en sous population du fait de l'efficacité de l'échange exogame, la valeur de F_{IT} sera égale à celle de F_{IS}, taux d'endogamie spécifique et général de l'espèce, soit 20 %.

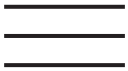
Question 4 : on admet l'absence d'effet Walhund si F_{IT} est égal à F_{IS}, soit 20 % plus ou moins 5 % de F_{IS}, soit F_{IT} compris entre 0,19 et 0,21. On peut donc considérer qu'il n'y a plus d'effet Walhund pour des populations réparties sur 400 km².

Exercice 4.10

On entreprend l'étude du polymorphisme d'un gène codant pour une enzyme d'intérêt dans une espèce végétale partiellement autogame ?

On sait que l'enzyme est un homodimère et présente deux variants électrophorétiques dont les mobilités sont différentes, appelés *R* pour rapide et *L* pour lent.

On prélève trois lots différents de végétaux dans trois localités de son aire de distribution, et on observe les résultats suivants :

			
Lot 1	**65**	**29**	**56**
Lot 2	**67**	**31**	**52**
Lot 3	**78**	**50**	**72**

Question 1 : combien d'allèles peut-on définir ?

Question 2 : y a-t-il une variation de la composition génétique de l'espèce d'un endroit à l'autre de son aire de répartition ?

Question 3 : en conséquence quelles sont les estimations des fréquences alléliques ?

Question 4 : peut-on admettre l'hypothèse de l'équilibre de Hardy-Weinberg ?

Question 5 : quelle solution proposez-vous ? Détaillez celle-ci en supposant que l'espèce est à l'équilibre attendu.

Solution

Question 1 : il y a deux allèles qui seront nommés *R* et *L*, aboutissant à trois phénotypes codominants. Chez l'hétérozygote *R/L*, il y a trois types d'homodimères dont un de mobilité intermédiaire.

Question 2 : non, un test d'homogénéité donne un χ^2 de valeur égale à 2,32 qui est inférieure à la valeur seuil de 9,49 associée à un risque de 5 %, pour un χ^2 à 4 degrés de liberté.

Question 3 : de ce fait la meilleure estimation des fréquences alléliques est celle réalisée sur la somme des lots, soit $f(R) = 0{,}60$ et $f(L) = 0{,}40$. (Attention : il y a 180 hétérozygotes !)

Question 4 : non, parce que les effectifs observés sont égaux à 210, 110 et 180 et que les effectifs attendus sous l'hypothèse de Hardy-Weinberg sont respectivement égaux à 180, 80 et 240 ce qui donne un χ^2 dont la valeur observée est égale à 31,25, largement supérieure à la valeur seuil de 3,84, associée à un risque de 5 %, pour un χ^2 à 1 ddl.

Question 5 : la population n'est évidemment pas panmictique puisqu'il s'agit d'une espèce partiellement autogame. Une proportion λ des croisements correspond à de l'autofécondation et une proportion $(1 - \lambda)$ des croisements correspond à une pollinisation aléatoire, panmictique.

Si la population est à l'équilibre, il est facile, connaissant la fréquence *H* des hétérozygotes, d'estimer le taux d'autogamie en appliquant la relation démontrée au chapitre 4, soit :

$$\boldsymbol{H = 4pq(1 - \lambda)/(2 - \lambda)}$$

d'où on tire que

$$\boldsymbol{\lambda = [2H - 4pq]/[H - 4pq]}$$

La valeur de λ est égale à 0,4 : il y a 40 % d'autofécondation. Ce taux d'autogamie aboutit à un excès d'homozygotes pour tous les gènes.

Exercice 4.11

Une population est constituée de 51 % de bruns et de 49 % de blonds.

On estimera, en première analyse que cette différence phénotypique dépend, dans cette population, d'un seul gène, le phénotype blond étant récessif, de génotype *a/a*, le phénotype brun étant dominant et de génotype *A/A* ou *A/a*. Afin de revenir à la pureté de la « race » un dictateur fasciste interdit les mariages entre bruns et blonds.

Question 1 : quel est le régime (de croisement !) imposé par ce dictateur ; comment évoluera la population ?

Question 2 : la mise au point d'un test biochimique permettant de distinguer, chez les bruns, les hétérozygotes des homozygotes, permet alors la signature d'un décret excluant les mariages entre génotypes différents.

a) Quel est ce nouveau régime (de croisement) ?

b) Quelle sera l'évolution génétique de la population ?

Question 3 : bien évidemment la résistance active ou passive permet à 20 % de la population d'échapper aux contrôles (moyennant l'utilisation massive de colorants ou en recourant, pour une fois, à la calvitie !). Quelle sera, dans ce cas, l'équilibre vers lequel pourrait tendre la population (si elle ne se débarrasse pas assez vite de son dictateur) ?

Solution

Question 1 : il s'agit d'homogamie phénotypique.

Bien évidemment, les blonds n'auront entre eux que des enfants blonds, mais les bruns, compte tenu de leurs génotypes respectifs pourront transmettre l'allèle *a*, sans qu'il se manifeste jamais, à partir du moment où il s'agira de couples $A/A \times A/a$.

À terme cependant, le groupe des bruns verra sa fréquence diminuer au fur et à mesure que les allèles *a* auront, au détour de la naissance d'un enfant blond, rejoints le groupe des blonds.

Du fait de la dominance, l'évolution de la composition génétique de la population sera plus lente sous l'homogamie phénotypique qu'elle ne la serait sous l'homogamie génotypique, dont l'évolution est semblable à celle de l'autofécondation (voir chapitre 4).

Question 2 : il s'agit maintenant d'homogamie génotypique totale. Chacun des génotypes ne se croisant qu'avec un génotype identique, les génotypes *A/A* se croisent entre eux et ne donnent que des descendants *A/A* ; de même les *a/a* ne donnent que des descendants *a/a*. Quant aux *A/a*, croisés entre eux, ils donnent 1/4 de *A/A*, 1/4 de *a/a* et 1/2 de *A/a*.

D'où les relations suivantes, entre les générations i et $i + 1$, si D, H et R sont les fréquences respectives des trois génotypes *A/A*, *A/a* et *a/a* :

$$D_{i+1} = D_i + H_i/4$$
$$H_{i+1} = H_i/2$$
$$R_{i+1} = R_i + H_i/4$$

Ainsi, de générations en générations, la fréquence des homozygotes va croître à mesure que celle des hétérozygotes diminue (de moitié à chaque génération). Formellement les relations de récurrence sont les mêmes que pour l'autofécondation et, à la limite, on obtient des « lignées pures ».

Il faut au moins 8 à 10 générations (voir chapitre 4), ce qui représente deux siècles et demi, et surtout on n'aura acquis la « pureté » que pour le gène *A* uniquement, puisque le critère de choix des conjoints ne porte que sur les génotypes pour ce gène ; dans le cas de l'autofécondation ou des croisements frères-sœurs l'homozygotie porte sur tout le génome.

> **Remarque :** l'homozygotie s'étendra cependant aux gènes physiquement proches du gène *A*, par un effet d'auto-stop (voir chapitre 7). Cependant cette homozygotie sera partielle puisque le déséquilibre gamétique généré par la liaison physique et la sélection sur le gène *A* aura tendance à disparaître avec le temps.

Question 3 : il existe, dans ce cas, une fraction λ des individus se croisant par homogamie et une fraction $(1 - \lambda)$ se croisant de manière panmictique. On a établi les relations suivantes (voir 4.4.2) :

$$\boldsymbol{D_{i+1} = \lambda [D_i + H_i/4] + (1 - \lambda) p^2}$$
$$\boldsymbol{H_{i+1} = \lambda [H_i/2] + (1 - \lambda) 2pq}$$
$$\boldsymbol{R_{i+1} = \lambda [R_i + H_i/4] + (1 - \lambda) q^2}$$

Si H_e est la fréquence des hétérozygotes à l'équilibre, on doit avoir la relation suivante :

$$H_e = \lambda [H_e/2] + (1 - \lambda) 2pq$$

d'où on tire que $H_e = 4pq(1 - \lambda)/(2 - \lambda)$

ou $\boldsymbol{H_e = 2pq - 2pq\, \lambda/(2 - \lambda)}$

Ici encore, le résultat est semblable à celui qu'on obtient pour une autofécondation partielle, à la différence qu'il concerne un seul gène et non tout le génome.

Si $\lambda = 0{,}2$, sachant que la population était panmictique, avant la dictature, et que les fréquences alléliques restent inchangées (d'où $p = 0{,}3$ et $q = 0{,}7$) on aura alors $\boldsymbol{H = 37{,}3\ \%}$.

Ce résultat (37 % d'hétérozygotie) montre la pression énorme d'isolement génétique que la nature impose à certaines populations ou que l'homme impose à certaines variétés animales ou végétales afin que le concept de « race » puisse correspondre à la « pureté génétique » qu'il sous-tend.

Chapitre **5**

La dérive génétique

5.1 INTRODUCTION

Le modèle de Hardy-Weinberg suppose que l'effectif des populations est suffisamment grand pour être considéré comme infini. Cette condition est évidemment irréaliste, mais, avant de définir un seuil à partir duquel cette condition serait en pratique acceptable, il convient d'établir ce qu'il advient de la composition génétique des populations quand l'effectif est limité.

On montre que la limitation de l'effectif d'une population conduit à un phénomène appelé « dérive génétique » parce qu'il est caractérisé par une fluctuation aléatoire, d'une génération à l'autre, des fréquences alléliques.

On montre aussi que le même phénomène peut être caractérisé par une augmentation récurrente de la consanguinité.

5.2 FLUCTUATION DES FRÉQUENCES ALLÉLIQUES

5.2.1 Approche intuitive de la dérive génétique

Que cela corresponde ou non à la réalité biologique de l'espèce, tout se passe, dans une population panmictique, comme si les individus qui s'unissent mettaient un même nombre (il n'y a pas de sélection) de gamètes dans une urne ; chaque individu de la génération suivante étant le résultat d'un double tirage dans cette urne.

Les fréquences *p* et *q* des allèles *A1* et *A2* d'un gène di-allélique chez les parents sont aussi les fréquences des gamètes porteurs de *A1* et *A2* dans l'urne et ce sont les probabilités de tirages de ces gamètes.

Quand la population est de grand effectif, le nombre de tirages est tellement élevé que la fréquence des gamètes tirés ne sera pas beaucoup, voire pas du tout différente de leur probabilité de tirage, soit p et q. C'est la même chose, quand on tire à pile ou face, la fréquence des piles ne sera pas très différente de 1/2 si le nombre de tirage est très élevé (loi des grands nombres).

De ce fait les fréquences des génotypes réalisés par les doubles tirages sont dans les proportions de Hardy-Weinberg, soit p^2, $2pq$ et q^2.

Au contraire, dès que l'effectif sera assez petit (on verra plus tard à quoi correspond concrètement le « assez »), les fréquences des allèles *A1* et *A2*, après le tirage peuvent différer considérablement de leurs probabilités respectives de tirages. C'est la même chose, quand on tire dix fois à pile ou face, et que le nombre de piles observés est égal à six ou sept. Dans ce cas, la fréquence des piles, 0,6 ou 0,7, n'est pas du tout impossible et diffère beaucoup de la probabilité de tirage 0,5 (fréquence observée si on réalise un grand nombre de tirages). Dans une petite population, les fréquences alléliques vont donc fluctuer au gré des variations aléatoires de tirages, et ce, d'une génération à l'autre. Cette fluctuation des fréquences est un peu comme celle de la trajectoire d'une bouteille lancée à la mer, soumise aux aléas des vents, des chocs et des courants, et c'est la raison pour laquelle elle a été appelée « dérive génétique ».

5.2.2 Formulation mathématique de la dérive génétique

Si on considère une population d'effectif constant N, on conçoit qu'il faut réaliser $2N$ tirages à chaque génération, dans l'urne gamétique des parents.

Si les deux allèles *A1* et *A2* d'un gène di-allélique, ont pour fréquence p_i et q_i à la génération i, quelles seront leurs valeurs à la génération suivante ?

En fait, il suffit de s'intéresser à la fréquence d'un seul allèle, celle de l'autre étant le complément à 1.

La probabilité de tirage de l'allèle *A1* dans l'urne gamétique est égale à la fréquence p_i de l'allèle *A1* chez les parents.

Si on fait $2N$ tirages dans l'urne gamétique le nombre d'allèles *A1* tirés peut être compris entre 0 et $2N$. En fait, ce nombre est une variable aléatoire X_{i+1} = « nombre d'allèles *A1* tirés dans l'urne, pour réaliser la génération $i + 1$ », dont la valeur est comprise entre 0 et $2N$.

On sait que toute les valeurs de cette variable sont possibles mais que les probabilités d'observer chacune de ces valeurs ne sont pas les mêmes. On sait plus précisément que chacune des valeurs de cette variable aléatoire X_{i+1} a une probabilité d'observation donnée par la loi binomiale $B(p_i , 2N)$.

Cette loi permet de définir les paramètres de la distribution de X_{i+1}

– son espérance (appelée aussi moyenne) est égale à : $E(X_{i+1}) = 2N. p_i$
– sa variance est égale à : $V(X_{i+1}) = 2N. p_i.(1 - p_i)$

En fait, ce qui nous intéresse n'est pas tant le nombre tiré d'allèles *A1* que leur fréquence f_{i+1}, c'est-à-dire le rapport $X_{i+1}/2N$. La distribution de probabilités de la

fréquence f_{i+1} suit la même binomiale, mais les paramètres de distribution sont différents puisqu'il s'agit d'une variable aléatoire X multipliée par le scalaire $1/2N$. Dans ces conditions on sait que l'espérance est multipliée par le scalaire $1/2N$, et sa variance par $(1/2N)^2$, soit :

$$E(f_{i+1}) = p_i$$
$$V(f_{i+1}) = p_i.(1 - p_i)/2N$$

On montre bien ainsi que, dans une grande population idéale (hypothèse du modèle de Hardy-Weinberg), quand N tend vers l'infini, la variance de la fréquence f_{i+1} est nulle ; de ce fait la fréquence f_{i+1} de *A1* ne peut qu'être égale à son espérance p_i, soit la fréquence de *A1* à la génération précédente.

Au contraire, dès que l'effectif est limité, la variance n'est plus nulle, et la fréquence f_{i+1} à la génération $i + 1$, peut prendre une valeur différente de la valeur p_i à la génération précédente. Cette variation de valeur entre générations ne dépend que du hasard d'échantillonnage. Comme la variation de la fréquence de l'allèle *A1*, de génération en génération est totalement imprévisible, elle fluctue de manière aléatoire ou chaotique.

Cette fluctuation allélique peut être illustrée par la figure 5.1, où trois populations, au départ identiques dans leur composition génétique (même fréquence initiale f_0), voient leur diversité évoluer de manière chaotique et divergente en raison de la dérive génétique.

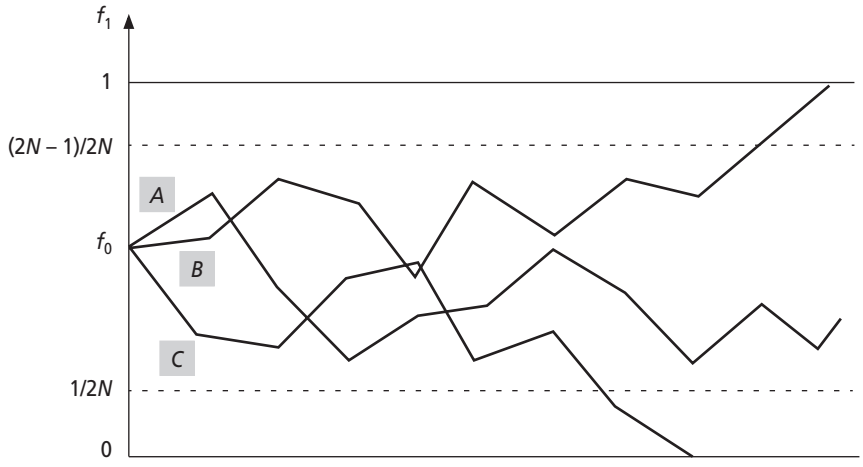


Figure 5.1

Les fréquences alléliques des trois populations *A*, *B* et *C* fluctuent de manière totalement aléatoire entre 0 et 1, les deux valeurs extrêmes possibles de la fréquence, à chaque génération.

Il existe cependant deux situations limites qui finiront tôt ou tard par être atteintes, celle où l'allèle *A1*, passant la barre de la valeur $2(N - 1)/2N$ prend la valeur 1 (population *B*), ou, symétriquement, passant la barre $1/2N$, prend la valeur

zéro (population *C*). Dans ces deux situations la fréquence allélique prend une valeur définitivement stable 1 ou 0, l'allèle *A1* est fixé ou éliminé. Ces deux états limites possibles sont appelés absorbants parce qu'on y reste quand on y tombe[1].

5.2.3 Conséquences génétiques de la dérive sur la diversité génétique

Toute population en dérive génétique finira tôt ou tard par atteindre un état absorbant pour tel ou tel allèle de tel ou tel gène : c'est inéluctable et ce n'est qu'une question de temps.

La dérive génétique aboutit donc à une réduction du polymorphisme génétique des populations par la perte, pour certains gènes, de tous les allèles sauf un, celui qui est fixé.

On peut même imaginer que le temps aidant (on suppose toujours pour l'instant qu'il n'y a pas de mutations) on obtiendrait, par dérive, une souche pure. Les conséquences génétiques de cette réduction du polymorphisme du point de vue évolutif sont développées un peu plus loin.

> ***Remarque :*** la dérive génétique touche la diversité allélique et partant la diversité génotypique alors que les écarts à la panmixie, dans une grande population ne touchent que la diversité génotypique sans modifier la diversité allélique (voir chapitre 4).

5.2.4 L'effet fondateur

L'effet fondateur est une variation d'échantillonnage affectant la composition génétique d'une population, en une occasion particulière unique.

Un premier exemple d'effet fondateur est celui qui touche les petits groupes essaimant d'une population. En raison de la taille de ce groupe, sa composition génétique peut différer fortement de celle de sa population d'origine. Cette variation d'échantillonnage induite par la scission de population est appelée effet fondateur. À la limite, la dérive génétique peut être considérée comme un effet fondateur récurrent à chaque génération.

Un deuxième exemple est celui du goulot d'étranglement démographique (*bottle neck*). Certaines espèces ou populations dont la taille est assez importante pour ne pas donner prise à la dérive et garder une composition génétique stable, subissent, occasionnellement ou périodiquement, des goulots d'étranglement démographique, en

1. Le billard américain donne une image très suggestive de la dérive génétique et de ses effets. Dans ce jeu, il s'agit de mettre, dans un ordre précis, des boules de couleurs différentes, dans certains des six trous existant aux quatre angles et au milieu des grands cotés. Supposons qu'un joueur fasse n'importe quoi et tape n'importe comment dans les boules, au hasard et dans n'importe quel sens, il finira toujours, tôt ou tard, par avoir envoyé toutes les boules sauf une dans les trous. Ceux-ci sont au billard ce que les états absorbants sont à la dérive, et les trajectoires chaotiques des boules simulent celles des fréquences alléliques. Par hasard, il y aura bien quelques trajectoires qui élimineront une boule du jeu, un allèle du patrimoine ; au bout du compte, il ne restera qu'une seule boule, c'est-à-dire un seul allèle.

raison d'un abaissement brutal des ressources nutritives ou d'un accroissement brutal de la fréquence d'un prédateur (ou d'une guerre chez l'homme !). Cet abaissement brutal de l'effectif peut générer une variation échantillonnage car l'effectif restant en vie peut être assez faible pour ne pas être représentatif de la composition d'origine. Par ailleurs, il peut aussi y avoir quelques générations de dérive génétique avant que l'effectif ait récupéré une taille compatible avec une absence de dérive et une composition génétique stable. Cet épisode d'étranglement démographique peut réaliser un effet fondateur.

Un troisième type d'effet fondateur correspond à une fusion de population formant une nouvelle population, avec un nouveau stock génique et des déséquilibres gamétiques, mais il ne correspond pas, comme les deux précédents à une variation d'échantillonnage (voir chapitre 3).

L'effet fondateur a joué un rôle important dans l'histoire génétique de l'espèce humaine, compte tenu des modalités de son expansion géographique par essaimage de petits groupes, notamment vers les continents américain et océanien, puis par les fusions de populations.

5.3 AUGMENTATION RÉCURRENTE DE LA CONSANGUINITÉ

5.3.1 Approche intuitive

Dans une grande population, les unions entre apparentés résultent d'un choix, car la taille de la population est telle que la probabilité de rencontre, par hasard, de deux apparentés est nulle. Au contraire, dans une petite population, même si il y a panmixie, cette probabilité n'est pas nulle. Des unions panmictiques entre apparentés surviennent et la parenté moyenne des couples conduit à une consanguinité moyenne à la génération suivante.

Comme le même phénomène se reproduit à chaque génération, on conçoit que tous les individus d'une petite population finissent par être, à des degrés divers, apparentés entre eux, et, qu'avec les générations, les enfants sont toujours de plus en plus consanguins. À la limite la population devrait tendre vers une lignée pure avec un taux de consanguinité de 100 % pour chaque gène.

5.3.2 Formulation mathématique de l'augmentation récurrente de la consanguinité résultant de la limitation de l'effectif

Afin de bien marquer leur différence d'origine, la consanguinité qui résulte d'un choix du conjoint en fonction de la parenté, au sein d'une grande population, et la consanguinité qui survient au sein d'une petite population malgré la panmixie, sont notées de manière différente. Dans le premier cas, elle a été notée F (la valeur de F, moyenne des taux de parenté, variant à chaque génération en fonction de la fraction des unions entre apparentés), dans le deuxième cas elle est notée α_g, g étant l'indice de génération.

On montre que la consanguinité d'un individu tiré au hasard, à la génération *g*, dans une petite population désignée par le paramètre α_g, s'accroît indéfiniment, au fil des générations.

Une telle population a un effectif *N* qui se décompose en N_1 individus de sexe masculin et N_2 individus de sexe féminin.

La consanguinité d'un individu tiré au hasard, à la génération *g*, est désignée par le paramètre α_g. La parenté entre deux individus tirés au hasard, à la génération g, est désignée par le paramètre β_g.

Du fait des définitions respectives des coefficients de parenté et de consanguinité, on peut écrire que :

$$\alpha_g = \beta_{g-1}$$

Si on considère que les effectifs sont stables d'une génération à l'autre, la figure 5.2 schématise le transfert des gènes au sein de la population, chaque effectif N_1 et N_2 assurant la formation des effectifs N_1 et N_2 de la génération suivante.

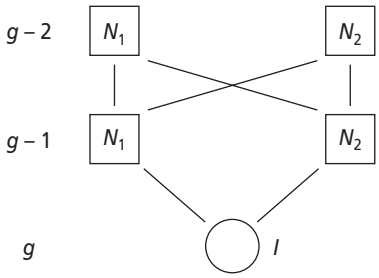


Figure 5.2

On se propose d'établir la valeur α_g de la consanguinité d'un individu *I* de la génération *g* en fonction de l'origine possible de ses gènes, et des valeurs des coefficients de consanguinité ou de parenté de ses ascendants.

Pour tout gène, les deux exemplaires présents chez I, peuvent, en fonction de la relation de parenté possible entre ses parents, venir :

– soit d'un même grand-père ;
– soit d'une même grand-mère ;
– soit de deux grands-parents différents (les deux grands-pères ou les deux grands-mères ou un grand-père et une grand-mère).

Ces trois cas rassemblent de manière exhaustive toutes les origines possibles des deux exemplaires d'un gène, présents chez *I* (voir à ce sujet le calcul de l'évolution de la consanguinité dans les croisements systématiques frère-sœur, 4.2.3.c).

1. Dans le premier cas (même grand-père), la probabilité d'identité par ascendance est égale à $(1/2 + f_{GP}/2)$ où f_{GP} est le coefficient de consanguinité du grand père, mais f_{GP} est aussi égal à α_{g-2}, le coefficient moyen de consanguinité à la génération $g - 2$.

2. Dans le deuxième cas (même grand-mère), la probabilité d'identité par ascendance est égale à $(1/2 + f_{GM}/2)$ où f_{GM} est le coefficient de consanguinité de la grand-mère, mais f_{GM} est aussi égal à α_{g-2}.
3. Dans le troisième cas (grands-parents différents), la probabilité d'identité par ascendance est égale, par définition, au coefficient de parenté entre deux grands-parents, soit β_{g-2}. Et on sait que $\beta_{g-2} = \alpha_{g-1}$.

On a donc défini de manière exhaustive les origines possibles des deux exemplaires d'un gène de *I*, et dans chacun des trois cas, la probabilité que ces deux exemplaires soient identiques par ascendance. Il ne reste plus qu'à calculer la probabilité des trois cas en question.

Avec quelle probabilité les deux exemplaires d'un gène de I viennent-ils d'un même grand-père ?

Prenons l'un des deux exemplaires d'un gène chez I, il a évidemment une chance sur deux de venir de son grand-père paternel ou maternel. De même, le deuxième exemplaire a aussi une probabilité 1/2 de venir de son grand-père paternel ou maternel.

Les deux exemplaires viennent donc, l'un du grand-père paternel, l'autre du grand-père maternel, avec une probabilité 1/4.

Ces deux grands-pères paternel et maternel peuvent être une seule et même personne avec la probabilité $1/N_1$. En effet, sachant que les deux exemplaires de I viennent de grands-pères (événement de probabilité 1/4), la probabilité que le deuxième exemplaire vienne du même grand-père que le premier est égale à $1/N_1$ puisqu'il y a N_1 grands-pères possibles. Cette probabilité est évidemment nulle dans une grande population panmictique mais pas dans une petite.

Les deux exemplaires d'un gène de *I* viennent donc d'un même grand-père avec une probabilité égale à$1/4 \times 1/N_1 = 1/4N_1$

Par un raisonnement identique, on conclura que les deux exemplaires d'un gène de *I* viennent de la même grand-mère avec une probabilité égale à $1/4N_2$.

Enfin les deux exemplaires d'un gène de I viennent de deux grands-parents différents avec une probabilité égale à 1 moins la somme des deux autres, soit $(1 - 1/4N_1 - 1/4N_2)$.

On peut alors établir la relation permettant l'estimation de α_g, en fonction de α_{g-1} et de α_{g-2} :

$$\alpha_g = [1/4N_1]\,(1/2 + \alpha_{g-2}/2) + [1/4N_2]\,(1/2 + \alpha_{g-2}/2) + (1 - 1/4N_1 - 1/4N_2)\,\alpha_{g-1}$$

Si on définit un effectif théorique, appelé effectif efficace, N_e tel que

$$1/N_e = 1/4N_1 + 1/4N_2$$

soit

$$N_e = 4\mathrm{N}1\ N_2/(N_1 + N_2)$$

L'équation de récurrence pour la consanguinité devient :

$$\boldsymbol{\alpha_g = [1/N_e]\,(1/2 + \alpha_{g-2}/2) + [1 - 1/N_e]\,\alpha_{g-1}}$$

La valeur limite de la consanguinité, notée α_e, est telle qu'elle vérifie la récurrence, soit :

$$\alpha_e = [1/N_e]\,(1/2 + \alpha_e/2) + [1 - 1/N_e]\,\alpha_e$$

d'où on tire que

$$\boldsymbol{\alpha_e = 1}$$

Cette formule de récurrence permet de calculer la consanguinité à la génération g, sachant les valeurs obtenues aux deux générations précédentes. On peut vérifier qu'avec les croisements frère-sœur où on a *N1* = 1 et *N2* = 1, cette formule donne bien les valeurs déjà calculées (figure 4.8).

Le calcul itératif n'est cependant pas aisé si l'on souhaite calculer la consanguinité après 50 ou 500 générations. C'est pourquoi il est utile de « résoudre une formule de récurrence » en lui substituant une fonction algébrique équivalente où l'indice de la variable est devenu variable elle-même.

Cela est facile, et a été déjà vu à plusieurs reprises, quand l'équation de récurrence ne possède qu'un seul terme (exemple *Hn* en fonction de *Hn* – 1, avec ou sans scalaire additionnel) ; mais ici, il s'agit d'une récurrence à plus d'un terme et la résolution algébrique est un problème classique de mathématique qui ne sera pas abordé.

On montre que la solution algébrique de l'équation de récurrence :

$$\alpha_g = [1/N_e\,](1/2 + \alpha_{g-2}/2) + [1 - 1/N_e]\,\alpha_{g-1}$$

est

$$\alpha_g = \mathbf{1 - e^{-g/2Ne}}$$

Remarque 1 : quand *g* augmente, et tend vers l'infini, le terme exponentiel tend vers zéro et la consanguinité tend bien vers 1.

Remarque 2 : les valeurs données par la récurrence et sa solution algébrique sont peu différentes au bout de quelques générations, mais il est préférable et plus précis d'utiliser la récurrence pour les premières générations.

5.3.3 Limite du processus d'augmentation récurrente de la consanguinité

L'augmentation de la consanguinité se traduit par une homozygotie croissante, c'est-à-dire une diminution croissante de la diversité génétique. À la limite, quand la consanguinité est égale à 1, les individus sont, pour chaque gène, homozygotes pour un même allèle, et constituent une souche pure.

Bien que sous une autre forme, on retrouve ici le même résultat final sur la diversité que celui résultant de la fluctuation des fréquences alléliques. Fluctuation des fréquences alléliques et augmentation récurrente de la consanguinité sont, comme face et pile d'une même pièce, deux aspects différents d'un même phénomène, la dérive génétique qui s'accompagne d'une chute de la diversité allélique

5.3.4 Vitesse du processus d'augmentation récurrente de la consanguinité

La vitesse de la dérive est plus facile à appréhender mathématiquement à travers l'évolution de la consanguinité.

L'accroissement de la consanguinité dans une population d'effectif limité est donné par la fonction $\alpha_g = 1 - e^{-g/2Ne}$. Cette fonction peut être écrite sous la forme $(1 - \alpha_g) = e^{-g/2Ne}$ qui exprime la décroissance de l'écart entre la consanguinité et sa

valeur limite égale à 1. La décroissance exponentielle de l'écart entre la consanguinité et sa limite permet alors d'estimer la vitesse d'évolution de la consanguinité.

En effet, tout processus de décroissance exponentielle est caractérisé par une période, un intervalle de temps constant, noté *T,* à l'issue duquel l'écart à la limite est réduit de moitié ; au bout d'un temps *T* l'écart initial est divisé par deux, au bout de 2*T,* il est divisé par quatre, au bout de 3*T,* par huit etc.

L'exemple le plus connu de période associée à une décroissance exponentielle est la période des éléments radioactifs ; ainsi le phosphore 32, dont la période est égale à 15 jours voit sa radioactivité divisée par deux tous les 15 jours alors que les plutonium 238 et 239 ont des périodes de 86,4 et 24 390 ans et que les uranium 238 et 235 ont des périodes 4,5 et 7,2 milliards d'années.

La décroissance de l'écart entre la consanguinité et sa limite est donnée par la figure 5.3, où la période *T* peut facilement être estimée en considérant :

– sa définition : *T* est telle que $1 - \alpha_T = (1 - \alpha_0)/2$;
– la relation liant la consanguinité au temps : *T* est telle que $1 - \alpha_T = e^{-T/2Ne}$

En identifiant ces deux égalités, sachant que $\alpha_0 = 0$, à la génération initiale, on tire que :

$$1/2 = e^{-T/2Ne}$$

d'où, en passant en Log : $-\text{Log}2 = -T/2N_e$

et $$\boldsymbol{T = 2\text{Log}2\ N_e = 1{,}4\ N_e}$$

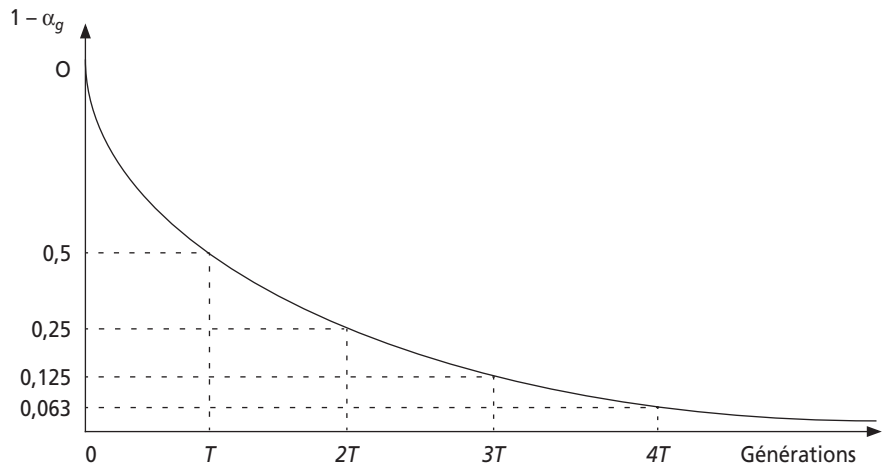


Figure 5.3

On peut évidemment, par symétrie, en déduire (figure 5.4) le graphe de la fonction d'accroissement de la consanguinité, avec une même période *T,* temps nécessaire à l'accroissement de moitié de l'écart entre la consanguinité et sa limite égale à 1.

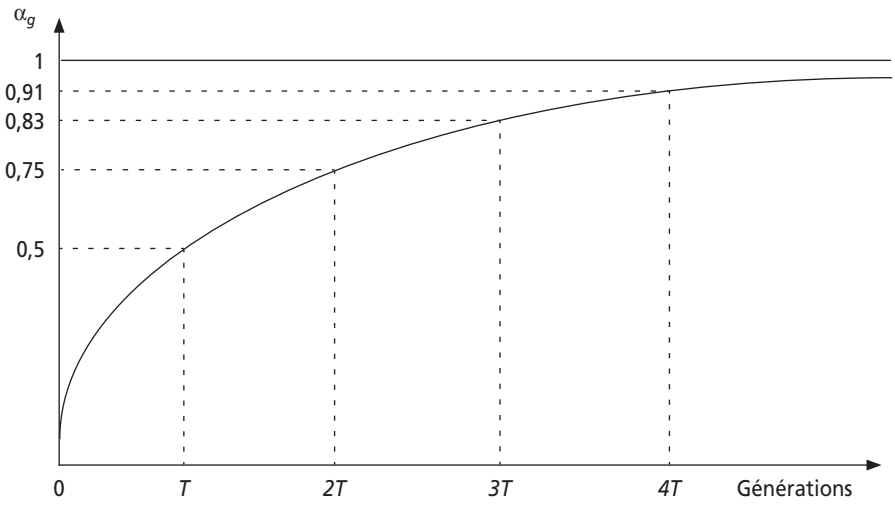


Figure 5.4

La période ne dépend que de l'effectif efficace N_e. Plus l'effectif efficace est petit (plus la population est « petite »), plus rapide est l'accroissement de la consanguinité, c'est-à-dire la dérive ; plus l'effectif efficace est élevé, plus lent est cet accroissement.

L'estimation de la vitesse de la dérive par le calcul de la période associée à l'effectif efficace permet de montrer à quel point la dérive est un processus « lent », du moins à l'échelle de la vie humaine ou même de l'histoire, y compris quand l'effectif efficace est très faible.

Par exemple, une population d'effectif efficace égal à 50 (25 hommes et 25 femmes) a une période de 70 générations : sa consanguinité passera de 0 % à à 50 % après 70 générations, soit 2000 ans et il faudra encore 2000 ans pour la voir monter à 75 %, puis encore 2 000 ans pour passer à 82,5 %.

5.3.5 Signification de l'effectif efficace

L'effectif efficace a été défini par la formule $N_e = 4N_1 .N_2./(N_1 + N_2)$ où N_1 et N_2 sont les effectifs de chacun des sexes.

En effet, l'effectif démographique $N = N_1 + N_2$ n'est pas un paramètre pertinent pour la dérive génétique, car l'évolution de la consanguinité dépend non seulement des individus participant **effectivement** à la reproduction, mais aussi de l'importance relative de chacun des sexes, le sexe ratio, dans cette reproduction.

Si le rapport est équilibré (participation équilibrée de chaque sexe à la reproduction), on montre que $N_e = N = N_1 + N_2$, mais dès que le sexe ratio est déséquilibré, l'effectif efficace peut devenir très inférieur à la somme $N = N_1 + N_2$. Dans ce cas l'effectif efficace N_e représente l'effectif d'une population théorique, équivalente du

point de vue de la dérive, mais qui aurait un sexe ratio équilibré, avec un effectif égal à $N_e/2$ dans chacun des deux sexes.

S'il est vrai que le sex-ratio n'est jamais très déséquilibré dans les populations humaines, ce n'est pas le cas dans nombre de populations naturelles d'espèces animales ou végétales.

Le tableau 5.1 donne une illustration de l'effet de ce sex-ratio sur l'effectif efficace, et à travers lui, sur la vitesse et l'efficacité de la dérive.

Tableau 5.1

N_1	N_2	$N = N_1 + N_2$	N_e	T
500	500	1 000	1 000	1400
400	600	1 000	960	1344
200	800	1 000	640	896
100	900	1 000	360	504
10	990	1 000	40	56
1	999	1 000	4	6

Il est évident que si un seul mâle participait à la reproduction (dernière ligne du tableau 5.1), tous les descendants seraient déjà demi-germains, ce qui explique la vitesse de la dérive en de telles circonstances. La réalité peut ne pas être si éloignée de cette situation dans certains groupes animaux avec un mâle dominant, ou dans certains cas d'insémination artificielle.

5.3.6 Effectif efficace et variance de la fréquence allélique

L'étude de l'accroissement de la consanguinité a permis d'introduire le paramètre d'effectif efficace. Mais cet accroissement continu de la consanguinité n'est qu'une des manières d'envisager la dérive génétique, la variation aléatoire des fréquences alléliques étant l'autre façon de concevoir ce phénomène. Dans ce dernier cas aussi, on a montré que la puissance de la dérive dépendait de l'effectif puisque la variance d'une fréquence allélique, $V(f) = p(1 - p)/2N$ (voir 5.2.2), est d'autant plus grande que N est petit.

On doit noter que, dans cette formule, l'effectif est l'équivalent de l'effectif efficace N_e introduit dans l'étude de la consanguinité, puisqu'il s'agit ici, non de l'effectif démographique mais du nombre de gamète réellement engagé dans la reproduction. La variance peut donc s'écrire :

$$V(f) = p(1 - p)/2N_e$$

Et l'effectif efficace ainsi défini est désigné comme *effectif efficace de variance*. Cette définition peut être très utile quand on souhaite estimer N_e dans des cas où la consanguinité n'a pas de sens, notamment pour les gènes du X chez l'homme et les autres espèces hétérogamétiques, ou dans les espèces où l'un des sexes est haploïde (insectes sociaux). Ces deux cas sont traités sous forme d'exercice (voir exercice 6.4).

5.3.7 Variation de l'effectif efficace dans le temps

Les effectifs des populations naturelles, même quand il y a stabilité démographique, ne sont jamais invariants, il est donc contestable, sauf si les variations sont minimes de garder un même effectif efficace sur plusieurs générations.

Si des variations de l'effectif efficace surviennent, on peut logiquement penser que les périodes d'effectif faible auront un effet plus important du point de vue de la chute de la diversité et que la recherche d'un effectif « moyen » sur un ensemble de générations devrait conduire à une formule privilégiant le poids de ces périodes d'effectifs faibles.

Par un développement mathématique qui ne sera pas présenté, on montre que si on dispose, sur n générations d'effectifs efficaces différents N_{ei}, où i est l'indice de générations, la population dérive comme si elle avait un effectif efficace invariant égal à N_e, dont l'inverse est la moyenne des inverses des N_{ei}, soit la moyenne harmonique, qui s'écrit :

$$1/Ne = (1/n)\Sigma[1/N_{ei}]$$

où i varie de l'indice de première génération, soit 0 à l'indice de la nième, soit $(n - 1)$.

5.4 RÔLE DE LA DÉRIVE DANS L'HISTOIRE GÉNÉTIQUE DES POPULATIONS

La dérive génétique conduit à une réduction de la diversité génétique de la population par la perte de nombreux allèles, allant éventuellement et progressivement jusqu'à la fixation d'un seul d'entre eux. Cette réduction de la diversité est souvent jugée comme un «appauvrissement génétique» sur la base du postulat selon lequel la diversité génétique constitue, au regard de l'évolution, un capital adaptatif où l'espèce pourra trouver les moyens, c'est-à-dire les gènes et les allèles, susceptibles de lui offrir une voie de survie en cas de bouleversement écologique.

Cela n'est qu'en partie vrai car si une espèce est divisée en plusieurs populations indépendantes, celles-ci, en dérivant indépendamment, n'élimineront ni ne fixeront le même allèle, puisque la dérive est un processus aléatoire. Aussi le polymorphisme de l'espèce, dans sa globalité, ne sera pas altéré, mais seulement réparti. La dérive génétique aura réduit la diversité génétique intra-populationnelle et accru la diversité inter-populationnelle.

Ce faisant, la dérive participe alors à la différenciation génétique entre populations d'une même espèce, prélude à la spéciation.

On peut certes arguer que la dérive n'a d'effet notable qu'au bout d'un temps « significativement long » et si l'effectif efficace est « significativement petit ».

Mais justement l'histoire de la vie a été assez longue et celle des espèces et de leurs populations a connu des épisodes de réduction suffisante des effectifs pour qu'on puisse donner à la dérive le statut de moteur de l'évolution qu'on attribue parfois trop exclusivement à la sélection naturelle.

5.4.1 Dérive et différenciation ethnique chez l'homme

À l'échelle de l'histoire, des 25 siècles qui nous séparent de Périclès ou des 5 000 ans qui nous séparent de l'antiquité égyptienne ou mésopotamienne, la dérive peut paraître sans effet. Mais l'humanité est beaucoup plus ancienne que cette si proche antiquité et les effectifs des populations qui l'ont constituée n'ont jamais été avant cette époque ce qu'ils furent après.

L'homme moderne est apparu entre l'Arabie et l'Est africain il y a quelques 100 000 à 200 000 ans, et, reprenant ou dépassant les traces de ces prédécesseurs a envahi l'ensemble de la planète.

Durant 95 % de cette période, jusqu'à la découverte de l'agriculture et de l'élevage vers –10 000, ce n'étaient que groupes de chasseurs-cueilleurs dont les conditions de ressources et de vie limitaient fortement la taille démographique. Des groupes de quelques dizaines ou même quelques centaines d'individus, sont « assez petits » si pendant 90 000 à 190 000 ans, c'est-à-dire 4 500 à 9 500 générations, ils sont restés isolés les uns des autres. Au bout d'un tel temps, la dérive génétique peut, pour une partie même faible du génome, avoir fixé assez d'allèles différents entre ces groupes pour pouvoir être considérée comme l'un des agents de la différenciation ethnique au même titre que la sélection naturelle.

La dérive est, par exemple, la cause la plus vraisemblable de l'absence du groupe sanguin B et de la réduction de la diversité génétique pour le complexe majeur d'histocompatibilité, dans la plupart des populations d'Amérique et d'Océanie.

La révolution néolithique (domestication des espèces végétales et animales) a permis un accroissement considérable des ressources alimentaires. Il s'en est suivi une urbanisation et une inflation des effectifs démographiques qui a assez vite limité les effets de la dérive. Puis le développement économique local, régional, continental, aujourd'hui mondial a conduit, par le commerce, les guerres, les expéditions coloniales, les crises économiques à de multiples mouvements de populations et des mélanges génétiques qui, d'une manière ou d'une autre, induisent depuis trois millénaires un lent mouvement inverse de dédifférenciation ethnique.

Cependant, des études plus récentes et plus fines à partir de marqueurs moléculaires de l'ADN et l'utilisation d'outils mathématiques comme l'analyse de coalescence, ont montré que l'expansion démographique du néolithique (vers – 10 000) a été précédée d'une première expansion (figure 5.5), vers – 35 000 ans. De ce fait, il est vraisemblable que cette expansion au paléolithique récent a, d'une part « figer » en l'état la différenciation acquise depuis l'origine du fait de la puissance de la dérive, d'autre part laisser peu de champ à la dérive dans l'interphase entre les deux expansions démographiques puisque les effectifs étaient déjà conséquents.

5.4.2 Dérive et spéciation

La différenciation entre populations, qu'elle soit induite de manière aléatoire par la dérive, ou de manière dirigée par la sélection (voir plus loin), peut aller à terme jusqu'à la spéciation. Celle-ci est acquise dès que les individus de deux populations ne sont plus interféconds, par suite de modifications majeures affectant leur physiologie, leur morphologie ou leur comportement.

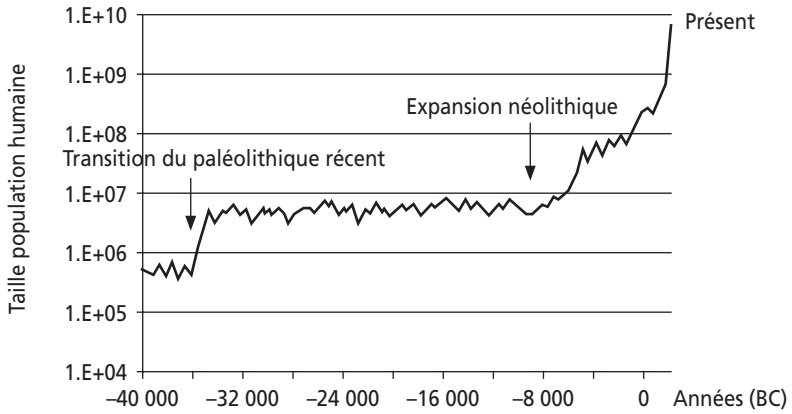


Figure 5.5 Modèle de croissance démographique humaine (Biraben, 1979. D'après Laurent Excoffier).

5.4.3 Dérive et migrations

Tant qu'elle n'a pas conduit à la spéciation, la différenciation génétique induite entre des populations par la dérive est réversible par l'effet des migrations et des mélanges entre ces populations.

Pour estimer l'effet des migrations sur la dérive, nous allons supposer qu'une population d'effectif efficace Ne reçoit à chaque génération une fraction *m* d'individus extérieurs.

Quand on tire un gène dans cette population, une alternative est posée : le gène est d'origine interne avec une probabilité égale à $(1 - m)$ ou externe avec la probabilité m.

Si on reprend les calculs qui ont conduit à la relation de récurrence

$$\alpha_g = [1/N_e]\,(1/2 + \alpha_{g-2}/2) + [1 - 1/N_e]\,\alpha_{g-1}$$

en tenant compte de cette alternative, on obtient :

$$\alpha_g = \mathbf{[(1-m)^4/N_e](1/2 + \alpha_{g-2}/2) + (1-m)^2\,[1 - 1/N_e]}\,\alpha_{g-1}$$

Après quelques simplifications, on montre que la valeur d'équilibre limite de la consanguinité n'est, comme on pouvait s'y attendre, plus égale à 1, mais égale à

$$\alpha_e = \mathbf{1/(\,1 + 4m.N_e\,)}$$

Or $m.N_e$ représente le nombre n_i d'immigrants à chaque génération, d'où :

$$\boldsymbol{\alpha_e = 1/(\,1 + 4n_i)}$$

Ce résultat montre qu'il suffit d'un seul migrant par génération pour réduire de 100 % à 20 % la limite de la consanguinité résultant de la dérive dans une petite population.

Il peut paraître paradoxal qu'un tel résultat soit indépendant de l'effectif efficace et que l'effet d'un seul migrant soit le même quand l'effectif efficace est de 10 ou de 1 000. En fait, si le nombre de migrants conditionne la valeur de la limite, cette

limite est atteinte plus ou moins vite en fonction de la période qui dépend toujours, quant à elle, de l'effectif efficace comme cela est illustré dans la figure 5.6, pour trois populations de période *T1*, *T2* = 2*T1* et *T3* = 3*T1*.

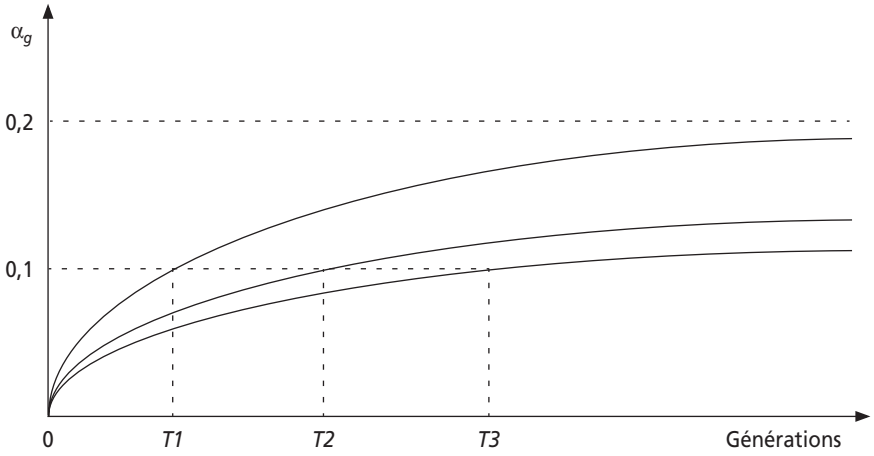


Figure 5.6

RÉSUMÉ

Lorsque l'effectif d'une population est limité, il s'instaure un processus de dérive génétique. Celui-ci est caractérisé par :

– une fluctuation aléatoire des fréquences alléliques ;
– une augmentation récurrente de la consanguinité, malgré la panmixie, en raison de la restriction de l'effectif.

L'étude de ces deux phénomènes montre que la dérive conduit, avec le temps, à une réduction progressive de la diversité génétique, au niveau allélique et, partant génotypique, par la fixation d'un allèle, et l'élimination de tous les autres, pour un nombre croissant de gènes. À terme la population devrait être constituée d'une lignée pure d'individus homozygotes pour tous les gènes.

La vitesse du processus de dérive et d'homogénéisation génétique de la population dépend de l'effectif efficace qui tient compte de la taille de la population mais aussi du sexe ratio entre les effectifs de chacun des sexes prenant effectivement part à la reproduction.

Ce processus est d'autant plus rapide que l'effectif efficace est réduit et que le temps de dérive est long ; il peut être très ralenti, soit par un accroissement de l'effectif, soit par des entrées récurrentes de migrants.

Compte tenu de la taille réelle des populations naturelles et des temps d'isolement à l'échelle de l'évolution, la dérive génétique a pu jouer un rôle dans la différenciation ethnique chez l'homme, et, plus généralement dans la différenciation des espèces, sans pour autant négliger le rôle dévolu classiquement à la sélection naturelle.

EXERCICES

Exercice 5.1

Une petite population, totalement isolée sur le plan migratoire, est constituée en moyenne à chaque génération de 80 organismes de sexe féminin et de 80 organismes de sexe masculin, dont 20 ont une fécondité nulle, parce qu'ils sont exclus du cercle matrimonial.

Calculez le nombre de générations nécessaires pour qu'un tel groupe voit sa consanguinité moyenne atteindre la valeur de 50 %.

Solution

Il y a en fait 80 femmes et 60 hommes qui participent « effectivement » au processus de reproduction, d'où un effectif efficace

$$N_e = 4N_f.N_m/(N_f + N_m)$$

soit $N_e = 137$

Le nombre de générations demandé est égal à la période T, soit

$$T = 2\text{Log}2.N_e$$

soit $T = 192$ générations

Cela fait environ 4000 ans chez l'homme, mais seulement deux siècles chez le chat ou le chien et 50 ans chez la souris.

Exercice 5.2

Les trois figures suivantes (empruntées au site web du cours de Laurent Excoffier, université de Genève) présentent les résultats de 10 simulations de dérive génétiques, initiées avec une fréquence allélique $p_o = 0{,}5$, pour trois populations d'effectif efficace égal à 10, 50 et 500 individus.

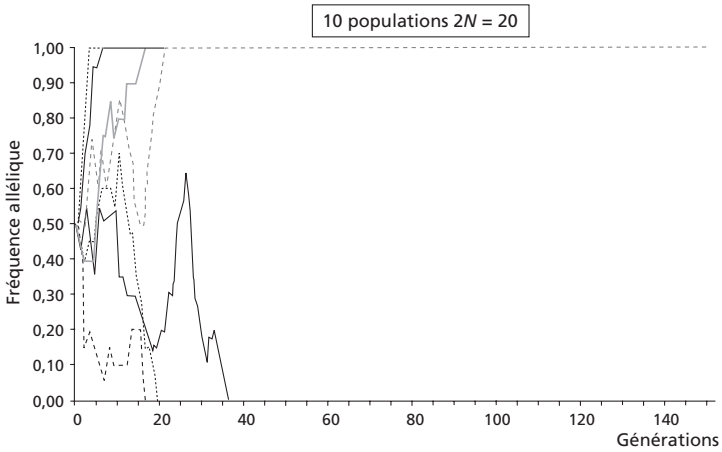
Question 1 : commentez ces résultats du point de vue de la variance allélique entre deux générations.

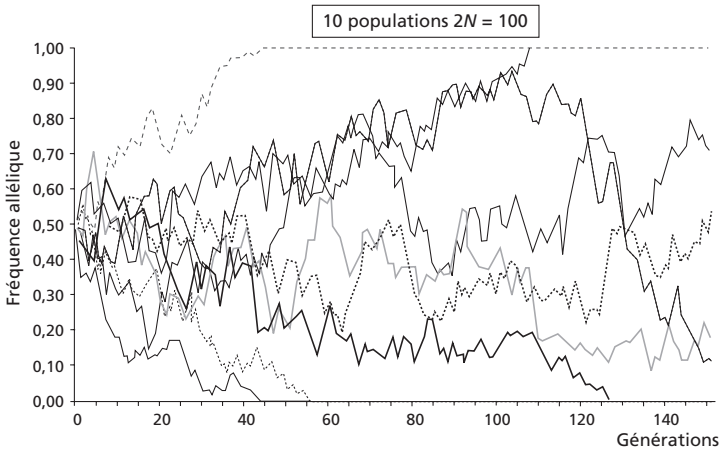
Question 2 : commentez les résultats, du point de vue du nombre de populations ayant atteint, à l'équilibre, la valeur 0 ou 1.

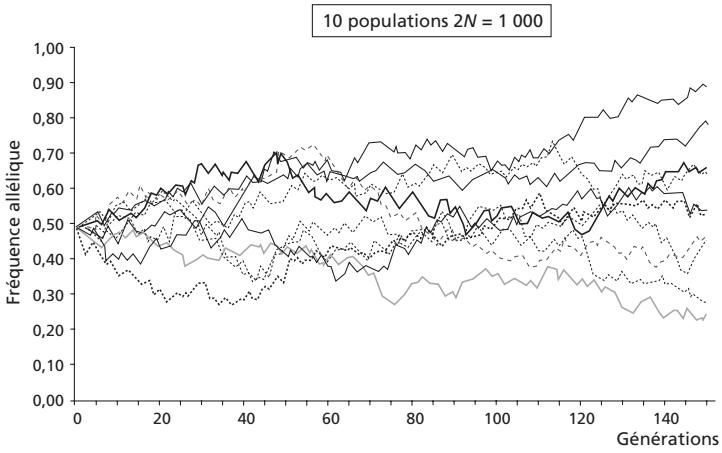
Question 3 : commentez ces résultats du point de vue de la diminution de la vitesse de la dérive génétique

Solution

Question 1 : la variance de la fréquence est $V(p) = p(1 - p)/2Ne$ et est donc d'autant plus grande que Ne est petit : on remarque bien que les changements de fréquence d'une génération à l'autre sont le plus souvent, mais pas systématiquement (hasard d'échantillonnage) de plus grande amplitude quand Ne est plus petit.







Question 2 : parmi les populations de 10 individus ($Ne = 20$), six ont atteint la fixation et quatre ont atteint l'élimination, ce qui est très près des résultats attendus de 5 pour 5 puisque la fréquence initiale est égale à 0,5. Parmi les populations de 50 individus, les résultats observés sont ceux attendus, soit 2 populations ayant fixé l'allèle *A* et trois l'ayant perdu.

Question 3 : malgré la fluctuation des fréquences alléliques, la diversité est maintenue d'autant plus longtemps que l'effectif efficace est élevé : après 20 générations, la diversité a disparu (pour ce gène) des 10 populations de 10 individus et d'aucune des deux autres ; après 120 générations, cinq des 10 populations de 50 individus ont fixé ou perdu l'allèle *A*, et pas une seule des 10 populations de 500 individus.

Exercice 5.3

Quel est le taux de polygamie dans une population où l'effectif efficace est égal à l'effectif d'un des sexes ?

Solution

$N_e = 4N_1.N_2/(N_1 + N_2)$, où N_1 et N_2 sont les effectifs de chacun des deux sexes.

La question posée revient à écrire que

$N_e = 4N_1.N_2/(N_1 + N_2) = N_2$

D'où on tire que $4N_1.N_2 = N_2.(N_1 + N_2)$, et $N_2\,(3N_1 – N_2) = 0$

ce qui signifie qu'il y a trois individus d'un sexe pour un seul de l'autre sexe dans le processus de reproduction.

Exercice 5.4

Pour les gènes localisés sur le chromosome X, dans les espèces hétérogamétiques (on supposera ici qu'il s'agit du mâle) et pour les espèces dont l'un des sexes est haploïde (insectes sociaux par exemple, ici aussi les mâles), la définition et la mesure de l'effectif efficace ne peuvent pas être approchées par l'analyse de la consanguinité qui n'a pas de sens dans ces cas là. Il est alors utile de faire référence à l'*effectif efficace de variance* (5.3.6), défini dans la situation classique d'un gène autosomique et/ou d'une espèce où les deux sexes sont diploïdes.

On rappelle que celui –ci est défini par la variance de la fréquence d'un allèle *A*, soit

$V(p) = p(1 – p)/2N_e$

On note p_f et p_m les fréquences respectives de l'allèle *A* dans les tirages conduisant aux descendants de sexes femelle et mâle.

On note N_f et N_m les effectifs de chacun des deux sexes et on précise que le nombre de gamètes utiles à leur formation est égal à $2N_f$ dans le premier cas mais seulement N_m dans le second cas.

Question 1 : écrire les variances de la fréquence de *A*, dans chacun des sexes, en fonction de p_f et p_m et de N_f et N_m.

Question 2 : quelle est, en supposant un sexe ratio équilibré, la fréquence moyenne p, tous sexes confondus, de l'allèle *A* dans l'effectif total ?

Question 3 : combinez les résultats précédents pour déterminer $V(p)$, la variance de p, fréquence de l'allèle *A*, dans l'effectif total, tous sexes confondus (on négligera la covariance).

Question 4 : que devient cette variance, à l'équilibre, quand les fréquences alléliques p_f et p_m sont égales entre les sexes, c'est-à-dire égales à p ? Tirez-en, par analogie avec l'effectif efficace de variance, l'effectif efficace pour un gène du X, présent en un exemplaire chez les mâles et en deux exemplaires chez les femelles.

Question 5 : rappelez la formule classique de N_e en fonction de N_f et N_m et précisez la valeur de *Ne* quand le sexe ratio est équilibré (effectifs des sexes égaux à $N/2$), dans ce cas et dans le cas d'un gène lié à l'X ; tirez en les conséquences en termes d'efficacité de la dérive.

Question 6 : rappelez la formule classique de N_e en fonction de N_f et N_m et précisez la valeur de *Ne* quand le sexe ratio est déséquilibré, avec un seul représentant d'un des sexes, et précisez quelle la situation pour les insectes sociaux.

Solution

Question 1 : dans le sexe femelle, on aura $V_f(f) = p_f(1 - p_f)/2N_f$
puisqu'on tire $2N_f$ gamètes avec une probabilité p_f d'avoir l'allèle *A*
Dans le sexe mâle, on aura $V_m(f) = p_m(1 - p_m)/N_m$
puisqu'on tire N_m gamètes avec une probabilité p_m d'avoir l'allèle *A*

Question 2 : $p = 2p_f/3 + p_m/3$ (voir 3.3.1)

Question 3 : $V(p) = V(2p_f/3 + p_m/3)$

d'où $V(p) = 4/9[V(p_f)] + 1/9[V(p_m)]$

soit $V(p) = 4/9[p_f(1 - p_f)/2N_f] + 1/9[p_m(1 - p_m)/N_m]$

ou en simplifiant $V(p) = 2/9[p_f(1 - p_f)/N_f] + 1/9[p_m(1 - p_m)/N_m]$

Question 4 : $V(p) = 2/9[p(1 - p)/N_f] + 1/9[p(1 - p)/N_m]$

soit $V(p) = pq\,[2/9N_f + 1/9N_m]$

Par analogie avec $V(p) = pq/2N_e$

on tire que $1/2Ne = 2/9N_f + 1/9N_m$

soit $\boldsymbol{Ne = 9N_f.N_m/(2N_f + 4N_m)}$

Question 5 :

Relation classique : $Ne = 4N_f.N_m/(N_f + N_m)$

Si on a $N_f = N_m = N/2$ alors $N_e = N$

Avec la relation $Ne = 9N_f.N_m/(2N_f + 4N_m)$

Si on a $N_f = N_m = N/2$ alors $N_e = 3/4\ N$

La dérive est donc plus efficace pour les gènes liés à l'X car l'effectif efficace est plus faible pour ces gènes que pour les gènes autosomiques.

Question 6 :

Relation classique : $Ne = 4N_f.N_m/(N_f + N_m)$

Si on a, par exemple $N_m = 1$ alors $N_e = 4$

La dérive est équivalente à celle d'une population formée de deux mâles et deux femelles.

Chez les insectes sociaux, non seulement un sexe est réduit à un individu diploïde, la reine, mais les mâles sont tous haploïdes, c'est-à-dire comme si tous leurs gènes étaient dans la situation des gènes liés à l'X. Il convient donc de prendre la formule

$$N_e = 9N_f.N_m/(2N_f + 4N_m)$$

avec $N_f = 1$ alors $N_e = 9N_m/(2 + 4N_m)$

soit, si 2 est négligeable devant Nm : $N_e = 9N_m/4N_m \# 2$

La dérive est donc encore plus efficace chez les insectes sociaux puisque deux facteurs jouent dans le même sens, l'haploïdie des mâles et le déséquilibre du sexe ratio ; elle est équivalente à celle de croisements frères-sœurs systématiques (voir 4.2.3.c).

Chapitre **6**

Mutations et migrations

6.1 INTRODUCTION

Le modèle de Hardy-Weinberg suppose qu'on puisse, sur quelques générations, négliger l'effet des mutations ou des migrations. Mais il est évidemment irréaliste, à long terme, de négliger les mutations qui ont généré l'évolution des gènes, des génomes, donc du vivant, et souvent à court terme, les migrations qui modèlent chez beaucoup d'espèces les variations du patrimoine génétique des populations.

Il est donc utile de pouvoir mesurer l'effet des mutations et des migrations sur la composition génétique des populations, ne serait-ce que pour justifier de négliger leurs effets sur quelques générations, quand c'est possible, ou bien pour évaluer dans quel sens et à quel rythme les mutations et les migrations peuvent modifier cette composition génétique.

Ce chapitre se contentera de présenter deux cas simples, renvoyant le lecteur intéressé à des ouvrages[1] débordant, par leur difficulté, le cadre que nous nous sommes fixé.

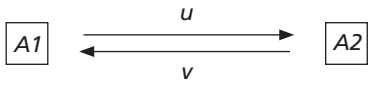
6.2 MUTATIONS RÉCIPROQUES

6.2.1 Définition et approche intuitive

Les mutations sont des changements de l'information génétique codée par l'ADN. Ce changement peut affecter la séquence d'un gène et altérer sa fonction, mais une mutation peut aussi n'induire qu'un polymorphisme sans conséquence fonctionnelle immédiate, que cette mutation touche ou non la séquence d'un gène.

1. *Génétique et évolution*, Solignac, Perriquet, Anxolabérère et Petit, Hermann, Paris, 1995.

Formellement, sans aucune considération fonctionnelle, la mutation d'un gène diallélique peut s'écrire ainsi :



où u et v représentent respectivement les probabilités ou taux de mutations de *A1* vers *A2* et réciproquement de *A2* vers *A1*.

On conçoit bien qu'avec des taux u et v non nuls, la diversité génétique sera maintenue, avec des fréquences alléliques qui devraient être fonction de u et v. En effet si $u = v$, on devrait, à terme, avoir des allèles équifréquents. Mais quelle sera la limite exacte des fréquences alléliques si u et v sont différents, et avec quelle vitesse ces valeurs limites seront-elles atteintes ? Ces questions n'ont de réponse qu'à travers une formulation mathématique du problème.

6.2.2 Formulation mathématique

Si, à la génération i, les fréquences des allèles *A1* et *A2* sont respectivement égales à p_i et q_i, quelles seront leurs valeurs à la génération suivante $i + 1$?

Dans les conditions du modèle de Hardy-Weinberg, la panmixie permet d'appliquer le schéma de l'urne gamétique. Les fréquences alléliques à la génération suivante $i + 1$, sont égales aux fréquences gamétiques qui, en l'absence de mutations, sont elles-mêmes égales aux fréquences alléliques à la génération i.

En présence de mutations, la composition de l'urne gamétique dépendra d'une part des fréquences alléliques p_i et q_i à la génération i, mais aussi des taux u et v de mutations.

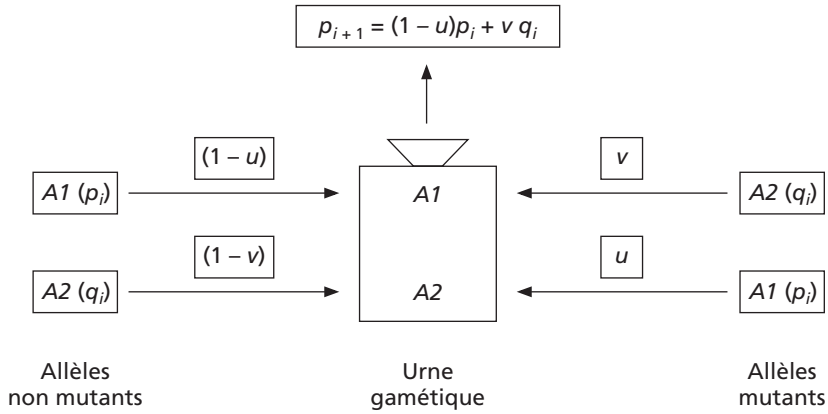


Figure 6.1 Effet des mutations réciproques.

Ainsi « tirer un allèle *A1* dans l'urne » correspond (figure 6.1) au tirage d'un allèle *A1* de la génération précédente (de fréquence p_i) qui n'aurait pas muté (probabilité égale à $1 - u$) ou à un allèle *A2* (de fréquence q_i) qui aurait muté en *A1*

(probabilité *v*). La probabilité de tirage de ce gamète *A1* est égale à sa fréquence à la génération $i + 1$, d'où :

$$p_{i+1} = (1 - u)\, p_i + v\, q_i$$

et $$q_{i+1} = (1 - v)\, q_i + u\, p_i$$

Si p_e est la fréquence d'équilibre, p_e vérifie la récurrence de manière telle que

$$p_e = (1 - u)\, p_e + v\,(1 - p_e)$$

d'où $$p_e = v/(u + v)$$

et $$q_e = u/(u + v)$$

Le même résultat peut être obtenu en écrivant l'équation de l'écart des fréquences entre deux générations, soit :

$$\Delta p = p_{i+1} - p_i = (1 - u)\, p_i + v\, q_i - p_i$$

d'où $$\Delta p = v - (u + v)\, p_i$$

À l'équilibre, on a $\Delta p = 0$ et $p_e = v/(u + v)$

6.2.3 Limite du processus et conséquences génétiques

Il existe donc un équilibre polymorphe de la composition génétique de la population. Les mutations, quand elles sont réciproques, ne suppriment pas le polymorphisme mais le modèlent en fonction des valeurs des taux respectifs de mutation d'un allèle vers un autre. À l'équilibre on a $\Delta p = -\Delta q$, ce qui signifie une égalité des flux : le nombre d'allèles *A1* transformés en allèles *A2* est égal, par unité de temps, au nombre d'allèles *A2* transformés en *A1*. C'est pourquoi $p = q$, si $u = v$, tandis que $p = 2q$, si $u = v/2$.

Si l'un des taux de mutations est nul, par exemple *v*, l'un des allèles disparaîtra, dans ce cas *A1*.

Enfin de nouveaux allèles A3, A4, etc. peuvent apparaître et augmenter le polymorphisme génétique. Ce faisant les nouveaux allèles définissent un nouvel équilibre polymorphe induisant une nouvelle évolution des fréquences alléliques, si celles-ci étaient à l'équilibre.

6.2.4 Vitesse du processus

Quand un facteur déterministe, ou aléatoire (voir la dérive), induit une variation des fréquences alléliques, il ne suffit pas de définir la valeur limite des fréquences alléliques à l'équilibre, il convient aussi d'estimer la vitesse avec laquelle la composition génétique de la population tend vers cet équilibre.

Pour ce faire, on a écrit la relation de récurrence sous la forme d'un écart entre la variable de cette relation et la valeur limite de cette variable.

Dans le cas des mutations réciproques, la fréquence allélique tend vers une valeur limite égale à $v/(u+v)$. On va donc étudier l'écart p_i à sa limite $v/(u + v)$ qui s'écrit :

$$p_i - v/(u + v) = (1 - u)\, p_{i-1} + v\, q_{i-1} - v/(u + v)$$

soit $$p_i - v/(u + v) = (1 - u)\, p_{i-1} + v\,(1 - p_{i-1}) - v/(u + v)$$

d'où, en mettant p_{i-1} en facteur : $p_i - v/(u+v) = (1-u-v)\, p_{i-1} + v - v/(u+v)$

puis $v/(u+v)$ en facteur : $p_i - v/(u+v) = (1-u-v)\, p_{i-1} + [u + v - 1]\, v/(u+v)$

et $(1-u-v)$ en facteur : $p_i - v/(u+v) = (1-u-v)\, [p_{i-1} - v/(u+v)]$

L'écart entre la fréquence allélique et sa valeur limite est réduit d'un facteur $(1-u-v)$ à chaque génération. Sous cette forme, la récurrence admet une solution algébrique simple :

$$\boldsymbol{p_i - v/(u+v) = (1-u-v)^i\, [p_0 - v/(u+v)]}$$

qui permet d'estimer la vitesse avec laquelle on tend vers la valeur d'équilibre.

On définit un temps T, tel que l'écart à la limite soit réduit de moitié ; par définition T (équivalant à la période) est tel que :

$$p_T - v/(u+v) = [p_0 - v/(u+v)]/2$$

Par ailleurs T vérifie la relation de récurrence, d'où :

$$p_T - v/(u+v) = (1-u-v)^T\, [p_0 - v/(u+v)]$$

En identifiant les deuxièmes membres de ces deux égalités, on écrit que

$$[p_0 - v/(u+v)]/2 = (1-u-v)^T\, [p_0 - v/(u+v)]$$

d'où $1/2 = (1-u-v)^T$

et, en passant en Log : $-\mathrm{Log}2 = T.\mathrm{Log}\,(1-u-v)$

Sachant que $\mathrm{Log}\,(1+\varepsilon)$ et $\mathrm{Log}(1-\varepsilon)$ sont respectivement peu différents de ε et $-\varepsilon$, on peut écrire alors que :

$$-\mathrm{Log}2 = -T(u+v)$$

d'où $T = (\mathrm{Log}2)/(u+v)$

soit $T = 0{,}7/(u+v)$

Les taux de mutations sont toujours des paramètres très faibles, de l'ordre de 10^{-5} pour les plus forts, à 10^{-10}, pour les plus faibles, ce qui donne des vitesses d'évolution extrêmement lentes.

Exemple 6.1

Considérons des taux médians comme $u = 3 \,.\, 10^{-6}$ et $v = 2 \,.\, 10^{-6}$. Les valeurs limites des fréquences allélique à l'équilibre seront $p = 3/5$ et $q = 2/5$, avec une vitesse telle que l'écart à la limite sera réduit de moitié toutes les T générations, T étant égal à $0{,}7\ (5.\ 10^{-6})$ soit 140 000 générations. Pour la drosophile cela représente environ 14 000 ans, mais pour l'homme moderne cela représente 2,8 millions d'années, soit de dix à vingt fois son âge !

La vitesse d'évolution des fréquences alléliques sous l'effet des taux de mutations est si faible qu'il est tout à fait légitime de négliger ce facteur sur quelques générations dans l'application du modèle de Hardy-Weinberg.

On peut aussi considérer qu'à l'échelle de l'évolution des espèces l'équilibre est rarement atteint, non seulement parce que la vitesse est excessivement lente, mais surtout parce que de nouveaux allèles apparaissent, qui définissent un nouvel équilibre de fréquences alléliques que les populations n'auront jamais le temps d'atteindre !

Exemple 6.2

Considérons le cas où l'un des taux est nul, par exemple : $u = 3 . 10^{-6}$ et $v = 0$: des allèles *A1* peuvent être mutés en *A2* et jamais le contraire. La valeur limite des fréquences alléliques sera logiquement $p = 0$ et $q = 1$.

Cependant le temps T nécessaire à la réduction de moitié de l'écart à la limite est égal à $0,7/3 . 10^{-6}$. Si la l'allèle *A1* est au départ très fréquent, voire unique (p = 1), sa fréquence sera égale à 1/2 après 233 000 générations, soit 4,6 millions d'années chez l'homme. C'est un exemple dont on se souviendra lors de l'étude de l'effet dysgénique de la médecine (chapitre 8).

6.3 MIGRATIONS UNIDIRECTIONNELLES

6.3.1 Définition et approche intuitive

Les migrations mélangent entre elles des populations pouvant présenter des différences génétiques plus ou moins importantes. De ce fait les populations diffèrent moins entre elles après les migrations qu'avant.

Cependant la formulation mathématique de ces phénomènes est très complexe, car il ne s'agit pas simplement de mélanges de type pot-pourri (*melting-pot*) où la panmixie s'instaure après le mélange.

On peut, par exemple, avoir un mélange faible mais récurent à la zone de contact entre deux populations, et étudier la diffusion des gènes extérieurs dans l'ensemble de chacune des populations à partir de cette zone de contact.

Ce problème de diffusion est aussi lié au fait que la panmixie n'est possible que dans un espace compatible avec les déplacements de l'individu et que, même en l'absence de toute contrainte sociale ou culturelle, des individus séparés par une grande distance ont moins de chance de former un couple que deux individus peu éloignés, ce qui aboutit à une structuration un peu particulière de la population, dans l'espace et dans le temps, conduisant sur l'ensemble de l'aire de répartition à une effet de type Walhund (chapitre 4).

On se contentera ici de l'étude d'un modèle simple appelé « modèle de l'île » parce qu'une population (l'île) reçoit à chaque génération des migrants d'une autre population (le continent), sans qu'il y ait de migrations en sens opposé. Du fait de ces migrations unidirectionnelles, la population continentale, à l'équilibre de Hardy-Weinberg, conserve sa composition génétique, et la population de l'île, qui reçoit à chaque génération des gènes du continent finira par avoir la même composition génétique que la population continentale. La question qui reste sans réponse, en l'absence de formulation mathématique est : au bout de combien de temps ?

6.3.2 Formule de récurrence

On considère une population continentale où les fréquences des allèles *A1* et *A2* sont respectivement égales à *P* et *Q*.

Si, dans la population de l'île, à la génération i, les fréquences des allèles *A1* et *A2* sont respectivement égales à p_i et q_i, quelles seront leurs valeurs à la génération suivante $i + 1$, sachant qu'à chaque génération une proportion m d'individus ont migré depuis le continent ?

Dans les conditions du modèle de Hardy-Weinberg, la panmixie permet d'appliquer le schéma de l'urne gamétique. Les fréquences alléliques à la génération suivante $i + 1$, sont égales aux fréquences gamétiques qui, en l'absence de migrations, sont elles-mêmes égales aux fréquences alléliques à la génération i.

En présence de migrations, la composition de l'urne gamétique dépendra d'une part des fréquences alléliques p_i et q_i à la génération i, pour les gamètes fournis par les individus de l'île, mais aussi des fréquences P et Q, pour la proportion m des migrants nouvellement arrivés et contribuant à l'urne gamétique,.

Ainsi « tirer un allèle *A1* dans l'urne » correspond (figure 6.2) au tirage d'un allèle *A1*(de fréquence p_i) fourni par un individu de l'île à la génération précédente (dans une proportion $1 - m$) ou à un allèle *A1* (de fréquence P) fourni par un migrant (dans une proportion m).

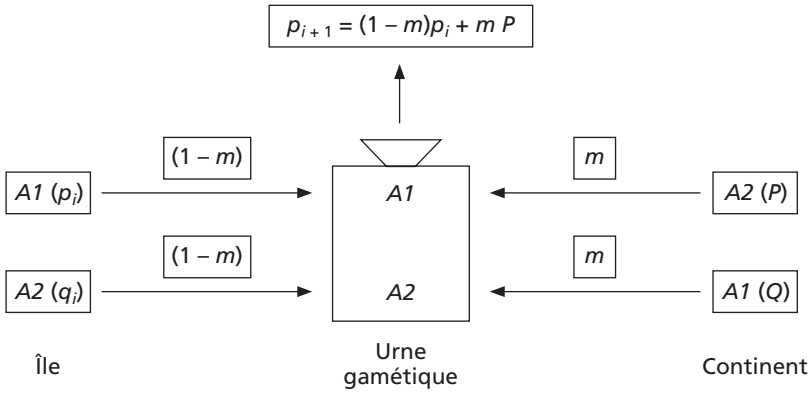


Figure 6.2 Effet des migrations (modèle de l'île).

La probabilité de tirage de ce gamète *A1* est égale à sa fréquence à la génération $i + 1$, d'où :

$$p_{i+1} = (1 - m)\, p_i + mP$$

et

$$q_{i+1} = (1 - m)\, q_i + mQ$$

Si p_e est la fréquence d'équilibre, p_e vérifie la récurrence de manière telle que

$$p_e = (1 - m)\, p_e + mP$$

d'où

$$\boldsymbol{p_e = P}$$

et

$$\boldsymbol{q_e = Q}$$

Le même résultat peut être obtenu en écrivant l'équation de l'écart des fréquences entre deux générations, soit :

$$\Delta p = p_{i+1} - p_i = (1 - m)\, p_i + mP - p_i$$

d'où

$$\Delta p = -\, m\,(p_i - P)$$

À l'équilibre, on a $\Delta p = 0$ et $p_e = P$

6.3.3 Limite du processus et conséquences génétiques

À l'équilibre, comme on pouvait s'y attendre, la composition génétique de l'Île est devenue identique à celle du continent, et il est très important de faire la différence entre ce processus et le résultat d'un mélange entre les deux populations.

Dans un mélange les différences entre populations disparaissent car les deux populations fusionnent en une seule ; tous les gènes et leurs allèles sont mis en commun pour fonder un nouveau patrimoine génétique (voir l'exemple de fusion de populations au chapitre 3).

Dans le processus de migrations unidirectionnelles les différences entre populations disparaissent, sans que, au moins dans un premier temps, les deux populations fusionnent en une seule. Dans ce premier temps, avant que les deux populations ne soient devenues presque génétiquement identiques, les gènes et leurs allèles n'auront pas été mis en commun pour fonder un nouveau patrimoine génétique et de nombreux allèles de la population de l'île auront disparu au profit d'allèles continentaux.

Une population qui contient quatre allèles *A1*, *A2*, *A3* et *A4* et qui reçoit des migrants d'une population qui ne contient que des allèles *A2* et A4, verra inexorablement la fréquence des allèles *A1* et *A3* diminuer jusqu'à devenir nulle. Les allèles *A1* et *A3* auront disparu ce qui ne serait pas advenu si les deux populations s'étaient mélangées.

C'est en ce sens qu'on peut dire que les migrations unidirectionnelles aboutissent à « l'éradication du particularisme génétique insulaire », dont la population noire américaine constitue un exemple (voir 6.3.5).

6.3.4 Vitesse du processus

Comme dans le cas des mutations réciproques la relation de récurrence est réécrite sous la forme d'un écart entre la variable de cette relation et la valeur limite de cette variable. Dans le cas des migrations, la fréquence allélique tend vers une valeur limite égale à *P*. On va donc écrire que

$$p_i - P = (1 - m)\, p_{i-1} + mP - P$$

soit :

$$p_i - P = (1 - m)\, p_{i-1} - (1 - m)P$$

d'où, en mettant $(1 - m)$ en facteur :

$$p_i - P = (1 - m)\, [p_{i-1} - P]$$

L'écart de la fréquence allélique à sa valeur limite est réduit d'un facteur $(1 - u - v)$ à chaque génération. Sous cette forme, la récurrence admet une solution algébrique simple :

$$\boldsymbol{p_i - P = (1 - m)^i\, [p_0 - P]}$$

qui permet d'estimer la vitesse avec laquelle on tend vers la valeur d'équilibre.

On définit un temps *T*, tel que l'écart à la limite soit réduit de moitié ; par définition *T* (équivalant à la période définie pour la dérive) est tel que :

$$p_T - P = [p_0 - P]/2$$

Par ailleurs, *T* vérifie la relation de récurrence, d'où :

$$p_T - P = (1 - u - v)^T\, [p_0 - P]$$

En identifiant les deuxièmes membres de ces deux égalités, on écrit que

$$[p_0 - P]/2 = (1 - u - v)^T [p_0 - P]$$

d'où $$1/2 = (1 - m)^T$$

et, en passant en Log : $$-\text{Log}2 = T.\text{Log}(1 - m)$$

Sachant que Log (1+ε) et Log(1 – ε) sont respectivement peu différents de ε et – ε, on peut écrire alors que :

$$-\text{Log}2 = -T(m)$$

d'où $$T = (\text{Log}2)/(m)$$

soit $$\boldsymbol{T = 0{,}7/m.}$$

On remarquera la symétrie des formules traitant des migrations et des mutations réciproques. Mais les taux de migrations sont des paramètres beaucoup plus élevés que les taux de mutations, de l'ordre de 10^{-1} pour les plus forts à 10^{-5}, pour les plus faibles, ce qui donne des vitesses d'évolution beaucoup moins lentes (tableau 6.1).

Tableau 6.1 Temps *T* de réduction de moitié de l'écart à l'équilibre, en générations et en années, pour la drosophile et l'homme, et pour divers taux de mutations ou de migrations.

Paramètre de	Migrations : *m*				Mutations : *u* + *v*		
Valeur des paramètres	10^{-1}	10^{-2}	10^{-3}	10^{-4}	2.10^{-5}	2.10^{-6}	2.10^{-7}
Valeur de T (générations)	7	70	700	7 000	35 000	350 000	3 500 000
Équivalent années (drosophile)	1	7	70	700	3 500	35 000	350 000
Équivalent années (homme)	140	1 400	14 000	140 000	700 000	7 000 000	70 000 000

Le tableau 6.1 montre que les migrations ont pu jouer un rôle déjà considérable dans l'histoire génétique de l'humanité, contrairement aux mutations. Pour les insectes, comme la drosophile, à la fois parce qu'ils sont plus anciens sur la terre et que leur temps de génération est plus court, mutations et migrations ont participé à l'évolution de leur stock génique.

6.3.5 L'exemple de la population noire des États-Unis

Les États-Unis d'Amérique ne sont pas un melting pot mais une agrégation de communautés pour ne pas dire de ghetto, qui conservent, malgré une adhésion plus ou moins forte au drapeau et aux valeurs communes qui fondent la nation, une culture, une langue, une endogamie et par là-même leurs gènes.

Le modèle de l'île s'applique à la population noire des États-Unis car, en raison de la discrimination raciale, qui existe non en droit mais en fait, tout enfant issu (volontairement ou non) d'un couple mixte est considéré comme appartenant à la communauté noire. De ce fait, le transfert de gènes ne peut s'effectuer que des blancs vers les noirs et jamais en sens inverse !

La conclusion du paragraphe 6.3.3 s'applique ici : le transfert unidirectionnel de gènes devrait aboutir, à termes, à la perte de la spécificité génétique de la population noire. En d'autres termes le comportement sociologique correspondant à la discrimination raciale a, pour conséquence génétique, une disparition génétique de cette population non par fusion des patrimoines mais par remplacement des gènes « noirs » par des gènes « blancs ».

Évidemment, on peut et on doit imaginer, qu'avec le temps le comportement sociologique peut se modifier et qu'une fusion puisse se faire à travers une panmixie généralisée. Mais, à ce moment, une partie importante de l'apport africain aura été éliminé, malgré la différence du taux de fécondité.

On peut maintenant essayer d'estimer le taux *m* d'apport migratoire blanc dans la population noire et d'en tirer une estimation de la vitesse *T* de remplacement des gènes.

Cette étude célèbre a été faite par Glass et Li en 1953 (*Am. J. Hum*. Genet. 1953, 5, 1-20). Ils ont choisi d'étudier l'allèle *Ro* du système rhésus car il est très rare en Europe alors qu'il est très fréquent en Afrique, ce qui facilite l'estimation de *m*.

L'équation de récurrence du modèle de l'île, sous sa forme algébrique, s'écrit :

$$p_i - P = (1 - m)^i \, [p_0 - P]$$

où :

- p_0 est la fréquence de l'allèle *Ro* dans la population noire, à l'origine ;
- p_i est la fréquence de l'allèle *Ro*, dans la population des États-Unis (l'île), *i* générations après la génération initiale ;
- *P* est la fréquence de l'allèle *Ro* dans la population blanche ;
- *m* est le taux de migration, ou de remplacement, dont on souhaite estimer la valeur.

En 1953, la fréquence de l'allèle *Ro* est égale à **0,446** dans la population noire des États-Unis et fournit une estimation de $\boldsymbol{p_i}$

La valeur de ***i*** peut être estimée à **10** générations, la traite des esclaves remontant au début du XVIII^e^ siècle, quelques 250 ans avant 1953.

La valeur $\boldsymbol{p_0}$ de la fréquence de *Ro* à l'origine n'est pas connue. Mais on peut considérer que la fréquence de cet allèle n'a pas changé depuis le XVIII^e^ siècle dans les diverses les populations d'origines des esclaves, où sa valeur moyenne actuelle est égale à **0,63**.

Enfin la fréquence ***P*** de l'allèle *Ro* dans la population blanche (le continent) peut être considérée comme constante dans le temps car cet allèle est très rare et pratiquement équifréquent dans toutes les populations européennes qui ont migré vers les États-Unis. La fréquence actuelle de *Ro* en Europe ou chez les blancs américains, invariante depuis le XVIII^e^ siècle, est égale à **0,028**.

D'où l'équation, après remplacement des paramètres par leurs valeurs respectives :

$$0{,}446 - 0{,}028 = (1 - m)^{10}\,[0{,}63 - 0{,}028\,]$$

dont on tire $\boldsymbol{m = 3{,}6\ \%}$.

Ceci signifie qu'à chaque génération, en moyenne 3,6 % des gènes de la population noire sont remplacés par des gènes de la population blanche.

La vitesse du processus, mesurée par le temps nécessaire pour réduire de moitié l'écart à la limite est égal à

$$T = 0{,}7/m$$

soit **T = 20 générations**, c'est-à-dire 5 à 6 siècles.

En effet, il a suffi de presque trois siècles pour passer de 0,63 à 0,446, ce qui montre l'efficacité des migrations dans la variation de la composition génétique des populations. Cependant 20 générations ou cinq siècles sont encore importants, non à l'échelle de l'histoire, mais à celle de la vie humaine. Par ailleurs, il demeure impossible de savoir ce qu'il adviendra de la réalité du modèle de l'île pour la population noire des États-Unis ; les deux populations blanches et noires peuvent finir par fusionner avant que la population noire n'ait génétiquement disparu. Mais, en tout état de cause, au moment de la fusion, la population noire aura déjà « perdu » une bonne partie de sa spécificité génétique d'origine.

RÉSUMÉ

Les migrations réciproques maintiennent le polymorphisme génétique des populations. Les fréquences d'équilibre dépendent exclusivement de la valeur des taux de mutation.

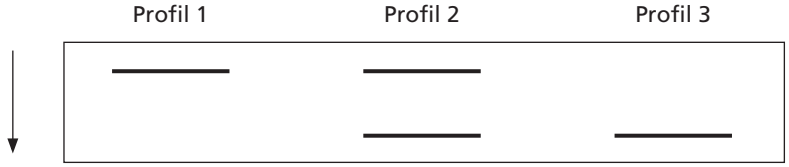
L'équilibre est atteint avec une vitesse très lente parce qu'elle dépend des taux de mutations qui sont eux-même très petits. De ce fait l'équilibre peut ne jamais être atteint en pratique, d'autant, qu'avec le temps, de nouveaux allèles apparaissent et définissent un nouvel équilibre.

Une population qui reçoit, de manière répétée, des migrants d'une autre population, sans que des migrations aient lieu en sens inverse, verra sa spécificité génétique disparaître progressivement. À terme, sa composition génétique sera identique à celle de la population d'origine des migrants. La vitesse de ce processus, bien que plus rapide que celui des mutations, est encore assez lent à l'échelle de la vie humaine ou de l'histoire pour rendre peu probable un maintien permanent des migrations unidirectionnelles et de leurs conséquences génétiques.

EXERCICES

Exercice 6.1 Drosophile des caves et drosophile des champs

Chez *Drosophila melanogaster*, on s'intéresse à l'étude des variants électrophorétiques relatifs au gène codant pour l'alcool deshydrogénase. Cette enzyme est codée par un gène autosomal et intervient dans la détoxification des alcools. Cette étude a révélé le polymorphisme suivant :



L'analyse de deux populations de Bourgogne, l'une située près d'une cave, l'autre dans un champ a donné les résultats suivants :

	Profil 1	Profil 2	Profil 3
Population Cave	140	200	60
Population Champ	60	140	200

Question 1 : interprétez les différents phénotypes de mobilité électrophorétique. Estimez les fréquences génotypiques et les fréquences alléliques.

Question 2 : les fréquences génotypiques sont-elles les mêmes dans les deux populations ?

Justifiez votre réponse par un test statistique. Que peut-on imaginer comme mécanisme pour expliquer ce résultat ?

Question 3 : les fréquences observées dans chacune des populations sont elles conformes à l'hypothèse de Hardy-Weinberg ? Justifiez votre réponse par un test statistique.

Question 4 : par techniques de marquage/recapture, des drosophiles marquées dans la population de cave ont été recapturées dans les champs. L'inverse n'a pas été trouvé. Comment peut-on interpréter ce résultat ? Que peut-on tirer de ce résultat pour expliquer la réponse à la question 3. Aucune des réponses ne nécessite de formules mathématiques.

Solution

Question 1 : Les phénotypes de mobilité électrophorétique mettent en évidence au moins deux espèces moléculaires (plus si certaines ont un même coefficient de migration et donnent des bandes confondues), de même fonction, puisque la révélation de la protéine dans le gel est spécifique de sa fonction (le gel contenant évidemment des centaines de protéines différentes), mais différant par leur mobilité électrophorétique.

Cette observation définit pour le gène codant la protéine étudiée deux allèles. La protéine semble être monomérique car, dans le cas contraire, on attendrait un profil avec trois bandes pour l'hétérozygote.

Ces allèles seront nommés *Er* (pour rapide) et *El* (pour lent). Les profils 1, 2 et 3 correspondent respectivement aux homozygotes *El*/*El*, *Er*/*El* et *Er*/*Er*.

Les fréquences génotypiques sont égales aux fréquences phénotypiques (phénotypes codominants) et les fréquences alléliques s'en déduisent directement :

	Profil 1 ***El*/*El***	**Profil 2** ***El*/*Er***	**Profil 3** ***Er*/*Er***	**Fréquence de l'allèle *El***	**Fréquence de l'allèle *Er***
Population Cave	0,35	0,5	0,15	0,60	0,40
Population Champ	0,15	0,35	0,5	0,325	0,675

Question 2 : homogénéité des deux populations

On construit un tableau de contingence ci-dessous pour réaliser le test homogénéité (voir chapitre 2), où les effectifs théoriques sont en italiques

La variable de χ^2 définie ici a $(n - 1)(m - 1) = 2$ degrés de liberté.

La valeur d'un χ^2 à 2 ddl, qui n'est dépassée que 5 fois sur 100 est égale 5,99.

	Profil 1 ***El*/*El***	**Profil 2** ***El*/*Er***	**Profil 3** ***Er*/*Er***	**Sommes marginales**
Population Cave	140 *100*	200 *170*	60 *130*	400
Population Champ	60 *100*	140 *170*	200 *130*	400
Sommes marginales	200	340	260	800

La valeur observée du χ^2 est égale 117,97 ! Elle est très supérieure à 5,99, ce qui signifie qu'elle avait une probabilité très inférieure à 5 % d'être observée par le simple hasard échantillonnage, si l'hypothèse nulle d'homogénéité était vraie.

On peut donc rejeter l'hypothèse nulle, avec un risque très inférieur à 5 %, et conclure que les deux échantillons n'étant pas homogènes, les populations de cave et de champ ont des compositions génétiques très différentes.

Il est assez difficile d'imaginer un mécanisme sans avoir étudié la variation de la composition génétique de chacune des populations dans le temps.

D'abord, elles peuvent être ainsi, sans autre raison que le hasard des fondatrices qui les ont reconstituées après l'hiver.

Si il y a un effet sélectif, il peut jouer dans l'une ou l'autre ou les deux populations. Par exemple, sachant que l'alcool déshydrogénase joue un rôle dans la détoxification, on peut supposer (ce qui reste à démontrer) qu'un avantage sélectif joue en faveur de l'allèle *El*, dans un environnement « alcoolique ». Mais dans ce cas (voir

chapitre sur la sélection), on pourrait s'attendre à la disparition de l'allèle défavorable, ce qui est peut-être en cours. Pour le savoir, il faudrait suivre, dans le temps, l'évolution de la composition génétique de ces populations, d'autant que des migrations peuvent modifier ce que la sélection aurait réalisé.

Question 3 : test de l'équilibre de Hardy-Weinberg

En multipliant les fréquences p^2, $2pq$ et q^2 calculées dans chaque échantillon par la taille de celui-ci (400), on obtient les effectifs attendus sous l'hypothèse de l'équilibre panmictique de Hardy-Weinberg (en italiques)

La variable de χ^2 définie pour chacune des populations a 1 degré de liberté.

La valeur d'un χ^2 à 1 ddl, qui n'est dépassée que 5 fois sur 100 est égale 3,84.

	Profil 1 *El*/*El*	Profil 2 *El*/*Er*	Profil 3 *Er*/*Er*	Fréquence de l'allèle *El*	Fréquence de l'allèle *Er*	Valeur du χ^2 observé
Population Cave	140 144	200 192	60 64	0,60	0,40	0,694
Population Champ	60 42,25	140 175,5	200 182,25	0,325	0,675	16,36

La valeur observée du χ^2 est égale à 0,694 dans la première population et à 16,36 dans la seconde, ce qui signifie que l'hypothèse de l'équilibre de Hardy-Weinberg est largement acceptable dans la population de cave et inacceptable dans celle des champs.

Ainsi la population de cave semble être à l'équilibre de Hardy-Weinberg, ce qui signifierait notamment, soit l'absence de sélection, soit une sélection avec avantage de l'hétérozygote et une composition génétique déjà en situation d'équilibre final (voir chapitre 7 sur la sélection).

La population de champ n'est pas à l'équilibre, ce qui signifie qu'au moins une des conditions de Hardy-Weinberg n'est pas réalisée.

Question 4 : cette observation montre l'existence d'un flux migratoire à sens unique des caves vers les champs, ce qui permet d'expliquer le résultat précédent d'absence d'équilibre de Hardy-Weinberg pour la population des champs.

Exercice 6.2

On étudie une espèce végétale allogame dont la floraison d'été dépend à la fois de l'ensoleillement (quantité minimale de lumière reçue et accumulée pour induire la floraison) et d'un gène bi-allélique dont dépend le système d'induction de la floraison.

Les organismes *A/A* et *A/a* exigent une quantité plus faible de lumière que les *a/a* et ont donc une floraison plus précoce, de deux semaines en moyenne, à ensoleillement constant, phénotype noté [A], le phénotype de floraison retardée des *a/a* étant noté [a].

Cependant les plantes de même génotype poussant au soleil ou à l'ombre présenteront aussi un décalage de floraison, également de deux semaines, puisque celles qui poussent à l'ombre reçoivent moins de lumière que celles qui poussent au soleil. D'où le tableau suivant :

Génotype	A/A	A/a	a/a
Floraison au soleil	Début juillet	Début juillet	Mi-juillet
Floraison à l'ombre	Mi-juillet	Mi-juillet	Fin juillet

On étudie deux populations *P1*, poussant en plein champ, au soleil, à la lisière de la forêt, et une population *P2*, poussant dans la forêt ; la pollinisation entre individus de la même population ou entre populations étant assurée par des insectes. On supposera que les graines demeurent dans leur population d'origine. Enfin, on précise que si les plantes sont allogames, elles ne sont fécondes qu'au moment de la floraison de sorte que des plants qui fleurissent avec 15 jours d'écarts ne peuvent échanger de gamètes.

Question 1 : montrez que les flux d'allèles sont différents d'une population à l'autre, l'une pouvant envoyer les deux types d'allèles A ou a vers l'autre, alors que celle-ci ne peut en envoyer qu'un seul type (vous préciserez lequel).

Question 2 : quelles conséquences peut-on en attendre ?

Pour la sous-population des plantes fleurissant début juillet, en justifiant par un calcul montrant la variation des fréquences alléliques d'une génération à l'autre.

Pour la sous population des plantes fleurissant fin juillet.

Pour les deux sous populations de plantes fleurissant à la mi-juillet.

Question 3 : en quoi la différenciation génétique des deux populations n'est pas de la sélection ?

Solution

Question 1 : compte tenu du décalage de floraison seules les plantes *a/a* de *P1* qui fleurissent en même temps que les plants *A/A* ou *A/a* de *P2* peuvent échanger des gamètes.

Il y a donc un flux possible d'allèles A ou a de *P2* vers *P1* mais uniquement un flux d'allèles *a* de *P1* vers *P2*.

Question 2 :

a) La sous population fleurissant début juillet est constituée des plants *A/A* et *A/a* de la population *P1* exposée au soleil. Les plants *a/a* sont donc exclus de cette sous population qui va par panmixie former les trois génotypes, dont les *a/a* seront de nouveau exclus, de sorte qu'on doit attendre une fixation de A dans ce groupe, qui, par ailleurs ne reçoit aucun gène des autres groupes puisqu'il fleurit seul, en premier.

Si on considère que les allèles *A* et *a* ont une fréquence égale à p_i et q_i, à la génération *i*, les génotypes *A/A*, *A/a*, *a/a* formés à la génération suivante seront, par panmixie, dans les proportions p_i^2, $2p_iq_i$ et q_i^2.

Mais les fréquences alléliques de *A* et *a*, parmi les plants fleurissant début juillet, les *a/a* étant exclus seront égales à

$$f(A) = p_{i+1} = (p_i^2+p_iq_i)/(p_i^2+2p_iq_i) = (p_i + q_i)/(p_i + 2q_i) = 1/(1 + q_i)$$

et $$f(a) = q_{i+1} = (p_iq_i)/(p_i^2+2p_iq_i) = (q_i)/(p_i + 2q_i) = q_i/(1 + q_i)$$

On voit bien que la fréquence de l'allèle *a* diminue de générations en générations et tend vers zéro (avec une vitesse qui est précisée en 7.2.4).

Cependant, on doit considérer que des individus *A/a* viennent rapporter des allèles *a* dans ce sous groupes, individus issus de la pollinisation des plants *a/a* par du pollen A venant, à la mi-juillet, de la population *P2*. La question est de savoir s'il peut exister un équilibre stable.

b) La sous population fleurissant fin juillet est constituée exclusivement des *a/a* de la population *P2* à l'ombre ; elle ne reçoit aucun allèles et demeurent un groupe génétiquement homogène.

c) À la mi-juillet, fleurissent d'une part les *a/a* de la population *P1* et les *A/A* et *A/a* de la population *P2* qui peuvent donc échanger leurs allèles, *P1* recevant A et a, *P2* ne recevant que a.

Pour le sous groupe de la population *P2*, la panmixie va générer des graines *a/a* qui vont venir s'ajouter à la génération suivante, au groupe de floraison tardive, il va donc voir la fréquence de l'allèle *a* diminuer mais, par ailleurs, il reçoit exclusivement des allèles *a* en provenance de *P1*. Or, on a établi par le modèle de l'île, que la structure génétique d'une population recevant un flux unilatéral de gènes d'une autre population tendait à devenir identique à la structure d'origine du flux migratoire, en l'occurrence les *a/a* de la population *P1*. Donc *a* va progressivement envahir la population *P2*. En conséquence, le flux migratoire de *P2* vers *P1* va se tarir à mesure que les génotypes *A/A* et *A/a* se raréfient.

Pour le sous groupe de la population *P1* fleurissant à la mi-juillet, le flux migratoire venant de *P2* apporte des allèles *A*, de sorte que des graines formées à la mi-juillet, de génotype *A/a* viendront rejoindre le groupe de floraison précoce.

À partir du moment où la fréquence de l'allèle A diminue dans le groupe mi-juillet de *P2*, le flux migratoire de A de *P2* vers *P1* se tarît, le groupe de floraison mi-juillet de *P1* tend lui aussi vers l'homogénéité *a/a* et de ce fait le groupe de floraison précoce tend vers l'homogénéité génétique *A/A*. À la limite on peut envisager une évolution vers une population *P2*, de génotype *a/a*, à floraison tardive et une population *P1* formée de génotypes *A/A* à floraison précoce et *a/a* à floraison médiane.

Question 3 : il ne s'agit pas de sélection car il n'y a pas de signification adaptative de cette différenciation au sens que certains génotypes seraient mieux adaptés à tel ou tel milieu.

La différenciation vient simplement du fait que l'environnement est différent d'une population à l'autre et qu'on s'intéresse à un caractère qui dépend d'un gène dont l'action sur le caractère est dépendante de cette différence d'environnement.

Il y a sélection (voir chapitre 7) si des génotypes différents donnent des phénotypes différents, par l'espérance de vie ou la fertilité, à environnement constant ; ici les

mêmes génotypes donnent des phénotypes différents en fonction du milieu dans lequel ils se développent.

La différenciation génétique entre populations peut donc dépendre des contingences locales pour les gènes dont l'expression est modulée par le milieu et les capacités de flux migratoires entre populations.

Chapitre **7**

La sélection

7.1 INTRODUCTION

Le modèle de Hardy-Weinberg suppose que l'effet de la sélection est négligeable sur quelques générations, une condition évidemment irréaliste sur une grande échelle de temps puisque la sélection, avec d'autres forces y fut responsable de l'évolution du vivant.

Charles Darwin fut le premier des naturalistes à concevoir la sélection naturelle comme une force active de la transformation (l'évolution) des espèces. La pensée darwinienne est populationelle car l'évolution des espèces est conçue non comme une adaptation ou une transformation des individus sous les contraintes de l'environnement, mais comme le tri, au sein des populations, par la sélection naturelle, des individus qui sont le mieux adaptés à ses contraintes, et qui, de ce fait, laissent un plus grand nombre de descendants. Cette fécondité différentielle induit une modification de la diversité (génétique) des populations vers une augmentation de la fréquence des types les mieux adaptés, et aboutit, après une longue période, à une transformation de l'espèce.

À la même époque, un autre naturaliste, Alfred Russel Wallace, vivant aux Indes, avait, indépendamment de Darwin, exprimé les mêmes conceptions. Il s'en était d'ailleurs ouvert à Darwin lui-même qui se dépêcha d'écrire un essai afin de publier ses conceptions conjointement avec celles de Wallace, dans le même numéro du *Journal of the Linnean Society.* Darwin poursuivit et approfondit sa réflexion et, par sa position en Angleterre, s'imposa comme le seul fondateur de la théorie de l'évolution par la sélection naturelle.

Le fait que les concepts développés par Darwin aient pu surgir à la même époque chez un autre naturaliste signifie tout simplement que les esprits étaient murs pour

reconsidérer complètement la problématique transformiste en dépassant les conceptions économiques, philosophiques et théologiques du XVIII^e siècle. Les conceptions darwiniennes apparaissent, pour tout historien des idées, comme un transfert, dans les sciences naturelles, de la pensée libérale, économique et philosophique, qui accompagna le développement industriel et commercial du début du XIX^e siècle. Que Darwin, à des passages clé de son ouvrage, fît explicitement référence à Malthus n'empêcha nullement Marx et Engels de considérer la théorie darwinienne comme de la plus grande importance sur le plan philosophique, révolutionnaire au sens Hégélien du terme.

Cependant la sélection telle que l'a conçue Darwin, ne résout pas vraiment toute la question de l'évolution des espèces. D'une part la formulation mathématique de la sélection, si elle rend compte de l'évolution darwinienne de la diversité des populations par la « sélection du plus apte », apporte aussi des résultats inattendus dans le cadre strict du darwinisme, dont la possibilité de maintenir cette diversité. D'autre part cette formulation montre qu'une population soumise à une contrainte sélective de l'environnement peut évoluer dans plusieurs directions, parfois opposées, ce qui pose le problème de la force qui « exerce le choix évolutif ». Enfin le coût de la sélection est incompatible avec l'importance de la diversité observée dans les populations naturelles, ce qui implique qu'une grande part de cette diversité est sélectivement neutre, mais que sa variation participe obligatoirement à la définition des espèces nouvelles ; la question se pose alors de définir les forces qui participent à la variation de la diversité sélectivement neutre.

Ce chapitre se contentera de présenter en détail le modèle général le plus simple ; des modèles plus réalistes, mais plus complexes, seront évoqués rapidement en renvoyant le lecteur intéressé à un ouvrage plus complet et plus difficile[1].

7.2 MODÈLE GÉNÉRAL DE SÉLECTION À COEFFICIENTS CONSTANTS

7.2.1 Définitions et approche intuitive

Il y a sélection pour un gène lorsque les différents génotypes relatifs à ce gène n'ont pas la même espérance de vie (par exemple un gène impliqué dans une maladie génétique ou la sensibilité à un facteur de l'environnement) ou la même fertilité (par exemple un gène impliqué dans la gamétogenèse ou la physiologie de la reproduction).

Que la sélection résulte d'une mortalité ou d'une fertilité différentielles entre les génotypes, il s'en suit une fécondité différentielle. Au sens darwinien du terme, le nombre de descendants moyens ne sera pas le même d'un génotype à l'autre, toutes choses étant par ailleurs égales quant aux autres paramètres biologiques, culturels ou aléatoires qui affectent la fécondité des individus.

1. *Génétique et évolution*, Solignac, Perriquet, Anxolabérère et Petit, Hermann, Paris, 1995.

On peut donc affecter à chacun des génotypes relatifs à ce gène un paramètre de fécondité, appelé valeur sélective (appelé aussi valeur adaptative ou coefficient de sélection). Les rapports entre les valeurs sélectives des génotypes expriment l'avantage ou le désavantage sélectif existant entre ces génotypes, du point de vue de leur fécondité.

Dans le cas d'un gène di-allélique, on peut définir les valeurs sélectives de la manière suivante :

Génotypes	*A1*/*A1*	*A1*/*A2*	*A2*/*A2*
Valeurs sélectives	$\sigma 1$	$\sigma 2$	$\sigma 3$

où $\sigma 1$, $\sigma 2$ et $\sigma 3$ sont trois paramètres égaux ou proportionnels à la fécondité relative des trois génotypes.

Dire qu'il n'y a pas de sélection revient à dire que les trois fécondités sont égales, soit $\sigma 1 = \sigma 2 = \sigma 3$.

Dire qu'il y a sélection revient à considérer que ces trois paramètres ne sont pas égaux.

Si l'allèle *A2* est défavorable et qu'il provoque une baisse de fécondité (parce qu'il diminue la viabilité ou la fertilité), le nombre de descendants de *A2*/*A2* sera inférieur à celui de *A1*/*A1* et on écrira que $\sigma 3 < \sigma 1$.

La valeur sélective $\sigma 2$ sera égale à $\sigma 1$ si l'allèle *A2* est récessif, du point de vue de la sélection (*A1A2* et *A1A1* ayant même fécondité) ; $\sigma 2$ aura une valeur égale à $\sigma 3$ si l'allèle *A2* a un effet dominant (*A1A2* et *A2A2* ayant même fécondité). La valeur de $\sigma 2$ sera intermédiaire si les allèles *A1* et *A2* ont des effets co-dominants du point de vue de la fécondité.

Il est facile de concevoir que la fréquence d'un allèle défavorable, par exemple *A2*, dont les porteurs laissent moins de descendants, devrait diminuer. À terme l'allèle *A2* devrait disparaître de la population. Ce phénomène sélectif qui voit la disparition de ce qui « est défavorable » (dans des conditions d'environnement donné) et la fixation de ce qui « est favorable » correspond à la vision darwinienne classique de la sélection.

Évidemment cela suppose que les valeurs sélectives soient constantes ; en effet, si la relation d'ordre entre ces valeurs s'inversait dans certaines circonstances, on pourrait alors maintenir le polymorphisme génétique, ce qui sera exposé au chapitre suivant.

Il existe par ailleurs deux cas où il n'est pas évident d'imaginer intuitivement la limite, sans formulation mathématique du processus évolutif ; il s'agit des cas où il y a avantage ou désavantage de l'hétérozygote.

Dans chacun de ces cas il n'y a pas d'allèle favorable ou défavorable mais un génotype favorisé ou défavorisé. Le traitement de ces cas aboutit à des conclusions absentes de la conception darwinienne stricte parce que la formulation mathématique de la sélection se fonde sur une conception mendélienne de l'hérédité, avec gènes, allèles et génotypes, absente du darwinisme.

7.2.2 Développement mathématique

a) Effet de la sélection sur les fréquences alléliques : composition de l'urne gamétique

Dans le cas d'un gène di-allélique, au sein d'une population panmictique, on peut écrire :

Génotypes	***A1/A1***	***A1/A2***	***A2/A2***
Valeurs sélectives	σ_1	σ_2	σ_3
Fréquences	p^2	$2pq$	q^2

où σ1, σ2 et σ3 sont les valeurs sélectives relatives aux trois génotypes, et p^2, $2pq$ et q^2, les trois fréquences génotypiques à la conception, à l'issue du tirage dans l'urne gamétique parentale, où les fréquences des allèles *A1* et *A2* étaient respectivement *p* et *q*.

La contribution des trois génotypes à la génération suivante dépendra de leur contribution à l'urne gamétique, puisque, sous la panmixie, la génération suivante est formée par l'union au hasard de gamètes tirés dans cette urne.

La contribution de ces trois génotypes à l'urne gamétique sera à la fois proportionnelle à leurs fréquences respectives (p^2, $2pq$ et q^2) mais aussi à leurs valeurs sélectives (σ1, σ2 et σ3). En effet si le génotype *A2*/*A2* est létal avant l'âge adulte, sa contribution sera nulle. La contribution des trois génotypes est donc proportionnelle au produit des fréquences de ces génotypes par leurs valeurs sélectives, et les contributions normalisées ainsi obtenues (tableau 7.1) représentent les fréquences des différents génotypes après l'effet de la sélection, et leur contribution à l'urne des gamètes de la génération suivante.

Tableau 7.1 Contributions respectives des génotypes à l'urne gamétique.

Génotypes	*A1/A1*	*A1/A2*	*A2/A2*
Valeurs sélectives	σ_1	σ_2	σ_3
Fréquences	p^2	$2pq$	q^2
Contribution à l'urne gamétique	$\sigma_1 p^2$	$2\sigma_2.pq$	$\sigma_3 q^2$
Contribution normalisée*	$\sigma_1 p^2/\sigma$	$2\sigma_2.pq/\sigma$	$\sigma_3 q^2/\sigma$

Afin de ramener la somme des contributions respectives à la valeur 1, on peut, sans changer leurs rapports respectifs les diviser par un même nombre, $\sigma = \sigma_1 p^2 + 2\sigma_2.pq + \sigma_3 q^2$.

Les nouvelles fréquences gamétiques (celles des parents étaient *p* et *q*) s'écrivent :

$$f(A1) = p' = (\sigma_1 p^2 + \sigma_2 pq)/\sigma$$
$$f(A2) = q' = (\sigma_3 q^2 + \sigma_2 pq)/\sigma$$

Les fréquences *p*' et *q*' sont les fréquences alléliques de la génération suivante. À la conception les trois génotypes auront pour fréquences respectives p'^2, $2p'q'$ et q'^2, qui seront de nouveau modifiées par la sélection, conduisant à la génération suivante à des fréquences alléliques p'' et *q*'', et ainsi de suite (voir figure 7.1).

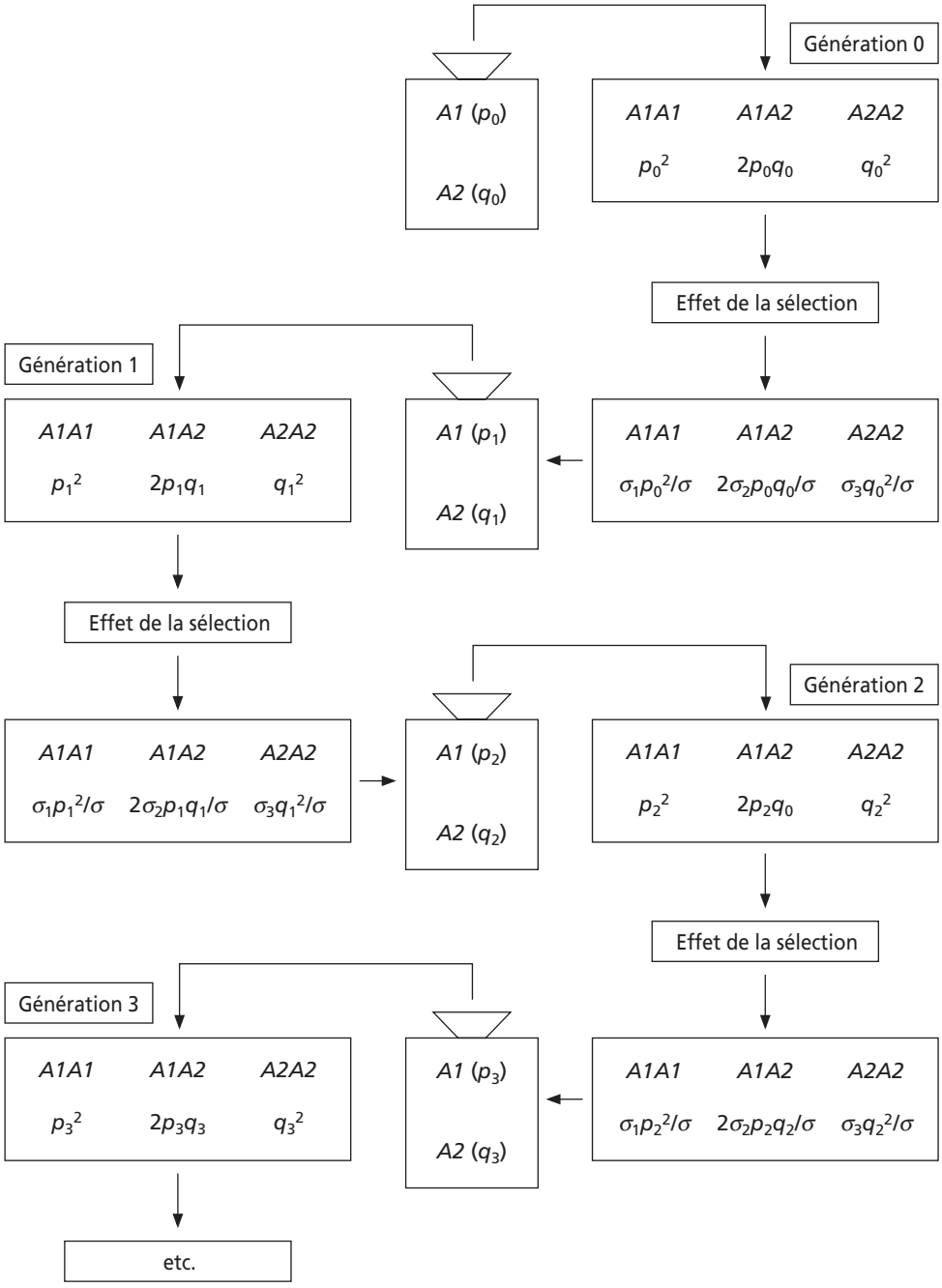


Figure 7.1 Évolution de la composition génétique d'une population panmictique, en fréquences allèliques et génotypiques, sous l'effet de la sélection (coefficients constants)

b) Variation des fréquences alléliques d'une génération à l'autre

Il est utile d'exprimer la variation Δp de fréquence allélique d'une génération à l'autre afin de pouvoir définir la limite du processus évolutif en résolvant l'équation $\Delta p = 0$.

Par définition :

$$\Delta p = p' - p$$

soit $\Delta p = (\sigma_1 p^2 + \sigma_2 .pq)/\sigma - p$

d'où, après réduction au même dénominateur et remplacement de σ par $\sigma_1 p^2 + 2\sigma_2.pq + \sigma_3 q^2$:

$$\Delta p = (1/\sigma)[\sigma_1 p^2 + \sigma_2 .pq - p(\sigma_1 p^2 + 2\sigma_2.pq + \sigma_3 q^2)]$$

on met p en facteur : $\Delta p = (p/\sigma)[\sigma_1 p + \sigma_2 .q - (\sigma_1 p^2 + 2\sigma_2.pq + \sigma_3 q^2)]$

on met $\sigma_1 p$ en facteur : $\Delta p = (p/\sigma)[\sigma_1 p(1-p) + \sigma_2 .q - 2\sigma_2.pq - \sigma_3 q^2]$

on met q en facteur $\Delta p = (pq/\sigma)[\sigma_1 p + \sigma_2 - 2\sigma_2.p - \sigma_3 q\ ]$

on peut écrire $\Delta p = (pq/\sigma)[\sigma_1 p + \sigma_2 - \sigma_2.p - \sigma_2.p - \sigma_3 q\ ]$

d'où $\Delta p = (pq/\sigma)[(\sigma_1 - \sigma_2)p + \sigma_2 (1 - p) - \sigma_3 q\ ]$

et, enfin $\boldsymbol{\Delta p = (pq/\sigma)[(\sigma_1 - \sigma_2)p + (\sigma_2 - \sigma_3)q\ ]}$

La variation de la fréquence allélique p d'une génération à l'autre sous l'effet de la sélection est une équation du troisième degré en p, qui admet donc trois racines telles que $\Delta p = 0$.

> ***Remarque 1 :*** l'équation Δp n'est pas modifiée si on multiplie tous les coefficients de sélection par un même nombre ; celui-ci sera mis en facteur au numérateur et au dénominateur et disparaîtra par simplification. C'est pourquoi on peut prendre comme valeurs sélectives, soit les fécondités respectives des génotypes, soit des nombres proportionnels à ces fécondités.
>
> ***Remarque 2 :*** quand les valeurs sélectives sont égales, on trouve bien $\Delta p = 0$; il n'y a pas de sélection et les fréquences restent inchangées d'une génération à l'autre.

c) Limite du processus sélectif

La sélection modifie les fréquences alléliques d'une génération à l'autre parce que les différents génotypes n'ont pas la même fécondité, et partant, le même nombre moyen de descendants.

Cette variation est exprimée par l'équation Δp, et la limite du processus sélectif sera atteinte lorsqu'un équilibre sera atteint pour certaines valeurs des fréquences alléliques.

Ces fréquences alléliques limites correspondent aux racines de l'équation pour lesquelles il y a équilibre puisqu'alors $\Delta p = 0$.

Ces racines ont deux valeurs banales 0 et 1, et une valeur fonction des coefficients de sélection :

- $p = 0$ (alors $q = 1$) ;
- $p = 1$ (alors $q = 0$) ;
- $p_e = (\sigma_3 - \sigma_2)/(\sigma_1 - 2\sigma_2 + \sigma_3)$ et $q_e = (\sigma_1 - \sigma_2)/(\sigma_1 - 2\sigma_2 + \sigma_3)$.

La dernière racine p_e est celle de l'équation entre crochet dans la formule de Δp et n'a la signification biologique d'une fréquence allélique que si elle est comprise entre 0 et 1.

7.2.3 Valeurs limites des fréquences alléliques sous l'effet de la sélection

a) Relations d'ordre entre valeurs sélectives

Il y a sélection quand les fécondités relatives à chacun des génotypes sont différentes et jouent le rôle de valeur sélective. Dans ce cas il peut exister quatre relations d'ordre entre ces coefficients de sélection (tableau 7.2).

TABLEAU 7.2

Génotypes	*A1/A1*		*A1/A2*		*A2/A2*
Relation d'ordre 1	$\sigma 1$	>	$\sigma 2$	>	$\sigma 3$
Relation d'ordre 2	$\sigma 1$	<	$\sigma 2$	<	$\sigma 3$
Relation d'ordre 3	$\sigma 1$	<	$\sigma 2$	>	$\sigma 3$
Relation d'ordre 4	$\sigma 1$	>	$\sigma 2$	<	$\sigma 3$

b) Allèles favorables et défavorables : relations d'ordre 1 et 2

Dans le premier cas, l'allèle *A1* est favorable, et *A2* défavorable. Les écarts $(\sigma_1 - \sigma_2)$ et $(\sigma_2 - \sigma_3)$ sont positifs, ce qui signifie que $\Delta p = (pq/\sigma)[(\sigma_1 - \sigma_2)p + (\sigma_2 - \sigma_3)q]$ est strictement positif, quelque soit p. La fréquence allélique, p, ne peut que croître et tend donc vers 1 ; l'allèle *A1* sera fixé et l'allèle *A2* éliminé.

Dans le deuxième cas, l'allèle *A1* est défavorable, et *A2* favorable. Les écarts $(\sigma_1 - \sigma_2)$ et $(\sigma_2 - \sigma_3)$ sont négatifs, ce qui signifie que $\Delta p = (pq/\sigma)[(\sigma_1 - \sigma_2)p + (\sigma_2 - \sigma_3)q]$ est strictement négatif, quelque soit p. La fréquence allélique, p, ne peut que décroître et tend donc vers 0 ; l'allèle *A1* sera éliminé et l'allèle *A2* fixé.

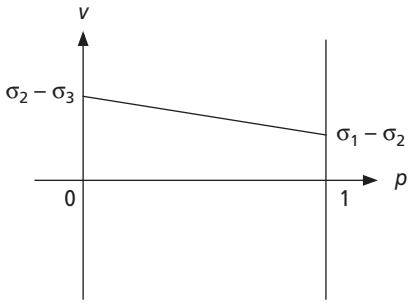
Il est intéressant de noter que dans chacun de ces deux cas, il existe bien une solution mathématique p_e de l'équation entre crochet $y = [(\sigma_1 - \sigma_2)p + (\sigma_2 - \sigma_3)q]$, pour laquelle $\Delta p = 0$, mais qu'elle est inférieure à 0 ou supérieure à 1 (figure 7.2) et qu'elle n'a donc pas la signification d'une fréquence allélique.

La figure 7.3 présente la variation Δp, pour les valeurs de p comprises entre $p = 0$ et $p = 1$, sachant que la racine p_e est inférieure à zéro ou supérieure à 1 (figure 7.2).

Il est utile, pour des discussions futures, d'observer l'aspect de la variation de $\sigma = \sigma_1 p^2 + 2\sigma_2 . pq + \sigma_3 q^2$, le coefficient moyen de sélection entre $p = 0$ et $p = 1$ (figure 7.4), où le signe de la dérivée permet de montrer que la pente est positive dans le premier cas et négative dans le second.

Ces deux cas correspondent à la vision strictement darwinienne de la sélection, connue comme un processus « purificateur » ou « normalisant » qui élimine le moins apte au profit du plus apte.

Relation d'ordre 1 : $\sigma_1 > \sigma_2 > \sigma_3$

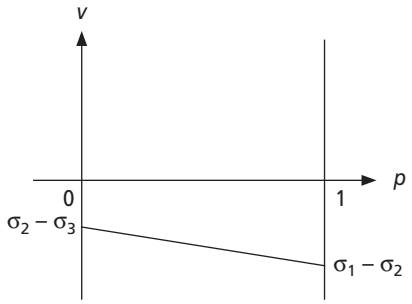


Avec cette relation d'ordre $(\sigma_1 - \sigma_2)$ et $(\sigma_2 - \sigma_3)$ sont positifs.

Si $(\sigma_1 - \sigma_2) < (\sigma_2 - \sigma_3)$, comme sur le graphe, alors p_e est supérieur à 1.

Si $(\sigma_1 - \sigma_2) > (\sigma_2 - \sigma_3)$, alors p_e est inférieur à 0.

Relation d'ordre 2 : $\sigma_1 < \sigma_2 < \sigma_3$



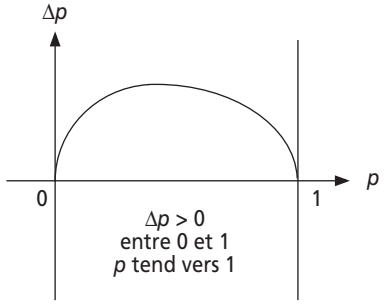
Avec cette relation d'ordre $(\sigma_1 - \sigma_2)$ et $(\sigma_2 - \sigma_3)$ sont négatifs.

Si $(\sigma_1 - \sigma_2) < (\sigma_2 - \sigma_3)$, comme sur le graphe, alors p_e est inférieur à 0.

Si $(\sigma_1 - \sigma_2) > (\sigma_2 - \sigma_3)$, alors p_e est supérieur à 1.

Figure 7.2

Relation d'ordre 1 : $\sigma_1 > \sigma_2 > \sigma_3$



Relation d'ordre 2 : $\sigma_1 < \sigma_2 < \sigma_3$

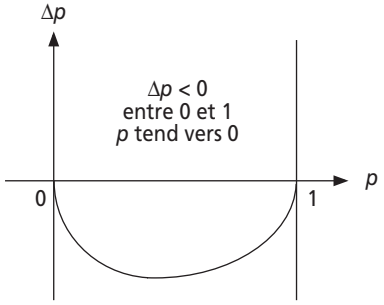
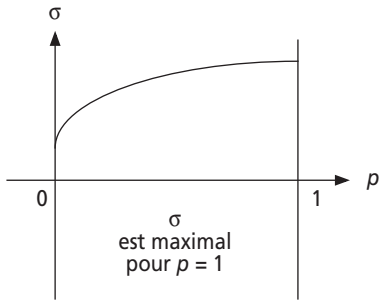


Figure 7.3

Relation d'ordre 1 : $\sigma_1 > \sigma_2 > \sigma_3$



Relation d'ordre 2 : $\sigma_1 < \sigma_2 < \sigma_3$

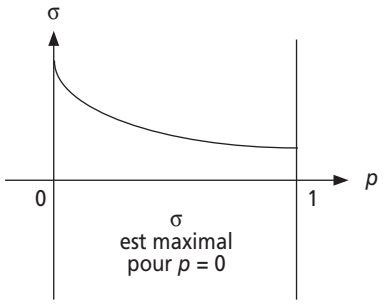


Figure 7.4

Cela correspond à la situation réellement rencontrée dans la nature pour les allèles d'un gène qui ont un effet défavorable, soit sur la viabilité (maladie génétique, sensibilité accrue à un facteur de l'environnement) ou la fertilité. On conçoit cependant que des mutations *de novo* maintiendront, à un niveau de fréquence très faible, ces allèles défavorables ; l'effet combiné de la sélection et des mutations et l'équilibre polymorphe qu'il engendre sera détaillé dans le chapitre suivant.

Pour Darwin, la définition du « plus apte » n'avait rien d'absolu ; au contraire ce qui était le plus apte dans un environnement pouvait être le moins apte dans un autre environnement. De la sorte, il est possible d'imaginer que des variations périodiques du milieu puissent maintenir la diversité génétique d'une population en inversant l'effet de la sélection. Il est surtout possible d'imaginer que deux populations de la même espèce, séparées l'une de l'autre et confrontées à des environnements différents évolueront vers la fixation de type différents. Cette divergence évolutive peut alors rendre compte à terme de la spéciation dès que les individus de populations originaires d'une population ancestrale, ne sont plus interféconds.

À cette conception de la sélection s'attache donc l'idée que les différences génétiques (héréditaires dans le langage prémendélien) sont faibles car la sélection tend à homogénéiser les individus vers la forme la plus apte ; c'est un processus purificateur et normalisant. Il s'ensuit que la seule place pour des différences génétiques au sein d'une espèce résident essentiellement entre les populations et non en leur sein.

Appliquée à l'anthropologie cette conception revenait à considérer l'espèce humaine comme un ensemble de populations séparées (ou races, si on veut les appeler ainsi) génétiquement différentes par le fait que la sélection avait depuis des temps immémoriaux, conduit ces « races » vers la norme correspondant à leur milieu propre.

Le racisme est fondé non seulement sur l'idée que de telles aptitudes raciales spécifiques existent, mais aussi et surtout sur l'idée que ces aptitudes permettent une hiérarchisation des populations (des races), une hiérarchie d'autant plus légitime qu'elle serait le produit de la sélection naturelle et non d'un choix de l'homme blanc. Le même discours appliqué aux classes sociales, à l'intérieur même des nations, a donné naissance au darwinisme social qui, se fondant avec une fraction du mouvement eugéniste, donna les pires déviations idéologico-politiques du XX^{e} siècle.

Ce type de déviations est encore illustré par les thèses sur l'infériorité génétique des noirs vis-à-vis des apprentissages cognitifs ou les tentatives de biologisation des problèmes sociaux ou culturels d'échec scolaire ou d'intégration, considérés par des courants d'extrême droite comme la traduction d'une incapacité génétique. Ces courants extrêmes tendent à contaminer les courants « libéraux » ou ultra-libéraux et leurs critiques purement économiques du « welfare state » et des politiques sociales d'assistance, pour les déborder par l'idée que leurs effets contrecarrent l'effet qu'aurait la sélection naturelle et conduiraient l'humanité à une dégénérescence génétique irrémédiable.

c) Avantage ou désavantage de l'hétérozygote, ou ce que le darwinisme n'avait pas prévu

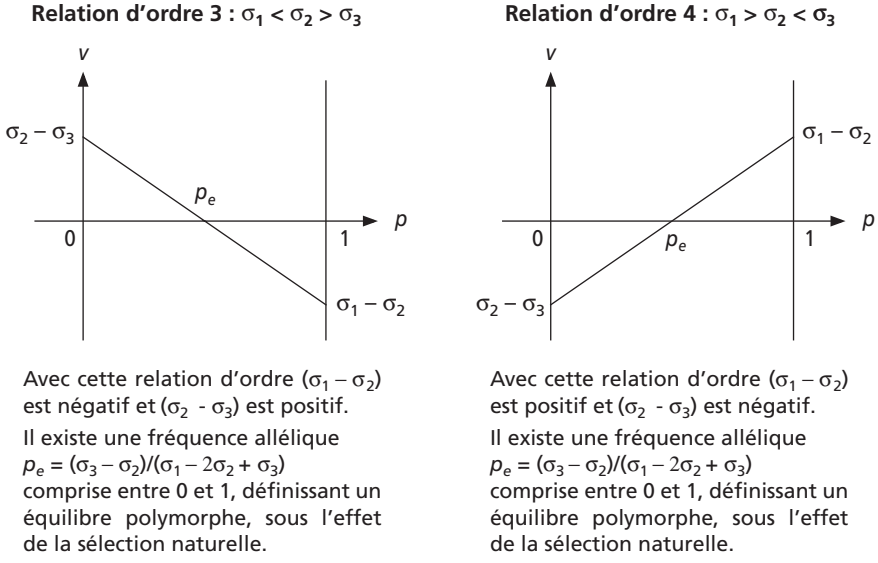
Dans chacun de ces cas, il n'y a pas d'allèle favorable ou défavorable puisque c'est l'hétérozygote qui est, selon le cas, avantagé ou désavantagé sur chacun des homozygotes. De telles circonstances étaient difficilement perceptibles par le darwinisme puisque, d'un point de vue scientifique, la génétique mendélienne et la diploïdie étaient inconnues, et que, d'un point de vue idéologique, il était assez difficile d'admettre l'idée que le « type idéal » favorisé par la sélection puisse être « impur » (hétérozygote).

Avec l'avantage ou le désavantage de l'hétérozygote vis-à-vis des deux homozygotes, l'un des écarts $(\sigma_1 - \sigma_2)$ ou $(\sigma_2 - \sigma_3)$ est positif, l'autre négatif, ce qui signifie que l'équation

$$\Delta \mathbf{p} = (pq/\sigma)[(\sigma_1 - \sigma_2)p + (\sigma_2 - \sigma_3)q]$$

peut prendre une valeur nulle entre $p = 0$ et $p = 1$.

Il existe bien alors une valeur p_e de la fréquence allélique pour laquelle un équilibre polymorphe existe puisque $\Delta p = 0$ et $0 < p_e < 1$ (figure 7.5).



Avec cette relation d'ordre $(\sigma_1 - \sigma_2)$ est négatif et $(\sigma_2 - \sigma_3)$ est positif.

Il existe une fréquence allélique $p_e = (\sigma_3 - \sigma_2)/(\sigma_1 - 2\sigma_2 + \sigma_3)$ comprise entre 0 et 1, définissant un équilibre polymorphe, sous l'effet de la sélection naturelle.

Avec cette relation d'ordre $(\sigma_1 - \sigma_2)$ est positif et $(\sigma_2 - \sigma_3)$ est négatif.

Il existe une fréquence allélique $p_e = (\sigma_3 - \sigma_2)/(\sigma_1 - 2\sigma_2 + \sigma_3)$ comprise entre 0 et 1, définissant un équilibre polymorphe, sous l'effet de la sélection naturelle.

Figure 7.5

La figure 7.6 présente la variation de Δp, pour les valeurs de p comprises entre $p = 0$ et $p = 1$ (la pente du graphe au point d'inflexion est donnée par le signe de la dérivée de Δp pour la valeur de la fréquence p_e).

Dans le cas de l'avantage de l'hétérozygote, l'équilibre polymorphe est stable : si la fréquence p de l'allèle *A1* est inférieure à p_e, Δp est positif et p va croître jusqu'à p_e ; au contraire si p est supérieure à p_e, Δp est négatif et p va décroître jusqu'à p_e. Une fois atteinte, la valeur p_e est stable puisque tout écart à droite est corrigé par un Δp négatif et tout écart à gauche par un Δp positif.

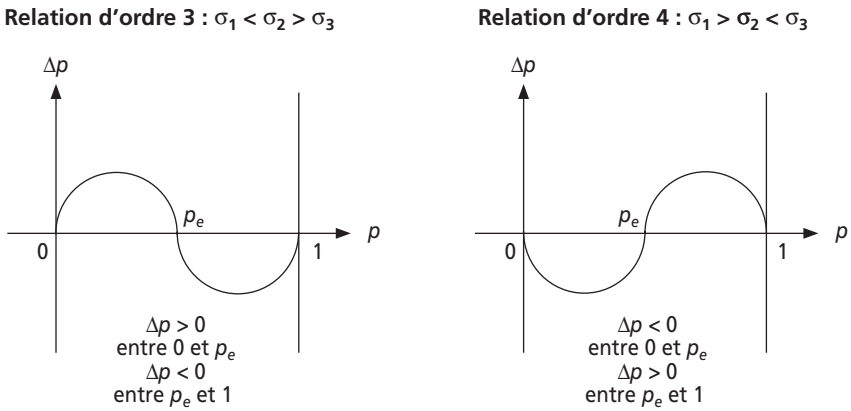


Figure 7.6

Au contraire, dans le cas du désavantage de l'hétérozygote, l'équilibre polymorphe est instable : si la fréquence p de l'allèle *A1* est inférieure à p_e, Δp est négatif et p va décroître jusqu'à 0 ; si p est supérieure à p_e, Δp est positif et p va croître jusqu'à 1 Si jamais la valeur de p est égale à p_e, l'équilibre est instable puisque tout écart à droite sera accentué par un Δp positif et tout écart à gauche par un Δp négatif.

Il est utile de noter, comme dans le cas précédent, que la sélection tend à donner à la composition génétique de la population, des fréquences alléliques telles qu'elles permettent de « maximiser » la valeur du coefficient moyen de sélection (figure 7.7).

En effet l'équation $\sigma = \sigma_1 p^2 + 2\sigma_2 . pq + \sigma_3 q^2$, expression du coefficient moyen de sélection entre $p = 0$ et $p = 1$, passe par un extremum qui est un maximum (stable) pour p_e dans le cas de l'avantage de l'hétérozygote et par un minimum (instable) pour p_e, dans le cas d'un désavantage, les deux maximum stables étant pour les valeurs $p = 0$ ou $p = 1$ (figure 7.7).

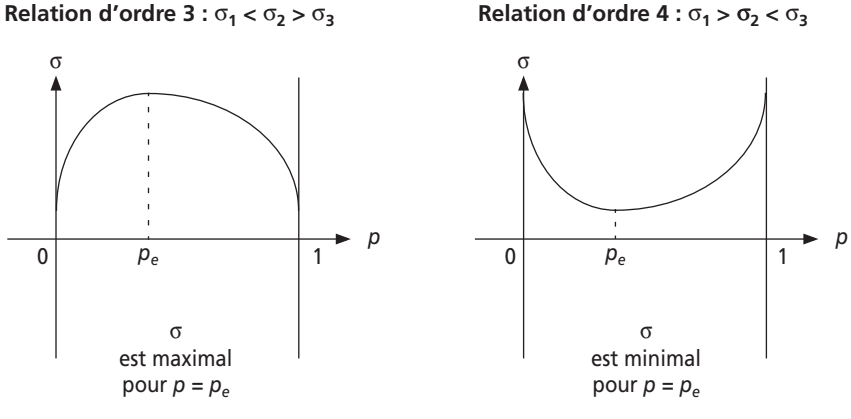


Figure 7.7

La valeur maximale de σ est unique pour les trois premières relations d'ordre, et correspond à la fréquence allélique $p = 1$, quand *A1* est favorable, $p = 0$, quand *A1* est défavorable et $p = pe$, quand il y a avantage de l'hétérozygote. Quand il y a désavantage de l'hétérozygote, la valeur de σ tendra vers un des deux maxima, correspondant à $p = 0$ ou $p = 1$. Le généticien des populations, Fischer, à qui l'on doit ces développements, a donné à ce résultat le nom de « théorème fondamental de la sélection naturelle » considérant que la sélection tend à maximiser la valeur adaptative moyenne de la population, ce qui n'est vrai que lorsque les valeurs adaptatives sont constantes et que la sélection n'agit que sur les génotypes d'un seul gène. On verra plus loin que la réalité est plus complexe.

d) La drépanocytose, exemple le plus évident d'avantage de l'hétérozygote

La drépanocytose est une hémoglobinopathie sévère et souvent létale. Elle est due à la présence d'une hémoglobine pathologique, dénommée hémoglobine S (HbS), chez les individus homozygotes pour une mutation du gène β de l'hémoglobine, mutation appelée drépanocytaire et notée β^S.

La séquence habituelle du gène β de l'hémoglobine présente un sixième codon GAG codant pour un acide glutamique dans la chaîne β. L'allèle normal non muté est noté β^A. L'allèle β^S muté présente un sixième codon GTG qui spécifie une valine. La substitution du glutamique par une valine dans les deux chaînes β de l'hémoglobine S lui donne une propriété pathologique, absente de l'hémoglobine adulte normale (HbA) chez les homozygotes β^A/β^A, celle de polymériser en concentration faible en oxygène.

Or cette situation est souvent rencontrée par les globules rouges circulant au sein du système artériel terminal, là où l'oxygène est pompé par l'activité des tissus. Dès que l'HbS polymérise l'hématie, celle-ci s'allonge en forme de faucille (d'où l'autre nom de la drépanocytose : l'anémie falciforme) et perd sa plasticité, ce qui peut provoquer l'obstruction du capillaire. On assiste alors à un processus autocatalytique car le blocage local de la circulation provoque une chute de la concentration en oxygène, ce qui accentue la falciformation des hématies et l'obstruction des capillaires déjà obstrués, et induit le même processus dans les capillaires adjacents. L'extension du blocage des capillaires provoque une anoxie locale du tissu qui peut lui être très dommageable. Par une réaction physiologique normale, l'organisme réagit en activant des enzymes sériques capables de lyser les hématies formant les bouchons.

Ces hémolyses récurrentes sont pathologiques pour l'organisme, car elles altèrent le fonctionnement de nombreux organes dont la rate, le foie, les reins puis le cœur. Périodiquement survient un épisode hémolytique plus grave, associant thromboses, chocs hépatiques et cardiaques qui peuvent être mortels.

La maladie est récessive car les hétérozygotes ne sont pas touchés : leurs hématies renferment en effet un mélange d'HbA et d'HbS, impropre à la polymérisation.

Les anthropologues et les médecins coloniaux avaient remarqué depuis longtemps que cette maladie était présente, à un haut niveau de fréquence malgré sa

létalité, dans des populations sans parenté proche (Afrique équatoriale ou Proche-Orient), mais toutes caractérisées par un même milieu de vie, la ceinture intertropicale impaludée.

C'est Allison qui, le premier, fit l'hypothèse d'un avantage de l'hétérozygote β^A/β^S en supposant que ceux-ci étaient avantagés, bien sur vis-à-vis des homozygotes β^S/β^S puisqu'ils n'étaient pas atteints de drépanocytose, mais aussi et surtout vis-à-vis des homozygotes β^A/β^A car ils seraient moins sensibles qu'eux à l'agent paludéen. En effet, les hématies des hétérozygotes sont légèrement fragilisées par la présence d'HbS, ce qui induit une légère hémolyse, insuffisante pour être pathologique, mais suffisante pour perturber le cycle vital du plasmodium dans les hématies.

Au Kenya, la fréquence da la mutation β^S est égale à 20 %, une valeur excessivement élevée qui ne pourrait pas se maintenir sans l'avantage de l'hétérozygote.

À la naissance les fréquences génotypiques sont les suivantes :

Génotypes :	β^A/β^A	β^A/β^S	β^S/β^S
Fréquences :	0,64	0,32	0,04
Valeurs sélectives :	$\sigma 1$	$\sigma 2$	$\sigma 3$
Ou écrites autrement :	$1 - t$	1	$1 - s$

Les valeurs sélectives $\sigma 1$, $\sigma 2$ et $\sigma 3$ étant définies à un coefficient de proportionnalité près (voir remarque 1, en 7.2.2.c), on adopte un changement de variable approprié au cas de l'avantage de l'hétérozygote. Sa valeur sélective est prise égale à 1 (il suffit de diviser toutes les valeurs sélectives par $\sigma 2$) ce qui permet d'écrire les valeurs $\sigma 1$ ou $\sigma 3$ comme égales à 1 moins le désavantage de chaque homozygote β^A/β^A ou β^S/β^S relativement à l'hétérozygote β^A/β^S. Avec ce changement de variable, les fréquences alléliques à l'équilibre s'écrivent :

$$p_e = (\sigma_3 - \sigma_2)/(\sigma_1 - 2\sigma_2 + \sigma_3) = \boldsymbol{s/(s + t)}$$
$$q_e = (\sigma_1 - \sigma_2)/(\sigma_1 - 2\sigma_2 + \sigma_3) = \boldsymbol{t/(s + t)}$$

Il est alors facile d'estimer le désavantage de l'homozygote par rapport à l'hétérozygote vis-à-vis du paludisme si l'on considère que la composition de la population est à l'équilibre.

Comme la maladie est létale, $s = 1$ ($\sigma_3 = 0$), d'où :

$$q_e = t/(1 + t)$$

par ailleurs $q_e = 0{,}20$

d'où $t = 0{,}25$

On doit donc considérer, si la population est à l'équilibre polymorphe, pour les allèles β^A et β^S, grâce à l'avantage de l'hétérozygote, que le paludisme ampute l'espérance de vie des homozygotes β^A/β^A de 25 % de celle des hétérozygotes β^A/β^S.

Cependant l'amputation d'espérance de vie observée dans la réalité n'est pas aussi forte. Il est vrai que la drépanocytose n'est pas non plus totalement létale ; aussi *s*, bien que faible, n'est pas nul, et t est inférieur à 25 %, mais *t* reste encore un peu élevé.

Il faut donc considérer que le maintien du polymorphisme allélique pour le gène β ne résulte pas du seul avantage de l'hétérozygote vis-à-vis de la drépanocytose et du paludisme. En Inde, il se trouve, par exemple associé à la persistance héréditaire d'hémoglobine fœtale (HbF) qui rend la drépanocytose (et la thalassémie) beaucoup moins grave chez les individus présentant un taux élevé d'HbF.

Bien que la situation soit dans la réalité plus complexe, le modèle théorique simple de l'avantage de l'hétérozygote est avéré et unanimement accepté pour les allèles du gène β, en zone impaludée.

Bien évidemment la disparition du paludisme entraînerait *ipso facto* celle de l'avantage de l'hétérozygote ; la sélection ne pourrait plus maintenir le polymorphisme, elle serait redevenue « darwinienne » et éliminerait inexorablement l'allèle β^S devenu défavorable, à une vitesse d'abord rapide puis plus lente (voir 7.2.4).

e) L'avantage de l'hétérozygote : aspects génétiques et philosophiques

Le maintien de la diversité génétique (de la « variation ») par la sélection naturelle dans des conditions stables d'environnement est une conception étrangère au darwinisme originel. La découverte de cette éventualité à travers la formulation mathématique du processus de la sélection a ouvert la voie à de nombreuses recherches pratiques et théoriques.

D'abord, en pratique, l'avantage de l'hétérozygote apporte une réponse au paradoxe de la vigueur hybride bien connue des agronomes. Dans le cadre d'un darwinisme étroit, les races étant bien adaptées à leur milieu d'origine, les hybrides entre races auraient dû être moins bien adaptés que leurs parents à chacun des deux environnements parentaux. Or on a observé depuis longtemps que le croisement de deux variétés ou deux races donnait des hybrides beaucoup plus vigoureux que les parents en termes de viabilité et de prolificité, ce qui semblait paradoxal.

L'avantage de l'hétérozygote permet aussi de comprendre, à condition d'en démontrer la réalité, comme dans le cas de la drépanocytose chez l'homme, le maintien à un niveau considérablement élevé de la fréquence d'allèles létaux qui auraient dû, sans cet avantage de l'hétérozygote, être éliminés depuis longtemps.

Enfin, l'avantage de l'hétérozygote a des conséquences sur un plan plus théorique et philosophique. En permettant le maintien de la diversité génétique à l'intérieur des populations, l'avantage de l'hétérozygote permet de ne plus confiner les différences génétiques au sein de l'espèce, à des différences entre populations. Il est un fait que l'étude du polymorphisme génétique chez de nombreuses espèces, notamment chez l'homme, à travers le typage des groupes sanguins et l'analyse biochimique de nombreux variants électrophorétiques pour de nombreux gènes codant pour des protéines sériques ou cellulaires, puis les séquences d'ADN, notamment leurs marqueurs polymorphes, a progressivement montré l'importance de la diversité intra-populationnelle et, dans le cas de l'homme, le caractère vraiment restreint des différences génétiques entre populations (voir à ce sujet le chapitre 1).

On a donc beaucoup fait appel à l'avantage supposé des hétérozygotes à la fois pour rendre compte de la diversité génétique au sein des populations, et pour contrer les conclusions idéologiques fondées sur des conceptions darwiniennes strictes attachées au rôle purifiant de la sélection (voir 7.2.2.b).

Mais l'adéquation entre la réalité de la diversité génétique et la théorie de l'avantage de l'hétérozygote reste cependant problématique, car cet avantage n'est pas facile à démontrer et son coût (voir plus loin le fardeau génétique) ne permet pas de l'invoquer pour un grand nombre de gènes.

Il n'en demeure pas moins que l'avantage de l'hétérozygote joue et a joué un rôle important dans l'histoire génétique des populations comme dans l'histoire de la génétique des populations !

f) Le désavantage de l'hétérozygote : conséquences génétiques et théoriques

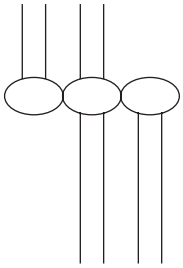
La perspective évolutive entrouverte par ce processus est encore plus étranger au darwinisme que le maintien de la diversité génétique par l'avantage de l'hétérozygote. En effet, dans ces conditions, un même effet de la sélection, au sein d'un même milieu, peut conduire à deux solutions finales opposées, puisque deux populations dont la fréquence de l'allèle *A1* serait inférieure à p_e pour l'une, et supérieure à p_e pour l'autre, verraient leurs trajectoires évolutives diverger diamétralement vers l'élimination de *A1* dans la première, et la fixation de *A1* dans la seconde. Or dans la conception darwinienne, la divergence évolutive suppose des conditions sélectives différentes ce qui n'est nullement le cas quand il y a désavantage de l'hétérozygote.

Ce type de sélection a sans doute largement opéré dans l'évolution des caryotypes associée à l'évolution des génomes et des espèces ; ce fut notamment le cas dans la divergence caryotypique entre l'homme et le chimpanzé. Leurs caryotypes se ressemblent beaucoup, ce qui n'est pas trop surprenant pour des espèces relativement proches sur le plan évolutif (plus de 95 % d'identité à l'échelle moléculaire du génome). Ils diffèrent cependant par neufs remaniements de taille majeure (translocations ou inversions) dont une translocation robertsonienne (fusion centrique) qui, chez l'homme associe en un chromosome métacentrique (centromère médian) deux chromosomes acrocentriques (centromère à une extrémité) et indépendants chez le chimpanzé. Aussi le caryotype humain standard présente 46 chromosomes contre 48 chez le chimpanzé. Et il est possible de supposer que cette fusion centromérique, survenue dans une population ancestrale aux deux espèces a été soumise ainsi que sa contrepartie (chromosomes indépendants) à une sélection du type « désavantage de l'hétérozygote », le chromosome fusionné ayant été fixé dans une sous population ayant conduit à *Homo,* alors qu'il a été éliminé dans une sous population à l'origine du chimpanzé.

En effet, la coexistence au sein d'une population d'une paire de chromosome et d'un chromosome fusionné correspond, à l'échelle chromosomique, à un désavantage du porteur d'un caryotype équilibré, assimilable à un désavantage de l'hétéro-

zygote, en raison de leur chute de fertilité, donc de fécondité (figure 7.8). Si on désigne par *A1* la présence d'un chromosome fusionné et par *A2*, celle des deux chromosomes indépendants, la méiose ne pose pas de problème chez les « homozygotes *A1*/*A1* ou *A2*/*A2*, alors qu'elle conduit (figure 7.8), à une proportion importante de gamètes déséquilibrés chez les « hétérozygotes » *A1*/*A2* réduisant ainsi leur fécondité par la formation d'embryons non viables en raison de leur caryotype déséquilibré.

Les individus « hétérozygotes » *A1/A2* sont viables et normaux car leur caryotype est équilibré et ne présente ni excès, ni déficit de gènes, mais les appariements à la méiose concernent six et non pas quatre paires de chromatides, trois et non pas deux centromères.

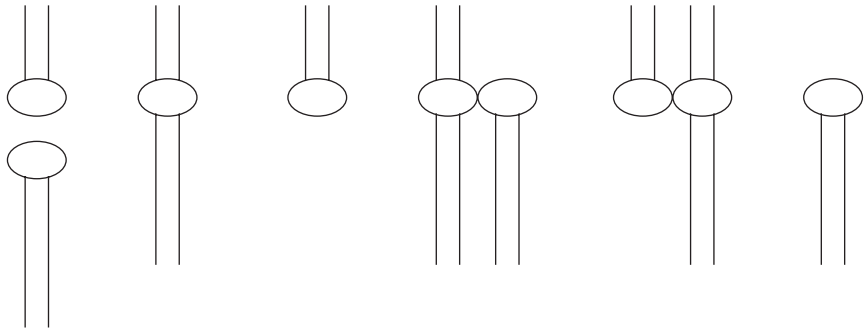


À l'anaphase de la méiose I, deux centromères migrent à un pôle et le troisième au pôle opposé. En fonction des deux centromères qui coségrègent, la méiose peut donner six types de gamètes, selon trois solutions :

Solution 1 :
Deux gamètes équilibrés

Solution 2 :
Un gamète nullisomique,
un disomique

Solution 3 :
Un gamète disomique,
un nullisomique



À la fécondation, après fusion avec un gamète équilibré (soit un chromosome métacentrique, soit deux chromosomes acrocentriques) les deux premiers gamètes donneront un caryotype équilibré et viable, les quatre autres donneront un caryotype déséquilibré (monosomie ou trisomie) non viable, réduisant ainsi la fertilité de l'individu porteur équilibré.

Figure 7.8

Le désavantage du porteur équilibré *A1/A2* définit une fréquence d'équilibre polymorphe instable p_e. Deux sous-populations issues de la population ancestrale ont pu, en raison d'un effet fondateur (voir chapitre 5), présenter une fréquence du chromosome métacentrique supérieure à la fréquence p_e dans l'une, et inférieure à p_e dans l'autre. À terme le chromosome métacentrique aura été fixé dans la première, vers *Homo*, et perdu dans la seconde, vers le chimpanzé.

Le désavantage de l'hétérozygote a donc pu jouer, à l'échelle génétique ou chromosomique, un rôle évolutif dans la spéciation. Il est important de noter que cette différenciation ne met pas seulement en jeu la sélection, à travers le désavantage de l'hétérozygote, mais également un facteur aléatoire, l'effet fondateur, qui positionne au départ les compositions génétiques des deux populations qui vont diverger, de part et d'autre de la fréquence p_e d'équilibre instable. La synergie entre les facteurs déterministes (sélection, mutations, migrations) et les facteurs aléatoires (effet fondateur, dérive) sera développée plus en détail au chapitre suivant.

Ce type de divergence évolutive sous l'effet d'un même mécanisme sélectif, au sein d'un même milieu, mais entre populations séparées, dont les compositions génétiques sont placées de part et d'autre d'un point d'équilibre instable doit amener les biologistes, notamment ceux qui s'intéressent à l'évolution, à beaucoup de prudence lorsqu'ils évoquent la « valeur adaptative d'un caractère » comme cause primitive de différenciation entre populations ou espèces.

Par exemple, dans la logique étroite du darwinisme, les naturalistes tendent à considérer que les différences entre populations, variétés ou espèces ne doivent rien au hasard et tout à la sélection, qu'en d'autres termes ces différences ont ou ont eu une signification adaptative. Dans cette ligne de pensée, faut-il admettre qu'il était plus avantageux pour le chameau d'Asie d'avoir deux bosses et pour le dromadaire d'Afrique de n'en avoir qu'une, et de la même manière pour les cornes du rhinocéros ?

Or un tel résultat peut parfaitement n'avoir aucune signification adaptative et n'avoir été que le plus pur effet du hasard. Imaginons simplement que le nombre de bosses soit déterminé par les deux allèles d'un gène et que l'hétérozygote soit, à un titre divers désavantagé. Selon la position des fréquences alléliques de départ par rapport à la fréquence de l'équilibre polymorphe instable, l'un des allèles sera fixé, et avec lui, le « phénotype une bosse », ou l'autre et avec lui, le « phénotype deux bosses » ! Dans ce cas, ce n'est pas la sélection qui aura choisi, mais le hasard de l'effet fondateur : la sélection n'ayant fait que fixer ce que le hasard avait choisi.

On peut même imaginer que la sélection ne porte pas sur le gène gouvernant le nombre de bosses, mais sur un gène dont le locus est très proche. Il suffit alors que l'information allélique « une bosse » soit génétiquement liée (en déséquilibre de liaison, voir chapitre 3) à l'un des allèles du gène soumis à la sélection, alors que l'allèle « deux bosses » est génétiquement lié à l'autre allèle de ce gène, pour que la fixation de tel ou tel des allèles du gène soumis à sélection entraîne la fixation de l'allèle « une bosse » ou de l'allèle « deux bosses », par un effet d'entraînement dû à la liaison physique des locus. Ce type d'effet d'entraînement, appelé aussi auto-stop (« hitch-hiking » par les auteurs anglo-saxons), doit amener à beaucoup de prudence dans l'évaluation du rôle adaptatif de certains caractères.

7.2.4 Vitesse du processus sélectif pour les maladies létales récessives

Il est facile, dans le cas d'une maladie récessive létale d'établir une relation de récurrence entre les fréquences alléliques de deux générations successives. Dès lors la solution algébrique de la récurrence permettra d'estimer la vitesse avec laquelle la sélection exerce son effet.

Dans le cas d'une maladie récessive létale, on a, à la naissance, la situation suivante, sachant que les allèles *N* et *m* avaient pour fréquence p_0 et q_0, chez les parents :

Génotypes	*N/N*		*N/m*	*m/m*
Fréquences	p_0^2		$2\,p_0\,q_0$	q_0^2
Valeurs sélectives	$\sigma 1$	=	$\sigma 2$	$\sigma 3 = 0$
soit	**1**		**1**	**0**

On a, après l'effet de la sélection, chez les adultes, à l'âge de la reproduction :

Génotypes	***N/N***	***N/m***	***m/m***
Fréquences	$p_0^2/(p_0^2 + 2\,p_0\,q_0)$	$2\,p_0\,q_0/(p_0^2 + 2\,p_0\,q_0)$	0
soit, en simplifiant par *p*	$\boldsymbol{p_0/(1+q_0)}$	$\boldsymbol{2\,q_0/(1+q_0)}$	**0**

> ***Remarque :*** ce résultat correspond aux formules générales développées plus haut (contributions normalisées $\sigma_1 p^2/\sigma$ et $2\sigma_2.pq/\sigma$) pour les valeurs sélectives du cas particulier considéré.

À la génération suivante la fréquence de l'allèle muté pathogène sera égale à la fréquence gamétique chez les parents de la génération 0, soit :

$$f(m) = q_1 = q_0/(1+ q_0)$$

Un même raisonnement appliqué à l'estimation de la fréquence à la génération suivante donnera :

$$f(m) = q_2 = q_1/(1+ q_1)$$

En remplaçant q_1 par sa valeur en fonction de q_0, on obtient :

$$q_2 = q_0/(1+ 2q_0)$$

De la même manière, on obtiendra :

$$q_3 = q_0/(1+ 3q_0)$$

Un raisonnement simple permet de montrer que la propriété de récurrence valable à l'ordre $i - 1$ est valable à l'ordre i, ce qui conduit à :

$$q_i = q_0/(1+ iq_0)$$

Quand *i* tend vers l'infini, la limite de q_i est bien égale à zéro ; l'allèle pathogène *m* finit par être éliminé.

Mais à quelle vitesse ?

Il suffit pour cela de calculer le temps *T* pour réduire de moitié l'écart à la limite, c'est-à-dire pour diviser par deux la fréquence de départ (test équivalent de la période).

Par définition, *T* est tel que $q_T = q_0/2$

Par relation, *T* est tel que $q_T = q_0/(1+ Tq_0)$

d'où on tire que $\boldsymbol{T = 1/q_0}$

Exemple 7.1

Supposons que le paludisme disparaisse d'Afrique équatoriale. L'avantage de l'hétérozygote disparaissant avec lui, l'allèle β^S deviendrait défavorable et serait éliminé. Sa fréquence, au départ égale à 0,20, serait d'abord rapidement puis plus lentement abaissée (tableau 7.3).

Tableau 7.3

Fréquence *q* de l'allèle pathogène à la fréquence *q*/2	**0,2 à 0,1**	**0,1 à 0,05**	**0,05 à 0,025**	**0,025 à 0,0125**	**0,0125 à 0,0062**	**0,0062 à 0,0031**	**0,0031 à 0,0016**
Temps de passage *T*, de la valeur *q* à la valeur *q*/2	5	10	20	40	80	160	320
Temps équivalent en années (chez l'homme)	1 siècle	2 siècles	4 siècles	8 siècles	16 siècles	32 siècles	64 siècles

Aujourd'hui la fréquence de l'allèle β^S est proche de 1 % dans la population égyptienne, alors qu'il n'y a pas de paludisme, mais on sait qu'il y en avait dans l'antiquité et qu'il y a disparu. Si la fréquence de l'allèle β^S était, au départ, de l'ordre de 0,2, il est mathématiquement cohérent qu'après 30 siècles, sa valeur soit encore de l'ordre du 1 % observé.

Remarque 1 : on notera que le temps double pour chaque réduction de moitié de la fréquence, ce qui signifie que l'allèle défavorable ne sera jamais éliminé totalement (sauf effet fondateur ou dérive).

C'est logique, car devenant très rare, l'allèle défavorable est toujours présent chez un porteur sain, rarement chez un homozygote (fréquence q^2 très petite), et n'est donc jamais en situation de donner prise à la sélection (voir à ce sujet la comparaison en 2.4.2.b des tableaux 2.4 et 2.5).

C'est pourquoi on doit considérer que les populations naturelles contiennent dans leur patrimoine génétique un grand nombre d'allèles défavorables, voire létaux, à des fréquences si faibles que la sélection n'a plus de prise sur eux, sauf quand un mariage entre apparentés génère un homozygote pour un tel allèle. On a calculé que chaque homme est porteur sain d'une à deux mutations défavorables très rares.

Remarque 2 : on notera que la fréquence de la maladie est divisée par 4, à chaque fois que la fréquence de l'allèle pathogène est divisée par 2 (à *q*/2 correspond $q^2/4$!).

Exemple 7.2

Si on considère que la mutation ΔF508 du gène *CFTR*, impliqué dans la mucoviscidose (voir chapitre 2) est la mutation la plus fréquemment rencontrée dans l'ensemble des populations européennes (50 à 90 % des exemplaires mutés y sont porteurs de ΔF508, alors qu'il existe plusieurs centaines d'autres mutations du gène), on doit admettre que la mutation principale ΔF508 est apparue dans une population ancestrale des populations européennes et qu'elle est donc assez ancienne. D'ailleurs, l'hypothèse d'une origine unique et ancienne est confortée par les analyses des polymorphismes moléculaires de l'ADN associés à la mutation ΔF508, montrant qu'elle n'est survenue qu'une fois et qu'elle est restée en déséquilibre de liaison étroit avec ses marqueurs polymorphes flanquants. Compte tenu des taux de recombinaison entre ces marqueurs ou des taux de mutations de ces marqueurs, l'âge de la mutation ΔF508 oscille entre 20 000 et 50 000 ans.

Comme un enfant sur 2 500 est atteint de mucoviscidose ce qui correspond à 1 porteur sain sur 25 individus et une fréquence des exemplaires pathogènes du gène égale à 2 % (voir chapitre 2), il suffirait alors de $T = 1/0{,}02 = 50$ générations (1 000 ans) pour réduire la fréquence allélique à 1 % et celle de la maladie à 1/10 000 (tableau 7.3).

Il devient alors difficile d'imaginer qu'une telle mutation, létale, puisse encore demeurer à un niveau de fréquence de 1 à 1,7 %, après avoir subi la sélection pendant si longtemps et il est difficilement imaginable qu'elle ait pu, même à la suite d'un effet fondateur ou de dérive être portée au départ, dans la population ancestrale, à un très haut niveau de fréquence car cela eut été mortel pour la population et la fréquence aurait de toute façon été très vite réduite (tableau 7.3). En effet, si aujourd'hui la fréquence de l'allèle pathogène est égale à 2 %, il faudrait supposer qu'elle était égale à 4 % il y a 25 générations, soit 500 ans, puis à 8 %, 250 ans plus tôt, à 16 %, 125 ans plus tôt ce qui devient vite une solution impossible. Car la dérive ne peut pas hisser la fréquence d'une mutation rare (au départ) à des valeurs comme 16 ou 20 % si la population n'est pas très petite, or on ne connaît aucune population européenne qui puisse avoir exister il y a 1 000 ans, dont la taille était suffisamment faible pour que la dérive puisse hisser la fréquence de ΔF508 à 16 % et qui soit aussi une population ancestrale de l'Europe. Remonter plus loin dans le temps est encore plus absurde car la fréquence de ΔF508 tendrait vite vers un.

Comme il est aussi impossible de concevoir une élévation plus récente et concomittante, par dérive au sein de chaque population européenne puisque la dérive est par nature aléatoire, il faut donc bien imaginer d'une part un avantage de l'hétérozygote pour expliquer le maintien de la fréquence des allèles pathologiques autour de 2 % et, en même temps un

effet fondateur ou de dérive en faveur de ΔF508, au sein de l'ensemble des exemplaires mutés, pour expliquer la prédominance de cette mutation vis-à-vis des autres, phénomène assez ancien pour que la mutation ΔF508 soit aujourd'hui présente et fréquente dans toute l'Europe.

7.2.5 Le fardeau génétique

Il existe plusieurs mécanismes par lesquels le polymorphisme génétique peut être maintenu, mais tout maintien de la diversité génétique a un coût appelé fardeau génétique. Le fardeau génétique peut être défini comme le taux de réduction de l'aptitude moyenne d'une population, pour des raisons génétiques.

Cette aptitude moyenne ou coefficient moyen de sélection a été défini plus haut comme :

$$\sigma = \sigma_1 p^2 + 2\sigma_2 . pq + \sigma_3 q^2$$

En absence de sélection, les trois valeurs sélectives sont égales ; étant définies à un coefficient près de proportionnalité, on peut leur donner la valeur 1 ; dans ce cas $\sigma = 1$.

En présence de sélection, les valeurs sélectives sont différentes ; mais étant toujours définies à un coefficient près de proportionnalité, on peut prendre la valeur 1 pour la plus grande des trois, les autres s'écrivant $1 - s$ et $1 - t$. Dans ce cas $\sigma < 1$. Le fardeau génétique L, défini comme la réduction de 1 à σ s'écrit

$$\boldsymbol{L = 1 - \sigma}$$

La fixation d'un allèle favorable a un coût appelé **fardeau de substitution**, quand on considère qu'un nouvel allèle apparu par mutation *de novo*, ou qu'un ancien allèle, à la suite d'un changement de milieu, se révélant plus « favorable » que les autres allèles, va, par effet de sélection, tendre vers la fixation en éliminant les autres, en se substituant à eux.

Le maintien du polymorphisme par avantage de l'hétérozygote a un coût appelé **fardeau de ségrégation**, car il concerne d'une manière ou d'une autre la ségrégation allélique chez l'hétérozygote. On a montré qu'on pouvait modéliser l'avantage de l'hétérozygote de la manière suivante :

Génotypes :	*A/A*	*A/a*	*a/a*
Valeurs sélectives :	$\sigma 1$	$\sigma 2$	$\sigma 3$
Ou écrites autrement :	$1 - t$	1	$1 - s$

Avec le changement de variable $1 - t/1/1 - s$, les fréquences alléliques à l'équilibre s'écrivent :

$$p_e = (\sigma_3 - \sigma_2)/(\sigma_1 - 2\sigma_2 + \sigma_3) = s/(s + t)$$
$$q_e = (\sigma_1 - \sigma_2)/(\sigma_1 - 2\sigma_2 + \sigma_3) = t/(s + t)$$

Et la valeur sélective moyenne s'écrit :

$$\sigma = \sigma_1 p^2 + 2\sigma_2 . pq + \sigma_3 q^2$$

soit

$$\sigma = (1 - t)p^2 + 2pq + (1 - s)q^2$$
$$\sigma = 1 - tp^2 - sq^2$$

ce qui donne à l'équilibre, avec $p_e = s/(s + t)$ et $q_e = t/(s + t)$:

$$\sigma = 1 - t.s/(s + t)$$

d'où la valeur du fardeau génétique :

$$\boldsymbol{L = ts/(s + t)}$$

Quand le désavantage des homozygotes vis-à-vis de l'hétérozygote est modeste, par exemple $s = t = 0{,}01$ (1 % de réduction de viabilité ou de fertilité), alors $L = 0{,}005$: la fécondité moyenne d'un adulte est égale à 99,5 %. Si chaque individu n'avait en moyenne qu'un descendant, la taille de la population diminuerait inexorablement.

La survie de l'espèce, dans ces conditions de maintien du polymorphisme par avantage de l'hétérozygote, exige que les adultes conçoivent un surplus de descendants au moins égal à 1/0,995 = 1,005 (une population de 10 000 individus devrait former 10 050 zygotes, dont 50 comme tribut à la sélection) afin que la sélection, dont le coût est $L = 0{,}005$, ne ramène pas le niveau net des descendants en dessous de 10 000.

Quand les valeurs sélectives sont élevées, le fardeau génétique relatif à l'avantage de l'hétérozygote est évidemment très lourd. Dans le cas de la drépanocytose (7.2.3.d), où on a évoqué des valeurs $s = 1$ et $t = 0{,}25$, le fardeau serait égal à $L = 0{,}20$, ce qui exigerait une fécondité minimale moyenne de 5 descendants par individu pour maintenir à la fois la population et son polymorphisme, ce qui devient tellement lourd qu'il est difficile d'imaginer que l'avantage de l'hétérozygote ne soit pas autre chose qu'un modèle trop simpliste.

La question est encore plus épineuse si on veut considérer qu'un grand nombre de gènes sont ou semblent maintenus à l'état polymorphe, avec un fardeau génétique pour chacun d'entre eux. Si l'action de la sélection est indépendante pour chacun de ces gènes, la valeur sélective globale est le produit des valeurs sélectives moyennes relatives à chacun des gènes ; elle devient très faible et le fardeau tend très vite vers 1. Alors chaque individu, pour assurer la pérennité de l'espèce et du polymorphisme devrait avoir des dizaines ou des centaines de descendants, ce qui est contradictoire avec la réalité des populations naturelles.

On peut certes arguer du fait qu'en raison des relations fonctionnelles entre certains gènes l'effet de la sélection sur l'un n'est pas indépendant de son effet sur l'autre et que le fardeau est moins intense quand les valeurs sélectives ne sont pas constantes mais variables (voir plus loin). On considère cependant que la sélection naturelle ne peut pas maintenir le polymorphisme pour un nombre élevé de gènes. Comme le nombre de gènes pour lequel une diversité allélique semble maintenue est conséquent, nombreux sont ceux qui ont conclu qu'il était maintenu par d'autres mécanismes, voire qu'il n'était pas maintenu du tout, qu'il était transitoire, sans valeur sélective (sélectivement neutre) et qu'on ne faisait qu'observer la composition à un instant *t*, pour des allèles dont les fréquences ne dépendaient que du hasard d'échantillonnage, lorsqu'il se manifeste, prélude à ce qui fut appelé la théorie neutraliste.

7.3 AUTRES MODÈLES DE SÉLECTION

7.3.1 Introduction

Alors même qu'il était formulé et discuté dans ses implications, le modèle de sélection à coefficients constants était remis en cause. Ce n'est pas tant la contradiction théorique relative au fardeau génétique (voir plus haut) que l'expérience pratique de la plupart des généticiens des populations qui amena ceux-ci à concevoir des modèles plus réalistes parce que plus proches des conditions écologiques concrètement rencontrées par les populations naturelles.

D'une part, on peut imaginer que les valeurs sélectives des génotypes ne dépendent plus seulement du génotype (*via* le phénotype qui est en fait la vraie cible de la sélection) mais aussi des fréquences des génotypes (des phénotypes), donc de celles des allèles, ou bien que l'environnement est hétérogène avantageant un allèle dans une niche écologique et un autre allèle dans la niche adjacente.

D'autre part la sélection s'exerce sur un organisme dans la globalité de sa relation au milieu (autres individus de l'espèce, prédateurs, parasites, ressources alimentaires, etc.) et sa viabilité ou sa fécondité est obligatoirement la résultante de l'effet sélectif de plusieurs gènes. Aussi est-il déjà plus réaliste de concevoir un modèle à deux gènes qu'un modèle à un gène. La généralisation à plus de deux gènes est mathématiquement impossible, mais le modèle à deux gènes réserve assez de surprises pour établir des conceptions théoriques plus réalistes.

7.3.2 Modèles à coefficients variables fonction des fréquences alléliques

On définit, dans ces conditions[1], une fonction « d'adaptabilité » $G(p)$ un peu plus complexe que la fonction Δp définie dans le cadre des coefficients constants, car le génotype désavantagé en deça d'une fréquence devient avantagé au delà. Mais la variation de $G(p)$ entre $p = 0$ et $p = 1$ a la même signification que celle de Δp, et les graphes de $G(p)$, quand les valeurs adaptatives sont fonction des fréquences alléliques, ont la même allure (figure 7.9) et présentent trois équilibres polymorphes dont au moins stable.

De la même façon, que dans le modèle avec désavantage de l'hétérozygote, la sélection va modifier les fréquences alléliques, vers des valeurs stables, en fonction de la valeur initiale de ces fréquences alléliques.

Ainsi, une population de composition génétique stable, à la valeur d'équilibre la plus faible de l'exemple 2 (figure 7.9), peut, à la suite d'un phénomène de goulot d'étranglement démographique, voir la fréquence de cet allèle passer au-dessus de la fréquence d'équilibre instable (deuxième fréquence d'équilibre de l'exemple 2).

1. Pour un exposé intéressant, clair et abordable, voir *Génétique et évolution* par Solignac, Periquet, Anxolabéhère et Petit, 1995, Hermann, Paris.

La sélection va conduire la fréquence de cet allèle vers la valeur d'équilibre stable la plus forte. L'action combinée de l'effet fondateur et de la sélection aura ainsi profondément modifié la composition génétique de la population.

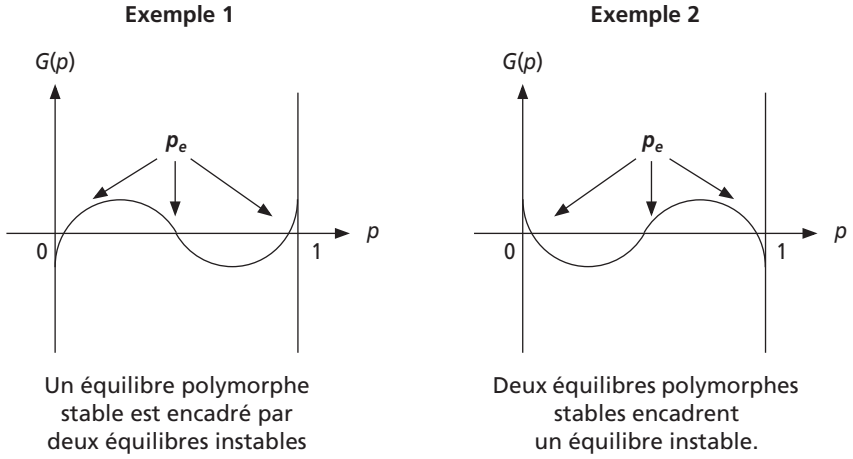


Figure 7.9

7.3.3 Modèles à niches écologiques multiples

Cette éventualité est citée pour mémoire et son traitement mathématique est clairement exposé ailleurs[1]. On conçoit qu'une population dont les génotypes sont répartis dans deux habitats différents par la température, l'hygrométrie ou tout autre paramètre écologique, un premier habitat où l'allèle *A1* est favorable, et un deuxième où l'allèle *A2* est favorable, devrait logiquement garder une composition génétique polymorphe. La réalité est plus complexe parce que le maintien de la diversité génétique dépend aussi de la proportion d'adultes issus de chacun des habitats, si la fécondité moyenne est beaucoup plus faible dans le premier que dans le deuxième, l'allèle *A1* sera globalement favorisé, à moins que des taux de migrations différents d'une sous population vers l'autre ne compensent cette différence de fécondité. On peut aussi définir des modèles plus complexes en associant la sélection fréquence-dépendante et l'environnement multiniche.

7.4 LE PAYSAGE ADAPTATIF

On arrive ici aux limites que l'ouvrage s'est fixé et le lecteur intéressé peut se reporter à l'ouvrage déjà cité[1] et à ses références bibliographiques.

Les modèles de sélection pour deux gènes di-alléliques sont amenés à traiter des matrices de neuf fréquences génotypiques et de neuf valeurs sélectives, en fonction

1. Pour un exposé intéressant, clair et abordable, voir *Génétique et évolution* par Solignac, Periquet, Anxolabéhère et Petit, 1995, Hermann, Paris.

des neuf génotypes possibles pour ces deux gènes (voire un dixième pour le double hétérozygote, si les gènes sont liés), soit :

	A1/A1	**A1/A2**	**A2/A2**
B1/B1	σ_{11}	σ_{21}	σ_{31}
B1/B2	σ_{12}	σ_{22}	σ_{32}
B2/B2	σ_{13}	σ_{23}	σ_{33}

Par ailleurs, la formulation mathématique de la variation des fréquences alléliques doit tenir compte :

– de l'indépendance ou de la dépendance des forces sélectives s'exerçant sur les deux gènes, et de la loi qui les combine dans ce dernier cas ;
– du taux *R* de recombinaison entre les locus des deux gènes ;
– de l'existence d'un déséquilibre gamétique, initial ou généré par la sélection.

L'étude analytique de plusieurs situations a conduit leurs auteurs à la constatation suivante que, même avec des modèles simples, où les coefficients sont constants et les effets sélectifs indépendants entre les gènes, plusieurs solutions d'équilibre polymorphe de la diversité génétique sont possibles. Ainsi, la sélection, bien qu'elle constitue un facteur totalement déterministe, agit, dans la réalité naturelle, avec d'autres facteurs, déterministes ou aléatoires, de sorte que plusieurs trajectoires évolutives peuvent s'offrir à la diversité génétique d'une population. La sélection participe à l'évolution comme moteur le long de la trajectoire évolutive mais ce sont peut être d'autres facteurs qui font, à l'occasion, le choix de la trajectoire, comme l'effet fondateur ou la dérive dans le modèle avec désavantage de l'hétérozygote.

Pour tenir compte de la multiplicité des trajectoires possibles et des solutions polymorphes, quand agissent simultanément des causes déterministes ou des facteurs aléatoires, on a développé une vision graphique à partir des graphes en trois dimensions obtenus dans des situations encore assez simples. La population, compte tenu de sa composition génétique est représentée par un point sur une surface aux reliefs (pics, vallées, plaines, etc.) dépendant des contraintes biologiques ou environnementales auxquelles sont soumis les gènes pris en compte. Cette surface dénommée « paysage adaptatif ».

Dans un espace à deux dimensions, ce paysage adaptatif peut se réduire aux graphes de la fonction σ, vue dans les modèles à coefficients constants (figure 7.7) ou variables (figure 7.9) ; dans les modèles de sélection à deux gènes on est conduit à un paysage à trois dimensions (figure 7.10) avec une fonction *W* équivalente, pour deux gènes, à la fonction σ pour un seul.

Dans le paysage adaptatif de la figure 7.10, les deux pics adaptatifs correspondent soit à la fixation de *A1* et *B2*, soit à la fixation de *A2* et *B1*. La position du point représentatif de la composition génétique de la population sera alors d'un côté ou de l'autre de la ligne joignant les dépressions et séparant les deux pics. Selon sa position,

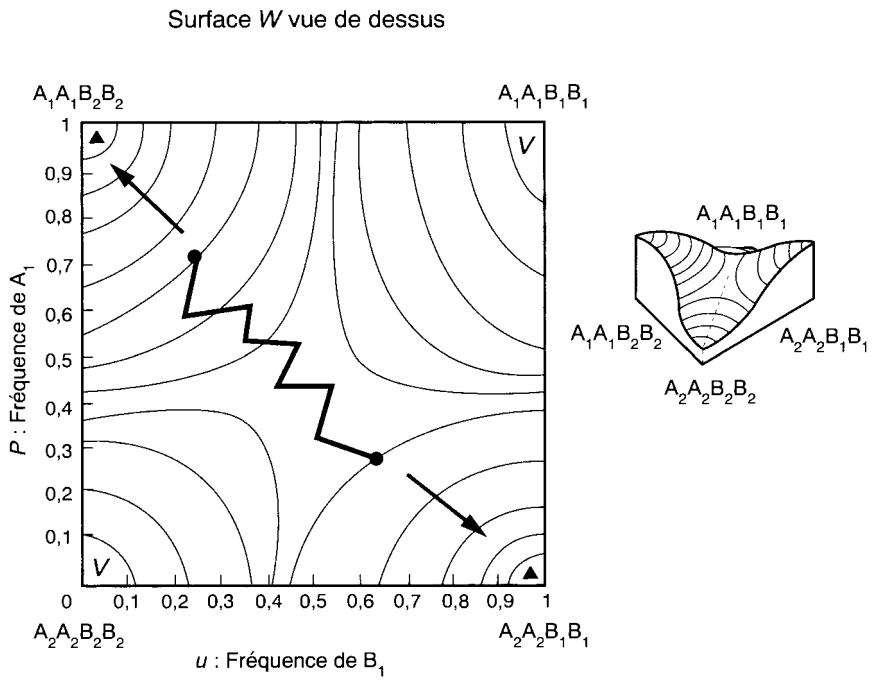


Figure 7.10

On distingue sur ce paysage adaptatif, vu en perspective, deux pics et deux dépressions respectivement symbolisés par ▲ et *V* sur la vue de dessus.

La population est représentée par un point, sur la surface de ce paysage, dont les coordonnées correspondent aux fréquences des allèles A1 et B1 de chacun des deux gènes et à la valeur de W.

Les courbes de niveaux sont les lignes d'égales valeurs de W, la valeur adaptative moyenne de la composition génétique de la population. L'étude de la variation des fréquences alléliques et le calcul des valeurs d'équilibre montre que la population tend à évoluer, sous l'effet de la sélection, vers les valeurs les plus hautes de W (comme pour σ sur les figure 7.7 et 7.9), c'est-à-dire vers les pics appelés logiquement pics adaptatifs.

sur le flanc d'un pic ou de l'autre, la sélection conduira le point vers le sommet correspondant. Mais si l'effectif de la population est assez faible pour que sa composition génétique soit sensible à la dérive, ou si un phénomène d'effet fondateur (perte massive d'effectif ou apport migratoire) survient, il est alors possible que le point représentatif de la population traverse la « vallée » (ligne brisée sur la vue aérienne) pour se retrouver sur l'autre flanc, en route vers l'autre sommet, si la sélection à nouveau devient le facteur principal agissant sur la composition génétique.

Un effet fondateur ou de dérive affectant une population proche de la dépression peut l'y amener ; dans ce cas exceptionnel, il peut s'y maintenir puisqu'elle constitue un « état absorbant », un allèle de chaque gène ayant été fixé (voir 5.2.2.b) ; cependant, il suffit que des néo-mutations surviennent pour que le point représentatif quitte la dépression et se retrouve sur l'un des deux flancs. Enfin, il est tout à fait possible

que l'effet combiné de la sélection, de la dérive et des mutations conduisent le point représentatif de la diversité génétique à suivre, pendant un temps, une trajectoire en partie erratique passant alternativement d'un flanc à l'autre d'un pic sans trop s'éloigner de la ligne de séparation, situation assez proche de ce qui sera désigné, quand il n'y a pas d'effet sélectif, comme « polymorphisme transitoire » (voir 8.3.3).

En conclusion, il est important de remarquer que, dans la figure 7.10 du paysage adaptatif, le choix évolutif, c'est-à-dire le choix du sommet à atteindre a été exercé par un facteur aléatoire (effet fondateur ou dérive) mais que c'est la sélection, avec les valeurs adaptatives des différents génotypes qui a défini la topographie du paysage adaptatif, c'est-à-dire les règles du jeu, la hauteur des pics, la vitesse de progression (quand les facteurs aléatoires n'ont plus assez d'effet pour contrecarrer ceux de la sélection).

Lorsque les phénomènes sélectifs deviennent complexes car nombreux et interactifs, le paysage va lui-même se modifier au cours du temps, comme le relief terrestre sous l'effet combiné de l'érosion et de la tectonique des plaques, de sorte que les « choix du hasard » vont jouer dans un contexte où les autres règles du jeu sont elles-mêmes variables.

Il est donc difficile après un tel exposé de considérer que l'évolution des espèces doive plus à la sélection ou la dérive. Au bout du compte, dans la mesure où le nombre de contraintes sélectives a toujours été très nombreux, et que ces contraintes ont elle-même varié au point de modifier le paysage adaptatif comme les tremblements de terre, les chutes de météorites, l'érosion ou la tectonique des plaques, ont pu modifier le paysage terrestre, il apparaît que le résultat de l'évolution peut bien être considéré comme un résultat réalisé parmi un grand nombre de résultats possibles au départ. De là à considérer que l'évolution ne fut que le résultat d'un coup de dés, c'est-à-dire du hasard, il n'y a qu'un pas que certains ont franchi sans complexes.

RÉSUMÉ

La sélection s'exerce sur un gène dès que les génotypes de ce gène n'ont pas la même fécondité, en raison d'une viabilité ou d'une fertilité différentielle entre ces génotypes.

L'étude de la sélection pour un seul gène, a montré que la composition génétique de la population évoluait selon la relation d'ordre entre les valeurs sélectives (fécondités) des génotypes. Dans le cas où ces valeurs sélectives sont constantes on montre qu'un allèle favorable est fixé et un allèle défavorable éliminé.

Par contre l'avantage de l'hétérozygote maintient la diversité génétique de la population. Le désavantage de l'hétérozygote conduit à la fixation d'un des allèles, mais le choix de l'allèle fixé dépend du hasard des conditions originelles (effet fondateur ou dérive) qui ont fixé les valeurs initiales des fréquences alléliques.

Cette action combinée des facteurs déterministes et aléatoires est encore retrouvée dans les modèles plus réalistes où les valeurs sélectives sont variables et où la sélection est envisagée pour deux gènes au moins. On montre alors que la

population, représentée par sa valeur adaptative, parcourt une surface, appelée paysage adaptatif. La sélection définit la topographie du paysage adaptatif et tend à faire progresser la valeur adaptative vers l'extremum, le pic, le plus proche. Cependant, un effet aléatoire peut faire passer la valeur adaptative de la surface d'un pic à celle d'un autre pic, conduisant alors la population, sous l'effet de la sélection, vers une composition génétique totalement différente de celle vers laquelle elle était primitivement orientée.

EXERCICES

Exercice 7.1

On a étudié dans une espèce végétale allogame, un gène qui affecte la fécondité (nombre efficace de graines) de sorte que les génotypes *A/A*, *A/a* et *a/a* ont respectivement des valeurs sélectives égales à 3/4, 1 et 0.

Question 1 : quelles sont les fréquences attendues des allèles ?

Question 2 : en observant les gousses formées par ces trois types de végétaux, et le nombre moyen de graines qu'elles renferment, on observe 44 végétaux aux gousses racornies dépourvues de graines, 318 végétaux aux gousses longues et pleines et 638 végétaux aux gousses remplies aux 3/4. Calculez les fréquences alléliques et testez la panmixie.

Question 3 : les résultats de la question précédente permettent-ils de dire que la structure génétique de la population est conforme au modèle de Hardy-Weinberg ?

Solution

Question 1 : il y a avantage de l'hétérozygote et l'équation vue en 7.2.3.d permet de montrer que la fréquence d'équilibre de l'allèle *A* est égale à $p_e = 4/5$ ($q_e = 1/5$)

Question 2 : les fréquences des allèles *A* et *a* sont respectivement égales à 0,797 et 0,203.

On teste la panmixie en reconstruisant un échantillon théorique sous l'hypothèse du modèle de Hardy-Weinberg et de la relation existant entre les fréquences alléliques (*p* pour *A* ; *q* pour *a*) et les fréquences génotypiques, soit pour les génotypes *A/A*, *A/a* et *a/a*, les fréquences p^2, $2pq$ et q^2.

Les effectifs théoriques sont respectivement égaux à 635,21 ; 323,59 et 41,20 ce qui donne un χ^2 égal à 0,299 ce qui permet sans grand risque d'accepter la panmixie.

Question 3 : bien évidemment non, le test précédent a permis de valider l'hypothèse panmictique, mais la structure génétique observée simule un équilibre de Hardy-Weinberg. En effet, on sait que le gène étudié est soumis à la sélection et on est simplement, avec cette structure génétique, dans la situation de l'équilibre polymorphe où les fréquences alléliques ne sont pas quelconques mais dépendantes, comme on l'a établi à la question 1, des valeurs sélectives.

Exercice 7.2

La fréquence élevée de la mucoviscidose, maladie récessive réduisant à zéro la fécondité des patients, ne peut être expliquée par un taux élevé de néo-mutations (voir chapitre 8). On est donc amené à postuler, sans l'avoir pour l'instant démontré, l'existence d'un avantage de l'hétérozygote. Quel devrait être l'ordre de grandeur de cet avantage si on considère que la maladie est aujourd'hui à sa fréquence d'équilibre (il y a un enfant sur 2 500 qui est atteint).

Solution

La population étant panmictique, pour ce gène, on peut appliquer la relation de Hardy-Weinberg et estimer la fréquence de *m*, l'allèle pathologique, en prenant la racine carrée de la fréquence des malades, soit $q = \sqrt{1/2\,500} = 1/50 = 2\ \%$.

Si on admet que les génotypes relatifs à la maladie sont soumis à un avantage de l'hétérozygote, on est amené à écrire le système suivant :

Génotypes	*N*/*N*	*N*/*m*	*m*/*m*
Valeurs sélectives	$1 - t$	1	$1 - s$

conduisant à un équilibre polymorphe où les fréquences alléliques stables ne dépendent que des valeurs sélectives selon les équations suivantes :

$f(A) = s/(s + t) = 0{,}98$

$f(a) = t/(s + t) = 0{,}02$

Le désavantage du patient atteint est connu, $s = 1$, puisque sa fécondité est nulle, ce qui permet de déduire la valeur qui serait celle de *t*, le désavantage de l'homozygote sain, dans le cas d'un équilibre par avantage de l'hétérozygote, soit $t = 0{,}0204$ dans ce cas.

On remarque donc que le désavantage de l'homozygote sain est à peu près égal à la fréquence de l'allèle pathogène.

Cette règle peut s'appliquer à tous les traits récessifs rares qui seraient maintenus par un avantage de l'hétérozygote. Quand l'allèle récessif est très rare, il suffit donc d'un très léger avantage de l'hétérozygote (ou désavantage de l'homozygote dominant) pour maintenir en équilibre les fréquences des deux allèles.

Dans le cas de la mucoviscidose, il suffirait d'un désavantage de 2 % de l'homozygote sain sur l'hétérozygote pour maintenir la fréquence de l'allèle pathogène et, partant, de la maladie. Hélas, un tel écart de fécondité n'est pas testable, sur le plan statistique, car cela supposerait des échantillons d'une taille impossible à réaliser, et on s'efforce de valider cette hypothèse par des recherches physiologiques attestant d'un avantage des porteurs sains.

Exercice 7.3

Question 3 de l'exercice 4.3 :

a) En supposant que la population devienne panmictique, quelle serait la fréquence de la maladie ?

b) Dans ces conditions, au bout de combien de générations la fréquence de l'allèle pathologique est elle divisée par deux ?

c) En supposant que la population soit panmictique et qu'un avantage de l'hétérozygote maintienne la fréquence de l'allèle pathologique à sa valeur actuelle, quelle devrait être le désavantage (noté s) de l'homozygote normal par rapport à l'hétérozygote ?

En effet, dans cette question on considérera que l'hétérozygote a une valeur sélective égale à 1 et que les deux homozygotes N/N et *m*/*m* ont respectivement les valeurs sélectives (1 – s) et (1 – *t*), *s* étant le désavantage sélectif de l'homozygote normal et *t*, celui des individus atteints.

Comparez les valeurs obtenues pour *q* et *s* et tirez en une règle générale qui admet d'ailleurs une démonstration algébrique en reprenant les équations utilisées.

Solution

a) Si la population est panmictique, la fréquence de la maladie sera égale à q^2,
soit $0{,}00746^2 = 0{,}00005565 = 1/17\,969$

b) Sous l'effet de la sélection, on a montré (voir 7.2.4) que le temps nécessaire pour diviser par deux la fréquence initiale d'un allèle récessif létal est égal à l'inverse de cette fréquence, soit $1/q = 134$ générations (entre 2 700 et 3 300 ans).

c) Dans ce cas, on peut considérer que l'allèle pathologique étant à sa fréquence d'équilibre, sa valeur vérifie l'équation $q = 1 - p$ où $p = [(1 - t) - 1]/[(1 - s) - 2 + (1 - t)]$
Soit $p = t/(s + t)$ sachant que $t = 1$ puisque l'allèle *m* est létal.
On arrive donc à $p = 0{,}99254 = 1/(1 + s)$ d'où $s = 0{,}0075$

> ***Remarque :*** on montre ainsi que pour maintenir la fréquence *q* de l'allèle létal à une valeur d'équilibre de 0,00746, il est nécessaire d'avoir un avantage de l'hétérozygote et un désavantage sélectif pour l'homozygote normal égal à 0,0075, soit la valeur de la fréquence de l'allèle pathologique, si on considère que la valeur sélective de l'hétérozygote est égale à 1.

En effet, on a écrit que $p = 1 - q = 1/(1+s)$ ce qui conduit à $s = q/(1 - q)$ et comme *q* est petit devant 1, à la conclusion que $s = q$.

Exercice 7.4

La fréquence de la phénylcétonurie (voir chapitre 2) est égale à 1/16 000.

Sans traitement, cette maladie conduit à une arriération mentale et à une fécondité nulle, mais les individus dépistés et traités recouvrent une fécondité normale.

Quelles sont les valeurs des fréquences de l'allèle pathologique, dans les deux conditions définies plus haut (sans ou avec traitement), au bout d'une génération ? Et la fréquence de la maladie ? Combien de générations pour diviser, ou multiplier, la fréquence de la maladie par quatre ?

On suppose que le taux de néo-mutations est élevé dans ce gène et égal à 2.10^{-4}.

Solution

L'application de la relation panmictique permet d'estimer la fréquence q de l'allèle pathologique :

$$R = 1/16\,000 = q^2$$

D'où $q = 0{,}8\,\% = 0{,}0079$

Sans traitement, à la génération suivante, on aura $q' = q/(1 + q)$ (voir 7.2.4)

Soit $q' = 0{,}00784$

Et la maladie aurait une fréquence égale à 1/16 253

Ce qui est logiquement plus faible puisque la fréquence de l'allèle pathologique est sensée avoir chuté ; il faudrait cependant 127 générations pour diviser la fréquence allélique par deux et celle de la maladie par quatre (carré de la fréquence allélique).

Avec le traitement, la fréquence est augmentée des néo-mutations soit

$$q' = (79 + 2).10^{-4}$$

Soit $q' = 0{,}0081$

Et la maladie aura une fréquence égale à 1/15 241.

Ce qui est légèrement plus fréquent puisque la sélection n'a pas opéré.

Pour multiplier la fréquence de la maladie par quatre, il faudrait multiplier la fréquence de la mutation par deux. En supposant, pour simplifier (dans les faibles valeurs) que l'apport des néo-mutations est constant, il faudrait n générations tel que $n \times (2.10^{-4}) = 0{,}0079$ soit $n = 40$ générations, soit un millénaire.

Chapitre **8**

Effet combiné de plusieurs facteurs déterministes et non déterministes

8.1 INTRODUCTION

Bien que le modèle de Hardy-Weinberg soit purement théorique, servant de modèle de référence dans une population idéale où les conditions semblent irréalistes, on a pu constater que ce modèle était utile et applicable dans des conditions limitées de temps (quelques générations), d'espace (population bien circonscrite ou définie) et pour des caractères dont la variation phénotypique dépendait des allèles d'un ou deux gènes.

Mais, à l'échelle de l'évolution, et parfois même au présent, il n'est pas conforme à la réalité et il faut considérer alors que mutations, sélection et dérive combinent leurs effets, auxquels s'ajoutent celui des migrations entre populations de la même espèce et, éventuellement les écarts à la panmixie au sein des populations. Ces différents facteurs qui combinent leurs effets se distinguent par le fait qu'ils sont :

– déterministes, quand il est possible, selon la nature et l'intensité de l'effet, de définir l'état final de la composition génétique et la vitesse exacte du processus (mutations, sélection et migrations) ;
– aléatoires (stochastique) quand le phénomène ou l'effet met en jeu une variation d'échantillonnage, ce qui permet de prévoir l'état final de la population seulement sous la forme d'un ensemble d'états, affectés d'une probabilité, dont l'un seulement sera réalisé, et pas forcément le plus probable. Par ailleurs, la vitesse des processus n'est, elle aussi définissable qu'en espérance (dérive génétique, effet fondateur).

Depuis leur origine et durant leur évolution génétique, toutes les espèces ont vu leurs compositions génétiques soumises à l'effet simultané de ces différents facteurs. La combinaison entre la sélection qui définit un paysage adaptatif et la dérive ou l'effet fondateur qui y positionne le point de départ de la trajectoire évolutive (voir chapitre précédent) était l'un des premiers exemples d'effet combiné d'un facteur déterministe et d'un facteur stochastique.

Il est aussi possible que la variation d'un grand nombre d'effets déterministes indépendants « simule le hasard de la dérive » dans la mesure où l'état final devient imprévisible si les paramètres déterministes sont inconstants et indépendants.

Ce chapitre se limitera à une présentation de quelques une des combinaisons, parmi les plus simples, entre plusieurs effets déterministes et ou stochastiques.

8.2 ÉQUILIBRES SÉLECTION-MUTATION

8.2.1 Définition et approche intuitive

La sélection peut maintenir la diversité génétique dans les cas exceptionnels d'avantage de l'hétérozygote. Mais pour la plupart des gènes, la sélection est finalement darwinienne et tend à éliminer les allèles défavorables pour fixer l'allèle le plus favorable.

Cependant le maintien de la diversité génétique est encore possible, sous une sélection darwinienne, si des variations du milieu modifient les valeurs sélectives (un allèle favorable devenant alors défavorable, et réciproquement) ou si les valeurs sélectives sont fréquence-dépendantes.

Enfin, les mutations sont toujours là pour empêcher, quel que soit le gène, la composition génétique d'une population, ou d'une espèce, de parvenir ou de demeurer au mono-allélisme. Pour tout gène, un équilibre polymorphe est attendu dans une population, quand le flux d'allèles défavorables éliminés par la sélection sera équilibré par un flux égal d'allèles défavorables apparus *de novo* par mutation. Comme les taux de mutations sont très faibles, on doit s'attendre à ce que les mutations *de novo* ne maintiennent la présence des allèles défavorables qu'à un très faible niveau de fréquence.

8.2.2 Changement de formalisme pour les valeurs sélectives

Un changement d'écriture des valeurs sélectives facilite le traitement algébrique des équilibres sélection-mutation. Dans le cadre du modèle sélectif à coefficients constants, les trois génotypes d'un gène bi-allélique sont affectés d'une valeur sélective, définie à un coefficient près de proportionnalité, ce qui permet d'opérer un changement de formalisme (tableau 8.1).

Dans ce changement d'écriture, la valeur sélective de l'un des deux homozygotes est prise comme référence et les valeurs des deux autres génotypes apparaissent comme la valeur de référence modifiée par un écart, fonction de deux paramètres h et s.

Tableau 8.1

Génotypes	*A1/A1*	*A1/A2*	*A2/A2*
Fréquences	p^2	$2pq$	q^2
Valeurs sélectives	σ_1	σ_2	σ_3
Ou écrites autrement	$1 - s$	$1 - hs$	1

➤ Le paramètre *s* est le paramètre sélectif

– il n'y a pas de sélection si $s = 0$; il y a sélection si s n'est pas nul ;
– il y a désavantage sélectif si s est positif (et inférieur à 1) ;
– il y a avantage sélectif si s est négatif.

➤ Le paramètre *h* est le paramètre de dominance (pour l'effet de la sélection)

Il situe le phénotype de l'hétérozygote vis-à-vis du phénotype des deux homozygotes.

– si l'hétérozygote *A1*/*A2* a la même fécondité que *A1*/*A1*, cela revient à écrire que $h = 1$;
– si l'hétérozygote *A1*/*A2* a la même fécondité que *A2*/*A2*, cela revient à écrire que $h = 0$;
– si l'hétérozygote *A1*/*A2* a une fécondité supérieure à celle des homozygotes, cela revient à écrire que h est négatif ;
– si l'hétérozygote *A1*/*A2* a une fécondité inférieure à celle des homozygotes, cela revient à écrire que h est positif (avec hs inférieur à 1).

Avec ce changement de formalisme, l'équation

$$\Delta p = (pq/\sigma)[(\sigma_1 - \sigma_2)p + (\sigma_2 - \sigma_3)q]$$

s'écrit

$$\Delta p = \left[pq/(1 - sp^2 - 2hspq)\right]\left[s(h-1)p - hsq\right]$$

8.2.3 Équilibre sélection-mutation pour un allèle défavorable à effet sélectif « dominant »

La situation générale des maladies autosomiques dominantes est décrite par le tableau 8.2.

Tableau 8.2

Phénotypes	[atteint]	[atteint]	[sain]
Génotypes	*M/M*	*M/n*	*n/n*
Fréquences	p^2	$2pq$	q^2
Valeurs sélectives	$1 - s$	$1 - hs$	1

La valeur de h est égale à 1 en cas de dominance totale, mais il arrive souvent, notamment pour les maladies dites dominantes chez l'homme que l'homozygote muté *M/M* soit en fait beaucoup plus atteint que le porteur d'une seule mutation *M/n* ; dans ce cas il y a dominance partielle ou codominance, mais on garde le terme de maladie dominante.

a) Effet de la sélection

Comme l'allèle *M* est défavorable (sélection darwinienne), sa fréquence va baisser sous l'effet de la sélection et ne sera équilibrée par celui des mutations *de novo* qu'au voisinage de $p = 0$, condition sous laquelle on pourra légitimement effectuer des simplifications algébriques.

La variation de fréquence allélique, sous l'effet de la sélection, noté Δp_S est égale à :

$$\Delta p_S = \left[pq/\left(1 - sp^2 - 2hspq\right)\right]\left[s(h-1)p - hsq\right]$$

mais, au voisinage de $p = 0$, l'équation se réduit à :

$$\boldsymbol{\Delta p_S = -\, hsp}$$

En effet :

• $pq = p(1 - p)$ est égal à p au voisinage de $p = 0$;

• $(1 - sp^2 - 2hspq)$ est égal à 1, au voisinage de $p = 0$, car sp^2 est infiniment petit, ainsi que $2hspq$, puisque s et h sont positifs et inférieurs à 1 ;

• $[s(h-1)p - hsq]$ est égal à hs, au voisinage de $p = 0$, car $s(h-1)p$ est infiniment petit devant hsq qui est proche de hs, q étant proche de 1.

b) Effet des mutations

La variation de fréquence allélique, sous l'effet des mutations, noté Δp_M est égale à :

$$\Delta p_M = v - (u + v)p \qquad \text{(voir chapitre 6)}$$

mais, au voisinage de $p = 0$,

on peut écrire que $\boldsymbol{\Delta p_M = v}$

c) Équilibre sélection-mutations de novo

À l'équilibre, on peut écrire que : $\Delta p = \Delta p_M + \Delta p_S = 0$

ce qui revient à écrire que : $\boldsymbol{\Delta p_M = -\ \Delta p_S}$

ce qui revient à écrire que le flux de gènes éliminés par la sélection équilibre celui des gènes apportés par les mutations *de novo*. On tire de cette dernière équation que :

$hsp = v$ (ou μ, écriture courante d'un taux de mutation), d'où, à l'équilibre

$$\boldsymbol{p = \mu/hs}$$

Si la maladie est totalement dominante, on a : $\boldsymbol{p = \mu/s}$

Et si la maladie est létale, on a : $\boldsymbol{p = \mu}$

D'un point de vue graphique (figure 8.1) les solutions algébriques $p = \mu/s$ ou $p = \mu$ correspondent à l'abscisse du point d'intersection de la courbe Δp_S avec la droite $-\Delta p_M$.

$\Delta p_S < 0$ entre 0 et 1
p tend vers 0
Au point d'abcisse p_e
$\Delta p_S = -\Delta p_M$
Alors $\Delta p = \Delta p_S + \Delta p_M = 0$
il y a maintien de la diversité génétique
à ce niveau de fréquences alléliques

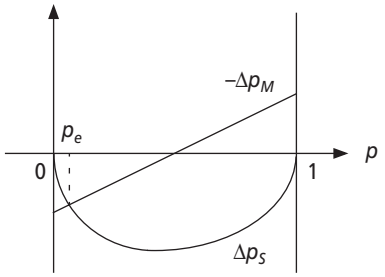


Figure 8.1 Relation d'ordre : $1 - s < 1 - hs < 1$.

d) Application à la mesure des taux de mutation

Les formules $p = \mu/hs$ et $p = \mu/s$ sont très utiles pour estimer les taux de mutations *de novo* dans les gènes impliqués dans quelques maladies dominantes où la valeur sélective *s* est estimable (tableau 8.3).

TABLEAU 8.3 EXEMPLES DE TAUX DE MUTATIONS CALCULÉS À PARTIR DE LA FRÉQUENCE *M* D'UNE MALADIE DOMINANTE ET DE LA RÉDUCTION DE FÉCONDITÉ QU'ELLE INDUIT.

Maladie	Fréquence *m* de la maladie	Fréquence *p* de la mutation	Valeur sélective (1 – *s*)	Taux μ de mutation
Aniridie	1/80 000	1/160 000	0,8	$1{,}25.10^{-6}$
Maladie de Huntington	1/50 000	1/100 000	0,8	$0{,}2.10^{-5}$
Achondroplasie	1/10 000	1/20 000	0,2	4.10^{-5}
Rétinoblastome (début du siècle)	1/36 000	1/72 000	0,1	$1{,}25.10^{-5}$

Pour ces maladies dominantes, *p* est très petit, les homozygotes sont inexistants et les seuls individus atteints sont hétérozygotes ; la fréquence *p* de l'allèle pathogène est donc égale à la moitié de la fréquence *m* des individus atteints (voir chapitre 2). Connaissant *p* et s, on peut en déduire μ.

d) L'effet dysgénique de la médecine

Ce sont les progrès décisifs de la médecine qui ont remis d'actualité, à partir de la fin du XIX^e^ siècle, des conceptions remontant à Platon dans l'antique Athènes. Pour le courant eugéniste, la médecine, en sauvant des individus autrefois condamnés par une maladie qu'on ne savait guérir, contrecarrerait l'effet de la sélection naturelle. Ce faisant la médecine permettrait à ces individus de propager leurs gènes, notamment leurs gènes délétères, de moindre résistance ou même pathologiques. Il s'en suivrait « un effet dysgénique » défini comme un accroissement de la fréquence des allèles délétères dans le pool génique. Au nom du bien de l'humanité et du respect

de la nature et des lois qui la gouvernent (sélection naturelle), il faudrait, si ce n'est renoncer à la médecine, tout au moins éviter qu'elle laisse se perpétuer, voire s'accroître, les éléments les plus délétères !

Un exemple simple de génétique des populations devrait remettre à leur juste place les délires fantasmatiques de ce courant de pensée obsédé par « l'envahissement inéluctable du stock génique par les tares génétiques » !

À la fin du siècle dernier, environ un individu sur 40 000 souffrait d'une forme de cancer de la rétine, le rétinoblastome. Vers 1890, de Graeff proposa l'énucléation des enfants ou des jeunes adultes (afin d'éviter les métastases). En sauvant ainsi quelques patients, on s'aperçut que certains rétinoblastomes se transmettaient selon un mode dominant, un enfant sur deux se trouvant atteint de rétinoblastome. L'analyse génétique puis moléculaire des formes héréditaires de rétinoblastome, dans les années 70-80 a permis d'identifier le gène impliqué : l'anti-oncogène *RB*.

➤ Calcul du taux de mutation

À la fin du XIX^{e} siècle, quand la maladie ne pouvait être soignée, la valeur sélective (1 – s) était égale à 0 (décès avant l'âge reproducteur), soit $s = 1$.

La fréquence F des individus atteints étant égale à 1/40 000, la fréquence p de l'allèle pathogène est donc égale à 1/80 000, moitié de la fréquence des atteints (pour une maladie dominante). On peut en tirer une estimation du taux de mutation *de novo* vers l'allèle pathogène :

$$p = \mu/s, \text{ avec } s = 1$$

d'où $\mu = p = \mathbf{1/80\ 000}$

➤ Mesure de l'effet dysgénique de la médecine

En l'absence de soins ($s_0 = 1$), la fréquence de l'allèle pathogène devait rester égale au taux de mutation, soit $p_0 = 1/80\ 000$.

Après 1890, grâce à la proposition d'énucléation de von Graeff, on obtient un taux de guérison d'environ 10 % des patients, soit une valeur $s_1 = 0{,}9$. Cette nouvelle valeur sélective induit un nouvel équilibre sélection-mutation, la nouvelle fréquence d'équilibre de l'allèle pathogène sera égale à $p_1 = \mu / s_1 = 1/80\ 000/0{,}9 = 1/72\ 000$

Certes la médecine a un effet dysgénique en augmentant la fréquence d'équilibre de l'allèle pathogène de la valeur 1/80 000 à 1/72 000 !

En supposant, avec optimisme, que le taux de guérison puisse bientôt atteindre 90 %, la valeur sélective s passerait alors à $s_2 = 0{,}1$, au lieu de 0,9.

Il s'ensuivrait une nouvelle fréquence d'équilibre $p_2 = \mu/s_2 = 1/80\ 000/0{,}1 = 1/8\ 000$.

Mais il convient de raison garder devant cet effet dysgénique :

– d'une part, même multipliée par dix, la fréquence d'équilibre de l'allèle pathogène reste très faible, alors que le taux de guérison est élevé ;
– d'autre part, la fréquence d'équilibre ne sera atteinte qu'après un temps très long. En effet, avec un taux de mutation μ égal à 1/80 000, le temps T nécessaire pour réduire de moitié l'écart entre la fréquence et sa valeur limite est égal à $0{,}7/\mu$ (voir

chapitre 6), soit 55 452 générations, environ 1,1 millions d'années ! Il y a certainement des problèmes plus urgents pour l'humanité que l'effet dysgénique de la médecine, pour le gène du rétinoblastome ou de toute autre maladie !

Il est donc utile de rappeler encore une fois le contexte idéologique du discours sur l'effet dysgénique de la médecine, ce délire eugéniste, visant à débarrasser le pool génique de tout allèle délétère parce qu'il se fonde sur le fantasme d'une relation si stricte entre gènes et phénotypes qu'il considère que tout est génétique et même héréditaire, de l'alcoolisme à l'adultère en passant par l'échec scolaire et la déqualification du travail. Derrière ce discours, se profile une politique extrémiste, totalitaire ou ultralibérale, toujours antisociale, prônant l'entrave à la « reproduction » de tous les individus « inaptes » comme un moyen légitime car assimilable à la sélection naturelle.

8.2.4 Équilibre sélection-mutation pour un allèle défavorable à effet sélectif « récessif »

La situation générale des maladies autosomiques récessives est décrite par le tableau 8.4. La valeur de *h* est ici égale à 0, seuls les homozygotes *m*/*m* présentant un désavantage sélectif égal à *s*.

Tableau 8.3

Phénotypes	[sain]	[sain]	[atteint]
Génotypes :	N/N	N/m	m/m
Fréquences :	p^2	$2pq$	q^2
Valeurs sélectives :	$\sigma 1$	$\sigma 2$	$\sigma 3$
Ou écrites autrement :	1	$1 - hs$	$1 - s$
Valeurs sélectives :	1	1	$1 - s$

Comme l'allèle *m* est défavorable, sa fréquence va baisser sous l'effet de la sélection et ne sera équilibrée par celui des mutations qu'au voisinage de $q = 0$ (ou $p = 1$) condition sous laquelle pourront être légitimement effectuées les simplifications algébriques.

a) Effet de la sélection

On a montré (chapitre 7) que la variation de fréquence allélique, sous l'effet de la sélection, notée Δp_S est égale à :

$$\Delta p_S = [pq/\,(\sigma)]\,[(\sigma 1 - \sigma 2)p + (\sigma 2 - \sigma 3)q]$$

soit, avec $h = 0$: $\Delta p_S = [pq/\,(1 - sq^2)]\,[sq]$

mais, au voisinage de $q = 0$ ou $p = 1$, on peut écrire que

$$\boldsymbol{\Delta p_S = sq^2}$$

En effet :

• $pq = q(1 - q)$ est égal à q au voisinage de $q = 0$,
• $(1 - sq^2)$ est égal à 1, au voisinage de $q = 0$, car sq^2 est infiniment petit.

Sachant que $p = q$, on en déduit que la variation de q est l'opposée à celle de p, soit : $\Delta q_S = -\Delta p_S$, d'où :

$$\boldsymbol{\Delta q_S = -sq^2}$$

b) Effet des mutations

La variation de fréquence allélique q, sous l'effet des mutations (voir chapitre 6), noté Δq_M est égale à :

$$\Delta q_M = u - (u + v)q$$

mais, au voisinage de $q = 0$, on peut écrire que

$$\boldsymbol{\Delta q_M = u}$$

c) Équilibre sélection-mutations

À l'équilibre, on peut écrire que : $\Delta q = \Delta q_M + \Delta q_S = 0$

ce qui revient à écrire : $\boldsymbol{\Delta q_M = -\Delta q_S}$

ce qui revient à écrire que le flux de gènes éliminés par la sélection équilibre celui des gènes apportés par les mutations *de novo*. On tire de cette dernière équation que :

$sq^2 = u$ (ou μ, écriture courante d'un taux de mutation)

d'où, à l'équilibre $q = \sqrt{\mu / s}$

Si la maladie est létale, on a : $q = \sqrt{\mu}$

D'un point de vue graphique, les solutions algébriques $q = \sqrt{\mu / s}$ ou $q = \sqrt{\mu}$ correspondent (figure 8.2) à l'abscisse du point d'intersection de la courbe Δq_S avec la droite $-\Delta q_M$.

$\Delta q_S < 0$ entre 0 et 1
q tend vers 0
Au point d'abscisse q_e
$\Delta q_S = -\Delta q_M$
Alors $\Delta q = \Delta q_S + \Delta q_M = 0$
il y a maintien de la diversité génétique
à ce niveau de fréquences alléliques

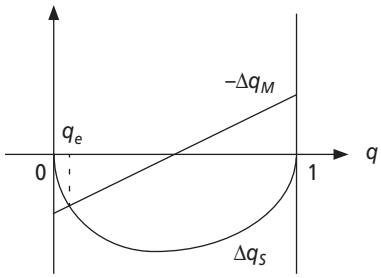


Figure 8.2 Relation d'ordre : $1 = 1 > 1 - s$.

d) Application aux maladies génétiques récessives chez l'homme

Le tableau 8.5 présente les valeurs supposées de taux de mutation pour des gènes impliqués dans des maladies récessives, compte tenu des valeurs sélectives connues pour ces maladies, en supposant que les fréquences alléliques sont à l'équilibre.

Tableau 8.5

Maladie	Fréquence *m* de la maladie	Fréquence *q* de la mutation	Valeur sélective (1 – *s*)	Taux μ de mutation
Mucoviscidose	1/2 500	1/50	0	4.10^{-4}
Mucoviscidose (sans ΔF508)	1/27 778	1/166	0	$3{,}6.10^{-5}$
Phénylcétonurie	1/10 000	1/100	0	10^{-4}
Galactosémie	1/40 000	1/200	0	$2{,}5.10^{-5}$

Sachant que les taux de mutations *de novo* sont rarement supérieurs à 10^{-5}, il est difficile d'imaginer que le multi-allèlisme des gènes impliqués dans ces maladies soit maintenu par un équilibre sélection-mutation, puisqu'un tel équilibre supposerait des valeurs bien trop élevées pour les taux de mutation.

Ce type d'équilibre peut être invoqué pour la majorité des maladies récessives atteignant moins de un individu sur 50 000, pas pour les maladies « fréquentes », et spécialement pas pour la mucoviscidose.

e) Les paradoxes de la mucoviscidose

La mucoviscidose est une maladie récessive, exceptionnellement fréquente dans les populations caucasoïdes (Europe, Maghreb, Proche-Orient), pour une maladie létale dans l'enfance. En effet, avec, en moyenne, un enfant atteint sur 2 500, la fréquence du gène muté y est égale à 1/50 et la fréquence des porteurs sains à 1/25.

1. Comment expliquer la fréquence élevée d'allèles létaux alors que la sélection est dans ce cas très efficace pour la diminuer (voir chapitre 7) ?

2. Comment expliquer que la mucoviscidose ne soit fréquente qu'en Europe ?

L'analyse moléculaire du gène a montré que la fréquence exceptionnellement élevée de la mucoviscidose est due à celle de la mutation principale du gène *CFTR* (*Cystic Fibrosis Transmembrane Regulator*) la mutation ΔF508. Si on ne tenait pas compte de cette mutation (tableau 8.5, deuxième ligne), qui représente 70 % des allèles mutés, les mille et une autres mutations du gène auraient globalement une fréquence égale à 0,006 (1/166), et la maladie serait dix fois plus rare, touchant environ un enfant sur 27 800 !

En excluant la mutation ΔF508, le multi-allélisme du gène *CFTR* pourrait être maintenu avec un taux de mutation encore élevé mais presque acceptable de $3{,}6.10^{-5}$. Et c'est sans doute la situation qui prévaut pour ce gène hors d'Europe.

Finalement les deux questions posées primitivement quant aux mutations responsables de la mucoviscidose dans son ensemble ne se posent donc en fait que pour la seule mutation ΔF508.

Les études des polymorphismes moléculaires, flanquant le gène *CFTR* ou internes à celui-ci, ont montré que la mutation ΔF508 n'était survenue qu'une fois dans l'histoire génétique des populations caucasoïdes et qu'elle est suffisamment ancienne, entre 20 000 et 50 000 ans, pour être survenue dans une population mère

de toutes les populations européennes. Cela explique sa présence de l'Europe de l'Ouest au Proche-Orient et au Maghreb.

Il semble difficile de ne pas admettre l'action d'un facteur stochastique comme la dérive génétique, dans cette population primitive, comme seule cause possible d'une élévation importante de la fréquence d'un allèle létal, malgré la sélection ; et c'est l'allèle ΔF508 qui aurait été « désigné » par la dérive !

Depuis cette époque la sélection contre cet allèle défavorable aurait dû abaisser efficacement sa fréquence à moins que le maintien de celle-ci n'ait été, dès ce moment, favorisé par un avantage de l'hétérozygote. Cette hypothèse est confortée sur le plan physiologique par le fait que les hétérozygotes présenteraient une sensibilité légèrement réduite aux toxines de la bactérie responsable du choléra (il s'agit toujours d'une hypothèse de travail qui n'est pas unanimement partagée par tous les spécialistes de la question).

Cependant, comme le choléra est au moins aussi fréquent en Afrique ou en Asie qu'en Europe, il est difficile de comprendre pourquoi l'avantage de l'hétérozygote n'aurait pas aussi promu, dans ces continents, le développement d'autres allèles mutés du gène *CFTR* et avec eux de la mucoviscidose ?

Tout simplement parce que l'augmentation de la fréquence des allèles mutés sous le seul effet de la sélection est très lente et que les populations humaines sont trop « jeunes » pour que l'effet de la sélection puisse y être perceptible, s'il n'est pas « aidé » par celui de la dérive. Il s'est sans doute trouvé que, par hasard, en Europe, l'un des allèles mutés du gène *CFTR*, la mutation ΔF508, s'est trouvé porté à un niveau de fréquence égal ou proche du niveau d'équilibre polymorphe et la sélection n'a plus eu qu'à maintenir ce niveau de fréquence, ayant été dispensée de l'établir.

Cette double hypothèse de la dérive affectant la fréquence de la seule mutation ΔF508, dans une population mère de l'Europe, et de l'avantage de l'hétérozygote affectant toutes les mutations du gène *CFTR* permet de rendre compte de l'ancienneté et de la localisation européenne de la mutation ΔF508, de la fréquence élevée de la mucoviscidose en Europe ainsi que de son absence des autres continents.

8.2.5 Équilibre sélection-mutation pour un gène « lié au sexe » : la règle de Haldane

Chez l'homme, les traits ou les maladies gouvernés par des gènes localisés sur le chromosome X sont dits « liés au sexe » parce que leur transmission héréditaire n'est pas indépendante de celle du sexe. Les mutations des gènes responsables d'une maladie récessive liée à l'X ont la particularité d'avoir un effet récessif chez les femmes, où l'hétérozygote est porteuse saine, mais de « simuler un effet dominant », du point de vue de la sélection, chez les hommes qui sont hémizygotes. Dans le sexe homogamétique, seuls les allèles mutés chez les homozygotes seront accessibles à l'effet de la sélection (tableau 8.6) ; chez les femmes hétérozygotes, l'allèle muté ne sera pas touché par la sélection, car il sera « protégé » de la sélection par l'allèle non muté ; au contraire, cet allèle sera toujours touché par la sélection chez les hommes, comme si c'était un allèle « dominant ». Cette différence aboutit à une situation particulière que décrit la « règle de Haldane ».

Tableau 8.6 Composition génétique d'une population, à l'équilibre des fréquences alléliques entre les sexes, pour un gène impliqué dans une maladie récessive liée au sexe.

	sexe féminin			sexe masculin	
Génotypes	*A/A*	*A/a*	*a/a*	A/Y	a/Y
Valeurs sélectives	1	1	1 – s	1	1 – *s*
Fréquences	p^2	$2pq$	q^2	*P*	q

Pour une maladie récessive (ou un trait rare) la fréquence des individus atteints est égale à q chez les hommes et q^2 chez les femmes.

Si q est très petit, les filles sont très rarement atteintes (q^2) et sont essentiellement conductrices (porteuses saines) avec une fréquence égale à $2pq$, soit $2q$ si q est petit.

La variation de fréquence, du fait de la sélection, est égale à $-sq^2$, chez les femmes, comme cela a été démontré pour les mutations récessives (voir plus haut).

La variation de q sera égale à $-sq$, chez les hommes, comme cela a été montré pour les mutations dominantes (voir plus haut).

La variation de fréquence sous l'effet de la sélection dans la population est la somme des deux variations, féminine et masculine, pondérée par la répartition 2/3 – 1/3, entre les deux sexes, des chromosomes X, porteurs des allèles mutés, soit :

$$\Delta q_s = [-sq^2 \times 2/3] + [-sq \times 1/3]$$

Or, au voisinage de $q = 0$, la variation de fréquence chez les femmes est négligeable devant la variation de fréquence chez les hommes, si bien que la variation de fréquence due à la sélection s'écrit en fait :

$$\Delta q_S = -sq/3$$

Pour une maladie récessive liée à l'X, l'effet de la sélection est donc semblable à celui exercé sur un allèle dominant ne concernant que le tiers des allèles mutés, ceux qui sont présents dans le sexe masculin, les deux tiers des allèles mutés féminins échappant à la sélection, en raison de leur récessivité.

Si le taux de mutation vers l'allèle pathogène a la même valeur μ dans chacun des sexes, la variation de fréquence due aux mutations *de novo* s'écrit :

$$\Delta q_M = \mu$$

À l'équilibre, on a : $\Delta q_S = \Delta q_M + \Delta q_S = 0$

D'où on tire que $\boldsymbol{\mu = sq/3}$

et $\boldsymbol{\mu = q/3}$ (pour les maladies létales où $s = 1$)

Cette relation dite « règle de Haldane » permet :

– d'estimer facilement les taux de mutations pour les maladies récessives liées au sexe, connaissant la fréquence q des garçons atteints ;
– de conclure qu'un tiers des garçons atteints ($q/3$) sont porteurs de mutations *de novo*.

Cette conclusion est d'une grande importance en médecine, pour des maladies comme la myopathie de Duchenne-Becker (tableau 2.7), puisqu'elle montre que les meilleures stratégies de dépistage et de diagnostic prénatal ne permettront jamais de prévenir la naissance d'au moins un tiers des garçons atteints, ceux qui sont porteurs d'une néo-mutation.

8.3 ACTION COMBINÉE DE FACTEURS DÉTERMINISTES ET STOCHASTIQUES

8.3.1 Approche intuitive

Dans la réalité, une population naturelle a de fait un effectif limité et subit un processus de dérive fonction de son effectif efficace N_e.

Si N_e est très petit la dérive sera très efficace, suffisamment pour contrecarrer l'effet déterministe des mutations ou de la sélection sur la composition génétique de la population. On peut imaginer, dans un tel cas, que l'élimination par la dérive, de tous les allèles d'un gène sauf un, puisse conduire exceptionnellement à la fixation d'un allèle moins favorable que certains de ceux qui ont été éliminés.

Au contraire, si N_e est assez grand, la dérive ne pourra empêcher les facteurs déterministes de peser de façon décisive sur l'évolution de la composition génétique de la population. Dans un tel cas, même si leurs trajectoires sont parfois un peu chaotiques en raison de faibles effets de la dérive, les fréquences alléliques évolueront vers les valeurs limites définies par les facteurs déterministes, mutations ou sélection.

Quand, pour un gène, l'effectif N_e de la population est tel que la dérive peut l'emporter sur l'effet des mutations et de la sélection, la population est dite « petite » !

Quand, pour un gène, l'effectif N_e de la population est tel que la dérive ne peut pas l'emporter sur l'effet des facteurs déterministes, la population est dite « grande » !

> ***Remarque 1 :*** les notions de « petite » et de « grande » population sont définies par un rapport de puissance ou d'efficacité entre facteurs stochastiques et déterministes dans la variation de la composition génétique de la population.
>
> ***Remarque 2 :*** il est très important de comprendre qu'une population, dont l'effectif efficace est le même pour tous les gènes, peut être « petite » pour un gène soumis à des facteurs déterministes de faibles valeurs, alors qu'elle sera « grande » pour un autre gène soumis à des pressions importantes de mutation ou de sélection.

Ces conclusions et ces remarques découlent des travaux mathématiques de Fisher, fondateur de la génétique des populations en 1918 et grand théoricien jusque dans les années 1950, et ceux de deux théoriciens plus jeunes, l'américain Crow et le japonais Kimura, au début des années 1960. L'exposé de ces travaux n'est pas envisageable dans cet ouvrage mais il est possible et intéressant de discuter des conséquences mathématiques de deux équations théoriques importantes, obtenues par ses auteurs.

8.3.2 Effet combiné dérive-sélection

Il a été démontré que la probabilité de fixation d'un allèle d'un gène soumis à la dérive et à la sélection peut s'écrire :

$$P = [1 - e^{-2s}]/[1 - e^{-4Nes}]$$

où N_e l'effectif efficace de la population et s l'avantage ou le désavantage sélectif conféré par l'allèle considéré (il y avantage si s est positif et désavantage si s est négatif, égal à – 1 en cas de létalité).

a) Dérive et fixation d'un allèle favorable

Si l'effectif efficace N_e est assez grand la probabilité de fixation se réduit à

$$P = 1 - e^{-2s}$$

ce qui permet de retrouver un résultat déjà démontré par Fisher :

– si l'allèle est défavorable (s négatif, supérieur ou égal à –1), la valeur de P est strictement négative, ce qui signifie qu'une mutation défavorable ne peut être fixée ;
– si l'allèle est favorable (s positif), alors la valeur de P est proche de $2s$ [en effet on pose que Log $(1 - P) = -2s$, d'où on tire que $P = 2s$, si s est petit], ce qui signifie qu'une mutation apportant un avantage de 1 % ou de 5 % n'a qu'une probabilité de 2 % ou 10 % d'être fixée, en raison des effets de la dérive.

Le fait que la sélection, malgré un effectif efficace élevé et la faiblesse de la dérive, ne puisse fixer toutes les mutations favorables apparaissant dans la population s'explique par un nouveau phénomène de variation d'échantillonnage, apparenté à la dérive.

En effet, quand une mutation favorable survient *de novo*, elle n'est présente qu'en un exemplaire unique chez son porteur. Celui-ci peut ne pas avoir de descendants, et s'il en a, à chaque fois, cette mutation n'a qu'une probabilité 1/2 d'être transmise. Comme le nombre de descendants est limité, il y a un risque que la néo-mutation ne soit pas transmise de la génération 0 à la génération 1, puis de la génération 1 à la 2, etc. Évidemment cette probabilité s'amenuise avec le nombre de générations mais elle est suffisamment forte pour que la mutation soit perdue dès les premières générations, même quand elle est très favorable. Si passant le cap des premières générations, une mutation favorable est présente en un nombre de copies assez élevé, son existence échappe alors à la possibilité de non-transmission et la sélection peut dès lors exercer son effet vers la fixation.

b) Petite population et fixation d'une mutation défavorable

Si l'effectif efficace N_e est faible, le dénominateur de P est inférieur à 1, ce qui signifie que la probabilité de fixation est telle que

$$P > 1 - e^{-2s}$$

Cette inégalité est d'autant plus importante que l'effectif efficace N_e est faible.

De ce fait, il existe une fourchette de valeurs négatives de s (correspondant à des allèles défavorables) pour lesquelles la valeur de P est comprise entre 0 et 1. Autre-

ment dit, malgré la sélection, des allèles pas trop défavorables (faiblement délétères) peuvent être fixés par la dérive (tableau 8.7). Bien évidemment les allèles favorables peuvent être fixés avec une probabilité légèrement supérieure à 2*s* puisque dérive et sélection peuvent alors « additionner » leurs effets.

La combinaison des effets entre la dérive et la sélection pour un allèle défavorable permet de placer la « frontière » entre une « grande » et une « petite » population. Selon que la dérive est ou n'est pas capable de contrecarrer l'effet déterministe de la sélection tendant à l'élimination d'un allèle moins favorable, la population sera, pour le gène et l'effet sélectif considéré, « petite » ou « grande ». Dans le tableau 8.7, en fonction du désavantage sélectif de l'allèle, le seuil entre « petite » et « grande » population sera compris entre 100 et 1 000 ou 1 000 et 10 000 individus.

Tableau 8.7 Probabilités de fixation d'un allèle défavorable en fonction de la valeur *s* du désavantage qu'il confère et de l'effectif efficace N_E qui, par la dérive, permet cette fixation (la probabilité est considérée comme nulle si elle est inférieure à 1 sur un milliard).

Effectif efficace / valeurs de *s*	$N_e = 50$	$N_e = 100$	$N_e = 1\,000$	$N_e = 10\,000$
– 0,00001	1/100	1/200	1/2 040	1/24 600
– 0,0001	1/100	1/200	1/2 500	1/268 000
– 0,001	1/110	1/245	1/26 800	0
– 0,01	1/316	1/2 650	0	0
– 0,1	0	0	0	0

Dès qu'un allèle défavorable confère un désavantage de l'ordre de 10 %, la dérive ne peut plus contrecarrer l'effet de la sélection fixer cet allèle et la population sera « grande » pour cet allèle, même avec un effectif efficace de 50 (tableau 8.7, dernière ligne). Mais pour un désavantage de 1 %, une population d'effectif efficace égal à 100 est encore une petite population alors qu'une population d'effectif efficace égal à 1 000 est déjà une grande.

8.3.3 Effet combiné dérive-mutation : le polymorphisme transitoire

Toute population voit apparaître régulièrement des mutations *de novo*, dont nous venons de voir qu'elles ont une probabilité assez élevée de disparaître, même quand elles sont favorables, a fortiori quand elles sont défavorables ou « sélectivement neutres ».

Cependant un certain nombre de générations peut séparer la survenue d'une mutation *de novo* et sa disparition sous l'effet de la dérive. Ces allèles ayant une existence transitoire plus ou moins longue dans la population, forment un fonds de polymorphisme génétique appelé « polymorphisme transitoire », qui peut être figuré par le schéma suivant.

Figure 8.3

À tout moment (par exemple au moment indiqué par le trait en pointillés sur la figure 8.3) la composition génétique de la population est polymorphe, mais ce polymorphisme est transitoire et ce ne sont pas obligatoirement les mêmes allèles ou les mêmes fréquences d'un moment à l'autre.

L'importance du polymorphisme transitoire dépend de l'effectif efficace, aussi bien pour le nombre moyen d'allèles présents dans la population que pour le temps moyen d'existence transitoire de ces allèles. Crow et Kimura ont montré que le nombre moyen d'allèles, appelé aussi « nombre efficace d'allèles » n_e, était donné par la formule suivante :

$$1/n_e = 1/(1 + 4\,N_e\,u)$$

soit

$$\boldsymbol{n_e = 1 + 4\,N_e\,u}$$

où N_e est l'effectif efficace et u le taux de mutation par génération.

Tableau 8.8 Nombre efficace d'allèles constituant le polymorphisme transitoire pour diverses valeurs de N_e et de u.

Valeur de N_e / Valeur de u	N_e = 100	N_e = 1 000	N_e = 10 000	N_e = 100 000
10^{-4}	1,04	1,4	5	41
10^{-5}	1	1,04	1,4	5
10^{-6}	1	1	1,04	1,4
10^{-7}	1	1	1	1,04

On peut remarquer la similitude de la formule avec la limite du taux de consanguinité d'une petite population en dérive qui reçoit quelques migrants à chaque génération (voir chapitre 5, § 5.4.3), ce qui est logique puisque migrations et mutations correspondent à une arrivée de gènes « neufs ».

Un polymorphisme transitoire réel (où $n_e > 1$) suppose d'une part un taux de mutation suffisamment élevé pour le gène considéré, d'autre part un effectif efficace assez élevé pour limiter l'effet de la dérive et laisser apparaître, transitoirement, ce polymorphisme (tableau 8.8).

8.4 CONCLUSION : DU DÉTERMINISME SUR UNE COURTE DURÉE AU HASARD SUR UN LONGUE DURÉE

La théorie du paysage adaptatif développée dans le chapitre précédent et les conclusions des modèles développés dans ce chapitre, peuvent permettre de tenter une vision synthétique du comportement évolutif de la composition génétique d'une population sous l'effet de tous les facteurs déterministes et stochastiques auxquels elle se trouve confrontée.

On peut considérer qu'il existe pour chacun des gènes, dans une population, un paysage adaptatif (figure 8.4) qui résulte notamment des contraintes sélectives exercées par l'environnement sur les divers génotypes de ce gène. La valeur adaptative de la population pour le gène considéré est un point sur la surface de ce paysage adaptatif (point *P*, figure 8.4).

La trajectoire évolutive de ce point va dépendre des facteurs déterministes qui tendent à le faire migrer vers une position correspondant aux valeurs limites d'équilibre des fréquences alléliques, tenant compte des effets des mutations, des migrations et de la sélection (figure 8.4, trajectoire a). Cette position limite correspond à un extremum, le pic du paysage adaptatif sur le flanc duquel le point représentatif de la valeur adaptative est situé à l'instant *t* (figure 8.4, sommet A). Il est important de noter que la situation limite qui sera atteinte n'est donc pas forcément le pic le plus élevé du paysage adaptatif (figure 8.4, sommet B), mais le pic sur le flanc duquel le point représentatif se trouvait situé (sommet A) tant que les effets stochastiques sont négligeables dans l'histoire de la population.

En effet la trajectoire du point représentatif de la population dans le paysage adaptatif est également soumise aux aléas des facteurs stochastiques (effets fondateurs ou dérive). Mais ces facteurs ne sont importants que si la population est « petite pour le gène considéré », et qu'ils arrivent à contrecarrer l'effet systématique des facteurs déterministes.

Dans ce cas, et dans ce cas uniquement, le point représentatif de la population peut, sous l'effet d'un aléa, traverser une vallée entre deux pics et passer du flanc d'un sommet au flanc d'un autre sommet (figure 8.4, trajectoire b).

Si, à ce moment, les effets stochastiques deviennent minimes, la trajectoire évolutive de la population la conduit vers ce nouveau sommet (figure 8.4, trajectoire c), où elle restera tant que le paysage demeurera inchangé et que les facteurs stochastiques resteront négligeables.

En fait, plutôt que de considérer un gène et le paysage adaptatif qui se rapporte à ses divers génotypes, il serait plus juste de considérer, à un niveau global, les différents génotypes relatifs à tout l'ensemble des gènes impliqués dans un même caractère global, dont les phénotypes constituent la réalité soumise par les organismes à la sélection (capacité de fuite ou de camouflage vis-à-vis d'un prédateur, fertilité, comportement alimentaire, etc.).

Cette vision, moins réductrice et plus proche de la réalité, ne change rien à ce qui a été développé quant au comportement évolutif d'une population dans le paysage adaptatif sous l'effet des facteurs déterministes et stochastiques.

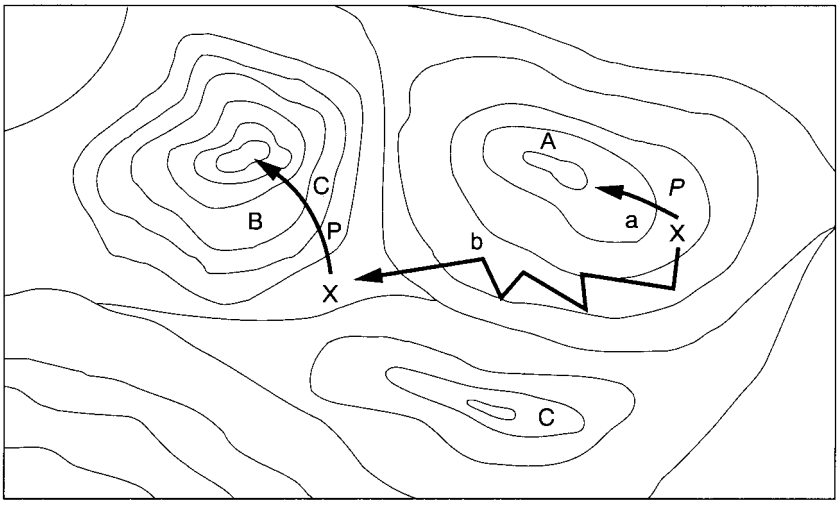


Figure 8.4

On pourrait alors considérer que le fait de prendre en compte l'action simultanée de plusieurs gènes dans les phénotypes d'un caractère conduirait simplement à une complexification du relief de ce paysage adaptatif.

Mais ce serait oublier que l'action des gènes étant, pour certains d'entre eux, interactive, les valeurs sélectives des génotypes d'un gène dépendant de génotypes pour d'autres gènes, le relief du paysage adaptatif sera alors, pour une part, soumis au « hasard » des recombinaisons génétiques et des fécondations.

Bien plus, de nombreux facteurs extérieurs aux organismes, comme le climat, les ressources nutritives, la fréquence des prédateurs ou des parasites, qui interagissent avec leur valeur adaptative, peuvent présenter des variations ou une évolution, et participent de ce fait à des variations du paysage adaptatif.

En conséquence, il faut considérer que le paysage adaptatif est un relief mouvant, aussi mouvant que le furent les reliefs terrestres sous l'effet de la tectonique des plaques, où se conjuguent les effets des variations des conditions extérieures, des facteurs déterministes s'exerçant sur les fréquences des allèles, des facteurs stochastiques dépendant de la taille de la population ou du hasard de la fécondation et de la recombinaison.

Comme l'ensemble de ces paramètres évoluent, pour une part de manière liée, mais pour une part de manière indépendante, cela induit une « mouvance » du paysage adaptatif telle que la population qui était sur le flanc ou au sommet d'un pic au temps t, peut se retrouver au fond d'une vallée ou sur un nouveau pic dans un nouveau paysage adaptatif au temps $t + \Delta t$. Et il convient aussi de discuter de cet intervalle de temps Δt.

Sur une courte durée, quand Δt est petit, la mouvance du paysage est imperceptible, le paysage semble présenter un état quasi stable. Alors les facteurs déterministes jouent localement leur rôle, plus ou moins contrecarré par celui des facteurs

stochastiques, comme cela a été développé au sein de la figure 8.4. À cette échelle, la vision du comportement dans le paysage adaptatif rejoint la vision darwinienne d'une évolution continue, modernisée par l'acceptation que des facteurs stochastiques puissent modifier et réorienter la trajectoire évolutive. À cette échelle de temps, cette vision de la population et de sa trajectoire sur le paysage adaptatif rend bien compte de la réalité et valide en cela ce qu'on a désigné comme théorie synthétique de l'évolution.

Mais sur une longue durée, quand Δt n'est pas infinitésimal, qu'il représente un écart de temps à l'échelle géologique, il ne peut plus être question d'ignorer la mouvance du paysage adaptatif et surtout le fait que cette variation du relief dépend d'un nombre si important de facteurs, dont certains aléatoires, qu'elle est imprévisible. Au bout du compte la trajectoire évolutive d'une population, ou d'une espèce, déjà soumise sur une courte durée, à l'effet de facteurs stochastiques, se trouve soumise sur une grande durée à la mouvance quasi chaotique du paysage adaptatif. De là à dire que l'évolution des espèces et que l'histoire du vivant n'est que le fruit du hasard (ou apparaît comme tel), il n'y a qu'un pas que certains ont franchi, avec le plaisir des iconoclastes et le goût du paradoxe.

En fait, il s'agit là d'une position extrémiste, car dire que l'évolution du vivant fut le fruit du hasard ne signifie nullement qu'elle ne fut que le fruit du hasard précisément parce que les contingences et les effets déterministes, sans lesquels il n'y aurait pas de vivant, jouent à plein sur les petites durées, tandis que le hasard joue de plus en plus quand les écarts de temps s'agrandissent. En d'autres termes, pour paraphraser un ouvrage célèbre, le hasard joue certes un rôle mais ne peut le faire qu'en raison de la nécessité qui joue le sien. Il est d'ailleurs intéressant de considérer une nouvelle fois que le « hasard » lui même peut consister en la combinatoire d'un grand nombre d'effets déterministes en partie indépendants de sorte que la complexité des possibles conduit à une multiplicité de choix confinant l'élu, l'espèce choisie ou éliminée, à devoir son destin au hasard puisqu'elle avait a priori plusieurs destins possibles.

RÉSUMÉ

L'effet combiné de la sélection et des mutations permet de maintenir à un niveau de fréquence faible les mutations défavorables. Les fréquences d'équilibre sont fonctions des valeurs des taux de mutations et de sélection.

Le fait que toute population concrète ait un effectif limité soumet celle-ci à un processus de dérive. Tant que celui-ci ne peut détourner efficacement la trajectoire évolutive de la composition génétique d'une population vers les fréquences limites définies par les facteurs déterministes (mutations, migrations et sélection), la population est définie comme « grande ». Dès que ce détournement est possible et que, par exemple un allèle défavorable puisse être fixé contre un allèle favorable, la population est définie comme « petite ».

Une population peut être « petite » relativement aux allèles d'un gène alors qu'elle sera « grande » relativement aux allèles d'un autre gène, pour lequel les facteurs déterministes sont plus intenses, et donc plus puissants pour contrecarrer les effets aléatoires de la dérive.

Sur une durée plus ou moins courte, on peut concevoir l'évolution génétique d'une population comme la trajectoire d'un point sur le relief d'un paysage adaptatif où les sommets correspondent à un ensemble de fréquences d'équilibre pour un bloc de gènes impliqués dans un ou des caractères important dans l'interface entre les organismes et leur environnement. La trajectoire évolutive est alors soumise à l'effet conjugué des facteurs déterministes et stochastiques, ces derniers n'ayant le pouvoir de réorienter la trajectoire que si leur effet est plus puissant que celui des facteurs déterministes.

Sur une longue durée, à l'échelle des temps géologiques, la mouvance du relief adaptatif accroît l'aspect chaotique des trajectoires de sorte que l'évolution des populations, des espèces, semble obéir plus au hasard qu'à l'effet permanent de facteurs déterministes.

Mais l'effet du hasard qui semble émerger à cette échelle de temps ne doit pas faire oublier l'effet de la contingence et les effets déterministes qui sont essentiels sur une courte durée. C'est même, au-delà des facteurs stochastiques, les effets déterministes qui, par leur nombre et leur diversité et leurs combinatoires, génèrent le hasard sur la longue durée des temps géologiques.

Tableau 8.9 Résumé des formalismes mathématiques et des simplifications algébriques conduisant aux équilibres Sélection-Mutation.

	Maladie dominante (M : allèle pathologique)			**Maladie récessive (m : allèle pathologique)**		
Génotypes	*M//M*	*M//n*	*n//n*	*N//N*	*N//m*	*m//m*
Fréquences	p^2	$2pq$	q^2	p^2	$2pq$	q^2
Valeurs sélectives	σ_1	σ_2	σ_3	σ_1	σ_2	σ_3
Équation de Δp_s	$\Delta p = (pq/\sigma)[(\sigma_1 - \sigma_2)p + (\sigma_2 - \sigma_3)q]$					
Changement d'écriture des valeurs sélectives	$(1 - s)$	$(1 - hs)$	1	1	1	$(1 - s)$
Équations de Δp_s	$\Delta p_s = [pq/(1 - sp^2 - 2hspq)]$ $[s\,(h - 1)p - hsq]$			$\Delta p_S = [pq/(1 - sq^2)]\,[sq]$		
Équations de Δp_s Au voisinage de	pour p proche de zéro $\Delta p_s = - hsp$			pour q proche de zéro $\Delta p_s = - sq^2$		
Équilibre avec $\Delta p_m = \mu$	$\Delta p_s + \Delta p_m = 0$ $p = \mu/hs$ $p = \mu/s$ si dominance totale $p = \mu$ si létalité			$\Delta p_s + \Delta p_m = 0$ $q = \sqrt{\mu/s}$ $q = \sqrt{\mu}$ si létalité		

EXERCICES

Exercice 8.1

Une femme sur 12 aura développé avant 65 ans, le plus souvent entre 55 et 65 ans, un cancer du sein et 5 % de ces cancers sont des formes héréditaires transmissibles comme une maladie dominante. Chez les femmes atteintes de cette forme héréditaire, la mortalité précoce diminue la période féconde d'environ 30 %, alors que chez les autres femmes la période féconde est terminée quand survient la pathologie.

Question 1 : en supposant que ces données correspondent à une situation d'équilibre, on vous demande de calculer le taux de mutation affectant le gène impliqué dans cette forme héréditaire de cancer.

On rappelle la formule liant le taux de mutation (μ) du gène, la fréquence (p) de la mutation (quand elle est dominante) et la diminution de fécondité (s) des phénotypes atteints : $p = \mu/s$

Question 2 : justifiez vos calculs et commentez vos résultats, sachant que la recherche a identifié deux gènes impliqués dans les formes mendéliennes (BRCA1 et BRCA2) et qu'il en existe au moins un troisième non encore localisé.

Solution

Question 1 : pour la forme héréditaire, on a $s = 0{,}3$.

La fréquence des formes héréditaires est égale à $1/12 \times 5/100 = 4{,}2.10^{-3}$

La forme héréditaire se transmettant de manière dominante, on peut considérer que l'allèle de susceptibilité a un effet dominant et que la plupart des femmes atteintes de la forme héréditaire sont en fait hétérozygotes et porteuses d'un seul exemplaire muté du gène.

La fréquence de cet allèle pathogène est donc égale à la moitié de la fréquence des femmes atteintes, soit $2{,}1.10^{-3}$.

En appliquant la formule $\mu = p.s$, on peut estimer μ, soit $\mu = p.s = 0{,}0021 \times 0{,}3 = 6{,}25.10^{-4}$.

Question 2 : cette valeur de μ peut sembler très élevée mais elle est surestimée si la fréquence allélique p est surestimée ; or c'est le cas puisqu'en fait, p est la fréquence des allèles pathologiques estimée sur au moins trois gènes, ce qui donnerait, par gène, une fréquence environ trois fois plus faible et un taux de mutation μ de l'ordre de 2.10^{-4}.

Exercice 8.2

On considère deux maladies de fréquence égale à 1/40 000, l'une récessive, l'autre dominante.

Question 1 : quelles sont, dans chaque cas, les estimations des fréquences de l'allèle pathologique (justifier le calcul en quelques mots) ? Comparez leurs valeurs et commentez.

Question 2 : calculez, dans chaque cas, le rapport [fréquences des hétérozygotes]/[fréquences des homozygotes atteints]. Quel est le sens de ce rapport et quelles conclusions en tirez-vous, notamment vis-à-vis des résultats de la question précédente ?

Question 3 : en supposant que ces fréquences sont à l'équilibre dans le temps, quels sont les deux modèles susceptibles d'expliquer cet équilibre ? Vous discuterez de la validité de ces modèles.

Question 4 : vous apprenez que vous avez négligé le fait que 8 % des mariages ont lieu entre cousins germains et 4 % entre doubles cousins germains. Montrez que cela ne change rien à votre estimation de la fréquence de l'allèle pathologique responsable de la maladie dominante, mais que cela change complètement celle de l'allèle responsable de la maladie récessive, fréquence que vous estimerez alors.

Solution

Question 1 : il convient de faire l'ensemble des calculs de cette question sous le modèle de Hardy-Weinberg. On ne sait pas si la maladie étudiée est soumise à la sélection mais on sait, même si la condition d'absence de sélection n'est pas valide, que l'effet de la sélection sur les fréquences alléliques et génotypiques est assez faible sur quelques générations pour pouvoir être négligé et se mettre ainsi sous le modèle de Hardy-Weinberg, notamment l'hypothèse panmictique qui permet d'estimer la fréquence des allèles récessifs.

Sous le modèle de H-W, la fréquence R des enfants atteints d'une telle maladie récessive est égale à q^2, où q est la fréquence de l'allèle pathologique.

En prenant la racine de R (0,000025), on en déduit la valeur de q, soit,

$q = 0{,}005 = 1/200$

Pour la maladie dominante, on peut se mettre sous l'hypothèse de H-W et considérer que la fréquence p de l'allèle pathologique est égale à $(1 - q)$ où q est celle de l'allèle normal correspondant à la racine carrée de la fréquence des phénotypes sains, soit (1 – 1/40 000) ; on peut aussi considérer que les individus sains sont tous porteurs d'un seul gène muté, la fréquence des homozygotes étant négligeable, ce qui conduit à considérer que

$f(M/n) = 1/40\,000$ et que $f(M) = f(M/n)/2 = 1/80\,000$ (0,0000125)

Commentaire : on note que pour deux maladies de même fréquence, 1/40 000, la fréquence de l'allèle pathologique est 400 fois plus élevée dans la maladie récessive que dans la maladie dominante.

Question 2 :

Pour la maladie récessive, la fréquence des hétérozygotes est celle des porteurs sains, soit :

$H = 2q(1 - q)$, et $H = 2q$ si q est petit, $\boldsymbol{H} = 0{,}01$ (**1/100**)

Le rapport [fréquences des hétérozygotes]/[fréquences des homozygotes atteints] est égal à $2/q$, soit 400 ; il exprime le nombre d'allèles pathologiques présents chez des porteurs sains pour deux allèles présents chez un atteint et donne un ordre de grandeur de la quantité des allèles pathologiques échappant à la sélection du fait de la « protection » par un allèle normal chez un porteur sain ; ceci explique pourquoi le niveau de fréquences des allèles pathologiques, dans les maladies récessives est nettement supérieur à celui observé dans les maladies dominantes, dans lesquelles les allèles pathologiques sont tous soumis à la sélection (voir la question **a**).

Pour la maladie dominante, la fréquence des hétérozygotes est égale à :

$H = 2p(1 - p)$, et H= $2p$ si p est petit, et on retrouve $\boldsymbol{H} = 0{,}000025$ (**1/40 000**)

Le rapport [fréquences des hétérozygotes]/[fréquences des homozygotes atteints] est égal à $2/p$, soit 40 000 ; il exprime le nombre d'atteints hétérozygotes pour un atteint homozygote ; la valeur observée justifie donc tout à fait de négliger les homozygotes dans le calcul de la fréquence d'un allèle pathologique responsable d'une maladie dominante (voir question **a)**.

Question 3 : si les fréquences alléliques sont à l'équilibre, cela signifie que l'effet sélectif qui tend à éliminer de la population des allèles pathologiques est contrebalancé par un autre effet.

Un premier modèle est l'équilibre sélection-mutation où l'effet de la sélection est équilibré par celui des néo-mutations.

Dans le cas d'une maladie dominante, on a $p = \mu/s$, où μ est le taux de néo-mutations et s le désavantage sélectif ($s = 1$ quand il y a létalité avant l'âge reproducteur), ce qui correspondrait ici à un taux de néo-mutation inférieur ou égal (si $s = 1$) à p, soit $1{,}25.10^{-5}$, ce qui est admissible.

Dans le cas d'une maladie récessive, on a $q = \sqrt{\mu/s}$, soit $q^2 = \mu/s$, ce qui correspondrait ici à un taux de néo-mutation inférieur ou égal (si $s = 1$) à q^2, soit $2{,}5.10^{-5}$, ce qui commence à être une valeur élevée pour un taux de mutation.

Un deuxième modèle, adapté aux maladies récessives, est l'équilibre par avantage des hétérozygotes porteurs sains, où l'effet de la sélection sur les homozygotes atteints est contre-balancé par la survie et/ou la fertilité accrue des porteurs sains qui, plus que les autres vont transmettre chacun de leurs allèles, le normal mais surtout le muté, ce qui conduit à un équilibre quand les allèles perdus par la sélection correspond au surplus fourni par les porteurs sains.

Question 4 : la population est consanguine puisque de nombreux mariages ont lieu entre apparentés, le taux moyen de consanguinité est égal à :

F= (1/16).8 % + (1/8).4 % = 0,01 = 1 %

Dans le cas de la maladie dominante, la fréquence de la maladie dominante est théoriquement plus faible que si la population était panmictique, mais l'écart est si faible qu'il est négligeable.

En effet, on a 1/40 000 qui correspond alors à :

$p^2 + 2pq - Fp(1 - p) = p^2 + 2p(1 - p) - Fp(1 - p)$

soit en négligeant les termes en p^2, $1/40\,000 = p\,(2 - F)$

Comme $F = 0{,}01$, négliger la consanguinité est sans conséquence sur l'estimation de p,

soit $1/40\,000 = 2p$ et $p = 1/80\,000$

Dans le cas de la maladie récessive, la fréquence de la maladie dominante est théoriquement plus grande que si la population était panmictique, et l'écart est ici important.

En effet, 1/40 000 correspond alors à :

$q^2 + Fq.(1 - q) = q^2 + Fq$, si on néglige le terme Fq^2.

Or, ici, le terme Fq est beaucoup plus élevé que le terme q^2 correspondant au seul risque panmictique. La fréquence des malades est donc égale au risque panmictique q^2 augmenté du surplus induit par la consanguinité, soit

$$1/40\,000 = q^2 + Fq$$

où $F = 0{,}01$

Après résolution de cette équation, on tire une estimation de la fréquence de l'allèle pathologique responsable de la maladie récessive, $q = 0{,}0021$ (1/476), soit plus de deux fois moins que l'estimation faite sous l'hypothèse de la panmixie ; on voit donc à quel point la négligence de la consanguinité, en négligeant le surplus d'homozygotes induits par celle-ci, aboutit à une surestimation importante de la fréquence de l'allèle pathologique.

Exercice 8.3

On entreprend l'étude génétique d'un oiseau dont le sexe est déterminé par un jeu d'hétérosomes, le sexe mâle étant X//X et le sexe femelle X//Y.

On identifie un gène dont la perte de fonction, récessive vis-à-vis de l'allèle sauvage, conduit à un phénotype de stérilité facile à identifier parce qu'il est associé à une tâche blanche entre les yeux chez les femelles.

Une étude collectant chez les éleveurs une information sur 100 000 individus permet de recenser 9 femelles présentant une tâche blanche et dont le phénotype stérile a été confirmé par croisement.

Question 1 : en supposant que les fréquences alléliques sont égales entre les sexes, quelles sont les fréquences dans chacun d'eux ? Quelle est la fréquence des mâles conducteurs ?

Question 2 : si on considère que le maintien de cet allèle de stérilité est assuré par des néo-mutations, quel est le taux de mutation ?

Solution

Question 1 : la fréquence de l'allèle muté est $q = 9/100\,000$ dans chacun des sexes. La fréquence des mâles conducteurs est égale à $2pq = 18/100\,000$

Question 2 : dans ce cas, la formule de Haldane conduit à $\mu = q/3$, soit 3.10^{-5}, valeur compatible avec les taux habituels.

Exercice 8.4

Une maladie C, dominante, touche, dans une population, un enfant sur 8 000 à la naissance.

Question 1 : vous calculerez la fréquence des allèles pathogènes impliqués dans cette maladie.

Question 2 : quel serait le taux de mutation vers l'allèle pathogène si on considère que la composition génétique de la population est stable et que la maladie est létale dans l'enfance ?

Question 3 : quelle serait le taux de mutation vers l'allèle pathogène si la mortalité était assez retardée pour ne réduire la fécondité que de 50 % ?

Question 4 : quelle serait la nouvelle fréquence d'équilibre si une thérapie nouvelle, en augmentant l'espérance de vie, limitait la réduction de la fécondité à 20 % ?

Solution

Question 1 : les génotypes respectifs, les phénotypes et leurs fréquences, sous l'hypothèse de Hardy-Weinberg sont donnés par le tableau suivant :

Phénotypes	Génotypes	Fréquences
Atteint	*C/C*	p^2
Atteint	*C/c*	*2pq*
Sain (7 999/8 000)	*c/c*	q^2

On en tire, sous l'hypothèse de Hardy-Weinberg, les estimations suivantes :

$f(c) = q = \sqrt{7\,999 / 8\,000} = 0{,}999937$

d'où $f(C) = p = 0{,}000063$ ($6{,}3.10^{-5}$)

> ***Remarque 1 :*** parce que l'effet pathogène de l'allèle n'est pas masqué chez les hétérozygotes, une maladie dominante ayant une fréquence comparable à celle d'une maladie récessive (voir exercice 8.2) résulte de l'effet d'un allèle pathogène beaucoup plus rare ($6{,}3.10^{-5}$) que l'allèle pathogène impliqué dans la maladie récessive (1,4 %).
>
> ***Remarque 2 :*** De ce fait, la plupart des individus atteints sont en réalité des hétérozygotes, ce qui permet d'estimer très directement la fréquence de l'allèle pathogène comme la moitié de la fréquence des atteints-hétérozygotes (la fréquence des homozygotes étant considérée comme nulle ou négligeable). Effectivement 1/16 000 est bien égal à $6{,}3.10^{-5}$.

Question 2 : l'équilibre sélection-mutation est régi, pour une maladie dominante, par l'équation $p = \mu/s$, où p est la fréquence, à l'équilibre, de l'allèle pathogène, μ, le taux de mutation vers cet allèle et s, le désavantage sélectif des individus atteints (en supposant qu'il est le même chez les homozygotes et les hétérozygotes).

Dans les conditions du problème, on a $p = 6{,}3.10^{-5}$, avec $p = \mu/s$,

Si $s = 1$, alors $\mu = p = 6{,}3.10^{-5}$

La valeur d'un tel taux de mutation est encore un peu élevée pour accepter l'hypothèse d'un équilibre sélection-mutation comme mécanisme du maintien du polymorphisme.

Question 3 : si la maladie n'est pas létale dans l'enfance mais permet une fécondité réduite de 50 %, on a toujours la même fréquence allélique ($p = 6{,}3.10^{-5}$), mais un équilibre régi par l'équation $p = \mu/s$, avec désormais une valeur $s = 0{,}5$. D'où un taux $\mu = 0{,}5xp = 3{,}2.10^{-5}$.

La valeur d'un tel taux de mutation est toujours un peu élevée pour accepter l'hypothèse d'un équilibre sélection-mutation comme mécanisme du maintien du polymorphisme, seules des maladies très rares peuvent entrer dans un tel cadre théorique.

Question 4 : dans les nouvelles conditions, on a maintenant :

$p' = \mu/s'$,

avec $\mu = 3{,}2.10^{-5}$ valeur définie à la question précédente,

Mais ici $s' = 0{,}2$

d'où $p' = 3{,}2.10^{-5}/0{,}2 = 1{,}6.10^{-4}$

La valeur d'équilibre de l'allèle pathogène est multipliée par 2,5 et la fréquence des individus atteints montera de 1/8 000 à 1/3 125.

Cependant le temps nécessaire à cette évolution est très long et la maladie est moins grave puisqu'une thérapie existe.

Il convient donc de juger avec beaucoup de mesure « l'effet dysgénique de la médecine » (voir aussi l'exemple du chapitre 7).

Exercice 8.5

Une séquence de quelques nucléotides répétée en tandem, neutre sur le plan sélectif, mute vers un autre allèle variant par le nombre de répétitions avec un taux μ égal à 10^{-5}. Ce phénomène est à l'origine de ce type de polymorphisme aussi appelé microsatellite (voir chapitre 1).

Un polymorphisme de type microsatellite peut-il s'établir et se développer dans des populations dont l'effectif efficace est respectivement égal à 100, 1 000, 10 000 et 100 000 ?

Solution

On a vu que le nombre effectif d'allèles neutres pouvant se maintenir malgré la dérive compte tenu du taux de mutation est égal à $ne = 1 + 4N_e\,\mu$.

Avec $\mu = 10^{-5}$, et $N_e = 100$ ou 1 000, il n'y a pas de polymorphisme possible pour une telle séquence. Avec un effectif de 100 000, on aura environ 5 allèles différents pour cette séquence, mais il s'agit ici d'une moyenne et on peut en avoir trois pour un site et dix pour un autre.

Avec un effectif de 10 000, il n'y a pas en moyenne de polymorphisme, mais là encore il ne s'agit que d'une moyenne ; cependant, la dérive est encore assez efficace pour, avec un tel effectif, réduire fortement ou limiter la diversité quand elle est générée avec un tel taux de mutation.

Index

49620 - (I) - (1,5) - OSB 80° - TRE - MER

Achevé d'imprimer sur les presses de
SNEL Grafics sa
rue Saint-Vincent 12 – B-4020 Liège
Tél +32(0)4 344 65 60 - Fax +32(0)4 341 48 41
mai 2006 — 37553

Dépôt légal : juin 2006

Imprimé en Belgique

SCIENCES SUP

Jean-Louis Serre

GÉNÉTIQUE DES POPULATIONS

La génétique des populations est à l'intersection de plusieurs champs scientifiques. Elle s'enseigne dans les cursus de biologie qui s'intéressent à l'évolution, la biodiversité, l'amélioration des plantes cultivées et la génétique humaine ou médicale.

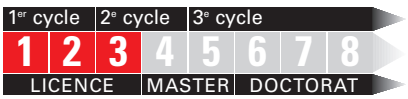
Dans ce manuel, les principaux concepts comme l'équilibre de Hardy-Weinberg, la consanguinité, le déséquilibre gamétique, la dérive génétique, la sélection ou l'effet des mutations sont d'abord introduits de manière totalement intuitive avant d'être présentés dans leur formulation théorique. Le principe et la réalisation des tests statistiques sont rappelés. L'ensemble est illustré par une soixantaine d'exemples et d'exercices avec corrigés détaillés.

Cet ouvrage s'adresse aux étudiants de Licence et de Médecine (PCEM 1 ou 2) et sera aussi utile aux candidats au CAPES ou à l'agrégation des sciences de la Vie et de la Terre.

JEAN-LOUIS SERRE

est professeur à l'université de Versailles-Saint-Quentin, membre du bureau de la Société Française de Génétique Humaine. Ses recherches portent sur l'analyse de la relation génotype-phénotype dans certaines maladies génétiques et, en génétique des populations humaines, sur la mesure et l'effet de la consanguinité, et son application à la cartographie des gènes.

MATHÉMATIQUES
PHYSIQUE
CHIMIE
SCIENCES DE L'INGÉNIEUR
INFORMATIQUE
SCIENCES DE LA VIE
SCIENCES DE LA TERRE



ISBN 2 10 049620 4

www.dunod.com

DUNOD