Polymorphisme de l'ADN mitochondrial et histoire du peuplement humain

Excoffier, Laurent Georges Louis

UNIVERSITÉ DE GENÈVE
Département d'Anthropologie
Laboratoire de Génétique et Biométrie

FACULTÉ DES SCIENCES

Prof. A. Langaney

POLYMORPHISME DE L'ADN MITOCHONDRIAL ET HISTOIRE DU PEUPLEMENT HUMAIN

THÈSE

PRÉSENTÉE À LA FACULTÉ DES SCIENCES DE L'UNIVERSITÉ DE GENÈVE POUR OBTENIR LE GRADE DE DOCTEUR ÈS SCIENCES BIOLOGIQUES

PAR

Laurent EXCOFFIER

DE

GENÈVE

THÈSE Nº 2323

GENÈVE

1988

UNIVERSITÉ DE GENÈVE
Département d'Anthropologie
Laboratoire de Génétique et Biométrie

FACULTÉ DES SCIENCES

Prof. A. Langaney

POLYMORPHISME DE L'ADN MITOCHONDRIAL ET HISTOIRE DU PEUPLEMENT HUMAIN

THÈSE

PRÉSENTÉE À LA FACULTÉ DES SCIENCES DE L'UNIVERSITÉ DE GENÈVE POUR OBTENIR LE GRADE DE DOCTEUR ÈS SCIENCES BIOLOGIQUES

PAR

Laurent EXCOFFIER

DE

GENÈVE

THÈSE Nº 2323

GENÈVE

1988

La faculté des sciences, sur le préavis de Messieurs A. LANGANEY, professeur ordinaire et directeur de thèse (Département d'anthropologie), S. EDELSTEIN, professeur ordinaire (Département de biochimie), B. MACH, professeur ordinaire (Faculté de Médecine - Département de microbiologie), G. PISON, docteur (Laboratoire d'anthropologie et biologie du Musée de l'Homme - Paris, France) et M. LATHROP, docteur (Centre d'étude du polymorphisme humain - Paris, France)

autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

GENÈVE, le 28 novembre 1988

Le Doyen

Jean-Pierre IMHOF

Thèse 2323

A Jacline et Georges

Lorsque, à une époque extrêmement reculée, les descendants d'un ancêtre commun ont revêtu des caractères distincts pour .former les races humaines, les différences entre ces races devaient être insignifiantes et peu nombreuses

Charles Darwin, The Descent of Man, 1871

RESUME

L'ensemble des données provenant du polymorphisme de l'ADN mitochondrial (ADN-mt) humain a été rassemblé et sa nature moléculaire réexaminée. Les phylogénies polarisées des types d'ADN-mt présentent une structure radiante des types actuels à partir d'un nombre restreint de types anciens retrouvés dans les différents groupes continentaux. Les populations caucasoïdes semblent posséder la constitution génétique la plus proche d'une population ancestrale d'hommes modernes à partir de laquelle auraient émergé tous les groupes continentaux. Les populations africaines, apparemment très différenciées, possèdent en fait une diversité moléculaire à long terme inférieure à celle de certaines populations caucasoïdes ou orientales. Un test de neutralité sélective montre l'existence de mécanismes évolutifs différentiels selon les populations. Ainsi, leurs apparentements ne peuvent être mesurés à l'aide de distances génétiques classiques et de nouvelles approches, ébauchées ici, sont nécessaires pour intégrer le polymorphisme de l'ADN-mt dans l'étude du peuplement humain.

TABLE DES MATIERES

REMERCIEMENTS

Ce travail a été réalisé grâce à toute une série d'éléments scientifiques et sociaux qui ont souvent interagi positivement pour favoriser la rédaction de cette thèse. Je tiens ici à exprimer ma gratitude à quelques personnes qui ont contribué à ce qu'elle prenne forme, et ceci dans un ordre qui ne respecte pas forcément l'ampleur de leur soutien.

J'aimerai tout particulièrement remercier le professeur André Langaney qui a réuni les conditions techniques et sociales qui ont permis que ce travail se réalise dans des conditions idéales. Qu'il soit également remercié pour son constant soutien scientifique et moral, son accueil, sa disponibilité et sa bonne humeur. Il nous a fait réaliser que le respect de la Science pouvait être incompatible avec celui des théories scientifiques dominantes.

Les professeurs Stuart Edelstein, directeur du Département de Biochimie de la Faculté des Sciences, et Bernard Mach, directeur du Département de Microbiologie de la Faculté de Médecine, nous ont fait l'honneur de faire partie du jury de thèse, ainsi que le Dr. Marc Lathrop et le Dr. Gilles Pison qui nous a fait visiter le pays Malinké au Sénégal Oriental.

Le professeur Michel Jeannet a été le promoteur d'échanges d'informations scientifiques ainsi que d'une étroite collaboration entre notre Département et la Faculté de Médecine.

Le professeur Alain Gallay, directeur du Département d'Anthropologie, a souvent fourni des hypothèses constructives concernant l'histoire du peuplement humain à partir des données fossiles. Sa vaste connaissance de l'histoire africaine nous a été précieuse pour vérifier le fondement de nos scénari génétiques.

Le professeur Pierre Moeschler, président de la Section de Biologie, nous a accordé un soutien permanent qui a permis l'achèvement de ce travail.

Le professeur Albert Jacquard a éveillé notre intérêt pour la génétique des populations et l'a entretenu par son enseignement et ses débats.

Le professeur L.-L. Cavalli-Sforza, président du Département de Génétique de l'Université de Stanford, nous a fourni de précieux conseils lors de discussions sur l'histoire du peuplement humain. Sa rigueur scientifique nous a servi de modèle dans la volonté de vérifier le bien fondé de divers scénari concernant l'origine de l'homme moderne.

Le Dr. Pierre Darlu nous a motivé à publier nos résultats à la suite de fructueuses discussions sur l'origine du polymorphisme de l'ADN-mt.

Le Dr. Walther Reith nous a prêté un temps précieux et nous a facilité l'accès à des programmes d'analyse de séquences d'ADN.

Béatrice Pellegrini nous a continuellement accompagné, soutenu et motivé pendant ces années de thèse. Sa haute compétence nous a été précieuse pour la confrontation d'hypothèses génétiques, linguistiques et historiques dans diverses régions du monde.

Alicia Sanchez-Mazas a été à la source de nombreuses hypothèses par sa connaissance du polymorphisme d'autres systèmes génétiques au niveau mondial.

Ninian Hubert van Blyenburgh nous a montré que certaines théories scientifiques devaient être remises en causes pour l'avancement des connaissances humaines.

David Roessli nous a prodigué de nombreux conseils dans le domaine de la programmation informatique. Nous avons apprécié sa disponibilité à tous les niveaux et son esprit rigoureux.

Georges Puissant a développé de nombreux programmes de représentation graphique qui ont permis de mettre en valeur certains de nos résultats.

Jean-Gabriel Elia a constamment été disponible pour assurer un travail de reproduction photographique et a contribué à la conception de nombreuses Figures.

Serge Aeschlimann et Yves Reymond n'ont pas ménagé leurs efforts pour la conception, la création et la mise en valeur de graphiques et autres dessins figurant dans ce travail.

Le Dr. Christian Simon nous a aidé à comprendre la qualité et la diversité des fossiles des premiers hommes.

Olga Petrovic, Marie-Noële Lahouze et Mirka Hausner nous ont aidé à collecter d'indispensables références bibliographiques.

Corinne de Haller nous a facilité de nombreuses démarches administratives et officielles. Elle a assuré un contact permanent de qualité avec le monde extérieur à notre Département.

Leila Gaude a eu la gentillesse de nous aider pour de multiples travaux de secrétariat.

Patrick Moinat et Markus Fischer nous ont souvent aidé à résoudre quelques petits problèmes informatiques.

Georgette Khalifa a saisi pour nous une quantité de données traitées au cours de la durée de notre assistanat. Elle a aussi fait preuve d'une patience souveraine lors du partage de certaines ressources informatiques.

Monsieur Denis Zaborszki nous a entretenu de fort intéressantes discussions jusqu'à des heures tardives et nous a souvent aidé à résoudre divers petits tracas techniques.

Les services informatiques de l'Université et le professeur Bernard Levrat ont mis à notre disposition un matériel informatique de qualité qui a grandement facilité l'élaboration de ce travail.

Enfin, beaucoup d'autres personnes du Département d'Anthropologie de Genève et du Musée de l'Homme de Paris nous ont rendu de multiples services et ont contribué à créer une atmosphère chaleureuse dans ces deux établissements.

Ce travail a pu être réalisé grâce au soutien du Fonds National Suisse de la Recherche Scientifique par les subsides 3.514.0.86 et 3.958.0.87.

POLYMORPHISME DE L'ADN MITOCHONDRIAL ET HISTOIRE DU PEUPLEMENT HUMAIN

INTRODUCTION

L'ADN mitochondrial est l'un des fragments d'ADN humain les mieux étudiés au niveau moléculaire, du point de vue de la variabilité entre les individus et les populations. Ce système génétique, potentiellement très riche en informations, nécessite, pour être bien interprêté, que l'on ait une bonne compréhension de sa structure, de son mode de transmission et de la façon dont il évolue au cours du temps. La connaissance de ces paramètres est d'ailleurs requise pour l'étude de tout système génétique au niveau moléculaire. Comme l'ADN mitochondrial diffère passablement des systèmes génétiques codés au niveau du noyau cellulaire, il nous a semblé utile de rappeler certaines de ses caractéristiques. Les lecteurs peu familiers avec les termes de la génétique moléculaire nous excuseront pour l'emploi de certains termes apparamment hermétiques qu'il était impossible de redéfinir dans le cadre de ce travail.

PRÉSENTATION DE LA MOLECULE D'ADN MITOCHONDRIAL

Les mitochondries possèdent une petite molécule d'ADN circulaire close qui se réplique et se transcrit de façon indépendante de l'ADN nucléaire. La longueur et la structure de l'ADN mitochondrial (ADN-mt) est bien conservée parmi tous les vertébrés (Brown, 1983, 1985), suggérant ainsi un important rôle fonctionnel. On estime qu'il existe plusieurs milliers de mitochondries par cellule (> 8'000 selon Bogenhagen and Clayton (1974)), et donc autant de molécules d'ADN-mt dans les cellules somatiques humaines. Ce nombre n'est pas exactement déterminé dans le cas des oocytes, mais il pourrait être supérieur (Piko and Matsumoto, 1976). L'ADN-mt humain a été le sujet d'un grand nombre d'études, mais certains paramètres essentiels concernant son mode de transmission, son polymorphisme intra-individuel ou la vitesse de son évolution ne sont pas encore bien connus. Le séquençage de ses 16'569 paires de bases (pb) (Anderson *et al.*, 1981) a permis de déterminer la structure du génôme mitochondrial et a fourni une séquence de référence très utile pour l'étude de son polymorphisme. Cette séquence, isolée à partir d'une culture de cellules HeLa sera dénommée par la suite "séquence de Cambridge".

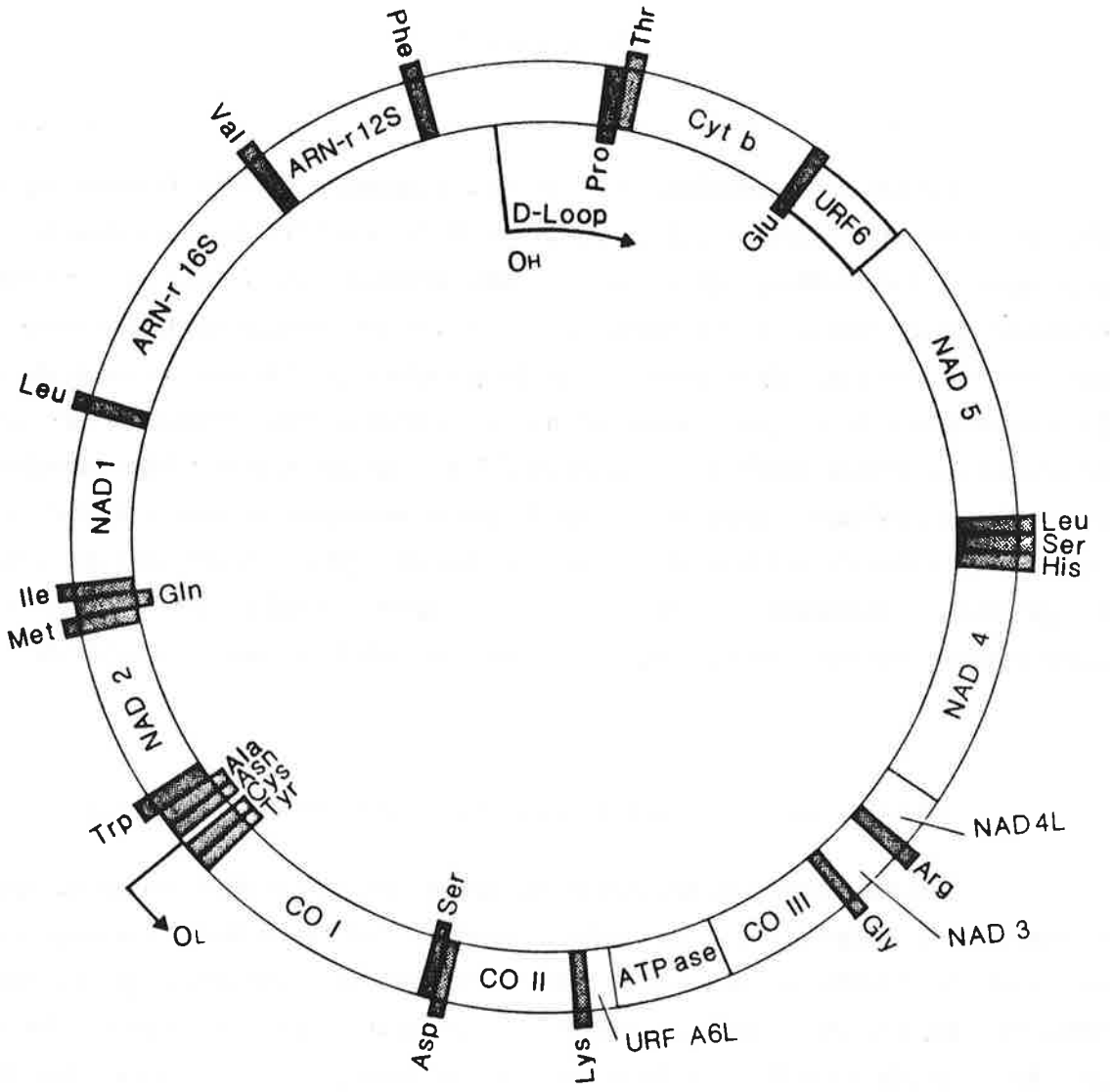

FIGURE 1.1 : Structure du génome mitochondrial humain. Le cercle extérieur représente le brin lourd (*"H-Strand"*) et le cercle intérieur, le brin léger (*"L-Strand"*). Les gènes codants pour les ARN-t des différents acides aminés sont représentés sous forme hachurée avec une protubérance émergeant du brin sur lequel ils sont codés. Les autres gènes sont tous codés sur le brin lourd à l'exception du gène URF 6. Les abrévations adoptées sont les suivantes : URF : *"Unidentified open Reading Frame"*; CO I, II et III : sous-unités I, II et III du complexe de la cytochrome oxydase; Cyt b : cytochrome *b*; ATPase : sous-unité 6 du complexe de l'ATPase mitochondriale; O_H: Origine de la réplication du brin lourd; O_L: Origine de la réplication du brin léger; NAD : Gènes du complexe de la NADH déhydrogénase.

La majeure partie de l'ADN-mt est transcrite. L'ARN résultant code pour 22 ARN de transfert (ARN-t), 2 types d'ARN ribosomique (ARN-r) (ARN 12S et 16S), ainsi que pour 13 protéines dont la plupart ont une fonction connue qui concerne la respiration cellulaire (Anderson *et al.*, 1981; Chomyn *et al.*, 1985) (Figure 1.1).

Seule une partie du génome mitochondrial, se trouvant entre les ARN-t de la proline et de la phénylalanine, n'est pas transcrite en ARN. Elle comprend la région entourant le site d'initiation de la réplication du brin lourd de l'ADN-mt ("*Heavy*" ou "*H-strand*") qui semble également être le site d'initiation de la transcription, d'où son nom de "région de contrôle". La forme principale de l'ADN-mt dans les mitochondries est une structure covalente circulaire avec une boucle de déplacement ("*D-Loop*") à l'origine de la réplication du brin lourd. Celle-ci est formée par la synthèse d'une petite séquence fille du brin lourd (aussi appelée "ADN 7S") qui reste associée au cercle parental et conduit à une structure triplex dite justement D-Loop. La fonction de cette D-Loop comme une amorce de la réplication n'a pas été prouvée. Un autre de ses rôles potentiels pourrait être de procurer un site d'association de l'ADN-mt avec d'autres molécules (protéines, ADN ...) ou d'exposer le brin lourd parental au processus de transcription (Clayton, 1982). La région de la D-Loop semble être la plus variable du génome mitochondrial et elle a été l'objet de la plupart des études de séquence de l'ADN-mt humain (Aquadro and Greenberg, 1983; Greenberg *et al.*, 1983; Walberg and Clayton, 1981). Hormis cette portion non codante, le reste de l'ADN-mt se présente sous la forme de portions codantes en quasi-continuité. Les différents gènes ne sont séparés les uns des autres que par quelques nucléotides ou même aucun dans certains cas. Cette absence des "*spacer*" fréquemment observés dans l'ADN nucléaire, alliée à une absence d'introns à l'intérieur des gènes de structure, conduit à un extrême compactage de l'ADN-mt.

L'ADN-mt semble être principalement, sinon uniquement, hérité de façon maternelle (Giles *et al.*, 1980) par l'intermédiaire des mitochondries qui sont présentes dans l'ovule en nombre beaucoup plus élevé que celles du spermatozoïde. Cette haploïdie apparente et la nature compacte du gémôme mitochondrial contribueraient à réduire considérablement les réorganisations de structure. Ainsi, aucun crossing-over n'a pu être mis en évidence jusqu'ici (Clayton *et al.*, 1974).

L'ADN mitochondrial semble évoluer 5 à 10 fois plus vite que l'ADN nucléaire (Brown *et al.*, 1979, 1982; Dawid, 1972), du moins chez les vertébrés (Vawter and Brown, 1986). Différents mécanismes ont été proposés pour rendre compte de cette

accélération évolutive: la polymérase γ, unique candidate pour la réplication de l'ADN-mt, semble être moins fidèle que la polymérase α qui est responsable de la réplication de l'ADN nucléaire (Kunkel and Loeb, 1981); il n'existe apparamment pas de mécanisme de réparation de l'ADN-mt (Backer and Weinstein, 1980; Clayton *et al.*, 1974); l'ADN-mt subirait plus de cycles de réplication que l'ADN nucléaire (Rabinowitz and Swift, 1970); certaines mutations défavorables pourraient être tolérées du fait du grand nombre de génomes par cellule.

Les mutations enregistrées sont principalement dues à des substitutions de nucléotides, ainsi que des délétions et des insertions (Cann and Wilson, 1983). Il est à noter que les transitions semblent être environ 25 fois plus fréquentes que les transversions (Aquadro and Greenberg, 1983), du moins dans la région de contrôle. Ces mutations ne semblent pas être réparties uniformément sur l'ensemble du génome mitochondrial, mais être accumulées préférentiellement (surtout pour les variations de longueur) dans les parties non codantes (D-Loop) (Cann and Wilson, 1983; Cann *et al.*, 1987; Greenberg, Newbold and Sugino, 1983; Whittam *et al.*, 1986).

De plus amples descriptions de la structure de l'ADN-mt des mammifères et en particulier de l'espèce humaine seront trouvées dans Anderson *et al.* (1981) ainsi que dans Brown (1983, 1985) et Clayton (1982).

PRÉSENTATION DES DONNÉES À DISPOSITION

Techniques d'analyse moléculaire de l'ADN

La structure moléculaire de l'ADN-mt a été étudiée au moyen de deux techniques conventionnelles de la biologie moléculaire. La première consiste à déterminer la séquence exacte des nucléotides constituant la chaîne d'ADN. On peut ainsi connaître la composition exacte des gènes dont on veut identifier la structure et/ou la variabilité. Cette technique, très complexe, et que nous ne décrirons pas ici, est coûteuse en temps et en matériel, ce qui l'a empêchée d'être souvent utilisée pour l'étude de séquences de plus de 1'000 pb et pour de grands échantillons.

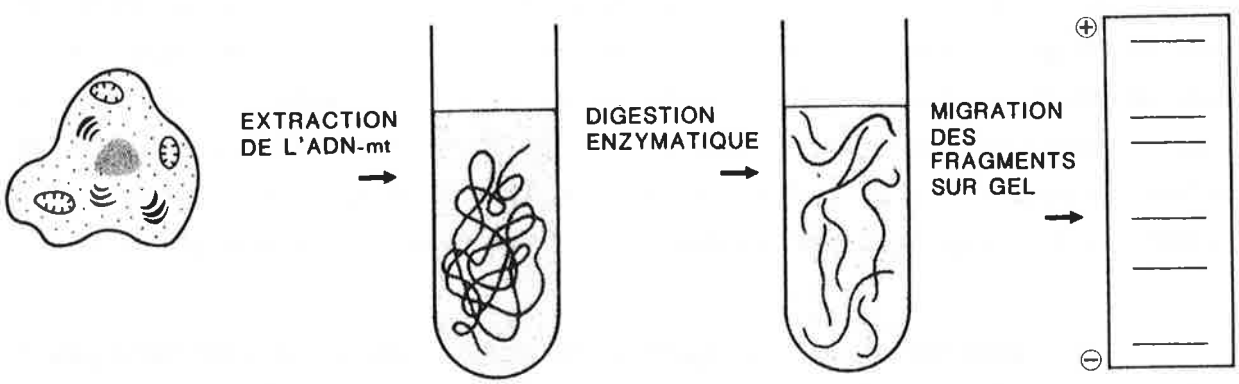

FIGURE 1.2 : Mise en évidence d'un profil de digestion de l'ADN-mt par électrophorèse, précédée de l'extraction et de la digestion de l'ADN.

La seconde technique consiste à déterminer la présence ou l'absence de certains sites sur l'ADN au moyen d'enzymes dits "de restriction" ou endonucléases, qui reconnaissent une séquence de nucléotides particulière et coupent la molécule d'ADN à l'intérieur de cette séquence ou à proximité. La molécule d'ADN-mt circulaire, mise en présence d'un de ces enzymes sera ainsi décomposée en autant de fragments qu'il existe de sites de reconnaissance. L'ADN étant une molécule chargée électriquement, les différents fragments obtenus pourront être mis en évidence en les faisant migrer sur gel dans un champ électrique et en les "colorant" de manière adéquate (voir Figure 1.2). Une certaine combinaison de présences-absences de sites de restriction et donc de différentes bandes sur le gel est appelée "morphe". Cette technique, moins complexe que la première peut s'appliquer sur un échantillon important de molécules de grande taille.

Ces deux techniques, appliquées à une même molécule, ne fournissent pas le même type de données. Le séquençage a l'avantage de pouvoir détecter toutes les différences accumulées entre deux génomes. Mais comme il ne peut pas encore être appliqué à grande échelle, il rend mal compte de la diversité d'une population. C'est néanmoins la technique de l'avenir qui permettra, lorsqu'elle sera automatisée et appliquée à de grands échantillons, de connaître très précisemment la constitution des gènes présents dans une population et par la suite de définir des liens existant entre un réseau de populations ou de grands groupes humains.

L'utilisation d'enzymes de restriction est actuellement la seule technique qui puisse être employée sur des échantillons de plusieurs dizaines d'individus (ce qui n'est souvent pas suffisant, nous le verrons, pour aborder précisément le problème des apparentements entre populations). Si un nombre important de gènes différents peuvent être étudiés, la quantité totale de pb surveillée est nettement inférieure à l'analyse par séquençage.

Pour pallier ce défaut, un même fragment d'ADN peut être étudié avec plusieurs enzymes de restriction reconnaissant chacun une séquence différente. La combinaison des morphes obtenus par les différents enzymes pour un certain génome mitochondrial est considéré comme un haplotype. Une difficulté peut surgir lorsqu'un enzyme reconnaît plusieurs séquences différentes. De ce fait, la précision de l'information apportée par la mise en évidence de la présence (ou de l'absence) de sites de restrictions, même nombreux, est moins bonne que celle apportée par le séquençage direct.

Avant d'aller plus loin dans l'analyse des principales études menées à ce jour, nous allons rapidement présenter l'ensemble des types de données disponibles.

Polymorphisme des séquences d'ADN-mt

Comme nous l'avons vu, la totalité de la molécule d'ADN-mt a été séquencée (Anderson *et al.*, 1981). Ce premier séquençage n'a pas apporté d'information sur le polymorphisme de la molécule, chose qui nous intéresse ici, mais il a fourni une source de référence, à la fois pour d'autres études de séquences, mais aussi pour des études sur le polymorphisme de longueur des fragments de restriction. Le travail monumental d'Anderson et de ses nombreux collaborateurs n'a plus jamais été répété sur l'ensemble de la molécule d'ADN-mt. Néanmoins, un petit nombre d'études ont été menée sur le polymorphisme de petits fragments du génome mitochondrial. Ces études ne concernent malheureusement pas le domaine de l'histoire du peuplement, puisqu'ils sont restreints à la mise en évidence de différences intra et inter-individuelles au niveau moléculaire.

Un dépouillement de la littérature nous a permis de déceler 3 principaux types de travaux. Premièrement, les travaux de Greenberg, Newbold et Sugino (1983) et ceux d'Aquadro et Greenberg (1983) qui se sont intéressés au polymorphisme d'environ 900 pb de la région de contrôle sur la base d'un échantillon de 7 individus.

Deuxièmement, les études de Monnat et Loeb (1985) et de Monnat et Raey (1986) qui ont travaillé sur le polymorphisme du gène codant pour CO III (voir la Figure 1.1 pour leur localisation) chez quelques individus seulement.

Enfin, le dernier type d'étude, effectué par Horai, Inoue et Matsunaga (1987) et Singh, Neckelmann et Wallace (1987) porte sur le séquençage de petits fragments d'ADN-mt chez quelques individus où des mutations de longueur avaient cru être mises en évidence par des analyses de fragments de restriction.

La simple énumération de ces quelques travaux démontre la nécessité d'accumuler d'autres études afin que les travaux de séquençage de l'ADN soient pleinement utilisables à des fins anthropologiques. Néanmoins, il nous a semblé important de mentionner leur résultat dans le cadre de ce travail, car nous verrons qu'ils ont permis de mettre en évidence les bases moléculaires exactes des polymorphismes de longueur des fragments de restriction.

POLYMORPHISME DE LONGUEUR DES FRAGMENTS DE RESTRICTION (PLFR)

Le taux de mutation élevé du génome mitochondrial conjugué à une extraction aisée de l'ADN-mt ont favorisé l'éclosion de nombreuses études sur le PLFR de cette molécule, permettant d'enregistrer des différences notables entre des génomes ayant divergé depuis peu de temps, à l'échelle de l'évolution humaine. Donc, contrairement aux études de séquençage de l'ADN-mt, qui ont plus cherché à révéler la nature exacte des mutations, les études de PLFR ont revêtu un caractère plus quantitatif. Elles ont aussi cherché à mettre en évidence des morphes ou des haplotypes propres à certaines populations ou grands groupes humains.

Dès ses débuts, l'étude des PLFR du génome mitochondrial (Brown, 1980; Denaro *et al.*, 1981) a été envisagée sous deux angles assez différents. Le premier courant, auquel se rattache le travail de Brown, tend à privilégier l'étude d'un grand nombre de sites de restriction aux dépends de la qualité de l'échantillon. Le second courant a une démarche inverse. Les échantillons sont généralement tirés de populations relativement bien définies, mais un petit nombre de sites sont surveillés. Il est clair que les deux types d'analyse fournissent des données de qualités très différentes. Vu la variabilité du génome mitochondrial, nous avons dans le premier cas presque autant d'haplotypes de restriction que d'individus dans l'échantillon. Il en résulte une certaine difficulté à définir des apparentements entre les populations dont ils sont tirés. Dans le second cas, un nombre restreint d'haplotypes (ou de morphes, si l'on n'utilise qu'un seul enzyme) seront définis, et leurs fréquences varieront selon les populations comme des fréquences alléliques dans un système génétique classique. Différents modèles de la génétique des populations s'appliquent bien à ce dernier type d'analyse, alors que dans le premier cas, nous abordons presque le problème de la variabilité inter-individuelle qui manque cruellement de bases théoriques.

TABLE 1.1 : Caractéristiques des principaux travaux portant sur le polymorphisme de longueurs des fragments de restriction (PLFR) de l'ADN-mt.

Etudes	Nombre de gènes dans l'échantillon	Nombre de populations étudiées	Nombre d'enzymes utilisés	Nombre d'haplotypes trouvés
Echantillons hétérogènes				
Brown (1980)	21	5	18	21
Cann *et al.* (1987)	147	5	12	132
Echantillons homogènes				
Denaro *et al.* (1981)	235	5	1	6
Johnson *et al.* (1983)[1]	200	5	5	35
Blanc *et al.* (1983)	116	2	1	8
Horai *et al.* (1984)	120	1	15	22
Wallace *et al.* (1985)[1]	74	1	6	8
Bonné-Tamir *et al.* (1986)[1]	81	2	5	18
Brega *et al.* (1986)[1]	91	1	5	13
Brega *et al.* (1986)[1]	229	2	6	29
Harihara *et al.* (1986)	122	2	3	11
Horai and Matsunaga (1986)	116	1	9	62

[1] Ces échantillons ont été analysés avec 5 enzymes équivalents et sont donc comparables entre eux.

La Table 1.1 résume quelques paramètres des principales études recensées à ce jour. L'appellation "d'échantillons hétérogènes" pour les deux premières études citées se justifie par le fait qu'elles portent sur des individus dont l'appartenance ethnique n'est définie que par les termes "Caucasian", "Oriental" ou "African", alors que, par exemple, dans le dernier cas, il s'agit de Noirs américains établis aux Etats-Unis depuis plusieurs générations. D'autre part, les individus constituant ces échantillons proviennent de localisations géographiques multiples. Lors d'un récent travail sur la diversité génétique des populations africaines, nous avons montré (Excoffier *et al.*, 1988) que le continent africain comprenait plusieurs groupes génétiquement bien différenciés. Il nous semble par conséquent difficile de classer l'ensemble des populations africaines sur la base de 18 Noirs américains d'origine africaine, d'un Nigérian et d'un Khoisan (Cann *et al.*, 1987). D'autres critiques similaires ont d'ailleurs été formulées récemment (Darlu et Tassy, 1987a, 1987b, 1987c). Cette hétérogénéité des échantillons provient peut être du fait que ces ADN-mt ont été extraits à partir de placentas fournis par des hôpitaux. Il est à noter que ceci est également le cas des travaux de Horai *et al.* (1984) et de Horai et Matsunaga (1986) qui portent néanmoins sur une population exclusivement japonaise.

Bien que le seul critère de la qualité d'un échantillon ne soit pas suffisant pour juger de la valeur d'une étude, il en constitue cependant une base indispensable. Comme notre travail a aussi pour but de dresser un bilan des recherches effectuées sur le polymorphisme moléculaire et de définir une base méthodologique solide pour de futures études, nous analyserons en détail des travaux provenant des deux types que nous avons définis. Ceci est aussi justifié par le fait que des études récentes (Cann *et al.*, 1987), bien que se situant dans le cadre de la variabilité interindividuelle, ont fourni des résultats ayant des répercussions dans l'étude de l'origine du peuplement humain.

De ce premier survol des données portant sur les PLFR, il ressort que les différentes études sont difficilement comparables entre elles. En effet, le choix des enzymes employés et donc des séquences de nucléotides surveillées est rarement identique pour tous les travaux. Nous nous trouvons devant un système génétique fractionné en autant de sous-systèmes qu'il existe d'enzymes de restriction avec lesquels on l'analyse. La situation est très différente de celle rencontrée dans l'étude des systèmes des groupes sanguins, ou tissulaires, ou des protéines, qui sont définis par des molécules provenant de l'expression de certains gènes, d'ailleurs pas toujours connus. Nous serons donc amenés à analyser la plupart des études de façon séparée.

UTILISATION DES DONNÉES MOLÉCULAIRES POUR L'ÉTUDE DE L'HISTOIRE DU PEUPLEMENT HUMAIN.

La génétique des populations a fourni certaines bases théoriques permettant de définir relativement clairement les critères requis pour que les études de polymorphisme génétique puissent être correctement interprétés et ceci dès les travaux de pionniers de Fisher (1930) et Wright (1931) sur les changements aléatoires de fréquences dans les populations de taille finie. L'étude de l'histoire du peuplement à travers les données génétiques suppose donc qu'un certain nombre de conditions soient remplies concernant le mode de différenciation des populations les unes par rapport aux autres, l'évolution du système génétique étudié et la représentativité de l'échantillon tiré de la population. En effet, s'il est possible d'analyser n'importe quelles données, la valeur des résultats dépendra avant tout de la qualité des échantillons et de la connaissance du système génétique.

CRITÈRES

La plupart des études génétiques globales sur les apparentements entre populations (Cavalli-Sforza and Edwards, 1964; Langaney, 1979; Nei, 1982; Nei and Roychoudury, 1974, 1982; Piazza, Menozzi and Cavalli-Sforza, 1981; Sanchez-Mazas and Langaney, 1988) découlent de l'analyse de fréquences géniques des allèles ou haplotypes de certains systèmes génétiques supposés neutres et étudiés dans des populations bien définies et à l'équilibre. Les différences de fréquences géniques sont censées refléter l'action de forces évolutives telles que la dérive génétique, les mutations ou les migrations (effet fondateur). Bien que ces modèles soient souvent de grossières simplifications de situations biologiques très complexes, ils se sont avérés indispensables pour échaffauder des hypothèses vérifiables sur l'apparentement génétique des populations. Ils reposent cependant sur un certain nombre de principes très généraux dont le non-respect peut considérablement biaiser les résultats obtenus. Il convient peut être de les rappeler ici brièvement :

- les échantillons doivent être suffisamment grands pour comprendre la grande majorité des types alléliques présents dans la population et pour que les fréquences géniques puissent être correctement estimées;

- les types alléliques ne doivent pas être trop fortement sélectionnés, pour éviter des phénomènes de convergence ou que des populations exposées à des environnements différents ne subissent des pressions sélectives non uniformes;

- les haplotypes doivent être définis de manière non équivoque;

- les haplotypes similaires trouvés dans différentes populations doivent être identiques par ascendance;

- le système génétique doit être suffisamment informatif pour pouvoir différencier les populations étudiées.

La réunion de tout ces critères peut parfois s'avérer difficile, voire impossible pour un système génétique particulier, étant donné, d'une part la difficulté de la récolte du matériel génétique lui-même et, d'autre part, la complexité des analyses moléculaires. Certaines incohérences peuvent donc apparaître après l'étude d'un système génétique. Celles-ci peuvent être corrigées par la confrontation de résultats provenant d'autres systèmes ou même de données extra-génétiques (histoire, linguistique, paléoanthropologie, archéologie ...).

L'emploi de méthodes d'analyse mathématiques ou techniques sophistiquées ne doit pas suffire à occulter les problèmes inhérant à la récolte des données. Si celles-ci sont faussées au départ, le biais se propagera dans les résultats et leur interprétation. Ces simples constatations nous ammènent donc à porter une attention toute particulière à la constitution des échantillons et à la définition moléculaire des systèmes génétiques considérés.

IMPORTANCE DE LA QUALITÉ DES ÉCHANTILLONS

La qualité d'un échantillon dépendra tout d'abord de son homogénéité ethnique et géographique. Il est bien connu que la collection d'individus provenant de différentes populations peut singulièrement biaiser les estimations de fréquences alléliques dans de tels échantillons. A ce critère qualitatif, il convient d'ajouter un critère quantitatif dont il a déjà été fait mention. Les fréquences alléliques seront d'autant mieux estimées que la taille des échantillons sera élevée. Ceci est particulièrement vrai dans le cas des systèmes génétiques possédant des types alléliques de faibles fréquences dans la population, ce qui est le cas pour le polymorphisme de l'ADN au niveau moléculaire (séquences, PLFR). Ce dernier type d'étude requiert souvent qu'une majorité des allèles soient retrouvés lors de la procédure d'échantillonnage. Or, on constate que le nombre de types alléliques attendus dans un

échantillon dépend à la fois du taux de mutation au locus considéré et de la taille de l'échantillon (voir Annexe A, formules A.12 et A.14), d'où l'importance considérable d'avoir un écantillon de taille importante.

TABLE 2.1 : Caractéristiques quantitatives d'échantillons étudiés pour différents systèmes génétiques

Système génétique[1]	Nombre d' échantillons étudiés[2]	Taille moyenne des échantillons	Nombre moyen d'haplotypes	θ moyen[3] (estimé)
Rhésus (7 hapl.)	30	439.4 (322.6)[4]	4.5 (0.86)	0.6 (0.13)
Gm (10 hapl.)	27	439.6 (447.6)	5 (1.41)	0.8 (0.38)
Hla-A (15 spéc. +blanc)	10	204.4 (57.3)	12 (0.82)	2.7 (0.28)
Hla-B (15 spéc. +blanc)	10	204.4 (57.3)	13.3 (1.25)	3.2 (0.56)
beta^A-globine (5 sites)	16	53.5 (27.96) 16-111[5]	4.5 (1.21)	1.2 (0.55)
ADN-mt (68 sites)	10	64.2 (33.2) 40-185[5]	11.2 (3.16)	4.3 (1.72)

[1] La détermination des présence ou absence de types alléliques, pour ces systèmes, a été effectuée par la mise en évidence d'un nombre limité d'allotypes ou de spécificités. Il en résulte que le nombre moyen d'haplotypes ou d'allèles effectivement présents dans les échantillons peut être sous-estimé (il en va de même pour θ).

[2] Les échantillons de chaque système génétique ont été analysés pour les mêmes spécificités (HLA), les mêmes allotypes (Rhésus, Gm) et les mêmes enzymes de restriction (beta^A-globine, ADN-mt). Ce nombre ne constitue donc pas un maximum possible, car nous nous sommes limités à l'étude de quelques travaux strictement comparables.

[3] La définition de θ n'est pas très précise lorsque plusieurs populations sont rassemblées. Elle fait appel à une moyenne conceptuelle des valeurs de 4Nu (ou Nu pour l'ADN-mt) qui ne peuvent être habituellement utilisées pour des tests statistiques. Elles permettent cependant d'évaluer les tailles d'échantillons requises pour collecter une majorité de types alléliques dans la population pour chaque système considéré (voir Annexe A).

[4] Les nombres entre parenthèses représentent des écarts-types.

[5] Tailles d'échantillons extrêmes

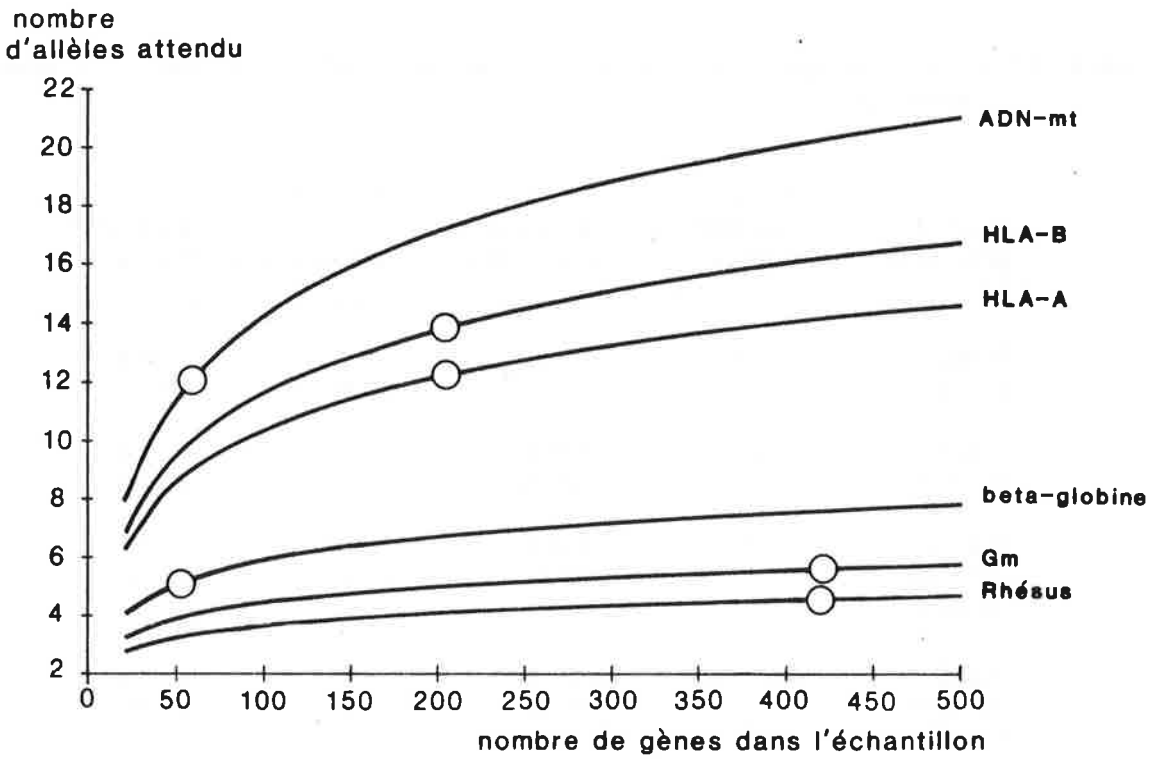

FIGURE 2.1 : Nombre attendu d'allèles, d'haplotypes ou de types d'ADN en fonction de la taille d'un échantillon de n gènes tirés d'une population stationnaire. Les courbes ont été calculées au moyen de l'équation (A.12) sur la base des valeurs θ moyennes reportées dans la Table 2.1.

Paradoxalement, il apparaît que la taille de la majorité des échantillons accumulés à ce jour dans les études de PLFR est comparativement plus faible que celle d'autres marqueurs génétiques classiques (Gm, Rhésus, HLA) (Fig. 2.1, Table 2.1). Donc, bien que, théoriquement, plus un système est polymorphe (valeur de θ élevée), plus la taille des échantillons doit être importante, on remarque que dans les faits le phénomène inverse est observé.

Il serait idéalement souhaitable de connaître la constitution génétique de la majorité des populations humaines avant de dresser un scénario de leur différenciation. Nous sommes malheureusement contraints de constater qu'un tel travail est irréalisable actuellement. De plus, le choix des populations étudiées est rarement dicté par une stratégie d'échantillonnage planifiée, mais plutôt par des circonstances liées au hasard ou à la disponibilité du matériel génétique. De ce fait, les populations analysées sont rarement représentatives d'un groupe plus important ou même d'un ensemble continental. L'étude du polymorphisme de l'ADN humain venant à peine de débuter, il s'ensuit que très peu d'échantillons comparables peuvent être rassemblés et analysés globalement, comme cela a été le cas dans d'autres systèmes à l'échelle mondiale (Sanchez-Mazas, 1986; Sanchez-Mazas and Langaney, 1988) ou continentale (Excoffier *et al.*, 1987; Sanchez-Mazas, 1988).

LIMITES D'INTERPRÉTATION

Dans cette étude, il ne nous sera donc pas possible de dresser des rapports exacts entre les populations humaines à partir des seules données d'un certain type de polymorphisme moléculaire (séquences, PLFR), vu le nombre limité des populations étudiées, mais nous pourrons néanmoins préciser comment quelques groupes de populations se sont différenciées du point de vue moléculaire. Il sera ensuite intéressant de confronter les résultats d'analyses de PLFR entre eux et à ceux obtenus pour d'autres systèmes génétiques afin d'examiner dans quelle mesure ils sont compatibles.

Une incompatibilité pourrait bien évidemment surgir si certaines conditions générales citées plus haut n'étaient pas vérifiées. Or, les difficultés sont potentiellement nombreuses dans le cas de l'ADN-mt où nous avons déjà vu que les tailles des échantillons étaient très réduites. De plus, l'ADN-mt, du fait de sa transmission maternelle, peut conduire à un schéma de différenciation des populations qui soit différent de celui d'un système génétique codé au niveau du noyau (Poulton, 1987).

Enfin, au cas où certains des allèles de l'ADN-mt seraient sélectionnés (Johnson *et al.*, 1983; Whittam *et al.*, 1986), les différences observées pourraient être dues à une hétérogénéité des conditions de sélection plutôt qu'à un processus évolutif neutre.

L'avantage des systèmes génétiques moléculaires réside avant tout dans la précision de la définition des haplotypes et dans la connaissance des événements moléculaires possibles ayant conduit à leur différenciation. Ce gain d'information par rapport à des systèmes génétiques classiques (groupes sanguins, tissulaires ou de protéines) est quelque peu balancé par la nécessité de collecter des échantillons plus grands. Un équilibre doit être trouvé entre ces deux qualités aujourd'hui antagonistes, mais qui pourront et devront être alliées à l'avenir. Dans le cas contraire, si les techniques de séquençage étaient appliquées à grande échelle sans se soucier de la qualité de l'échantillonnage, nous serions confrontés à des collections de gènes tous différents entre eux et entrerions dans le domaine de l'étude de la variabilité interindividuelle qui n'a pas de fondement en génétique des populations. Ceci n'aurait guère plus d'intérêt que l'autre cas extrême qui consisterait à étudier de grands échantillons pour un seul site de restriction et qui classerait toute l'espèce humaine en deux catégories.

DONNÉES PROVENANT DU SÉQUENÇAGE DE L'ADN MITOCHONDRIAL

L'analyse des séquences d'ADN-mt provenant d'individus non apparentés a porté, nous l'avons vu, sur de petits échantillons et n'entre donc pas dans le cadre de l'étude des populations (au sens biologique du terme). Néanmoins, ces travaux ont permis de mieux connaître la structure et la nature des variations moléculaires de l'ADN-mt. Ces précisions nous seront utiles pour mieux cerner certains problèmes d'interprétation des données portant sur les PLFR.

VARIABILITÉ DE LA PORTION NON CODANTE DE L'ADN-MT

Les travaux de Greenberg *et al.* (1983) et d'Aquadro et Greenberg (1983) ont porté sur une même région, principalement non codante, de l'ADN-mt (D-Loop) (Figure 3.1). Cette région comprend le lieu de l'initiation de la réplication et de la transcription d'un des deux brins d'ADN. Au niveau interspécifique, elle semble être la portion la moins bien conservée de l'ADN-mt (Anderson *et al.*, 1981; Upholt and Dawid, 1977; Walberg and Clayton, 1981). La comparaison des séquences provenant de 7 individus différents et d'une longueur de 899 pb ont permis de préciser plusieurs points importants.

Tout d'abord, Greenberg *et al.* (1983) ont montré que cette région n'était pas homogène du point de vue du taux de substitution des nucléotides. Deux régions hypervariables ont été mises en évidence aux extrémités de la D-Loop. Celles-ci s'étendent approximativement des nucléotides 146 à 263 et des nucléotides 16124 à 16362 selon la nomenclature d'Anderson *et al.* (1981) que nous utiliserons tout au long de ce travail. Le taux de substitution moyen entre individus a été calculé selon plusieurs procédures (voir plus loin pour leur développement) qui conduisent à des résultats légèrement différents, mais qui montrent toutes que ce taux est considérablement plus élevé que celui qui avait été estimé sur la base d'études menées avec des enzymes de restriction sur l'ensemble du génome mitochondrial. Ainsi, Brown (1980) et Ferris *et al.* (1981) estiment celui-ci à respectivement 0,36% et 0,30%, alors qu' Aquadro et Greenberg (1983) trouvent un chiffre 5,5 fois plus élevés (1,8%) (voir Table 3.3) pour

cette région non codante. Il faut toutefois noter que les régions hypervariables sont presque seules responsables de cet écart (Greenberg *et al.*, 1983).

Délétions et insertions

Sur les 7 séquences, 56 sites de mutations ont été observés. Parmi ceux-ci, 5 mutations de longueur ont été détectées (délétions ou insertions). Elles se présentent généralement à l'intérieur de répétitions de mono ou dinucléotides et concernent 1 ou 2 bases seulement. De tels polymorphismes de longueur sont peu probables dans le reste du génome mitochondrial du fait de son extrême compactage et de sa nature codante. Le déplacement d'une ou deux positions du cadre de lecture conduirait en effet à l'élaboration de protéines non fonctionnelles.

Nature des substitutions

Les autres mutations sont dues à des substitutions de nucléotides (Figure 3.2). De façon surprenante, la grande majorité des substitutions observées est composée de transitions (49/51 ou 96,1%) (voir Tables 3.1 et 3.4). Si la substitution d'une base pour une autre était équiprobable, on s'attendrait à un taux de 1 transition pour 2 transversions (Figure 3.2). Au niveau interspécifique, Brown *et al.* (1982) ont montré que ce degré de biais transitionnel était corrélé négativement avec le temps de divergence entre espèces d'hominiens. Les transitions représentent encore plus de 90% des substitutions apparentes dans la comparaison de séquences d'ADN-mt entre des hominiens ayant divergé récemment (environ 5 millions d'années) (Brown *et al.*, 1982). On constate néanmoins que plus le temps de divergence entre espèce augmente, plus le taux de transversions est important par rapport aux transitions. Les travaux d'Aquadro *et al.* (1984) ont confirmé l'hypothèse selon laquelle cette augmentation serait uniquement due au biais transitionnel. Il semble donc bien qu'il y ait une accumulation préférentielle de transitions dans la région de contrôle de l'ADN-mt à un taux environ 25 fois plus important que les transversions. Deux transitions sur un même site passeront inaperçues alors que les quelques transversions qui se produisent seront visibles, même si elles subissent encore d'autres transitions. A long terme, ce phénomène peut expliquer la baisse du taux total de substitutions après 15 millions d'années de divergence, observée par Brown *et al.* (1979), mais interprétée alors comme une saturation des sites connaissant un taux de substitution élevé.

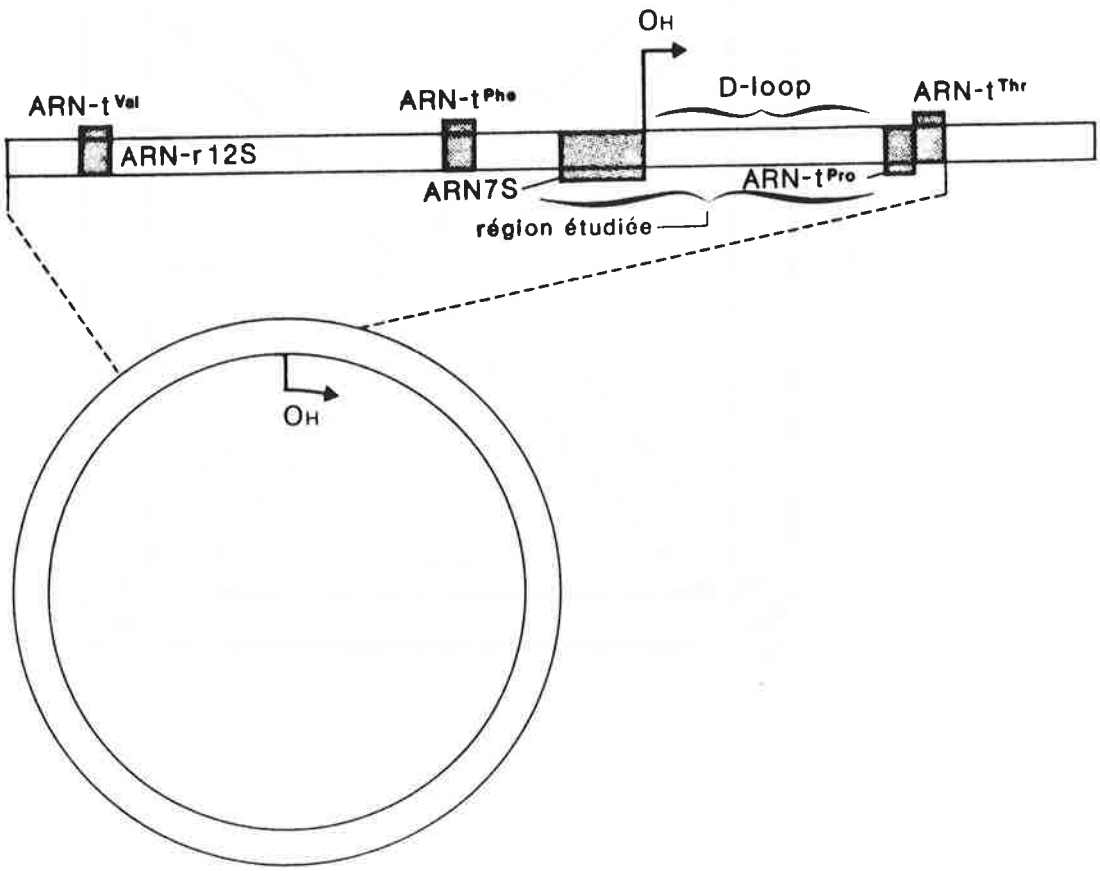

FIGURE 3.1 : Région de la molécule d'ADN-mt étudiée par Greenberg *et al.* (1983) et par Aquadro et Greenberg (1983). L'ARN de coefficient de sédimentation 7 S ne doit pas être confondu avec l'ADN 7 S qui est une ancienne appellation erronée du brin d'ADN de 570-655 pb, complémentaire du brin léger, synthétisé pour former la structure triplex D-Loop (Clayton, 1982).

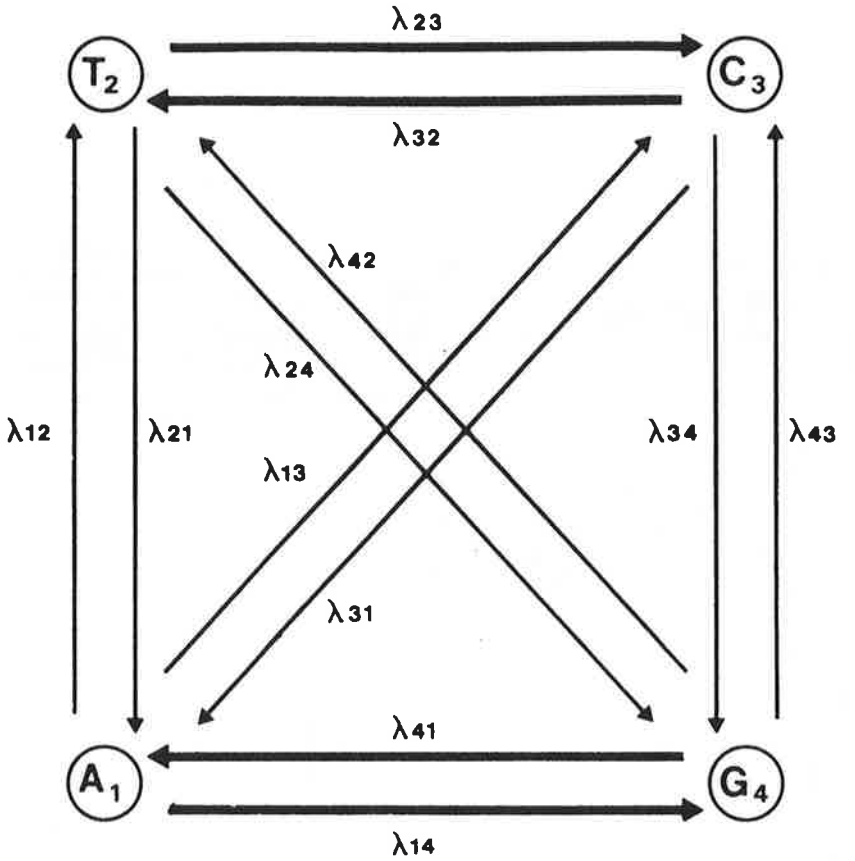

FIGURE 3.2 : Substitutions entre nucléotides. Les flèches en gras indiquent des transitions.

TABLE 3.1 : Origine des substitutions de nucléotides observées entre 7 séquences de la région de contrôle (d'après Aquadro et Greenberg (1983).

Type de substitutions	Nombre observé		
Transitions			
A → G	13	33	49
G → A	20		
T → C	9	16	
C → T	7		
Transversions			
A → T	0	2	
T → A	0		
C → G	0		
G → C	1		
A → C	0		
C → A	0		
T → G	1		
G → T	0		

La Table 3.1 nous montre également que la majorité des transitions (33/49) provient des passages A ↔ G. Ces substitution non aléatoires ont pour conséquence de fausser considérablement les estimations du taux de substitution du fait de l'occurrence de mutations parallèles et réverses.

Substitutions multiples

Aquadro et Greenberg (1983) ont construit un réseau phylogénique entre les 7 séquences étudiées (voir plus loin, pour la discussion de la validité des réseaux ou arbres phylogéniques) sur la base de 10 sites de substitution "informatifs". Notons qu'une mutation est dite "informative", pour une construction phylogénique, si elle est présente sur plus d'une séquence (Fitch, 1977). Dans le cas inverse, la mutation a pu se produire n'importe quand dans le processus évolutif, et n'apporte, par conséquent, pas d'information quant à l'ordre d'embranchement des séquences les unes par rapport aux autres. Le réseau est obtenu en minimisant le nombre d'événements mutationnels

nécessaire pour relier les 7 séquences. Si le réseau proposé par Aquadro et Greenberg (1983), et repris comme s'il s'agissait d'un arbre phylogénique par Cann *et al.* (1987), ne peut fournir d'inférences valables sur la différentiation des populations d'où sont tirés les individus séquencés, elle apporte néanmoins un enseignement précieux : il apparaît que 5 sites de substitution sur 10 ont été l'objet de transitions multiples (pour A ↔ G ou C ↔ T), et cela *quelle que soit la topologie choisie* pour relier les 7 séquences. Le nombre de substitutions observé entre 2 séquences est ainsi parfois fortement sous-estimé par rapport au nombre calculé sur la base du réseau phylogénique (Table 3.2).

Quelques méthodes d'estimation du nombre de substitutions

Si le temps de divergence est faible entre 2 séquences quelconques, le nombre de substitutions par site de nucléotide s peut être estimé par

$$\hat{s} = n_d/n, \tag{3.1}$$

où n_d et n sont respectivement le nombre de nucléotides qui diffèrent entre les 2 séquences et le nombre total de nucléotides comparés. La variance de s a été donnée par Kimura et Ohta (1972) et est égale à $s(1\text{-}s)/n$. Si s devient grand ou si le temps de divergence entre les séquences augmente, il a tendance à sous-estimer le nombre réel de substitutions par site du fait des mutations parallèles. Plusieurs auteurs ont proposés des modèles pour tenir compte de ce phénomène.

a) Méthode de Jukes et Cantor

Soit λ_{ij} le taux de substitution du $i^{ème}$ nucléotide pour le $j^{ème}$ nucléotide par année (i,j = 1, 2, 3, 4 correspondant à A, T, C et G respectivement). Le modèle de Jukes et Cantor (1969) suppose que les probabilités de substitution de n'importe quel nucléotide pour un des trois autres sont identiques, soit $\lambda=\lambda_{ij}$, $\forall$ i et j (voir Figure 3.2). Les deux auteurs estiment la proportion de nucléotides différents entre deux séquences ayant divergé t années auparavant par

$$\hat{p} = 3/4\ (1 - e^{-8\lambda t/3}) \tag{3.2}$$

Selon leur modèle, le nombre attendu de substitutions par site est égal à $s=2\lambda t$ et il peut être estimé par

$$\hat{s} = -3/4 \log_e(1 - 4p/3) \tag{3.3}$$

La variance due à l'échantillonnage est estimée par

$$V(\hat{s}) = \frac{9p(1-p)}{(3-4p)^2 n} \tag{3.4}$$

B) Méthode des 3 types de substitution de Kimura

Le modèle précédent estime *s* correctement si le taux de substitution est le même pour tous les nucléotides et si *t* est petit. Or, nous avons vu que les taux de transitions sont souvent différents des taux de transversions. Afin d'essayer de tenir compte de cette observation, Kimura (1981) a proposé un modèle dans lequel le taux de substitution par site par unité de temps est $\lambda = \alpha + \beta + \gamma$, où α représente le taux de transitions ($\lambda_{14}=\lambda_{41}=\lambda_{23}=\lambda_{32}=\alpha$), β le taux de transversion pour les nucléotides A ↔ T et C ↔ G ($\lambda_{12}=\lambda_{21}=\lambda_{34}=\lambda_{43}=\beta$), et γ le taux de transversion pour les nucléotides A ↔ C et G ↔ T ($\lambda_{24}=\lambda_{42}=\lambda_{13}=\lambda_{31}=\gamma$). Si l'on note par x_{ij} la proportion de paires de nucléotides homologues *i* et *j* observée entre 2 séquences et que $P = x_{14} + x_{23}$, $Q = x_{12} + x_{34}$ et $R = x_{13} + x_{24}$. Le nombre de substitutions par site peut ainsi être estimé par

$$\hat{s} = -1/4 \log_e[(1-2P-2Q)(1-2P-2R)(1-2Q-2R)] \tag{3.5}$$

Dans le cas où $\beta = \gamma$, c'est-à-dire lorsque toutes les transversions connaissent le même taux de substitution,

$$\hat{s} = -1/2 \log_e[(1-2P-Q')(1-2Q')^{\frac{1}{2}}] \tag{3.6}$$

où $Q' = Q + R$ est la fréquence totale des transversions. Cette simplification de 3.5 est identique au modèle à 2 paramètres de Kimura (1980). La variance de *s* est donnée dans ce cas par

$$V(\hat{s}) = 1/n\,[a^2P + b^2Q' - (aP+bQ')^2] \tag{3.7}$$

où $a = 1/(1-2P-Q')$ et $b = 1/2\,[1/(1-2P-Q') + 1/(1-2Q')]$

c) Méthode de Gojobori *et al.*

L'observation des schémas de substitution des 4 nucléotides dans 3 gènes fonctionnels (α- et β-globine, ACTH) et dans 6 pseudogènes a conduit Gojobori *et al.* (1982) à élaborer un modèle complexe où 6 taux de substitutions différents sur les 12 possibles (voir Figure 3.2) sont autorisés. Selon ce modèle, $\lambda_{31}=\lambda_{41}=\lambda_{32}=\lambda_{42}=\alpha$, $\lambda_{13}=\lambda_{23}=\lambda_{14}=\lambda_{24}=\beta$, $\lambda_{21}=\alpha_1$, $\lambda_{43}=\alpha_2$, $\lambda_{12}=\beta_1$, $\lambda_{34}=\beta_2$.Dans ce cas, s peut être estimé par

$$\hat{s} = -yz \log_e[B_1/(pq)] - (2q_1q_2/y)\log_e[y(F_{12}-B_1+3E_{12}/B_1)/(3q_1q_2)] - (2q_3q_4/z)\log_e[z(F_{34}-B_1+3E_{34}/B_1)/(3q_3q_4)] \quad (3.8)$$

où les q_i représentent les fréquences, supposées à l'équilibre, des 4 nucléotides qui peuvent être estimées par $q_i=x_{ii}+\Sigma_{i\neq j} x_{ij}/2$ et

$y = q_1+q_2$,

$z = q_3+q_4$,

$B_1 = yz - (x_{13}+x_{14}+x_{23}+x_{24})/2$,

$E_{12} = [q_1z-(x_{13}+x_{14})/2]\,[q_2z-(x_{23}+x_{24})/2]$,

$E_{34} = [q_3y-(x_{13}+x_{23})/2]\,[q_4y-(x_{14}+x_{24})/2]$,

$F_{12} = x_{11}+x_{22}-x_{12}/2-y^2+3q_1q_2$, et

$F_{34} = x_{33}+x_{44}-x_{34}/2-z^2+3q_3q_4$.

Il faut noter que pour certaines fréquences de nucléotides, la formule 3.8 peut être inapplicable du fait de l'occurrence d'arguments négatifs dans le logarithme.

d) Méthode de Tajima et Nei

Cette méthode, plus simple que la précédente, assume que le taux de substitution du $j^{ème}$ nucléotide vers le $i^{ème}$ nucléotide est identique, $\forall\ i$. Ainsi, $\lambda_{12}=\lambda_{13}=\lambda_{14}=\alpha$, $\lambda_{21}=\lambda_{23}=\lambda_{24}=\beta$, $\lambda_{31}=\lambda_{32}=\lambda_{34}=\gamma$, et $\lambda_{41}=\lambda_{42}=\lambda_{43}=\delta$. Tajima et Nei (1984) ont estimé s par

$$\hat{s} = -b \log_e(1-p/b) \quad (3.9)$$

$$\text{où } b = (1 - \sum_{i=1}^{4} q_i^2 + p^2/h)/2 \text{ et } h = \sum_{i=1}^{4} \sum_{j=i+1}^{4} [x_{ij}^2/(2q_iq_j)]$$

La remarque énoncée pour la formule 3.8, et concernant l'inapplicabilité de cette formule dans certains cas, est également valable pour la formule 3.9

Comparaison des différentes méthodes

Ces 4 méthodes ne constituent pas une liste exhaustive de celles qui ont été proposées, mais elles figurent parmi les plus couramment employées (voir Li *et al.*, 1985 et Tajima, 1985 pour une liste plus complète). Tajima et Nei (1984) ont étudié le comportement de certains de ces estimateurs sur des séquences fictives pour différentes valeur de *s* (voir aussi Tajima, 1985 et Li *et al.*, 1985). Il apparaît que toutes les méthodes sont également valables pour des valeurs de *s* inférieures à 0,25. Pour des valeurs supérieures (et donc pour des temps de divergence plus élevés), il semble que la méthode de Gojobori *et al.* (1982) fournisse les estimations les mieux centrées, mais la formule 3.8 est souvent inapplicable, ce qui limite son utilité. Toutes ces méthodes présentent donc quelques défauts. La méthode de Jukes et Cantor (1969) est de loin la plus simple, mais elle n'est pas très réaliste et conduit à sous-estimer *s*, même quand il est faible, lorsque les taux de substitution pour les 4 nucléotides sont très hétérogènes (Nei, 1987, p. 73). Les autres méthodes, bien que tenant compte de taux de substitution différentiels, sont toutes dépendantes d'un certain schéma de substitution de nucléotides pour la région étudiée. Idéalement, il faudrait connaître ce schéma pour élaborer un modèle *ad hoc*. Mais comme, dans la plupart des cas, celui-ci n'est pas connu exactement, les estimations de *s* seront biaisées.

Application à un cas concret

L'examen des Tables 3.2 et 3.3 nous montrent que la formule 3.5 (Kimura, 1981) contribue effectivement à réhausser la valeur moyenne de *s* en augmentant le nombre estimé de substitutions par site pour toutes les comparaisons deux à deux. Par contre, si l'on tient compte du réseau phylogénique pour calculer n_d, la valeur de *s* est rehaussée d'une toute autre manière. On s'aperçoit qu'une augmentation du nombre de différences de nucléotides, dues à des substitutions parallèles du type A ↔ G ↔ A ou C ↔ T ↔ C, ne se produit que pour certaines comparaisons de séquences, alors que les autres ne sont pas affectées du tout.

TABLE 3.2 : Nombre observé de différences de nucléotides (n_d) pour la comparaison de 7 séquences[1] de 899 pb de l'ADN-mt (D-Loop) et corrections[2] ajoutées à n_d pour tenir compte des substitutions multiples (d'après Aquadro et Greenberg (1983)).

Séquences	1	2	3	4	5	6	7
1		5	20	21	7	10	4
2			21	24	10	13	7
3	*6 0,4*	*8 0,5*		28	23	28	22
4	*0,5*	*0,6*	*0,9*		18	23	17
5		*0,1*	*0,6*	*0,3*		9	5
6	*0,1*	*0,2*	*0,9*	*0,6*	*0,1*		8
7			*2 0,6*	*2 0,4*		*0,1*	

[1] La séquence Nº 1 provient de Anderson *et al.* (1981), les séquences Nº 2 à 6 proviennent de Greenberg *et al.* (1983) et la séquence Nº 7 est tirée de Walberg et Clayton (1981). La provenance "ethnique" des individus séquencés est la suivante : 1, 2, 5, ,7 : "Caucasoid"; 3, 4, 6 : "Negroid". Ces termes ethniques sont tirés de Aquadro et Greenberg (1983)et ne correspondent plus à l'état des connaissances sur la diversité génétique des populations humaines, mais restent des termes généraux encore souvent utilisés.

[2] Les nombres de différences de nucléotides observés entre les 7 séquences (n_d) sont reportés en dessus de la diagonale. En dessous de la diagonale sont reportées les corrections qu'il faut ajouter à n_d pour tenir compte de substitutions multiples. Les colonnes de gauche comprennent les corrections calculées sur la base du réseau phylogénique proposé par Aquadro et Greenberg (1983) et les colonnes de droite représentent les corrections apportées par la formule 3.5 pour la comparaison des séquences deux à deux.

TABLE 3.3 : Estimations du nombre de substitutions de nucléotide par site (s)[1] sur une séquence 899 pb de la région D-Loop de l'ADN-mt.

Méthode d'estimation	E(s)[2]	σ_s
Equation 3.1	0,0171	0,00432
Equation 3.2	0,0173	0,00443
Equation 3.5	0,0174	0,00429
Réseau phylogénique[3]	0,0181	0.00445

[1] Ces estimations sont calculées en prenant la moyenne des s sur toutes les paires de séquences homologues (21 comparaisons)

[2] La comparaison des valeurs de E(s) deux à deux par un test de Student révèle que ces valeurs ne sont pas significativement différentes au niveau α=0,05

[3] L'équation 3.1 est appliquée aux valeurs de n_d déduites du réseau phylogénique d'Aquadro et Greenberg (1983)

Comme ces formules de correction sont censées compenser les effets de taux de substitutions différentiels, elles donnent des valeurs estimées de *s* effectivement plus élevées, mais la différence n'est pas significative (voir Table 3.3) lorsque l'on emploie la méthode de Kimura (1981). D'autre part, elles ne font que renforcer les plus grandes

différences déja observables, ce qui est un ajustement peu réaliste. Le processus de différenciation des séquences homologues étant un phénomène historique (ne se déroulant d'ailleurs pas forcément de manière minimaliste du point de vue des événements mutationnels) et généalogique, une approche phylogénique nous semble, elle, plus réaliste. Cette approche est possible lorsque l'on compare un nombre restreint de séquences, mais devient une tâche parfois insurmontable quand on veut l'appliquer à un échantillon de taille plus importante (voir les tentatives de reconstruction d'arbres phylogéniques à partir de types provenant d'études de PLFR). Elle nous a néanmoins permis de montrer les limites des formules de correction pour l'estimation du taux de substitution par site, particulièrement dans la comparaison de deux séquences quelconques.

VARIABILITÉ D'UNE SÉQUENCE CODANTE DE L'ADN-MT

Bien que la variabilité interindividuelle des portions codantes de l'ADN-mt ait été principalement documentée par des travaux portant sur les PLFR (Brown, 1980; Cann, 1982; Cann *et al.*, 1983, 1984; Whittam *et al.*, 1986), des études restreintes ont néanmoins été menées à partir de séquences d'ADN. Dans un premier temps, Monnat et Loeb (1985) se sont attachés à mettre en évidence les divergences éventuelles de l'ADN-mt entre 5 individus non apparentés au niveau du gène de la sous-unité III de la cytochrome oxydase (CO III) (Figure 3.3). Ce gène code pour un composant essentiel de la chaîne respiratoire cellulaire. Il doit donc être soumis à une forte pression sélective favorisant l'élimination d'une partie des séquences mutantes.

Contrairement aux études de séquences de la région de la D-Loop (Aquadro and Greenberg, 1983; Greenberg *et al.*, 1983; Walberg and Clayton, 1981), Monnat et Loeb ont comparé les séquences de plusieurs clônes chez 5 individus afin d'essayer de mettre en évidence l'existence d'une variabilité intra-individuelle. Le principe de la technique de clônage utilisée par Monnat et Loeb (1985) est simple et peut être expliquée schématiquement de la façon suivante. L'ADN-mt provenant d'un tissu (ici, le sang périphérique) d'un donneur est sectionné par digestion enzymatique. Des séquences provenant de molécules d'ADN-mt, tirées au hasard parmi tout l'ADN extrait des mitochondries, sont insérées dans des vecteurs de clônage (ici des bactériophages M13 mp11). Les vecteurs ayant incorporé une séquence d'ADN sont utilisés pour transformer des bactéries *Escherichia coli* qui croissent ensuite dans un milieu nutritif. Une séquence particulière est isolée et amplifiée par purification sur plaque d'une

colonie bactérienne et une nouvelle croissance. L'ADN ainsi clôné est produit en assez grande quantité pour être séquencé.

Polymorphisme inter-individuel

Deux segments d'ADN ont été étudiés à partir du site de coupure de l'enzyme *Sac I* situé entre les nucléotides 9647 et 9648 (voir. Figure 3.3) au milieu du gène CO III. Le brin léger du segment 4 (selon la terminologie de Monnat and Loeb (1985)), long de 1360 pb a été séquencése ainsi que le brin lourd du segment 7 (608 pb). Ces deux segments n'ont malheureusement pas été séquencés en totalité, ni pour la même longueur dans tous les clônes, ce qui pose quelques problèmes de comparaison. Au total, Monnat et Loeb ont examiné 170 clônes du fragment 4 d'une longueur moyenne de 202,7 nucléotides, ainsi que 78 clônes du fragment 7 d'une longueur moyenne de 188,2 nucléotides.

De la comparaison de ces séquences, il ressort que seules 5 substitutions ont été mises en évidence. Parmi celles-ci, la transversion G → C à la position 9559 a été retrouvée sur tous les clônes du fragment 4. On peut supposer que la séquence de Cambridge porte une mutation à cet endroit ou qu'il s'agit d'une erreur de lecture. De plus, quatre autres transitions ont pu être mises en évidence (3 T → C et 1 G → A), dont deux chez un seul individu. Il est intéressant de noter que la transition G → A à la position 9547 n'a été retrouvée que sur un seul des 66 clônes séquencés provenant du même individu. Il s'agit de l'un des rares cas détecté de polymorphisme intra-individuel chez l'Homme. Les trois autres transitions ont été retrouvées chacune chez un seul individu, mais dans tous les clônes. Bien que les séquences comparées soient en nombre très restreint et de faible taille, il est possible de déterminer le taux de substitution pour cette séquence. En moyenne, le nombre de substitutions entre les 6 séquences connues pour cette région (en incluant la séquence de Cambridge) est de 1,33 pour la comparaison d'une moyenne de 391 nucléotides. Le taux de substitutions par site peut donc être estimé à 0,34%, ce qui est nettement plus faible que les valeurs trouvées pour la région de la D-Loop (Tables 3.3 et 3.5), mais comparable aux résultats des études de PLFR portant sur la totalité de l'ADN-mt (Brown, 1980; Ferris *et al.*, 1981).

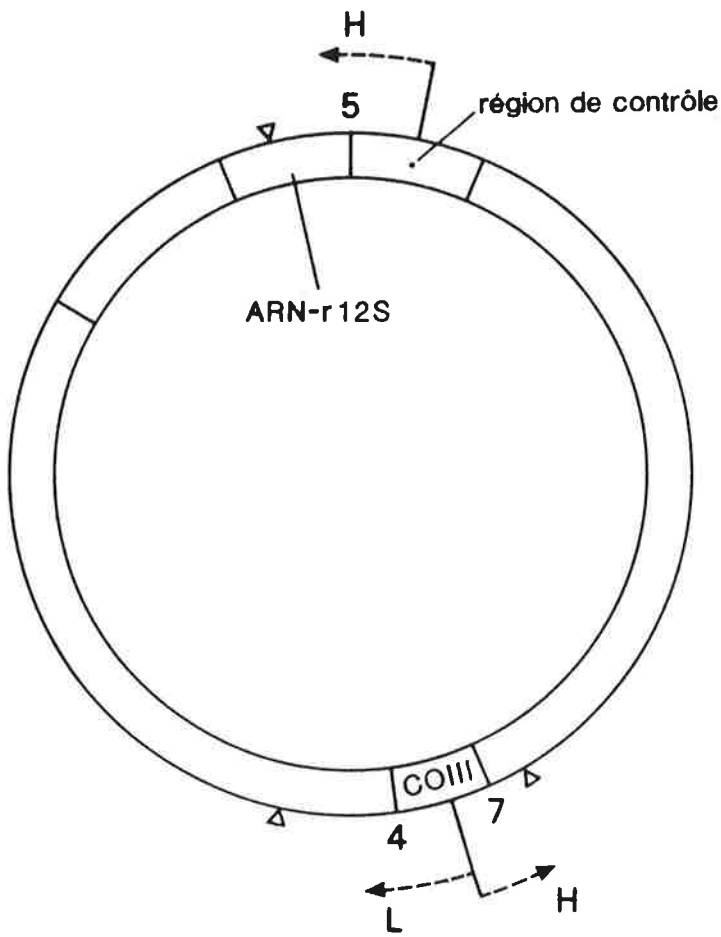

FIGURE 3.3 : Régions de la molécule d'ADN-mt étudiée par Monnat et Loeb (1985) et Monnat et Reay (1986) d'après Monnat and Reay (1986). Les séquences des clônes contenant le brin lourd ("*H-Strand*") des fragments 5 et 7 et le brin léger ("*L-Strand*") du fragment 4 ont été déterminées. Les différents fragments sont bornés par un site *Sac I* et un site *Xba I* représenté par un ▽ . Le nombre de pb séquencé est variable d'un clône à l'autre.

POLYMORPHISME INTRA-INDIVIDUEL

Afin de mieux cerner la problématique du polymorphisme intra-individuel, Monnat et Reay (1986) ont étudié les variations de séquence entre plusieurs clônes provenant de différents tissus (cerveau, coeur, foie, reins et muscle strié) chez deux individus. Les portions du génome mitochondrial étudiées sont identiques à celles des travaux de Monnat et Loeb (1985) avec l'addition d'une séquence (fragment 5) de la D-Loop (voir Figure 3.3). Malheureusement ces deux études ne sont pas comparables pour la détermination du taux de substitution par site, car le nombre moyen de nucléotides séquencé par clône est beaucoup plus faible dans la dernière étude. Sur 121 clônes séquencés pour une longueur moyenne de 145,77 nucléotides, seul le fragment 5 d'un des dix clônes isolés à partir du foie d'un patient a montré une transition (G → A) non observée dans le reste des clônes. Ceci constitue la troisième évidence d'un polymorphisme intra-individuel chez l'Homme (voir page précédente). En marge de leur étude des séquences de la région de la D-Loop, Greenberg *et al.* (1983) avaient également examiné les PLFR de cette région et avaient trouvé que 2 clônes d'un même individu différaient pour un site de reconnaissance *Hae III*. D'autre part, des travaux portant sur la généalogie des types d'ADN-mt d'une lignée maternelle de vaches Holstein (Hauswirth *et al.*, 1984; Olivo *et al.*, 1983) suggèrent qu'une ancêtre récente possédait au moins 4 types de séquences pour la région de la D-Loop, puisqu'on les retrouve dans sa descendance. D'autres cas de polymorphisme intra-individuel ont aussi été recensés chez le rat (Brown and DesRosiers, 1983), la mouche du vinaigre (Solignac *et al.*, 1983) et chez une espèce de lézard (Densmore *et al.*, 1985). Les polymorphismes intra-individuels observés pour l'ADN-mt humain (Greenberg *et al.*, 1983; Monnat and Loeb, 1985; Monnat and Raey, 1986) pourraient aussi être dus à au moins deux autres causes: une mutation post-natale avec ségrégation cellulaire pendant l'ontogénèse ou une mutation s'étant produite par erreur de réplication de la polymérase bactérienne dans le processus d'amplification des clônes. Il semble cependant plus vraisemblable que ce polymorphisme intra-individuel existe bel et bien, car il est théoriquement possible (Birky *et al.*, 1983) et même indispensable pour la transmission à la génération suivante d'une nouvelle mutation apparue dans le lignée germinale.

NATURE ET PHYLOGÉNIE DES SUBSTITUTIONS

Mise à part la transversion C → G observée pour la séquence de référence, toutes les substitutions rencontrées dans les fragments 4 et 7 du gène CO III sont des transitions (voir Table 3.4). Etant donné le nombre limité de séquences comparées, il serait vain de vouloir définir un taux de transversions/transitions, comme cela a été fait

pour la région de la D-Loop. On constatera simplement que le biais transitionnel observé pour la portion non codante du génome mitochondrial semble être également présent dans le gène CO III, mais sa valeur exacte reste à préciser. Le nombre total de substitutions étant très faible, un réseau phylogénique est facilement trouvé entre les différentes séquences à disposition pour cette région (Figure 3.4). Il montre une absence de mutations parallèles ou réverses. Ceci peut être expliqué de deux manières: le fort degré de parallélisme serait absent ou réduit pour les régions codantes ou bien, et cela semble plus vraisemblable, l'absence de parallélisme des substitutions pourrait être uniquement dû à un nombre de sites surveillés plus faible que dans le cas de la D-Loop.

TABLE 3.4 : Liste des substitutions observées dans les études de séquençage de l'ADN-mt.

Position	Substitution[1]	Groupe[2]	Source[3]
7	T → C	N	G
9	C → T	N	G
73	T → C*	C,C,C,C,N,N	G, MR
94	C → T†	C	MR
146	A → G	C	G
150	G → A	N	G
151	G → A	N	G
152	A → G*	C,C,N,N	G, MR
182	G → A	N	G
185	C → T	N	G
186	G → A	N	G
189	T → G	N	G
189	T → C	N	G
195	A → G	N	G
200	T → C	N	G
207	C → T	C	MR
236	A → G	N	G
247	C → T*	N,N	G
263	T → C	C,C,C,N,N,N	G
316	C → T	N	G
444	T → C	C	G
456	G → A	C	G
499	C → T	N	G
750	T → C	N	G
769	C → T	N	G
1190	G → A	C	G
1216	G → A	C	G
1226	G → A	N	G
9540	A → G	C	ML
9547	C → T†	C	ML
9559	C → G	C,C,C,C,C,C,C	ML, MR
9698	A → G	C	ML
9716	A → G	C	ML
9758	A → G	C	MR
16124	A → G	C	G

Position	Substitution[1]	Groupe[2]	Source[3]
16129	C → T	N	G
16134	G → A	N	G
16148	G → A	N	G
16163	T → C	N	G
16167	G → A	N	G
16172	A → G*	C,N	G
16187	G → A	N,N	G
16188	G → C	N	G
16189	A → G	N,N	G
16223	G → A	N,N	G
16224	A → G	C	G
16230	T → C	N	G
16242	G → A	N	G
16243	A → G	N	G
16278	G → A*	C,N	G
16280	T → C	N	G
16293	T → C	N	G
16294	G → A	N	G
16304	A → G	C	G
16311	A → G	C,N,N	G
16320	G → A	N	G
16356	A → G	N	G
16360	G → A	N	G
16362	A → G	N	G
16424	A → G	N	G
16519	A → G*	C,C,N,N	G

[1] La notation des substitutions s'effectue en prenant comme référence le brin lourd de la séquence de Cambridge.
[2] Les notations utilisées sont : C : "Caucasoid" ; N : "Negroid" d'après la terminologie employée par Greenberg *et al.* (1983). Une répétition de sigles indique que la substitution a été retrouvée sur plusieurs individus.
[3] G : Greenberg *et al.*, 1983; ML : Monnat and Loeb, 1985; MR : Monnat and Raey, 1986
* Site de substitution multiple
† Site de substitution polymorphe intra-individuel

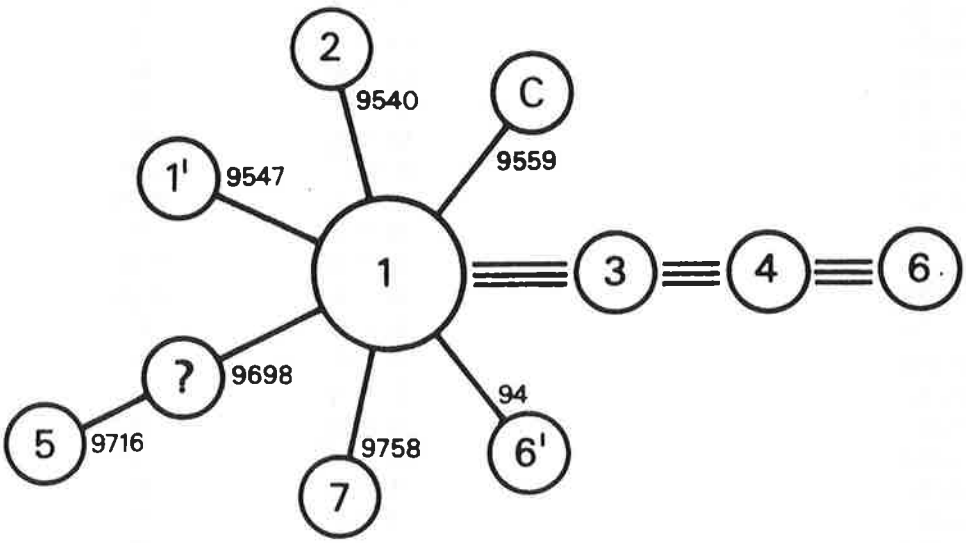

FIGURE 3.4 : Phylogénie des 9 séquences d'ADN-mt définies par Monnat et Loeb (1985) et Monnat et Reay (1986) sur les fragments 4 et 7 (voir Figure 3.3). Les types 1, 3, 4 et 6 sont identiques. Le type C correspond à la séquence de Cambridge. Les types 1' et 6' correspondent à des polymorphismes intraindividuels. Un type intermédiaire noté ? est nécessaire pour relier le type 1 au 5 par deux substitutions.

TABLE 3.5 : Résumé des études de séquençage de l'ADN-mt.

Référence	Taille échantillon	Constitution de l'échantillon[1]	Région étudiée	Taille séquence (pb)	Nombre de substitutions	Taux de transversion/ transitions	Nombre délétions-insertions	Taux de substitutions[2] moyen (x 10^3)
Aquadro and Greenberg (1983)	7	C,N	D-Loop	899	51	0,04	5	18,1 (4,45)
Monnat and Loeb (1985)	6	C	CO III	391	5	0,25	0	3,41 (2,96)

[1] Les abréviations pour les différents groupes ethniques sont les mêmes que celles de la Table 3.4
[2] Calculé au moyen de la formule de Jukes et Cantor (1969). Le chiffre entre parenthèse représente l'écart-type.

ANALYSE DES DONNÉES PROVENANT DU POLYMORPHISME DE LONGUEUR DES FRAGMENTS DE RESTRICTION

Etant donné la complexité des études de séquençage de l'ADN, la diversité du génome mitochondrial humain a été principalement étudiée avec une grande variété d'enzymes de restriction afin de couvrir la plus grande portion possible de l'ADN-mt. Avant d'aborder la description de ce polymorphisme, nous allons préciser quelques notions relatives aux séquences succeptibles d'être reconnues par les enzymes de restriction.

QUELQUES DÉFINITIONS

Nous appellerons une séquence de *r* nucléotides un "*bloc*". Un bloc considéré comme une séquence de reconnaissance pour un enzyme quelconque dans au moins un des génomes d'un échantillon sera dénommé "*site*". Un site sera considéré comme "*monomorphe*" dans un échantillon si tous les membres de l'échantillon ont des blocs identiques et reconnus par le même enzyme. Dans le cas contraire, le site est dit "*polymorphe*". Un "*morphe*" représentera une combinaison particulière de sites pour un enzyme quelconque sur une molécule d'ADN-mt. La combinaison de morphes définie par plusieurs enzymes pour une même molécule d'ADN-mt sera nommée "*haplotype*" ou "*type*". On considérera que la dynamique des haplotypes dans les populations suit la théorie du modèle des allèles infinis développée dans l'Annexe A.

ENZYMES EMPLOYÉS

Le principal intérêt d'utiliser plusieurs enzymes est d'obtenir une série d'haplotypes qui sont des marqueurs génétiques beaucoup plus précis que les morphes, souvent sujets au phénomène de convergence évoqué dans la partie relative au séquençage de l'ADN-mt. La Table 4.1 présente une liste des enzymes employés dans les différentes études de l'ADN-mt que nous avons répertoriées.

TABLE 4.1 : Enzymes utilisées dans les études de PLFR de l'ADN-mt.

Enzyme	Séquence de reconnaissance	Nombre de sites attendus par molécule d'ADN-mt	Etudes concernées
	4 pb		
Alu I	A G C T	51,7	B_{80},C_{82},Bo_{86}
Hpa II	C C G G	27,7	B_{80},C_{82},W_{85},H_{86}
Msp I	C C G G	27,7	J_{83},B_{86},Bo_{86}
Acc II	C G C G	27,7	H_{86}
Fnu D II	C G C G	27,7	C_{82}
Mbo I	G A T C	51,7	B_{80},C_{82}
Sau3A I	G A T C	51,7	H_{86}
Hha I	G C G C	27,7	B_{80},C_{82},H_{86}
Hae III	G G C C	27,7	B_{80},C_{82},H_{86}
Rsa I	G T A C	51,7	C_{82},H_{86}
Taq I	T C G A	51,7	B_{80},C_{82},H_{86}
	5 pb		
Dde I	C T N A G	51,7	C_{82}
Hinf I	G A N T C	51,7	B_{80},C_{82},H_{86}
Ava II	G G [T/A] C C	15,4	C_{82},J_{83},W_{85},B_{86},Bo_{86},H_{86},Ha_{86}
	6 pb		
Hind III	A A G C T T	3,9	B_{80},H_{84}
Stu I	A G G C C T	2,1	H_{84}
Sca I	A G T A C T	3,9	H_{84}
Pvu II	C A G C T G	2,1	B_{80},H_{84},Ha_{86}
Xho I	C T C G A G	2,1	B_{80},H_{84}
Pst I	C T G C A G	2,1	B_{80},H_{84}
Eco RI	G A A T T C	3,9	B_{80},H_{84}
Sac I	G A G C T C	2,1	B_{80},H_{84}
Eco RV	G A T A T C	3,9	H_{84}
Bam HI	G G A T C C	2,1	B_{80},J_{83},W_{85},B_{86},Bo_{86},H_{84}
Kpn I	G G T A C C	2,1	B_{80},H_{84}
Hinc II	G T Py Pu A C	12,7	B_{80},B_{83},W_{85},B_{86},Ha_{86},H_{84}
Hpa I	G T T A A C	3,9	B_{80},D_{81},C_{82},J_{83},W_{85},B_{86},Bo_{86}, H_{86},Ha_{86}
Xba I	T C T A G A	3,9	B_{80},H_{84}
Dra I	T T T A A A	7,4	H_{84}
Hae II	Pu G C G C Py	6,8	J_{83},H_{84},W_{85},B_{86},Bo_{86}

Pu : Purine (C ou T)
Py : Pyrimidine (A ou G)
N : A, C, G ou T

B_{80} : Brown, 1980; D_{81} : Denaro *et al.*, 1981; C_{82} : Cann, 1982 et Cann *et al.*, 1987; B_{83} : Blanc *et al.*, 1983; J_{83} : Johnson *et al.*, 1983; H_{84} : Horai *et al.*, 1984; W_{85} : Wallace *et al.*, 1985; B_{86} : Brega *et al.*, 1986a, 1986b; Bo_{86} : Bonné-Tamir *et al.*, 1986; H_{86} : Horai and Matsunaga, 1986; Ha_{86} : Harihara *et al.*, 1986

Nous avons calculé le nombre moyen de sites attendus par génome mitochondrial (m) pour chaque enzyme dans l'hypothèse que les fréquences des substitutions entre les 4 nucléotides sont identiques. Ce nombre est donné par

$$E(m) = m_T a_0 \qquad (4.1)$$

où m_T est le nombre de blocs de r nucléotides définissable sur un fragment d'ADN donné (ici m_T = 16'569 = longueur du génome mitochondrial circulaire) et a_0 est la probabilité qu'un bloc de r nucléotides soit un site. Cette probabilité dépend de la constitution de la séquence de reconnaissance de chaque enzyme, ainsi que de la fréquence des 4 nucléotides dans la portion du génome étudié. Dans le cas d'un enzyme qui reconnaît une séquence unique, a_0 est obtenu par

$$a_0 = g_A^{r_A} g_T^{r_T} g_C^{r_C} g_G^{r_G} \qquad \text{(Nei and Tajima, 1980)} \quad (4.2)$$

où g_A, g_T, g_C et g_G sont les fréquences des nucléotides A, T, C et G dans la séquence d'ADN étudiée et r_A, r_T, r_C et r_G sont les nombres d'occurrence respectifs des 4 nucléotides dans la séquence de reconnaissance. Dans le cas où un enzyme peut reconnaître plusieurs séquences de r nucléotides (*Ava II* par exemple), a_0 est donné par une formule quelque peu différente et plus générale :

$$a_0 = \prod_{i=1}^{r} f_i \qquad (4.3)$$

$$\text{avec } f_i = \sum_{k=1}^{j} g_k$$

où j est le nombre de bases possibles pour une position donnée de la séquence et g_k peut prendre les valeurs g_A, g_T, g_C ou g_G. Dans le cas d'*Ava II*, qui reconnait les deux séquences GGTCC et GGACC, $a_0 = g_G g_G (g_T + g_A) g_C g_C$. Les fréquences g_k ont été estimées par Anderson *et al.*(1981) sur l'ensemble de l'ADN-mt comme $\hat{g}_A$ = 0,309, $\hat{g}_C$ = 0,312, $\hat{g}_T$ = 0,247 et $\hat{g}_G$ = 0,131. Bien que les séquences d'ADN puissent varier selon les individus et les populations, on considérera que la composition en nucléotides ne varie guère d'une séquence à l'autre, et, en première approximation, on reprendra ces chiffres comme estimateurs des vraies valeurs de g_A, g_C, g_T et g_G.

Le nombre de sites attendus peut varier considérablement d'un enzyme à l'autre (voir Table 4.1). En admettant qu'il n'y ait pas de biais transitionnel, ce nombre ne sera pas forcément observé pour chaque séquence examinée, mais il constitue une limite vers laquelle la moyenne des observations devrait tendre. L'emploi d'enzymes reconnaissant un bloc de 4 pb permet de surveiller une plus grande portion du génome

mitochondrial, car on s'attend à observer beaucoup plus de sites qu'avec les enzymes reconnaissant 5 ou 6 pb. Notons que *Hinf I* et *Dde I* reconnaissent effectivement 4 pb spécifiques seulement, la nature de la 5^{e} étant indifférente. Les enzymes possédant des séquences de reconnaissance multiples peuvent théoriquement cacher une certaine proportion de substitution, notamment *Hinc II* et *Hae II* qui tolèrent des séquences éloignées d'une ou deux transitions. Ils pourront donc conduire à sous-estimer le nombre de types alléliques effectivement présents dans un échantillon et, par là même, le taux de substitution.

Comme la plupart des études de PLFR n'ont pas utilisé la même batterie d'enzymes, les haplotypes obtenus ne sont pas directement comparables entre eux. Pour tenter des comparaisons, il serait nécessaire de fragmenter les haplotypes en sous-haplotypes regroupant les morphes des seuls enzymes utilisés en commun dans certaines études. On se rend vite compte qu'un tel travail est fastidieux et qu'il demande que la liste des morphes de chaque enzyme soit tout d'abord standardisée, ce qui n'est pas encore le cas. Ainsi, les mêmes morphes d'enzyme donné peuvent porter des numéros différents selon les études. L'inverse a également été rencontré. Ceci nous a conduit à réexaminer complétement tous les morphes à disposition et à établir parfois notre propre nomenclature.

Dans un premier temps, nous allons nous concentrer sur un groupe d'études compatibles entre elles (voir Table 1.1) qui portent sur des échantillons relativement homogènes tirés de populations au sens biologique du terme. Nous nous intéresserons ultérieurement aux autres études de PLFR de l'ADN-mt.

ETUDES DE PLFR PORTANT SUR DES POPULATIONS

Dans un chapitre précédent, nous avons insisté sur l'importance de la qualité de l'échantillonnage des individus testés pour des marqueurs génétiques. Sur cette base, nous avons été obligé de classer les études de PLFR de l'ADN-mt menées à ce jour en deux catégories. Celles qui se basent sur une procédure d'échantillonnage grossièrement compatible avec les critères requis en génétique des populations et celles qui ne le font pas. Ce classement ne préjuge évidemment en rien de l'investissement en temps et en matériel, ainsi que des efforts considérables consentis par certains chercheurs.

L'étude de Denaro *et al.* (1981) fut la première à utiliser les PLFR pour évaluer des fréquences de morphes (*Hpa I*) de l'ADN-mt dans plusieurs échantillons, consacrant ceux-ci au rang de marqueurs génétiques utilisables pour des comparaisons de populations. Les travaux de Johnson *et al.*(1983) ont repris et étendu l'utilisation des fréquences des morphes dans les populations étudiées par Denaro *et al.* (1981) avec l'analyse de 4 autres enzymes (*Hae II*, *Ava II*, *Msp I* et *Bam HI*). Ils ont ainsi obtenu une série d'haplotypes possédant des fréquences différentes dans 4 populations humaines. Par la suite, plusieurs auteurs ont réutilisés cette batterie de 5 enzymes pour étudier d'autres populations, enrichissant ainsi la palette d'haplotypes trouvés et permettant des distinctions plus fines entre groupes continentaux sur une base d'haplotypes définis sur 642 individus testés. C'est vers cette série d'études regroupant 10 populations que nous allons nous tourner.

Constitution des échantillons

La composition des échantillons testés avec la même batterie d'enzymes est présentée dans la Table 4.2. Tous les individus sont en principe non-apparentés au premier degré. Les échantillons dénommés Caucasoïdes et Orientaux sont ceux qui sont le moins biens définis ethniquement et géographiquement. Nous les incluons *a priori* dans cette étude, tout en restant prudent quant aux éventuelles interprétations que nous en tirerons.

La taille des échantillons notée dans la Table 4.2 représente le nombre d'individus à disposition pour les différentes analyses enzymatiques. En consultant les Tables 4.5, 4.7, 4.9, 4.11 et 4.13, on s'aperçoit que cette taille n'est pas identique pour toutes les analyses, certains individus n'ayant pas été testés pour tous les enzymes. Précisons que l'échantillon de Sardes est principalement constitué de nouveaux-nés d'une clinique de Cagliari et seuls 134 individus sur 185 ont été testés pour les 5 enzymes communs aux 10 échantillons. Quoiqu'il en soit la taille brute des échantillons reste faible (71,6), surtout si on la compare à celle d'études d'autres marqueurs génétiques (voir Table 2.1).

TABLE 4.2 : Composition des échantillons

Echantillons[1]	Taille[2]	Localisation géographique[3]	Auteurs
Caucasoïdes	54	Etats-Unis, Europe	Johnson *et al.*, 1983
Orientaux	46	Taïwan, Chine, Japon	Johnson *et al.*, 1983
San	41	Bostwana	Johnson *et al.*, 1983
Bantous (dont 23 Zulu)	48	Afrique du Sud (Johannesburg)	Johnson *et al.*, 1983
Amérindiens (Pima, Papago, Hualapai)	74	Etats-Unis (Arizona)	Wallace *et al.*, 1985
Tharu	91	Népal (Chitwan)	Brega *et al.*, 1986a
Israéliens Arabes	41	Palestine (villages du centre et du nord)	Bonné-Tamir *et al.*, 1986
Israéliens Juifs (dont 35 Ashkénazes)	40	Palestine	Bonné-Tamir *et al.*, 1986
Romains	96	Rome (originaires du centre et du sud de l'Italie)	Brega *et al.*, 1986b
Sardes	185	Sardaigne (Cagliari)	Brega *et al.*, 1986b

[1] La dénomination des échantillons est reprise et traduite littéralement des articles originaux.
[2] La taille de l'échantillon est exprimée en nombre de gènes. Ce nombre correspond au nombre d'individus testés dans le cas de l'ADN-mt pour au moins un des cinq enzymes à disposition.
[3] La localisation géographique est réinterprétée.

La procédure de prélèvement de l'ADN-mt dans ces divers travaux est similaire à celle qui est effectuée pour d'autres marqueurs génétiques : le sang périphérique de certains individus, pris au hasard parmi une population relativement bien définie, est extrait pour être ensuite analysé. Cela contraste nettement avec d'autres procédures d'extraction de l'ADN-mt à partir de placentas (Brown, 1980; Cann, 1982; Cann *et al.*, 1987; Horai and Matsunaga, 1986; Horai *et al.*, 1984) ou de plaquettes sanguines (Blanc *et al.*, 1980) qui peuvent influer directement sur la constitution et l'homogénéité des échantillons, du fait de la collecte difficile de ces tissus. Le fait de prélever du sang périphérique n'assure pas en lui-même une meilleure qualité des échantillons, mais rend néanmoins possible une procédure d'échantillonnage correcte.

LOCALISATION DES SITES DE RECONNAISSANCE SUR L'ADN-MT

Grâce à la disponibilité d'une séquence complète de l'ADN-mt (Anderson *et al.*, 1981), il a été possible de préciser à deux exceptions près la position des 69 sites, polymorphes ou non, définis par les 5 enzymes cités plus haut. La position de 42 sites a été directement établie à partir de la séquence de référence d'Anderson *et al.* (1981). La localisation des autres sites polymorphes a été le plus souvent proposée par les auteurs eux-mêmes après avoir établi des cartes de restriction à partir des longueurs des fragments obtenus sur les gels d'électrophorèse pour chaque morphe. Des doubles digestions ont permis de préciser à environ 100 pb près (cela dépend de la taille des fragments) la position du site sur le génome mitochondrial. La confrontation de cette position approximative avec la séquence de référence a conduit à une localisation plus précise (voir Table 4.3).

Il est en effet possible de trouver des blocs ne différant d'un site que d'une substitution par rapport à la séquence de référence. Ces blocs sont aussi appelés "*sites potentiels*" (Adams and Rothman, 1982). Si plusieurs sites potentiels sont présents dans une même région d'ADN-mt, il n'est pas toujours aisé de déterminer avec exactitude l'emplacement du site en question.

On prendra comme exemple la localisation d'un site *Ava II* trouvé par Johnson *et al.* (1983) aux alentours des pb 8229 ou 8275. La séquence de référence nous montre qu'il existe des sites potentiels commençant aux pb 8165, 8186, 8212, 8248, 8249 et 8269. Le dernier site potentiel est le seul qui tombe dans une région non codante (8267-8294) de l'ADN-mt (c'est en fait un des rare *spacer* du génome mitochondrial), ce qui le prédisposerait à subir moins de contraintes fonctionnelles. Il serait donc le meilleur candidat pour avoir muté. Une autre étude, techniquement plus précises, a pu montrer qu'il pourrait s'agir plutôt du site débutant aux pb 8248 ou 8249 qui serait impliqué dans les passages entre morphes (Cann, 1982). Horai and Matsunaga (1986) ont confirmé que dans certains cas il s'agissait bien du site 8249 qui était impliqué, mais d'autres sites avoisinants l'étaient également (site 8269 probablement) de façon indépendante. Le problème consistera à identifier le nombre de mutations indépendantes et les passages entre types concernés. Pour notre part, nous adopterons la position 8269 pour ce site, tout en considérant que d'autres blocs ont pu également muter indépendemment dans cette région.

TABLE 4.3: Sites de restriction recensés parmi 61 types trouvés dans 10 populations.

Position[1]	Site[2] Nº	Enzyme	Région
104	1	*Msp I*	D-Loop
657	2	*Ava II*	ARN 12 S
931	3	*Msp I*	"
1113	4	*Hpa I*	"
1169	5	*Ava II*	"
2157	6	*Hpa I*	ARN 16 S
2268	8	*Ava II*	"
2621	9	*Ava II*	"
2776	10	*Ava II*	"
3077	12	*Msp I*	"
3164	13	*Hae II*	"
3246	14	*Msp I*	ARN-t Leu
3592	15	*Hpa I*	NAD 1
3881	16	*Ava II*	"
≈4308	17	*Ava II*	ARN-t Ile
4533	18	*Hae II*	NAD 2
4711	19	*Msp I*	"
4810	20	*Ava II*	"
4830	21	*Hae II*	"
4846	22	*Msp I*	"
5242	23	*Msp I*	"
5260	24	*Ava II*	"
5693	25	*Hpa I*	ARN-t Asn
5742	26	*Msp I*	O_L
5766	27	*Msp I*	ARN-t Cys
6262	28	*Msp I*	CO I
6571	29	*Msp I*	"
6688	30	*Msp I*	"
6850	31	*Msp I*	"
7204	32	*Msp I*	"
7855 ou 3131	33 ou 12	*Msp I*	CO II ou ARN 16 S
7973	34	*Msp I*	CO II
8112	35	*Msp I*	"
8150	36	*Msp I*	"
≈8269	37	*Ava II*	non codant
9056	38	*Hae II*	ATPase 6
9264	39	*Hae II*	CO III
9292	40	*Msp I*	"
9689	41	*Hae II*	"
10016	42	*Hpa I*	ARN-t Gly
11001	43	*Hae II*	NAD 4
≈11454	44	*Msp I*	"
11688	45	*Msp I*	"
12026	46	*Hpa I*	"
12123	47	*Msp I*	"
12191	48	*Ava II*	ARN-t His
12408	49	*Hpa I*	NAD 5
12629	50	*Ava II*	"
12815	51	*Msp I*	"

Position[1]	Site[2] Nº	Enzyme	Région
≈13100	52	*Msp I*	NAD 5
13181	53	*Hae II*	"
13364	54	*Msp I*	"
13366	55	*Bam HI*	"
13367	56	*Ava II*	"
13367	55/56	*Ava II/Bam HI*	"
13598	57	*Hae II*	"
13712	58	*Msp I*	"
14204	59	*Msp I*	"
14258	60	*Bam HI*	"
14862	61	*Hae II*	Cyt B
15006	62	*Hae II*	"
15487	63	*Ava II*	"
15497 ou 2540	64 ou 8	*Hae II*	Cyt B ou ARN 16 S
15503	65	*Msp I*	Cyt B
15885	66	*Ava II*	non codant
15925	67	*Msp I*	ARN-t Thr
16389	68	*Bam HI*	D-Loop
16390	69	*Ava II*	"
16390	68/69	*Ava II/Bam HI*	"
16453	70	*Msp I*	"

[1] Les sites soulignés sont polymorphes sur l'ensemble des échantillons
[2] Numéros utilisé dans notre nomenclature, indicateurs de l'ordre des sites de reconnaissance des divers enzymes sur le génome mitochondrial
O_L : Origine de réplication du brin léger; CO I, II et III : Cytochrome oxydase I, II et III; NAD 1 à 5 : NADH-déhydrogénase (Chomyn *et al.*, 1985).

Si l'on considère chaque enzyme indépendamment, aucune paire de site ne se chevauche. Il n'en est pas de même si l'on tient compte de plusieurs enzymes simultanément. En consultant la Table 4.3, on s'aperçoit qu'il existe deux paires de sites potentiellement chevauchant, à savoir les sites *Ava II* 56 (pb 13367) et *Bam HI* 55 (pb 13366), ainsi que les sites *Ava II* 69 (pb 16390) et *Bam HI* 68 (pb 16389). Ces sites ne sont pas indépendants et une substitution sur une des bases reconnue par les deux enzymes peut théoriquement entraîner un changement de site pour ces deux enzymes. L'haplotype résultant de cette unique substitution aura apparemment une différence de 2 substitutions par rapport à l'haplotype non mutant. Cet événement s'est effectivement produit dans les deux cas sus-mentionnés. Cela entraîne d'importantes modifications lorsque l'on veut construire un réseau phylogénique entre les haplotypes. Cela souligne aussi l'importance de la localisation précise des sites de reconnaissance des divers enzymes pour la détermination de la nature moléculaire des différences observées entre haplotypes.

DÉFINITION DES MORPHES

La digestion enzymatique de l'ADN-mt provenant d'un individu quelconque conduit à l'observation d'un certain profil de digestion mis en évidence par électrophorèse. Dans le cas du génome mitochondrial circulaire, on observe autant de fragments sur le gel qu'il existe de sites de coupure. La longueur des fragments obtenus est déterminée par la distance séparant deux sites de coupure adjacents. En connaissant la taille de tous les fragments sur le gel, il est possible de déterminer quels sites sont présents sur un génome particulier. Les morphes sont ainsi définis par une certaine combinaison de présences de sites. Par analogie avec des marqueurs génétiques classiques (Gm par exemple), on peut les considérer comme des allotypes.

Le nombre de morphes théoriquement observables (φ) dépend du nombre de sites identifiés (s) comme $\varphi = 2^s$. Ce nombre devient rapidement considérable lorsque s augmente, mais dans la pratique, on n'observe qu'une très petite partie des morphes possibles. Ceux-ci sont généralement issus les uns des autres par un seul événement mutationnel, de sorte qu'il est possible d'établir un réseau reliant tous les morphes. Ce réseau ne peut pas toujours être interprété comme un réseau "phylogénique" étant donné que dans certains cas un morphe peut découler de plusieurs autres par une seule mutation dans chaque cas. Cependant, il semble bien que les morphes observés aient été obtenus par des mutations successives à partir d'un ou plusieurs morphes très anciens. Il n'est pas pour autant possible de dire de façon certaine si celui ou ceux-ci étaient présent avant ou pendant le phénomène de la spéciation humaine, à moins de disposer de références extérieures à l'espèce humaine (certains hominiens par exemple) ou de pouvoir disposer d'estimations du temps nécessaire pour créer le polymorphisme observé actuellement (voir chapitres suivants).

Nous allons maintenant définir, pour chaque enzyme, la liste des morphes trouvés dans les échantillons et les rapports existant entre eux. La référence aux sites se fera principalement par l'intermédiaire des numéros que nous leur avons attribués dans la Table 4.3 et non par leur position en pb telle que décrite par Anderson *et al.* (1981).

Pour nos calculs, nous avons codé les morphes sous forme de vecteurs booléens avec, pour chaque site de reconnaissance, un 1 si le site est présent dans le morphe en question ou un 0 si le site y est absent. Cette représentation vectorielle est utilisée également dans les Tables 4.4, 4.6, 4.8, 4.10 et 4.12

Hpa I

La nomenclature de référence utilisée pour la définition des morphes *Hpa I* est celle décrite par Denaro *et al.* (1981). La liste des morphes, ainsi que les sites présents dans chacun de ceux-ci est indiquée dans la Table 4.4.

TABLE 4.4 : Codification des morphes *Hpa I*

Morphes	Sites[1]						
	4	6	15	25	42	46	49
1	0	0	0	1	1	0	0
2	0	0	0	1	1	0	1
3	0	0	1	1	1	0	1
4	0	0	0	1	1	1	1
5	1	0	0	1	1	0	1
6	0	1	0	1	1	0	0

[1] La notation des sites correspond à celle qui est définie dans la Table 4.3

Le polymorphisme de 5 des 7 sites observés génère 6 morphes dont les fréquences sont reportées dans la Table 4.5 pour 10 populations humaines. Notons qu'un autre échantillon de 44 Pygmées Aka (Denaro *et al.*, 1981) n'a pas été mentionné dans la Table 4.5, car nous ne possédions pas de données pour les 4 autres enzymes utilisés pour la formation des haplotypes. Les fréquences des morphes y sont néanmoins très comparables à celles des San (morphe 2 : 4,5 %; morphe 3 : 95,5 %).

TABLE 4.5 : Fréquence (%) des morphes *Hpa I* dans diverses populations humaines.

Morphes	Populations									
	Cauc.[1]	Rom[2].	Sard.[2]	Isr. J.[3]	Isr. A.[3]	Bant.[1]	San[1]	Orient.[1]	Tharu[4]	Am.[5]
	54	96	185	38	40	48	41	48	91	74
1	0	0	0	0	0	4,2	0	12,5	7,7	1,4
2	98,1	99,0	99,5	100	87,5	25,0	7,3	81,3	92,3	98,6
3	0	1,0	0	0	12,5	70,8	92,7	0	0	0
4	0	0	0,5	0	0	0	0	4,2	0	0
5	1,9	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	2,1	0	0

Sources : 1) Denaro *et al.*, 1981; 2) Brega *et al.*, 1986b; 3) Bonné-Tamir *et al.*, 1986; 4) Brega *et al.*, 1986a; 5) Wallace *et al.*, 1985

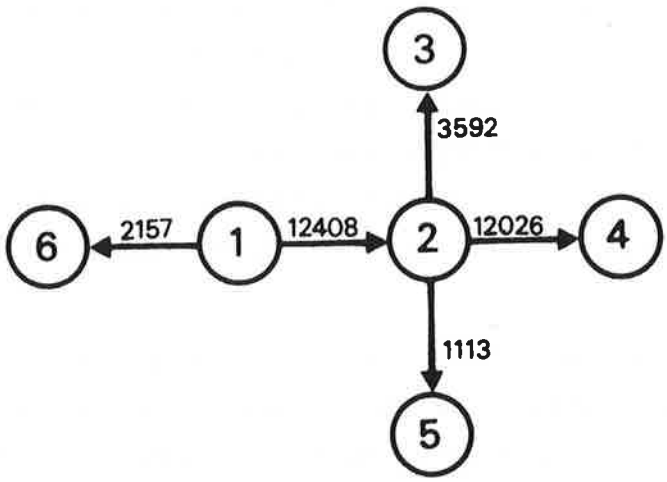

FIGURE 4.1 : Différenciation des morphes *Hpa I*. Le sens des flèches indique un gain de site.

L'examen de la Table 4.5 nous montre que le morphe 2 est retrouvé chez toutes les populations testées. Il est très courant chez les populations européennes, moyen-orientales ou orientales alors que dans les populations africaines, c'est le morphe 3 qui est largement prédominant. Ce dernier est retrouvé parmi les Arabes Israéliens avec une fréquence non négligeable (12,5 %) ainsi que parmi les Sardes, mais chez un seul individu. Le morphe 3, tout d'abord considéré comme un marqueur exclusivement africain (Denaro *et al.*, 1981) se retrouve donc également dans le bassin méditerranéen.

Le morphe 1 est trouvé principalement chez des populations d'origine orientale ainsi qu'assez curieusement chez des Bantous de Johannesburg. Cette présence dans 2 groupes continentaux suggère une origine ancienne de ce morphe. Cependant deux autres causes peuvent expliquer la présence du morphe 1 dans l'échantillon : une mutation indépendante ou un artefact de l'échantillonnage d'individus ayant une ascendance mixte (orientale et africaine). Ce morphe semble être dérivé du morphe 2 par une unique mutation (Figure 4.1). Une étude interspécifique de Ferris *et al.* (1981) a montré que l'on trouvait également ce morphe chez l'Orang-Outan alors que les individus testés chez le Gorille, le Chimpanzé ou le Gibbon présentaient des morphes différents de ceux de l'espèce humaine. La localisation géographique actuelle de l'Orang-Outan et le fait que ce morphe soit présent chez des populations d'origine asiatique a fait germer l'hypothèse d'une origine de l'homme en Asie (Denaro *et al.*, 1981), avec l'idée sous-jacente d'une liaison directe entre la lignée humaine et l'Orang-Outan. Cette hypothèse ne semble pas être confirmée par d'autres études de l'ADN-mt (Brown *et al.*, 1982; Hixson and Brown, 1986; Nei *et al.*, 1985), bien qu'elle ait été soutenue encore récemment à l'aide d'arguments morphologiques (Schwartz, 1984). L'occurrence de morphes similaires dans des espèces voisines n'implique pas forcément une origine commune, mais elle peut être due à des phénomènes de convergence dont l'importance est aujourd'hui débattue (Nei and Tajima, 1985, 1987; Templeton, 1983a, 1983b, 1987) (voir Figure 4.9).

D'autre part, dans le cas d'allèles ou d'allotypes neutres, la probabilité qu'un allèle soit le plus ancien dans un échantillon est égal à sa fréquence (Watterson and Guess, 1977). Ceci laisse penser qu'en l'absence de forces sélectives ou d'effets fondateurs importants, le morphe 2 serait vraisemblablement le plus ancien. Cette éventualité est renforcée par l'étude des liens existant entre les morphes au niveau moléculaire (Figure 4.1). Le morphe 2 pourrait ainsi être à l'origine des morphes 1, 3, 4 et 5. Le morphe 6 découlerait lui du 1, ce qui est en accord avec le fait qu'il est trouvé chez les Orientaux uniquement, où le morphe 1 atteint une fréquence de 12,5 %.

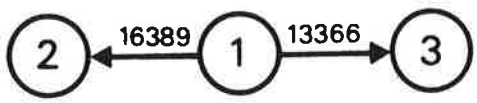

FIGURE 4.2 : Différenciation des morphes *Bam HI*. Le sens des flèches indique un gain de site.

Il est intéressant de constater que l'analyse de ces données, dans le cadre d'une hypothèse de neutralité sélective et de différenciation des populations par dérive génétique, est consistante avec l'hypothèse d'une population humaine originelle proche des populations européennes ou asiatiques actuelles, mais vraisemblablement pas des populations africaines.

Bam HI

Trois morphes ont été définis par Johnson *et al.* (1983) (Table 4.6) et retrouvés dans les populations d'origine européenne et méditerranéenne (Table 4.7). Au niveau moléculaire, le morphe 1 pourrait être à l'origine des deux autres morphes par un gain de site (55 et 68) dans les deux cas (Figure 4.2). Sa fréquence élevée dans toutes les populations atteste également de sa probable ancienneté. Les morphes 2 et 3 pourraient être absents des populations orientales et africaines, car on ne le retrouve pas dans les divers échantillons examinés provenant de ces régions. Toutefois, les effectifs des échantilons testés ne permettent évidemment pas d'affirmer l'absence d'un morphe donné.

TABLE 4.6 : Codification des morphes *Bam HI*

Morphes	Sites[1] 55	60	68
1	0	1	0
2	0	1	1
3	1	1	0

[1] La notation des sites correspond à celle qui est définie dans la Table 4.3

Le nombre limité de morphes et l'absence totale de polymorphisme dans les populations non-caucasoïdes limite quelque peu l'importance de l'emploi de cet enzyme dans des études de l'histoire du peuplement au niveau mondial. Il pourrait néanmoins s'avérer utile afin de différencier des populations d'origine européenne. Le morphe 3 semble être relativement fréquent en Italie (9,9 %) et en Sardaigne (15,5%) alors que le morphe 2 y est soit absent, soit rare (2 individus à Rome). Inversément, l'échantillon de Caucasoïdes présente un plus grand nombre d'individus portant le morphe 2 que le 3. Les individus de cet échantillons provenant en partie des Etats-Unis et donc

anciennement de divers pays d'Europe, cela suggère que le morphe 2 pourrait être encore plus fréquent dans des pays situés au Nord de l'Italie qui restent à déterminer.

TABLE 4.7 : Fréquence (%) des morphes *Bam HI* dans diverses populations humaines.

Morphes	Populations									
	Cauc.[1] 50	Rom[2]. 95	Sard.[2] 162	Isr. J.[3] 40	Isr. A.[3] 41	Bant.[1] 40	San[1] 34	Orient.[1] 46	Tharu[4] 91	Am.[5] 74
1	86,6	82,1	90,1	97,5	100	100	100	100	100	100
2	8,0	2,1	0	2,5	0	0	0	0	0	0
3	6,0	15,8	9,9	0	0	0	0	0	0	0

Sources : 1) Johnson *et al.*, 1983; 2) Brega *et al.*, 1986b; 3) Bonné-Tamir *et al.*, 1986; 4) Brega *et al.*, 1986a; 5) Wallace *et al.*, 1985

D'autre part, le fait que ce polymorphisme de longueur des fragments de restriction *Bam HI* ne se retrouve qu'en Europe peut sembler frappant. Ceci peut bien évidemment être dû à un phénomène de hasard historique ou à un problème d'échantillonnage. Néanmoins, jusqu'à preuve du contraire, ces polymorphismes pourraient constituer des marqueurs intéressant pour détecter d'éventuelles origines occidentales dans certaines populations.

Hae II

Les morphes 1 à 7 ont été définis par Johnson *et al.* (1983). Le morphe 8 a été trouvé par Bonné-Tamir *et al.* (1986) et le 9 par Brega *et al.* (1986b) qui l'avaient alors dénommé morphe "$11^{Sardinia}$". Cette autre appellation est due à une non-standardisation des numéros des morphes. Par souci de simplification, nous avons renuméroté par ordre croissant et continu les 9 morphes découverts dans les 10 populations que nous avons considérées. Par ailleurs, signalons que Horai *et al.* (1984) ont identifié 2 autres morphes correspondant aux numéros 10 et 11 de la Figure 4.3 et qui n'ont pas été retrouvés ailleurs.

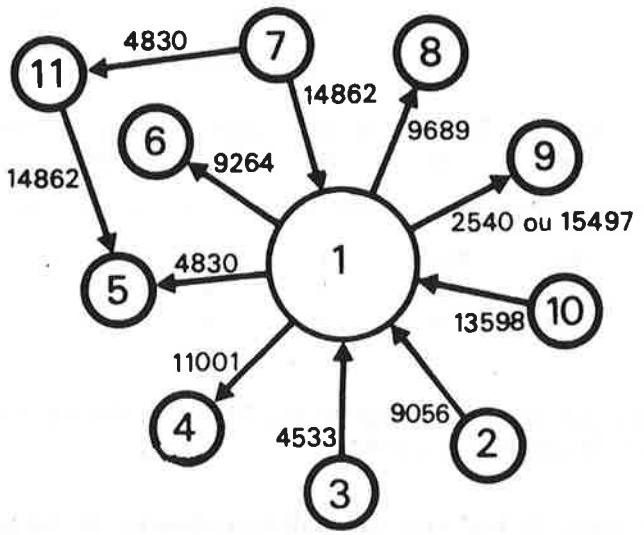

FIGURE 4.3 : Différenciation des morphes *Hae II*. Le sens des flèches indique un gain de site.

TABLE 4.8 : Codification des morphes *Hae II*

Morphes	Sites[1] 8	13	18	21	38	39	41	43	57	61	62	64
1	0	1	1	0	1	0	0	0	1	1	1	0
2	0	1	1	0	0	0	0	0	1	1	1	0
3	0	1	0	0	1	0	0	0	1	1	1	0
4	0	1	1	0	1	0	0	1	1	1	1	0
5	0	1	1	1	1	0	0	0	1	1	1	0
6	0	1	1	0	1	1	0	0	1	1	1	0
7	0	1	1	0	1	0	0	0	1	0	1	0
8	0	1	1	0	1	0	1	0	1	1	1	0
9	1	1	1	0	1	0	0	0	1	1	1	1

[1] La notation des sites correspond à celle qui est définie dans la Table 4.3

Le nouveau site responsable de la différenciation du morphe 9 par rapport au 1 n'a pas été localisé précisément par une double digestion enzymatique. Il s'ensuit que 2 sites potentiels sont possibles : le site 8 (pb 2540) ou le site 64 (pb 15497). Etant donné cette ambiguité, les 2 sites ont été reportés dans la Table 4.8. Des études ultérieures devront pouvoir préciser la position exacte de ce nouveau site polymorphe. Ceci ne revêt pas une très grande importance dans le cas présent, car aucun site défini par un autre enzyme n'est potentiellement chevauchant avec l'une des 2 positions possibles.

TABLE 4.9 : Fréquence (%) des morphes *Hae II* dans diverses populations humaines.

Morphes	Populations Cauc.[1] 50	Rom[2]. 96	Sard.[2] 134	Isr. J.[3] 40	Isr. A.[3] 41	Bant.[1] 40	San[1] 34	Orient.[1] 46	Tharu[4] 91	Am.[5] 74
1	77,0	88,5	91,0	57,5	95,1	97,5	100	80,4	69,2	92,0
2	14,0	7,3	4,5	37,5	2,4	2,5	0	13,0	5,5	2,7
3	6,0	2,1	0	2,5	0	0	0	0	0	4,1
4	0	0	1,5	2,5	0	0	0	2,2	0	1,4
5	0	2,1	1,5	0	0	0	0	4,3	25,3	0
6	2,0	0	0	0	0	0	0	0	0	0
7	2,0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	2,4	0	0	0	0	0
9	0	0	1,5	0	0	0	0	0	0	0

Sources : 1) Johnson *et al.*, 1983; 2) Brega *et al.*, 1986b; 3) Bonné-Tamir *et al.*, 1986; 4) Brega *et al.*, 1986a; 5) Wallace *et al.*, 1985

L'examen de la Table 4.9 nous montre la prédominance du morphe 1 dans toutes les populations étudiées. Le morphe 2 semble également important et atteint des fréquences supérieures à 10 % chez les Israéliens Juifs, les Caucasoïdes et les Orientaux.

Il semble difficile de considérer un morphe particulier comme marqueur d'un grand groupe continental. Le morphe 3 pourrait être plutôt considéré comme propre au groupe d'origine européenne s'il n'était également retrouvé chez les Nord-Amérindiens. Le morphe 5 possède une fréquence importante chez les Tharu népalais (25,3 %), mais on l'observe aussi parmi les échantillons Romains et Sardes. Les morphes 6 à 9 n'ont été identifiés que chez 1 ou 2 individus d'un échantillon donné. Ils sont par conséquent considérés comme rares et n'interviennent pas dans les distinctions inter-populations. Les morphes 2, 3, 4 et 5 se retrouvent à la fois dans certaines populations orientales et occidentales, ce qui pourrait laisser supposer qu'ils étaient présents chez ces 2 groupes avant leur séparation. Les hypothèses alternatives d'une apparition postérieure par des mutations convergentes ou des échanges de gènes par migration ne sont pas à négliger et pourraient être tranchées par l'examen des haplotypes. Une autre difficulté provient du fait que les sites *Hae II* ne sont pas déterminés de façon unique, mais que l'enzyme peut reconnaître comme équivalentes 4 séquences différentes. Cette ambiguité sur la nature moléculaire exacte des morphes nous empêche pour l'instant d'avoir un discours trop précis à propos des différences inter-populations sur la seule base des fréquences des morphes *Hae II* trouvés.

Selon les liens trouvés entre les différents morphes (Figure 4.3) et leurs fréquences observées, le morphe 1 serait à l'origine de tous les autres morphes, excepté le morphe 11, trouvé dans une population japonaise (Horai *et al.*, 1984), qui est plus probablement issu du morphe 5. Cette conclusion se base sur le fait que ce dernier possède une fréquence relativement élevée dans une autre population orientale (Tharu), et qu'il est également présent dans l'échantillon de japonais. Il est en effet logique qu'une mutation ait plus de chance de se produire sur un morphe fréquent (>5%) que sur un morphe rare (<5%). Une phylogénie des morphes se basant sur ce postulat aura donc tendance à privilégier les liens unissant un morphe fréquent à un autre morphe fréquent ou non, plutôt qu'un lien entre deux morphes rares.

Msp I

Cet enzyme n'a, en fait, été utilisé que pour l'étude de 9 des 10 populations répertoriées ici. L'échantillon d'Amérindiens a en effet été analysé par *Hpa II*, un enzyme reconnaissant la même séquence que *Msp I* (CCGG), mais qui est sensible à la méthylation des cytosines dans l'ADN. C'est à dire que *Msp I* reconnaîtra un site, que celui-ci ait une nucléotide méthylée ou non, alors que *Hpa II* ne le fera qu'en l'absence de toute méthylation. Apparamment, le polymorphisme des sites de restriction ne semble pas être dû à une méthylation différentielle chez les mammifères (Groot and Coon, 1979; Castora *et al.*, 1980) et chez l'homme (Johnson *et al.*, 1983). Ceci paraît justifier le fait que nous puissions regrouper les résultats de ces 2 isoschizomères.

Johnson *et al.* (1983) ont trouvé 5 types de morphes différents. Un doute subsistait quant à la nature de leur morphe 1 qui est devenu le morphe 6 selon notre propre nomenclature. En quelques mots, nous allons tenter de justifier ce changement d'appellation. En consultant la Figure 4.4, on remarque qu'il faut 2 mutations pour passer du morphe 1 au morphe 2 (disparition des sites 35 (8112) et 36 (8150)). Le morphe 1 correspond à la séquence de référence d'Anderson *et al.* (1981). Le morphe intermédiaire, nécessaire pour passer du morphe 1 au 2, n'a en fait pas été retrouvé par Johnson et ses collaborateurs dans leurs échantillons. Les 2 sites incriminés par cette apparente double mutation étant très proches et la résolution de leur gel d'électrophorèse ne permettant pas de trancher entre la présence d'un des deux sites seulement (morphe 6) ou la présence des deux à la fois (morphe 1), Johnson *et al.* (1983) ont finalement postulé que le morphe correspondant à la séquence de référence n'avait surement pas été observé. Plus récemment, Bonné-Tamir *et al.* (1986) ont découvert, dans un échantillon d'Israéliens Arabes, un morphe où le site 35 (8112) était interprété comme absent. Après réexamen de leur données, nous arrivons à la conclusion, et en accord avec Horai et Matsunaga (1986) qui l'ont également observé, qu'il s'agit plutôt de la perte du site 36 (8150) qui conduit à la fusion des 2 fragments de 1142 et 38 pb en un seul fragment de 1180 pb. D'autre part, les travaux de Cann (1982) et d'Horai et Matsunaga (1986), effectué avec l'enzyme *Hpa II*, ont confirmé le fait que le morphe 1 possède bien les sites 35 et 36.

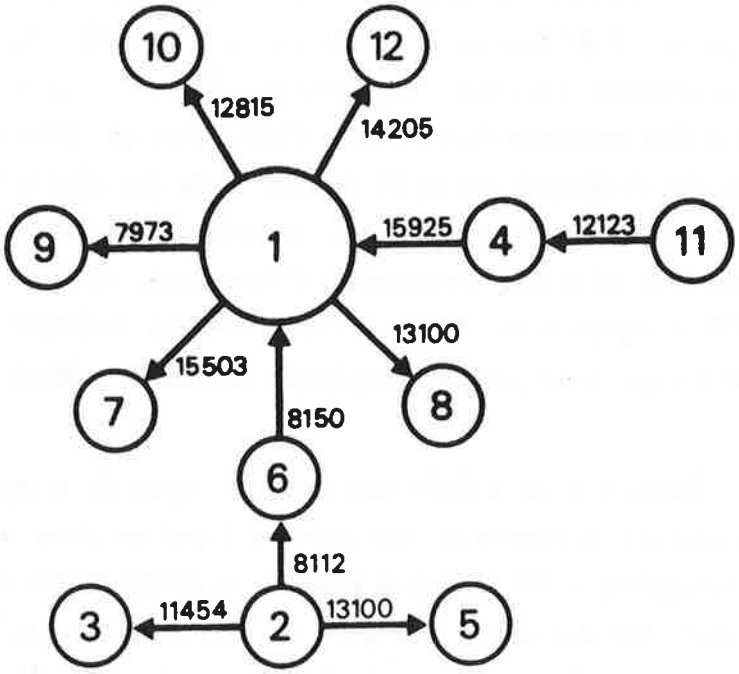

FIGURE 4.4 : Différenciation des morphes *Msp I*. Le sens des flèches indique un gain de site.

TABLE 4.10 : Codification des morphes *Msp I*

Morphes	1..32[2]	Sites[1] 34	35	36	40	44	45	47	51	52	54	58	59	65	67	70
1	1	0	1	1	1	0	1	1	0	0	1	1	0	0	1	1
2	1	0	0	0	1	0	1	1	0	0	1	1	0	0	1	1
3	1	0	0	0	1	1	1	1	0	0	1	1	0	0	1	1
4	1	0	1	1	1	0	1	1	0	0	1	1	0	0	0	1
5	1	0	0	0	1	0	1	1	0	1	1	1	0	0	1	1
6	1	0	1	0	1	0	1	1	0	0	1	1	0	0	1	1
7	1	0	1	1	1	0	1	1	0	0	1	1	0	1	1	1
8	1	0	1	1	1	0	1	1	0	1	1	1	0	0	1	1
9	1	1	1	1	1	0	1	1	0	0	1	1	0	0	1	1
10	1	0	1	1	1	0	1	1	1	0	1	1	0	0	1	1
11	1	0	1	1	1	0	1	0	0	0	1	1	0	0	0	1
12	1	0	1	1	1	0	1	1	0	0	1	1	1	0	1	1

[1] La notation des sites correspond à celle qui est définie dans la Table 4.3
[2] Les sites 1, 3, 11, 14, 19, 22, 23, 24, 25, 26, 27, 28, 29, 30,31 et 32 sont monomorphes

Les morphes 7 et 8 ont été découvert par Bonné-Tamir *et al.* (1986) dans un échantillon d'Israéliens Arabes. Les morphes 9 à 12, selon notre nomenclature, ont été trouvé par Brega *et al.* (1986a) dans l'échantillon népalais. Cette liste de morphes ne semble pas exhaustive, car 9 autres morphes dérivés du morphe 1 ont été décrits par Cann (1982) (morphes 17 à 21) et par Horai et Matsunaga (1986) (morphes 13 à 16). Les liens existants entre les 12 premiers morphes sont présentés dans la Figure 4.4.

Le morphe 1 est nettement prédominant dans toutes les populations testées, excepté chez les San où il s'agit du morphe 2 (Table 4.11). Dans les populations occidentales et orientales, le morphe 4 est retrouvé avec des fréquences non négligeables, surtout chez les populations européennes. Les morphes 5 à 11 ne sont retrouvés que chez un seul individu d'une des populations testées. Le morphe 12, quant à lui, a été retrouvé chez 3 individus de la population Tharu, mais ne peut être considéré comme un marqueur important.

TABLE 4.11: Fréquence (%) des morphes *Msp I* dans diverses populations humaines.

Morphes	Populations									
	Cauc.[1]	Rom[2].	Sard.[2]	Isr. J.[3]	Isr. A.[3]	Bant.[1]	San[1]	Orient.[1]	Tharu[4]	Am.[5]
	50	95	136	40	40	39	34	46	91	74
1	92,0	84,2	88,2	100	89,7	87,5	17,6	97,8	90,1	100
2	0	0	0	0	0	12,5	52,9	0	0	0
3	0	0	0	0	0	0	26,5	0	0	0
4	8,0	14,7	11,8	0	2,6	0	0	2,2	3,3	0
5	0	0	0	0	0	0	2,9	0	0	0
6	0	0	0	0	2,6	0	0	0	0	0
7	0	0	0	0	2,6	0	0	0	0	0
8	0	1,1	0	0	2,6	0	0	0	0	0
9	0	0	0	0	0	0	0	0	1,1	0
10	0	0	0	0	0	0	0	0	1,1	0
11	0	0	0	0	0	0	0	0	1,1	0
12	0	0	0	0	0	0	0	0	3,3	0

Sources : 1) Johnson *et al.*, 1983; 2) Brega *et al.*, 1986b; 3) Bonné-Tamir *et al.*, 1986; 4) Brega *et al.*, 1986a; 5) Wallace *et al.*, 1985

Les populations africaines testées se distinguent des autres groupes continentaux par la présence de morphes où les sites 35 et 36 sont absents, selon Johnson *et al.* (1983). A ce propos, il est intéressant de parler d'une autre controverse existant autour de ces sites. Les travaux de Cann (1982) et de Cann *et al.* (1984 et 1987) montrent que ces 2 sites semblent être présents chez la plupart des individus d'origine africaine (Noirs américains), ce qui contredirait les résultats obtenus par Johnson *et al.* (1983). Sans analyser d'autres échantillons, il ne nous est pas possible de trancher définitivement sur cette divergence. Néanmoins, une hypothèse conciliant les 2 résultats peut être envisagée. Il faut bien réaliser tout d'abord que les échantillons d'individus d'origine africaine de Cann (1982) et de Johnson *et al.* (1983) sont différents par leur homogénéité, d'une part, mais surtout par leur origine. En effet, 18 Noirs américains, 1 San et 1 individu provenant du Nigéria, analysés par Cann (1982), ne sauraient prétendre être représentatifs du même groupe de populations que des échantillons de Bantous de Johannesburg ou de San qui ont été moins impliqués dans les traites d'esclaves que des populations de la côte Atlantique de l'Afrique. Nous avons donc affaire à au moins 3 stocks génétiques potentiellement différents. Le premier stock (Noirs américains), bien qu'hétérogène dans sa constitution, ne présente pas d'individus où l'on constate l'absence des sites 34 et 35; le second (San) comprend une majorité d'individus où ils sont absents; le troisième (Bantou), quant à lui, comprend une minorité d'individus (12,5%) qui ont également perdu ces deux sites. Dans un précédent travail (Excoffier *et al.*, 1987), nous avons postulé que des populations bantoues de la

côte est du sud de l'Afrique avaient visiblement assimilés, dans un passé relativement récent (moins de mille ans), le gène $Gm^{1,17;;13,15}$ de populations Khoisan établies dans cette région. Par un phénomène similaire, le morphe *Msp I* 2 aurait pu être introduit dans les populations dont sont issus les Bantous de Johannesburg. Dans cette optique, l'absence des sites 34 et 35 serait plutôt une caractéristique des San qui ne serait pas partagée par les autres populations africaines. Sans parler de phénomènes de sélection, on peut imaginer que les San aient été l'objet d'un effet fondateur qui aurait bouleversé les fréquences géniques de certains marqueurs (Excoffier *et al.*, 1988).

De ce fait, on a une nouvelle preuve que le choix des populations représentatives d'un continent est extrêmement délicat. Dans ces conditions, les particularités apparentes de l'ensemble du "groupe africain" pour les fréquences des morphes *Msp I* sont à relativiser, en ayant conscience qu'elles peuvent être dues à l'octroi d'une trop grande importance à un isolat non représentatif de l'ensemble de la diversité des populations africaines.

Le schéma de différenciation des morphes de la Figure 4.4, ainsi que leurs fréquences tendent à montrer que le morphe 1 doit être relativement ancien pour être à la source d'un grand nombre d'autres morphes. Les morphes 2 et 4 qui atteignent des fréquences parfois importantes, et dont sont issus d'autres morphes, sont également candidats à une certaine ancienneté. En l'absence d'une phylogénie polarisée des types, il n'est pas possible de savoir si ces 3 morphes ont été présents simultanément dans une population originelle (donc à ce moment polymorphe), ou s'ils sont apparus progressivement au cours de l'évolution humaine.

Il est intéressant de noter que le morphe 6, assurant le lien moléculaire entre les morphes 1 et 2, a été retrouvé dans une population du Proche-Orient, suggérant peut-être que cette région ait pu conserver les traces de la différenciation du morphe 2 retrouvé principalement chez les San.

La diversité des morphes *Msp I* est telle qu'elle nous permet d'observer que certains sites ont pu être l'objet de mutations parallèles ou convergentes, et cela sans devoir recourir à l'examen des haplotypes. En comparant les Figures 4.4 et 4.23, on remarque que le site localisé aux alentours de la position des pb 11454 aurait été acquis par le morphe 3 et le morphe 15. Si Horai et Matsunaga (1986) ont identifiés le gain de ce site sans ambiguité sur le morphe 5, Johnson *et al.* (1983) ont été plus prudents en précisant que cette région du génome mitochondrial contenait plusieurs sites potentiels

(situés aux pb 11436, 11454 et 11475) et qu'il était difficile de localiser le site responsable de la formation du morphe 3 par rapport au 2.

Un problème analogue se pose pour le site localisé approximativement aux pb 13100, et qui apparaît chez les morphes 5 et 8. Ce dernier morphe, trouvé par Bonné-Tamir *et al.* (1986) chez un Israélien Arabe a également été observé chez un américain d'origine européenne (Cann, 1982) et chez un Japonais (Horai and Matsunaga, 1986). La localisation du site responsable de la création du morphe 5 à partir du morphe 2 a été située aux alentours des pb 13070 par Johnson *et al.* (1983). Il existe en fait 4 sites potentiels autour de cette position (13028, 13060, 13100 et 13119). De ce fait, il semble difficile de définir lequel est concerné, étant donné la faible résolution de la taille des fragments obtenus par électrophorèse. Nous avons néanmoins choisi d'adopter le même site que celui qui est responsable de la formation du morphe 8 pour minimiser le nombre de sites impliqués dans les passages entre morphes. En fait, que le même site soit responsable de la création de ces morphes est secondaire, si l'on reconnait que ces 2 mutations sont des évènements indépendants. Il en découle que 4 substitutions sont nécessaires pour passer du morphe 8 au morphe 5, et non pas 2 comme calculé en se contentant de compter les différences de sites de la Table 4.10 (Une procédure similaire s'applique au calcul des différences entre les morphes 3 et 15).

Ava II

Cet enzyme a généré un nombre important de morphes différents, d'un ordre de grandeur comparable à certains loci HLA. Cette extraordinaire diversité rend son étude complexe, mais pourrait en faire un objet très précieux pour la comparaison de populations.

Les morphes numérotés de 1 à 11 ont été définis par Johnson *et al.* (1983), non sans quelques ambiguités dans la localisation exacte des sites. Nous nous trouvons en effet confrontés aux même types de problèmes qu'avec *Msp I*, à savoir que certains sites dont la localisation est imprécise semblent impliqués plusieurs fois dans des passages entre morphes (voir Figure 4.5).

Les morphes 12 à 15 ont été découverts par Bonné-Tamir *et al.* (1986) et correspondent à leur propre numérotation. Cependant, une étude de leurs données nous a conduit à proposer des relations quelque peu différentes entre deux de ces nouveaux morphes et les 11 autres décrits par Johnson *et al.* (1983). Le morphe 12 est en effet

dérivé du morphe 3 par la perte du site 50 (pb 12629), et non du morphe 1 comme ces auteurs l'avaient interprété. Le morphe 13 n'est pas dérivé du morphe 12, mais plutôt du morphe 9, également par la perte du site 50 (voir Figure 4.5).

Les morphes 16 et 17 ont été décrits par Brega *et al.* (1986a) et correspondent respectivement à leurs morphes 17^{Tharu} et 18^{Tharu}. Ce décalage dans la nomenclature est dû à la comptabilisation d'un nouveau morphe trouvé par Horai et Matsunaga (1986), dont nous n'avons pas tenu compte à ce niveau de la discussion, mais qui sera répertorié ultérieurement. Ces 2 morphes dérivent du morphe 1 par une unique substitution.

Les morphes 18 à 22 ont été découverts par Brega *et al.* (1986b), chacun chez un seul individu, et correspondent respectivement aux morphes 19^{Italy} et 20^{Italy} et aux morphes $21^{Sardinia}$ et $22^{Sardinia}$. Leurs relations avec les autres morphes n'ayant pas été établies par ces auteurs, nous allons proposer un schéma de différenciation de ces morphes compatible avec la taille des fragments de restriction observée sur leurs gels d'électrophorèse.

Le morphe 18 semble dériver du morphe 5 par le gain d'un site (12 ou 33), qui coupe un fragment de 5,4 Kilobases (Kb) du morphe 5 en 2 fragments de 5 et 0,4 Kb. Une double digestion aurait vraisemblablement permis de préciser quel site était impliqué. Le morphe 19 peut dériver des morphes 15 ou 6, qui sont tout deux présents dans le même échantillon de Romains. Dans les deux cas, cela provoquerait une mutation parallèle du site impliqué. Le morphe 20, quant à lui, pourrait découler du morphe 15 par le gain du site 37 ou d'un site voisin, ou encore du morphe 2 par le gain du site 17. L'échantillon Sarde ne comprenant pas le morphe 15, mais le 2, il serait tentant de privilégier le site 17 pour la responsabilité de l'apparition du morphe 20 dans cet échantillon. Néanmoins, vu la faible fréquence du morphe 15 dans l'échantillon Sarde, son absence de l'échantillon Romain peut résulter d'un simple processus aléatoire d'échantillonnage et ne permet pas d'être certain de son absence dans la population. Le morphe 21 semble être dérivé du morphe 13 par le gain du site 37 (à nouveau) ou d'un site proche qui scinde un fragment de 13,5 Kb du morphe 13 en 2 fragments de 8,1 et 5,4 Kb (Brega *et al.*, 1986b).

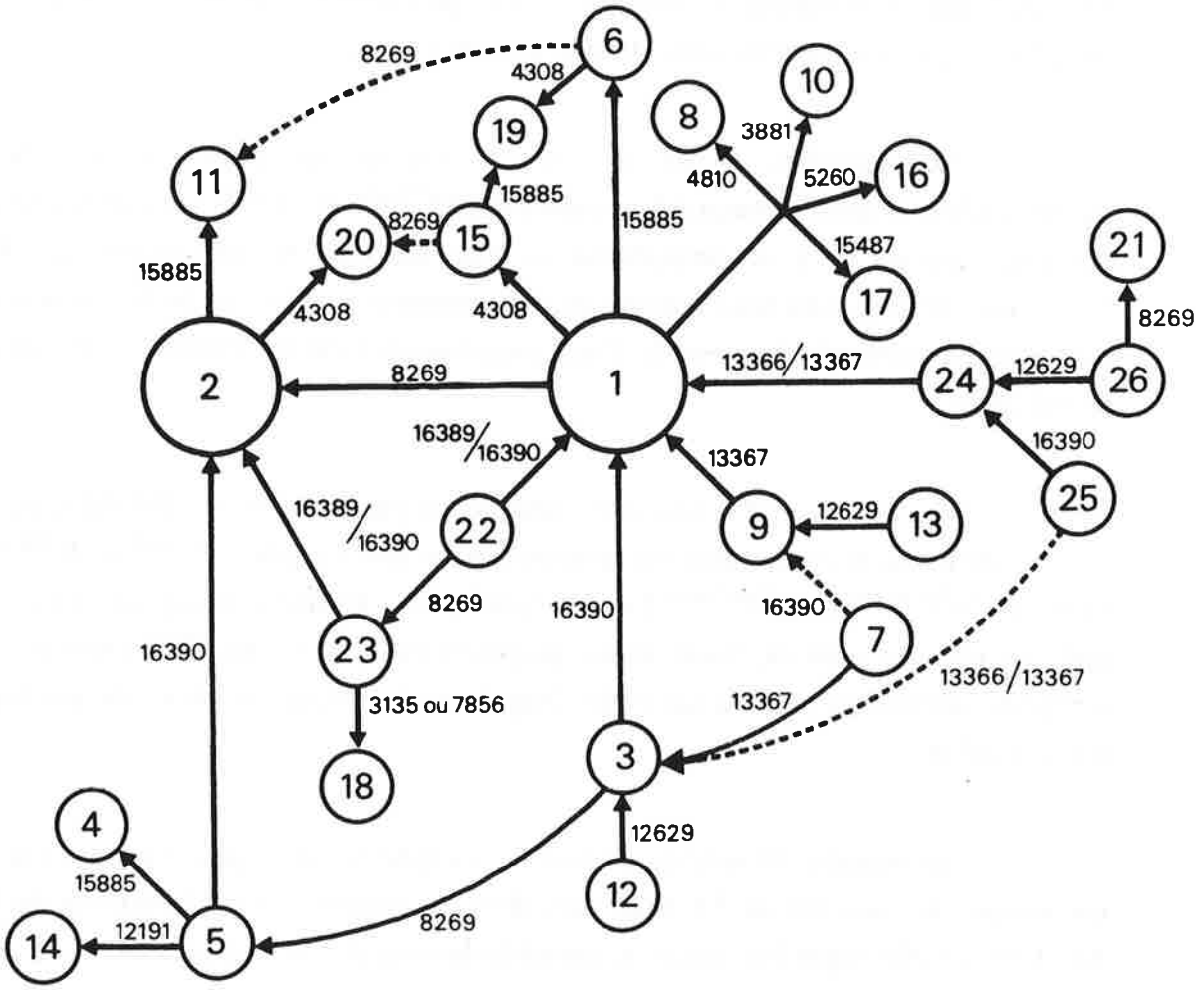

FIGURE 4.5 : Différenciation des morphes *Ava II*. Le sens des flèches indique un gain de site.

TABLE 4.12 : Codification des morphes *Ava II*

Morphes	2..10[2]	Sites[1] 12	16	17	20	24	33	37	48	50	5655/56	63	66	6968/69	
1	1	0	0	0	0	0	0	0	0	1	1	1	0	0	1	1
2	1	0	0	0	0	0	0	1	0	1	1	1	0	0	1	1
3	1	0	0	0	0	0	0	0	0	1	1	1	0	0	0	1
4	1	0	0	0	0	0	0	1	0	1	1	1	0	1	0	1
5	1	0	0	0	0	0	0	1	0	1	1	1	0	0	0	1
6	1	0	0	0	0	0	0	0	0	1	1	1	0	1	1	1
7	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1
8	1	0	0	0	1	0	0	0	0	1	1	1	0	0	1	1
9	1	0	0	0	0	0	0	0	0	1	0	1	0	0	1	1
10	1	0	1	0	0	0	0	0	0	1	1	1	0	0	1	1
11	1	0	0	0	0	0	0	1	0	1	1	1	0	1	1	1
12	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1
13	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1
14	1	0	0	0	0	0	0	1	1	1	1	1	0	0	0	1
15	1	0	0	1	0	0	0	0	0	1	1	1	0	0	1	1
16	1	0	0	0	0	1	0	0	0	1	1	1	0	0	1	1
17	1	0	0	0	0	0	0	0	0	1	1	1	1	0	1	1
18	1	1	0	0	0	0	1	1	0	1	1	1	0	0	1	0
19	1	0	0	1	0	0	0	0	0	1	1	1	0	1	1	1
20	1	0	0	1	0	0	0	1	0	1	1	1	0	0	1	1
21	1	0	0	0	0	0	0	1	0	0	1	0	0	0	1	1
22	1	0	0	0	0	0	0	0	0	1	1	1	0	0	1	0
23	1	0	0	0	0	0	0	1	0	1	1	1	0	0	1	0
24	1	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1
25	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1
26	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1

[1] La notation des sites correspond à celle qui est définie dans la Table 4.3
[2] Les sites 2, 5, 7, 9 et 10 sont monomorphes

Nous avons vu que 2 sites *Ava II* étaient potentiellement chevauchant avec des sites *Bam HI* et qu'une seule substitution pouvait induire la modification d'un morphe pour les 2 enzymes. En anticipant quelque peu sur la description des haplotypes, nous allons distinguer, pour ces 2 sites, 2 morphes *Ava II* différents selon que la substitution aura également provoqué un changement de morphe *Bam HI* ou non. Les 4 cas possibles seront : 1) la disparition du site 6 accompagnée de l'apparition du site 55 dans un morphe donné (noté $55^+/56^-$); 2) la disparition du site 56 sans modification du site 55 ($-/56^-$); 3) la perte du site 69, accompagnée du gain du site 68 ($68^+/69^-$); 4) la seule perte du site 69 ($-/69^-$).

L'apparition du site *Bam HI* 68 est toujours accompagnée de la disparition du site *Ava II* 69 dans les types 11, 23 et 58 définis dans la Table 4.14, alors que de nombreux autres types ont perdu le site 69 sans qu'il y ait création de site *Bam HI*. Il nous semble justifié de créer 2 nouveaux morphes *Ava II*. Ainsi, le morphe 22 a subi l'événement $68^+/69^-$ et sera l'homologue du morphe 3 qui n'a subi que la perte du site 69. De même, le morphe 23 sera l'homologue du morphe 5. Il est à noter que le morphe 18 a également été l'objet de ce double événement apparent. Il découle donc simplement du morphe 23.

Les haplotypes 18, 19, 20, 57 et 61 ont subi le double événement $55^+/56^-$. Ceci nous amène à constater que la perte du site *Bam HI* 55 est toujours liée à la perte du site 56, alors que l'inverse n'est pas vérifié pour les types 17, 35 et 36 qui ont subi l'événement simple $-/56^-$. Cela nous oblige à considérer 3 nouveaux morphes *Ava II* : le morphe 24 qui a subi l'événement $55^+/56^-$ et qui est l'homologue du morphe 9; le morphe 25 qui est l'homologue du morphe 7; le morphe 26 qui est l'homologue du morphe 13. Les relations au niveau moléculaire entre ces morphes sont décrites par la Figure 4.5. On remarque aussi que le morphe 21 découle du morphe 26 et non du 13. Nous examinerons plus loin les implications de ces liens sur les relations entre haplotypes.

Ces distinctions sur la nature moléculaire des substitutions nous ont conduits à recalculer les fréquences des morphes dans les différentes populations, pour les cas où les haplotypes concernés par les apparentes doubles mutations y étaient présents.

Afin d'être plus complet, nous devons ajouter que 7 autres morphes Ava II ont été reconnus dans d'autres populations. Ainsi, Cann (1982) a identifié 3 nouveaux morphes dérivant du morphe 1 et se caractérisant par des gains de sites aux pb 5984, 8722 et 10933. D'autre part, Horai et Matsunaga (1986) ont trouvé 3 autres morphes également dérivés du morphe 1. Le premier se différencie par la perte du site 50, le second par le gain d'un site débutant aux pb 16503, et le troisième par l'apparition du site 37 associée à la perte d'un site d'un autre enzyme (voir Figure 4.22). Enfin, un dernier morphe a été identifié par Harihara *et al.* (1986), qui découle du morphe 3 par le gain d'un site aux alentours des pb 6384 (voir Figure 4.31). Ceci suggère que bon nombre d'autres morphes restent à découvrir pour cet enzyme, ce qui sera probablement fait lorsque la taille des échantillons permettra d'identifier la majorité des allèles présents dans les populations.

TABLE 4.13: Fréquence (%) des morphes *Ava II* dans diverses populations humaines.

Morphes	Populations									
	Cauc.[1] 50	Rom[2]. 95	Sard.[2] 134	Isr. J.[3] 38	Isr. A.[3] 39	Bant.[1] 40	San[1] 34	Orient.[1] 46	Tharu[4] 91	Am.[5] 74
1	74,0	73,7	85,1	76,3	69,2	40,0	11,8	95,7	93,4	100
2	4,0	3,2	1,4	0	0	12,5	58,8	0	0	0
3	0	1,1	0,7	0	5,1	37,5	2,9	0	2,2	0
4	0	0	0	0	0	0	5,9	0	0	0
5	2,0	1,1	0,7	10,6	20,5	5,0	20,6	0	0	0
6	0	1,1	0	0	0	0	0	4,3	0	0
7	0	0	0	0	0	2,5	0	0	0	0
8	2,0	0	0	0	0	0	0	0	0	0
9	2,0	0	0	2,6	0	0	0	0	0	0
10	2,0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	2,5	0	0	0	0
12	0	0	0	0	2,6	0	0	0	0	0
13	0	0	0	2,6	2,6	0	0	0	0	0
14	0	0	0	2,6	0	0	0	0	0	0
15	0	2,1	0	2,6	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	3,3	0
17	0	0	0	0	0	0	0	0	1,1	0
18	0	1,1	0	0	0	0	0	0	0	0
19	0	1,1	0	0	0	0	0	0	0	0
20	0	0	0,7	0	0	0	0	0	0	0
21	0	0	0,7	0	0	0	0	0	0	0
22	2,0	0	0	0	0	0	0	0	0	0
23	6,0	1,1	0	2,6	0	0	0	0	0	0
24	4,0	13,7	7,5	0	0	0	0	0	0	0
25	2,0	0	0	0	0	0	0	0	0	0
26	0	1,1	3,0	0	0	0	0	0	0	0

Sources : 1) Johnson *et al.*, 1983; 2) Brega *et al.*, 1986b; 3) Bonné-Tamir *et al.*, 1986; 4) Brega *et al.*, 1986a; 5) Wallace *et al.*, 1985

L'examen conjoint de la Table 4.13 et de la Figure 4.5 nous montre qu'il existe une bonne consistence entre le schéma de différenciation des morphes et leurs fréquences. On s'aperçoit que les morphes atteignant des fréquences élevées (2, 3, 5 et 24) sont tous situés à une substitution près du morphe 1, qui semble donc central et ancien dans la majorité des populations humaines. D'une manière générale, il semble exister un gradient de fréquence allant du morphe 1 aux morphes périphériques ayant accumulé plusieurs mutations. L'échantillon San représente une exception car le morphe 2 y est majoritaire, mais la règle reste valable si l'on remplace le morphe 1 par le morphe dont la fréquence est la plus élevée dans l'échantillon. Ce principe est observé dans l'échantillon Bantou avec l'axe des morphes 1 → 3 → 7 et 1 → 2 → 11, dans l'échantillon San avec l'axe 2 → 5 → 4, et dans l'échantillon Sarde avec l'axe 1 → 24 → 26 → 21.

D'une manière plus générale, les populations d'origine asiatique apparaissent moins polymorphes que les populations africaines et, surtout, que celles d'origine européenne. Ces dernières ont en effet accumulés un nombre important de mutations ayant conduit à la différenciation de nouveaux morphes. On peut dénombrer pas moins de 11 morphes différents dans l'échantillon Romain, ainsi que 8 chez les Sardes et 7 chez les Israéliens Juifs. Ceci contraste avec les 4 morphes recensés dans l'échantillon de Tharu parmi 91 individus. Les 2 populations africaines apparaissent comme intermédiaires de ce point de vue.

La proximité génétique des populations asiatiques et européennes est fortement due aux fréquences élevées du morphe 1 dans ces 2 groupes de populations. Les populations africaines possèdent 3 autres morphes qui ont des fréquences supérieures à 20 %, et qui sont retrouvés avec des fréquences non négligeables dans plusieurs populations européennes et moyen-orientales, mais pas dans les populations asiatiques (à l'exception de 2 individus possédant le morphe 3 chez les Tharu). Comme pour le cas des morphes *Msp I*, l'échantillon San tend à être particulier du point de vue de ses fréquences géniques. Il faut quand même remarquer qu'il ne présente qu'un morphe (le 4) qui ne soit pas retrouvé dans d'autres populations. Nous nous devons de réitérer ici les réserves formulées quant à sa représentativité de l'ensemble des populations africaines.

Le fait de retrouver des allèles (morphes) avec des fréquences peu élevées en commun entre plusieurs populations, bien que cela ait peu d'importance dans les calculs de distances génétiques classiques, est néanmoins un bon indicateur de relations anciennes ou présentes, lorsque l'on connaît suffisammant la nature moléculaire des allèles et de leurs différences pour pouvoir discerner entre isoaction et identité par ascendance. Dans le cas présent, nous nous trouvons limité quant au degré de précision atteint dans la caractérisation des allèles observés. L'observation d'un seul allèle dans 2 populations différentes n'est pas suffisante pour être sûr d'une parfaite identité, étant donné le réseau complexe de mutations liant les morphes (voir Figure 4.5). De ce fait, il faut tenir compte de l'ensemble des morphes présents dans un échantillon et des liens qui les unissent pour savoir si l'on a affaire à un même groupe de morphes ou non.

Nous avons représenté toutes les relations pouvant exister entre les 26 morphes *Ava II* dans la Figure 4.5. Le réseau ainsi formé tend à être extrêmement complexe, mais il n'est pas certain que toutes les voies possibles aient été empruntées. Sans essayer de créer un réseau correspondant au principe de maximum de parcimonie, qui minimiserait le nombre total de substitutions pour définir tous les morphes observés,

nous avons tenté de simplifier le réseau en notant en gras les liens réunissant 2 morphes dont les fréquences obéissent au principe de gradient à partir du morphe le plus fréquent dans une population quelconque.

De ce fait, certaines branches du réseau s'isolent des autres, ce qui a tendance à rallonger le nombre total de mutations nécessaires pour passer d'un morphe à un autre. Comme exemple, nous prendrons l'absence probable de relations entre les morphes 15 et 20, qui n'ont jamais été observés ensemble dans un échantillon, ou l'isolation supposée des branches 1 → 3 → 7 et 1 → 9 → 13, les morphes 7 et 9 ne figurant jamais simultanément dans un échantillon. Malgré tout, certaines difficultés subsistent concernant les relations entre les morphes 1, 2, 3 et 5, ainsi que la branche 1 → 22 → 23 → 18, où il impossible de définir un seul chemin évolutif.

A ce stade du raisonnement, on constate que certains sites ont pu être l'objet de substitutions récursives. Ainsi, le site 37 (pb 8269) a pu connaître 6 substitutions indépendantes, le site 69 (pb 16390) a pu faire l'objet de 5 substitutions indépendantes, le site 66 (pb 15885) a pu en subir 4, les sites 50 (pb 12629) et 56 (pb 13367) 3 chacun et le site 17 (pb 4308) 2. Si l'on songe que l'enzyme *Ava II* peut reconnaître 2 séquences différentes, il est probable que ces chiffres constituent une sous-estimation du nombre total de mutations parallèles.

DÉFINITION ET FRÉQUENCE DES TYPES D'ADN-MT

En combinant les différents morphes identifiés pour chaque enzyme, 61 types d'ADN-mt différents ont été mis en évidence. Ce nombre restreint de types ne correspond pas à une combinaison aléatoire des morphes des 5 enzymes. En effet, en admettant que les morphes enzymatiques soient indépendants, on pourrait s'attendre à observer 6 x 3 x 9 x 12 x 26 = 50'544 types possibles. Il existe donc des associations préférentielles entre les morphes de plusieurs enzymes qui reflètent simplement l'histoire évolutive des types d'ADN-mt en l'absence de recombinaison. Ce processus historique était déja suggéré au niveau des relations entre morphes d'un même enzyme, mais apparaît encore plus clairement au niveau des types.

Les 61 types d'ADN-mt sont définis dans la Table 4.14 et leurs fréquences dans les 10 populations sont reportées dans la Table 4.15 et la Figure 4.6. La numérotation de la Table 4.14 correspond à celle qui a été élaborée par les auteurs qui ont découvert les différents types. Elle reflète uniquement la chronologie de leur définition, et aucune échelle de valeur ne saurait lui être associée. Le type 55 (2-1-1-1-3) défini par Brega *et al.* (1986b) est en fait identique au type 47 trouvé par Brega *et al.* (1986a), sauf erreur d'impression dans l'article original.

La majorité des types ne sont trouvés qu'à l'intérieur d'une population ou d'un groupe continental, à quelques exceptions près qui méritent de plus amples commentaires. En effet, en première approximation, des types retrouvés dans plusieurs groupes continentaux seront considérés comme des types qui ont existé avant la séparation de ces populations et qui se sont perpétués jusqu'à nos jours. Ceci repose sur l'hypothèse que 2 types similaires retrouvés dans 2 populations sont identiques par ascendance et exclue l'occurrence de mutations parallèles aboutissant à un type identique. Ces types sont donc des candidats pour être des types originaux. Ces assertions seront rediscutées dans le cadre de la tentative de reconstruction d'une phylogénie des types d'ADN-mt.

TABLE 4.14 : Définition des types.

Types	Morphes *Hpa I Bam HI Hae II Msp I Ava II*					Sites polymorphes[1]
1	2	1	1	1	1	
2	3	1	1	1	3	*3592*^{p} 16390^{a}
3	3	1	1	2	2	*3592*^{p} 8112^{m} 8150^{m} *8269*^{a}
4	3	1	1	3	2	*3592*^{p} 8112^{m} 8150^{m} *8269*^{a} *11454*^{m}
5	3	1	1	2	5	*3592*^{p} 8112^{m} 8150^{m} *8269*^{a} 16390^{a}
6	2	1	2	1	1	9056^{h}
7	3	1	1	1	1	*3592*^{p}
8	1	1	1	1	1	12408^{p}
9	1	1	2	1	1	9056^{h} 12408^{p}
10	3	1	1	1	2	*3592*^{p} *8269*^{a}
11	2	2	3	1	23	4533^{h} *8269*^{a} *16389*^{b}/16390^{a}
12	4	1	1	1	1	*12026*^{p}
13	2	1	5	1	1	*4830*^{h}
14	3	1	1	2	4	*3592*^{p} 8112^{m} 8150^{m} *8269*^{a} *15882*^{a} 16390^{a}
15	2	1	1	1	8	4810^{a}
16	2	1	2	1	10	*3881*^{a} 9056^{h}
17	2	1	1	1	9	13367^{a}
18	2	3	1	4	24	*13366*^{b}/13367^{a} 15925^{m}
19	2	3	1	4	25	*13366*^{b}/13367^{a} 15925^{m} 16390^{a}
20	2	3	7	4	24	*13366*^{b}/13367^{a} 14862^{h} 15925^{m}
21	2	1	1	1	2	*8269*^{a}
22	2	1	1	1	5	*8269*^{a} 16390^{a}
23	2	2	1	1	22	*16389*^{b}/16390^{a}
24	2	1	1	4	2	*8269*^{a} 15925^{m}
25	5	1	1	1	1	*1113*^{p}
26	2	1	6	1	1	*9264*^{h}
27	2	1	4	1	6	*11001*^{h} *15882*^{a}
28	2	1	1	4	1	15925^{m}
29	6	1	2	1	6	*2157*^{p} 9056^{h} 12408^{p} *15882*^{a}
30	3	1	1	1	11	*3592*^{p} *8269*^{a} *15882*^{a}
31	3	1	1	1	5	*3592*^{p} *8269*^{a} 16390^{a}
32	3	1	1	5	1	*3592*^{p} 8112^{m} 8150^{m} *13100*^{m}
33	3	1	1	2	3	*3592*^{p} 8112^{m} 8150^{m} 16390^{a}
34	3	1	2	1	3	*3592*^{p} 9056^{h} 16390^{a}
35	3	1	1	1	7	*3592*^{p} 13367^{a} 16390^{a}
36	2	1	1	1	13	12629^{a} 13367^{a} 16390^{a}
37	2	1	1	1	14	*8269*^{a} *12191*^{a} 16390^{a}
38	2	1	1	1	15	*4308*^{a}
39	2	1	4	1	1	*11001*^{h}
40	2	1	1	6	1	8150^{m}
41	2	1	1	7	1	*15503*^{m}
42	2	1	1	8	1	*13100*^{m}
43	3	1	1	1	12	*3592*^{p} 12629^{a}
44	2	1	1	4	3	15925^{m} 16390^{a}
45	2	1	8	1	1	*9689*^{h}
46	2	1	3	1	1	4533^{h}
47	2	1	1	1	3	16390^{a}
48	2	1	1	12	1	*14204*^{m}
49	2	1	1	9	1	*7973*^{m}
50	2	1	1	10	1	*12815*^{m}
51	2	1	1	11	1	12123^{m} 15925^{m}
52	2	1	1	1	16	*5260*^{a}
53	2	1	1	4	16	*5260*^{a} 15925^{m}
54	2	1	1	12	17	*14204*^{m} *15487*^{a}
55≡47	2	1	1	1	3	16390^{a}
56	2	1	1	1	6	*15882*^{a}
57	2	3	1	4	26	12629^{a} *13366*^{b}/13367^{a} 15925^{m}
58	2	2	3	1	18	*3131*^{a}(*7852*^{a}) 4533^{h} *8269*^{a} *16389*^{b}/16390^{a}
59	2	1	1	1	19	*4308*^{a} *15882*^{a}
60	2	1	1	1	20	*4308*^{a} *8269*^{a}
61	2	3	1	4	21	*8269*^{a} 12629^{a} *13366*^{b}/13367^{a} 15925^{m}
62	2	1	9	1	1	*2536*^{h}(*15497*^{h})

[1] Les sites polymorphes sont numérotés d'après la séquence de Cambridge. Les sites en italique ont été gagnés par rapport au type 1 et les sites en caractère non italique ont été perdus. Les sites indiqués entre parenthèse sont des localisations alternatives des sites déduits. Les sites séparés par une barre oblique indiquent un polymorphisme de deux sites de restriction causé par une seule substitution.Les enzymes sont identifiés de la manière suivante :a: *Ava II*; b: *Bam HI*; h: *Hae II*; m: *Msp I*; p: *Hpa I*

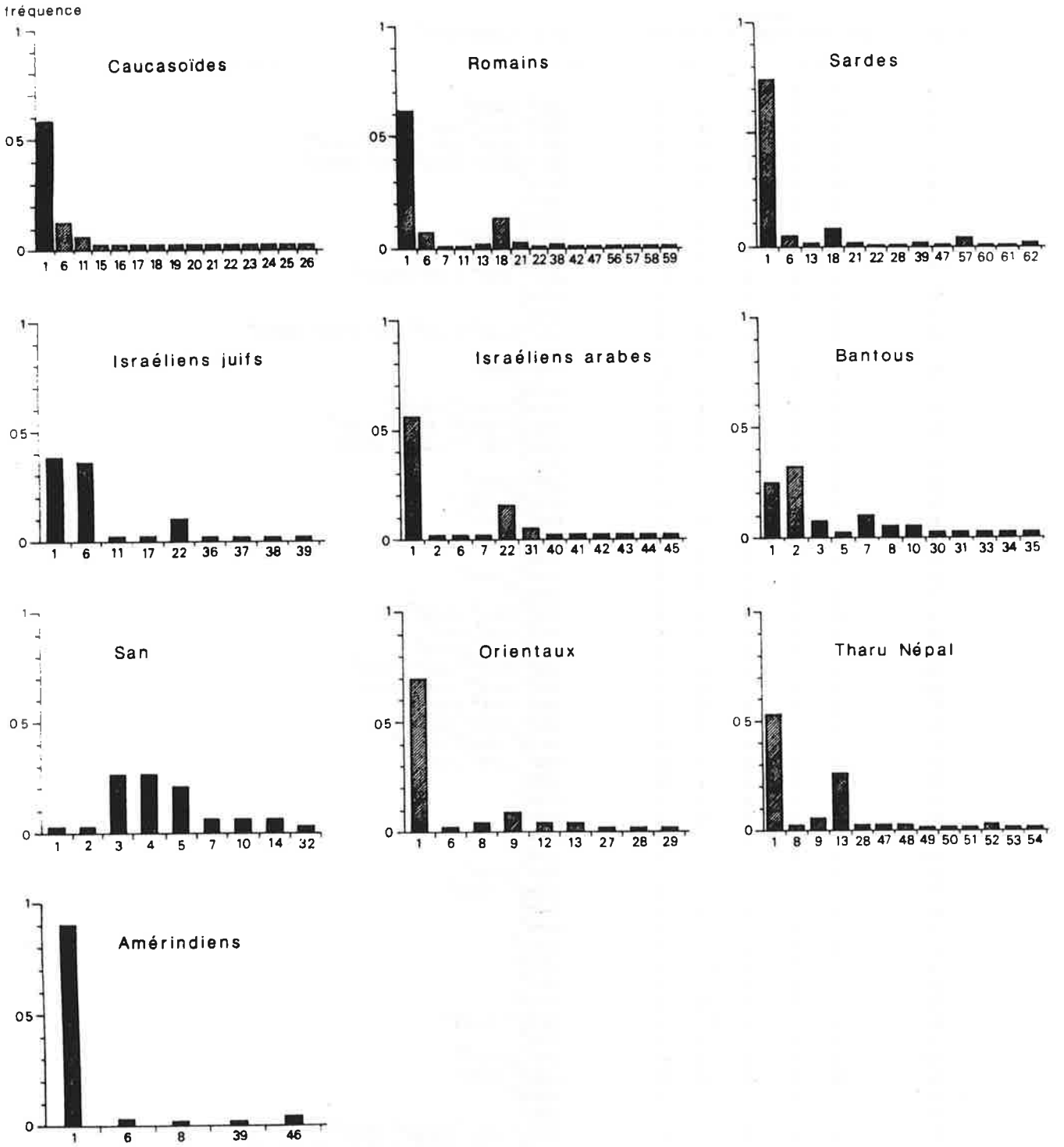

FIGURE 4.6 : Fréquence des types d'ADN-mt définis dans 10 populations.

TABLE 4.15: Fréquence (%) des types d'ADN-mt.

	Populations									
	Cauc.[1]	Rom[2].	Sard.[2]	Isr. J.[3]	Isr. A.[3]	Bant.[1]	San[1]	Orient.[1]	Tharu[4]	Am.[5]
Taille	50	95	134	38	39	40	34	46	91	74
Types										
1	58,0	62,1	75,4	38,5	56,4	25,0	2,9	69,6	52,7	90,5
2	0	0	0	0	2,6	32,5	2,9	0	0	0
3	0	0	0	0	0	7,5	26,5	0	0	0
4	0	0	0	0	0	0	26,5	0	0	0
5	0	0	0	0	0	2,5	20,6	0	0	0
6	12,0	7,4	4,5	35,9	2,6	0	0	2,2	0	2,7
7	0	1,1	0	0	2,6	10,0	5,9	0	0	0
8	0	0	0	0	0	5,0	0	4,3	2,2	1,4
9	0	0	0	0	0	0	0	8,7	1,1	0
10	0	0	0	0	0	5,0	5,9	0	0	0
11	6,0	1,1	0	2,6	0	0	0	0	0	0
12	0	0	0	0	0	0	0	4,3	0	0
13	0	2,1	1,5	0	0	0	0	4,3	25,3	0
14	0	0	0	0	0	0	5,9	0	0	0
15	2,0	0	0	0	0	0	0	0	0	0
16	2,0	0	0	0	0	0	0	0	0	0
17	2,0	0	0	2,6	0	0	0	0	0	0
18	2,0	13,7	7,5	0	0	0	0	0	0	0
19	2,0	0	0	0	0	0	0	0	0	0
20	2,0	0	0	0	0	0	0	0	0	0
21	2,0	3,2	1,5	0	0	0	0	0	0	0
22	2,0	1,1	0,7	10,2	15,4	0	0	0	0	0
23	2,0	0	0	0	0	0	0	0	0	0
24	2,0	0	0	0	0	0	0	0	0	0
25	2,0	0	0	0	0	0	0	0	0	0
26	2,0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	2,2	0	0
28	0	0	0,7	0	0	0	0	2,2	2,2	0
29	0	0	0	0	0	0	0	2,2	0	0
30	0	0	0	0	0	2,5	0	0	0	0
31	0	0	0	0	5,1	2,5	0	0	0	0
32	0	0	0	0	0	0	2,9	0	0	0
33	0	0	0	0	0	2,5	0	0	0	0
34	0	0	0	0	0	2,5	0	0	0	0
35	0	0	0	0	0	2,5	0	0	0	0
36	0	0	0	2,6	0	0	0	0	0	0
37	0	0	0	2,6	0	0	0	0	0	0
38	0	2,1	0	2,6	0	0	0	0	0	0
39	0	0	1,5	2,6	0	0	0	0	0	1,4
40	0	0	0	0	2,6	0	0	0	0	0
41	0	0	0	0	2,6	0	0	0	0	0
42	0	1,1	0	0	2,6	0	0	0	0	0
43	0	0	0	0	2,6	0	0	0	0	0
44	0	0	0	0	2,6	0	0	0	0	0
45	0	0	0	0	2,6	0	0	0	0	0
46	0	0	0	0	0	0	0	0	0	4,1
47	0	1,1	0,7	0	0	0	0	0	2,2	0
48	0	0	0	0	0	0	0	0	2,2	0
49	0	0	0	0	0	0	0	0	1,1	0
50	0	0	0	0	0	0	0	0	1,1	0
51	0	0	0	0	0	0	0	0	1,1	0
52	0	0	0	0	0	0	0	0	2,2	0
53	0	0	0	0	0	0	0	0	1,1	0
54	0	0	0	0	0	0	0	0	1,1	0
55	≡ 47									
56	0	1,1	0	0	0	0	0	0	0	0
57	0	1,1	3,0	0	0	0	0	0	0	0
58	0	1,1	0	0	0	0	0	0	0	0
59	0	1,1	0	0	0	0	0	0	0	0
60	0	0	0,7	0	0	0	0	0	0	0
61	0	0	0,7	0	0	0	0	0	0	0
62	0	0	1,5	0	0	0	0	0	0	0

Sources : 1) Johnson *et al.*, 1983; 2) Brega *et al.*, 1986b; 3) Bonné-Tamir *et al.*, 1986; 4) Brega *et al.*, 1986a; 5) Wallace *et al.*, 1985

Le type 1 est retrouvé dans toutes les populations recensées, et ceci avec des fréquences élevées, sauf dans les échantillons africains où d'autres types sont plus importants. Les types 2, 7 et 31 sont retrouvés à la fois dans les populations africaines et des populations moyen-orientales et européennes. Les types 6, 13, 28, 39 et 47 sont partagés entre des populations occidentales et des populations orientales. Enfin, le type 8 est retrouvé à la fois dans l'échantillon Bantou et les populations d'origine asiatique.

Mis à part le type 8, tous les types communs sont présents dans les populations méditerranéennes (Moyen-Orient et Italie), ce qui suggère que leur stock génétique possède un nombre important de types relativement anciens. Bien sûr, ces types communs sont souvent trouvés dans les échantillons avec de faibles fréquences, ce qui n'exclu pas qu'ils soient également présents dans d'autres populations, tout en n'ayant pas été identifiés par le processus d'échantillonnage. De meilleurs échantillons sont nécessaires pour cerner ce problème.

Le simple examen de la Table 4.15 et de la Figure 4.6 nous montre que seul un groupe très restreint de types possède des fréquences dépassant 10%, à savoir le type 1 bien évidemment, ainsi que les types 2, 3, 4, 5 et 7 chez les Africains et les types 6, 13, 18 et 22 dans d'autres populations. Les 51 autres types auront des fréquences relativement faibles et donc souvent mal estimées, vu la faible taille des échantillons.

Distances entre populations sur la base des fréquences des types dans les échantillons.

A partir des fréquences des types dans les différents échantillons, il est possible de calculer des distances entre les populations, en considérant chaque type comme un allèle d'un système génétique classique. Un nombre important de distances génétiques dérivées des fréquences géniques sont à la disposition des anthropologues et des généticiens (pour une revue, voir Jorde, 1980, 1985; Hedrick, 1985, pp.70-73; Nei, 1987, pp. 208-253). Pour notre part, nous avons utilisé, comme indice de similarité, le pourcentage (P) de fréquences géniques en commun entre 2 populations X et Y, à partir duquel une distance est facilement dérivée. Ce pourcentage est défini comme

$$P_{XY} = 100 \sum_{i=1}^{k} \min(f_{Xi}, f_{Yi}) \qquad \text{(Sanchez-Mazas } et\ al.\text{, 1986) (4.4)}$$

où k est le nombre d'allèles et f_{Xi} est la fréquence de l'allèle i dans la population X. Sur cette base, la distance entre les populations X et Y peut être définie par

$$D_{XY} = 1 - (P_{XY}/100) = \tfrac{1}{2} \sum_{i=1}^{k} |f_{Xi} - f_{Yi}| \qquad \text{(Sanchez-Mazas } et\ al.\text{, 1986) (4.5)}$$

qui est une distance métrique classique, pouvant être utilisée pour une analyse en coordonnées principales (Gower, 1966). Cette distance varie de 0 pour des populations ayant les mêmes fréquences géniques à 1 pour des populations qui ne possèdent pas d'allèles en commun. L'analyse en coordonnées principales est une méthode d'analyse multivariée donnant des résultats identiques à une analyse en composantes principales, quand elles sont appliquées aux mêmes données standardisées (Gower, 1967). La principale différence réside dans le fait que le premier type d'analyse est une "technique Q" qui repose sur l'étude d'attributs ou de caractères (ici des fréquences géniques) pour déterminer l'association entre populations (ou plus généralement entre des OTU's (*Operational Taxonomic Units*)), alors que le second type d'analyse est une "technique R" qui est l'étude des associations entre attributs pour toutes les populations (Sneath and Sokal, 1973).

Les données concernant les 10 populations ont été analysées avec ces 2 techniques. Dans le cas de l'analyse en coordonnées principales, nous avons tout d'abord calculé une matrice de distances (D_{XY}) entre les populations qui est reportée dans la Table 4.16. Ces distances ont ensuite été utilisées pour calculer les coordonnées principales permettant de représenter les populations comme des points dans un espace à 10 dimensions. Les 3 axes principaux ont été employés pour obtenir la Figure 4.7 qui synthétise 79% de l'information contenue dans la matrice de distances originale.

TABLE 4.16 : Distances[a] entre populations sur la base des fréquences des types d'ADN-mt.

Populations	Cauc.	Rom.	Sard.	Isr. J.	Isr. A.	Bant.	San	Orient.	Tharu
Romains	0,285								
Sardes	0,333	0,204							
Isr. J.	0,430	0,500	0,548						
Isr. A.	0,390	0,379	0,403	0,487					
Bantou	0,750	0,739	0,750	0,750	0,674				
San	0,971	0,960	0,971	0,971	0,919	0,732			
Orient.	0,398	0,336	0,260	0,594	0,414	0,707	0,971		
Tharu	0,451	0,419	0,421	0,593	0,451	0,706	0,949	0,352	
Amér.	0,393	0,352	0,206	0,575	0,410	0,736	0,971	0,269	0,437

[a]Ces distances sont calculées au moyen des formules (4.4) et (4.5) définies précédemment. Leur complément à 1 donne le pourcentage de fréquences géniques partagées entre 2 populations.

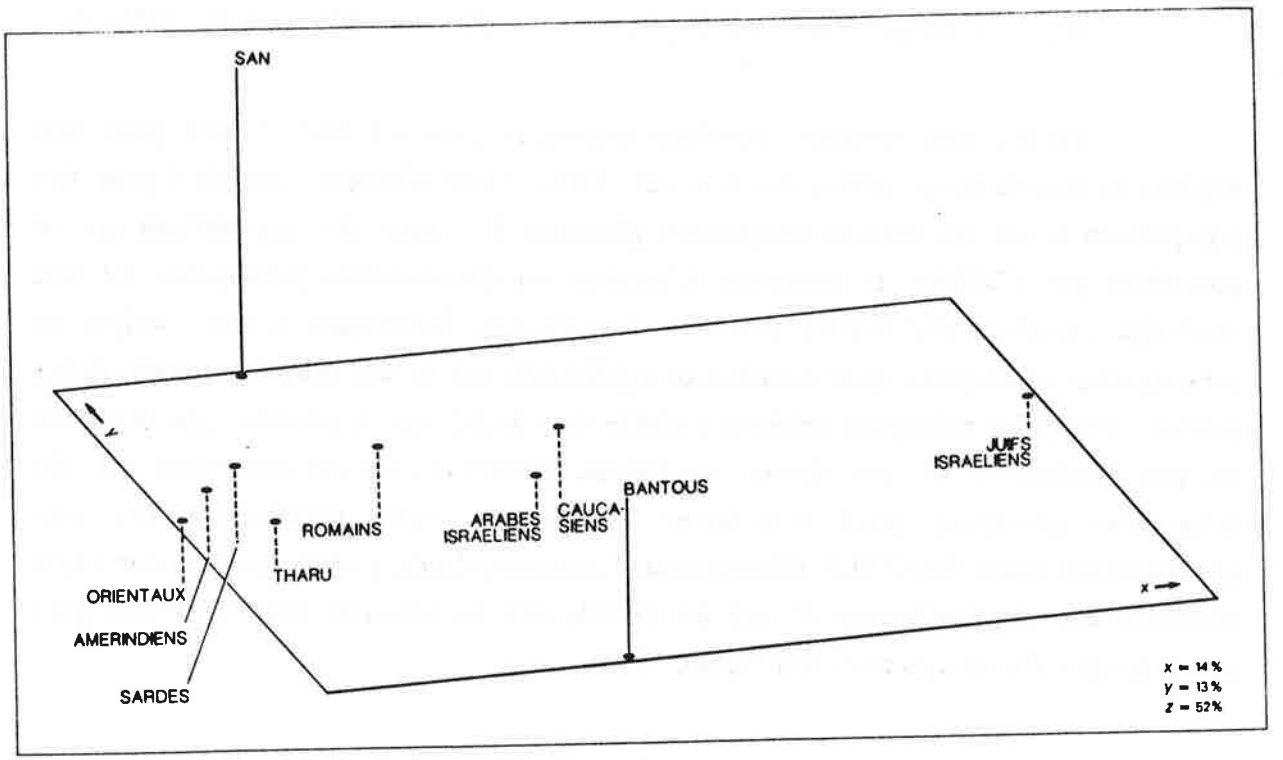

FIGURE 4.7 : Représentation des 3 premiers axes d'une analyse en coordonnées principales menée à partir de la matrice de distances entre populations de la Table 4.16.

Comme on pouvait s'y attendre, l'axe principal (Z) sépare nettement les population africaines des autres. Le second axe principal (X) oppose, quant à lui, les populations occidentales aux populations orientales. Le troisième axe principal (Y) sépare surtout les Bantous des San. Cette représentation, bien qu'elle traduise correctement la matrice des distances, ne fait que refléter l'information primordiale accumulée par notre mesure de distance génétique D_{XY}, à savoir le complément à 1 du pourcentage de fréquences géniques partagées entre 2 populations. D_{XY} est peu sensible aux types communs de faibles fréquences et est fortement dépendant des types les plus fréquents. On constate que dans le cas de l'ADN-mt, c'est le type 1 qui a une influence prépondérante sur notre distance, et sa faible représentation dans les échantillons africains est principalement responsable de leur séparation des autres populations. Nos populations peuvent être schématiquement classées en 2 groupes à partir du seul critère de la fréquence du type 1. Notre analyse n'explore donc pas toute la richesse informative des 61 types d'ADN-mt définis jusqu'à présent.

Afin de tenter de mieux cerner l'influence d'autres types d'ADN-mt, nous avons effectué une analyse en composantes principales directement à partir des fréquences (%) d'un nombre limité de types. Les types inclus dans cette analyse ont été choisi sur la base de leur ancienneté potentielle du fait de leur présence dans différents groupes continentaux, et/ou d'une fréquence supérieure à 10% dans une des populations au moins. Le résultat de cette analyse est présenté dans la Figure 4.8 où nous avons reporté la position des 10 populations selon les 3 composantes les plus importantes, qui représentent 70% de la variance des fréquences des 15 types étudiés. Pour interpréter correctement cette représentation tridimensionnelle, il convient d'essayer de définir la nature de l'information apportée par chaque composante. Ceci est possible grâce à la Table 4.17 où nous avons inscrit les coefficients de corrélation existant entre les 10 premières composantes principales et les types.

L'axe 1 (Z) est fortement corrélé avec les fréquences du type 1 et inversément corrélé avec les fréquences des types 2, 3, 4, 5 et 7 que l'on retrouve principalement en Afrique. Il n'est donc pas surprenant que celui-ci sépare les échantillons africains des autres, comme cela était le cas de l'axe principal dans l'analyse en coordonnées principales (Figure 4.7).

TABLE 4.17 : Corrélations[1] entre les 10 premiers axes de l'analyse en composantes principales et les 15 types d'ADN-mt retenus pour cette analyse.

Types	Axes									
	1	2	3	4	5	6	7	8	9	10
1	*0,83*	-0,08	0,06	0,14	-0,48	-0,04	0,21	0,00	0,04	0,00
2	*-0,54*	-0,21	0,46	*0,58*	0,28	0,13	0,10	-0,02	-0,07	0,00
3	*-0,93*	0,05	-0,31	-0,17	-0,04	0,09	0,02	0,03	0,00	0,00
4	*-0,80*	0,12	-0,45	-0,35	-0,12	0,03	-0,00	0,04	0,02	0,00
5	*-0,86*	0,09	-0,40	-0,28	-0,09	0,06	0,01	0,04	0,02	0,00
6	0,31	*0,67*	0,21	-0,25	0,36	0,33	-0,31	-0,07	0,05	0,00
7	*-0,84*	-0,10	0,20	0,41	0,24	-0,02	0,10	0,06	0,00	0,00
8	-0,20	*-0,68*	*0,56*	0,27	-0,10	0,21	-0,17	0,09	0,11	0,00
13	0,27	*-0,77*	-0,13	-0,42	0,35	-0,01	0,08	-0,11	0,02	0,00
18	0,39	0,13	*-0,67*	*0,52*	0,13	0,11	0,00	0,27	0,05	0,00
21	0,40	0,20	*-0,67*	*0,55*	0,08	0,03	-0,19	-0,09	-0,06	0,00
22	0,15	*0,56*	0,36	-0,22	0,32	*-0,59*	0,00	0,19	0,01	0,00
28	0,29	*-0,79*	0,05	-0,38	-0,10	0,06	-0,25	0,23	-0,10	0,00
39	0,34	*0,57*	0,28	-0,31	0,05	*0,54*	0,27	0,13	-0,04	0,00
47	0,39	*-0,60*	-0,44	-0,16	0,47	0,04	0,20	0,00	0,03	0,00

[1] Les coefficients de corrélation dont la valeur absolue dépasse 0,5 sont inscrits en italique.

L'axe 2 (Y) est, lui, corrélé positivement avec les types 6, 39 et 47 et négativement avec les types 8, 13 et 28. Il oppose ainsi l'échantillon Tharu à l'échantillon d'Israéliens Juifs, les autres populations prenant des valeurs intermédiaires et guères discriminantes, hormis l'échantillon d'Orientaux qui se rapproche des Tharu selon cet axe.

Le 3ème axe (X) considéré ici est surtout corrélé positivement avec le type 8 et négativement avec les types 18 et 21. Comme ces types ne sont pas présents dans toutes les populations, les types qui sont plus faiblement corrélés avec l'axe 3 doivent être aussi considérés pour pouvoir interpréter correctement la différenciation des populations selon cet axe. On constate que l'axe 3 sépare les échantillons Caucasoïdes, Romains, Sardes, San et Tharu des autres échantillons, sans qu'une signification claire puisse lui être attribuée. Notons simplement qu'il permet de distinguer l'échantillon Romain des Sardes qui possédaient tout deux des coordonnées très similaires pour les 2 premiers axes.

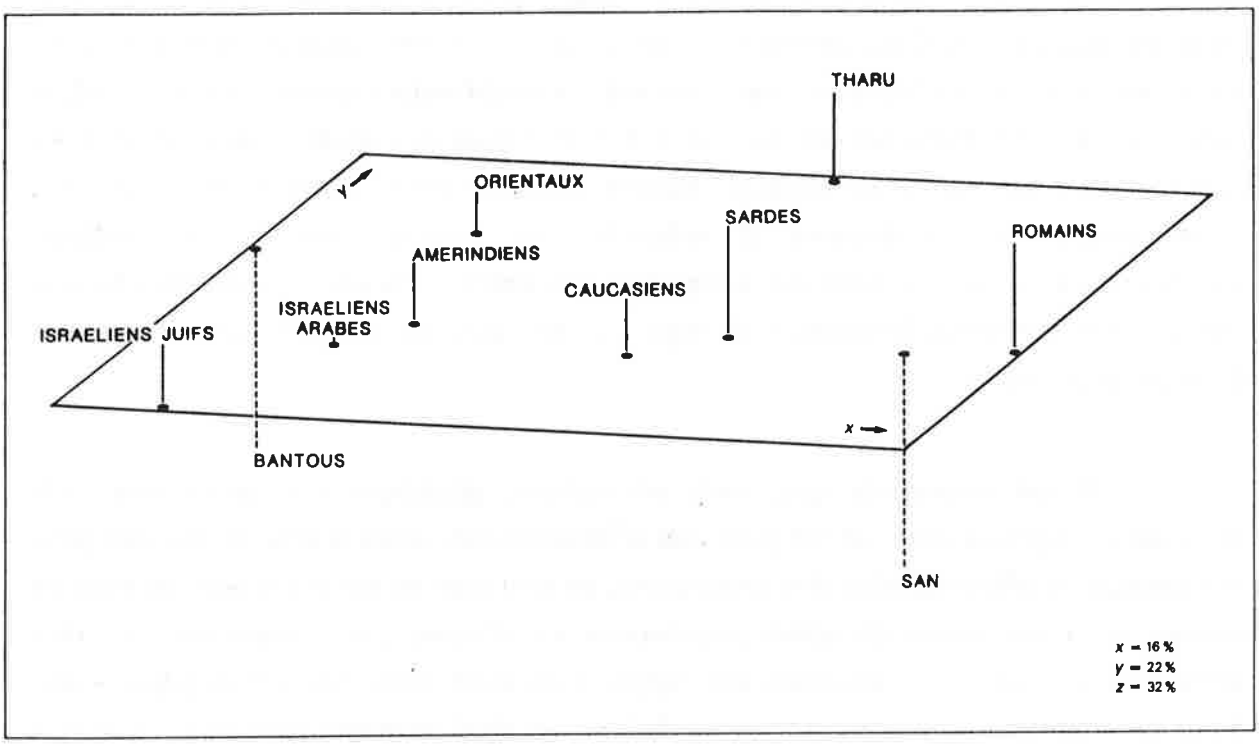

FIGURE 4.8 : Représentation des 3 premiers axes d'une analyse en composantes principales menée à partir des fréquences des types 1, 2, 3, 4, 5, 6, 7, 8, 13, 18, 21, 22, 28, 39 et 47 dans 10 populations (voir le texte pour le choix des types).

Si ces 2 types d'analyse multivariée ne présentent pas exactement les même résultats, car basés sur des données légèrement différentes, ils sont néanmoins cohérents entre eux. Ils mettent avant tout en relief l'opposition entre les échantillons africains et les autres populations. Cela semble être surtout dû à la prépondérance du type 1 parmi les populations extra-africaines pour l'analyse en coordonnées principales et le même facteur associé à la présence des types 2, 3, 4, 5 et 7 dans les échantillons africains pour l'analyse en composantes principales. Aucune structure claire concernant les rapports entre populations occidentales et orientales ne semble émerger des analyses multivariées effectuées à partir des fréquences des types. Cela peut également être une conséquence des fortes fréquences du type 1 et de l'absence d'autres types importants dans ces populations.

Il est surprenant que, dans un système génétique présentant une telle diversité de types et donc un tel potentiel d'informations aptes à être employées pour reconstituer la différenciation des populations, un seul type puisse d'une part imposer un clivage aussi net entre certaines populations et, d'autre part, constituer un filtre perturbateur pour la visualisation des rapports existant entre les autres populations. Nous ne sommes ainsi pas en mesure de déduire un résultat quelconque quant à la voie évolutive suivie pour arriver aux fréquences géniques observées actuellement.

A ce stade de notre étude, le polymorphisme de l'ADN-mt, considéré comme un système génétique classique et neutre, analysé à partir d'informations quantitatives, ne concrétise pas les espoirs que nous étions en droit d'attendre de lui. Une approche plus qualitative, portant sur la nature même des types et des rapports moléculaires qui les relient, pourrait s'avérer plus fructueuse pour la formulation ou le test d'hypothèses concernant l'histoire du peuplement humain.

Nous avons vu qu'il était possible de proposer un schéma de différenciation des morphes de chaque enzyme. Ce schéma ne correspond pas véritablement à une phylogénie des morphes car, d'une part, les morphes ancestraux ne sont pas connus avec certitude (tout morphe peut être ancestral avec une certaine probabilité) et, d'autre part, il existe des cas où l'on n'est pas sûr du chemin évolutif suivi pour aboutir à la diversité actuelle des morphes (un morphe peut souvent être dérivé de 2 autres morphes sans savoir lequel est ancestral). Néanmoins, la connaissance de ces évènements mutationnels, pour chaque analyse enzymatique, peut être mise en commun pour tenter de définir tous les liens possibles entre les types d'ADN-mt.

La détermination d'une phylogénie des types d'ADN-mt, en admettant qu'elle soit possible, est intéressante à plusieurs points de vue, bien qu'elle ne saurait constituer une fin en soi. Elle permet de préciser la nature des différences moléculaires entre les types d'une même population et vérifier dans quelle mesure la connaissance de la différenciation des types peut permettre de tenir un discours cohérent sur la différenciation des populations. D'autre part, le calcul de certaines distances génétiques entre populations (Nei and Tajima, 1981) basées sur le nombre moyen de substitutions entre 2 types tirés au hasard nécessite de connaître ce nombre avec précision.

Problèmes méthodologiques

La démarche que nous suivrons diffère quelque peu d'autres méthodes de reconstruction phylogénique à partir de cartes de restriction enzymatique, qu'il convient de décrire ici brièvement. Celles-ci peuvent être classées en 2 grandes catégories. La première repose sur l'établissement d'une distance entre 2 séquences d'ADN quelconques sur la base des sites communs entre 2 types. Plusieurs distances ont été proposées (Upholt, 1977; Gotoh *et al.*, 1979; Kaplan and Langley, 1979; Nei and Li, 1979; Engels, 1981; Kaplan and Risko, 1981; Nei and Tajima, 1983) qui reposent généralement sur l'estimation du nombre de substitutions par site de nucléotide s'étant produites entre 2 séquences. Nous prendrons ici comme exemple la méthode de Nei et Li (1979) qui est la plus répandue, relativement simple et, du moins, exemplaire des problèmes sous-jacents à ce type d'analyse.

Si l'on prend un bloc de r nucléotides dans 2 séquences homologues, 4 cas sont possibles: 1) Le bloc est un site pour les 2 séquences d'ADN. 2) et 3) Le bloc est un

site pour une seule des 2 séquences. 4) Le bloc est un site pour aucune des 2 séquences. On notera par m_X et m_Y les nombres respectifs de sites de restriction pour les séquences X et Y, et m_{XY} le nombre de sites partagés entre les 2 séquences. La probabilité (S) que X et Y partagent la même séquence de reconnaissance pour un site donné peut être estimée par

$$\hat{S} = 2m_{XY}/(m_X + m_Y) \qquad \text{(Nei and Li, 1979)} \quad (4.6)$$

On s'attend à ce que S varie avec le temps de divergence des 2 séquences. Formellement, Nei et Tajima (1983) montrent que

$$\hat{S} = (1\text{-}P)^2 + \Sigma_i w_i Q_i^2/a_0 \qquad (4.7)$$

où w_i est la probabilité qu'un bloc de r nucléotides soit différent d'un site par i nucléotides, P est la probabilité qu'un site présent au temps $t=0$ ne soit plus un site au temps t et Q_i est la probabilité qu'un bloc de r nucléotides différant d'un site par i nucléotides au temps $t=0$ devienne un site au temps t. Le terme $\Sigma_i w_i Q_i^2/a_0$ représente ici la probabilité que des sites communs soient apparus de façon indépendante depuis la divergence des 2 séquences.

En posant $(1\text{-}P) = e^{-r\lambda t}$ (où λ est le taux de substitution de nucléotide par site de nucléotide par année) et en négligeant les sites communs apparus de façon parallèle, Nei et Li (1979) redéfinissent S comme

$$S = e^{-2r\lambda t}, \qquad (4.8)$$

d'où l'on tire le nombre attendu de substitutions de nucléotides par site de nucléotide ($d = 2\lambda t$) en fonction de S comme

$$\hat{d} = -\log_e \hat{S}/r \qquad \text{(Nei et Tajima, 1983)} \quad (4.9)$$

et

$$V(\hat{d}) = (2\text{-}S)(1\text{-}S)/(2r^2\bar{m}S) \qquad \text{(Nei et Tajima, 1983)} \quad (4.10)$$

où $\bar{m} = (m_X + m_Y)/2$.

Nei et Tajima (1983) ont dérivé des formules équivalentes pour les cas, plus complexes, où un enzyme peut reconnaître plusieurs séquences et lorsque plusieurs enzymes reconnaissant des sites différents sont employés simultanément.

Toutes ces estimations sont basées sur des hypothèses relativement fortes, visant à simplifier considérablement le modèle utilisé. Elles supposent que la séquence étudiée est en équilibre au niveau moléculaire (cela implique que le nombre de sites doit être stable au cours du temps), que toutes les bases connaissent un taux de substitution équivalent, et que les sites communs apparus de façon indépendante ne soient pas pris en considération. L'incidence de ces 3 facteurs sur l'estimation de la distance de Nei et Li a été étudiée depuis lors. Adams et Rothmans (1982) ont montré que la distribution des sites de restriction n'était pas aléatoire dans le cas de l'ADN-mt et que le nombre de sites observés ne correspondait pas à celui qui était attendu pour la grande majorité des enzymes employés. Tajima et Nei (1982) ont montré que des taux de substitution inégaux pour les 4 nucléotides pouvaient introduire des biais considérables dans l'estimation de *S*. (Nous avons vu que les transitions étaient nettement prépondérantes par rapport aux transversions dans le cas de l'ADN-mt). Ce biais peut avoir pour conséquence une augmentation des substitutions parallèles qui contribuent d'une manière importante à des gains de sites homologues mais indépendants (Templeton, 1983a). Il est bon de rappeller que ce dernier cas a été jugé négligeable lorsque les séquences avaient divergé depuis peu (Nei and Li, 1979; Li, 1981; Nei, 1987).

Templeton (1983a et b) s'est attaché à étudier l'importance de ces substitutions indépendantes qui aboutissent à un gain ou à une perte de site commun chez 2 types différents. Il a ainsi distingué 4 cas possibles (Figure 4.9), selon que l'ancêtre commun aux 2 types présentait un site ou non à un endroit particulier du génome. En ignorant les cas impliquant plus de 2 mutations indépendantes, Templeton a montré que les pertes convergentes et les "gain-perte" avaient une probabilité non négligeable pour des valeurs du produit λt faibles (0,03), et, qu'à ce niveau, le nombre total de substitutions pouvait être sous-estimé de moitié (Templeton, 1983a). Il a ensuite développé une méthode non-paramétrique dérivée du principe de parcimonie pour reconstruire une phylogénie des séquences d'ADN étudiées au moyen des PLFR. Celle-ci privilégie les séries de substitutions convergentes les plus probables.

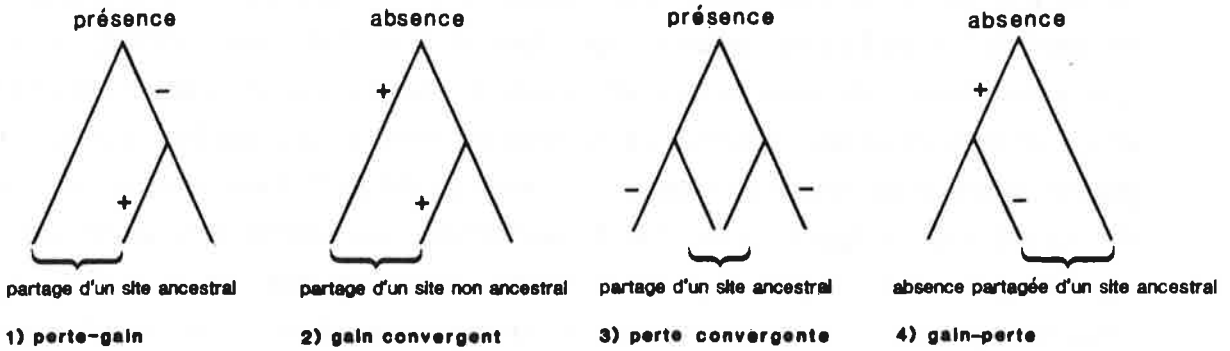

FIGURE 4.9 : Evolution convergente de présence ou absence de sites de restriction.

Nei et Tajima (1985, 1987) ont repris, étendu et critiqué les réserves de Templeton (1983 a et b), concernant le bien-fondé du critère de parcimonie dans la déduction d'une phylogénie des types lorsque le nombre de subtitutions par site (*d*) dépasse une certaine valeur, tout en défendant leur propre estimateur de *d*. En fait, la querelle idéologique de ces différents auteurs semble buter principalement, non pas sur la nature et l'importance du phénomène des substitutions parallèles, mais plutôt sur le fait de savoir à partir de quel seuil elles interviennent et perturbent l'estimation de *d* ou les contructions phylogéniques. Ce seuil dépend du produit λt qui comporte un paramètre temporel (temps de divergence entre les types) et un paramètre de mutation (taux de substitution par site par année) qui peuvent varier indépendamment suivant les séquences d'ADN étudiées.

Le débat n'étant pas encore clos (voir Templeton, 1987), nous nous garderons de prendre position, mais il nous semble important de préciser quels facteurs peuvent perturber de façon isolée ou conjuguée les méthodes actuelles de reconstruction phylogénique. Ceux-ci peuvent s'énoncer comme suit :

1) Temps de divergence élevé entre les séquences d'ADN.

2) Taux de substitution élevé.

3) Taux de substitution inégal pour les 4 nucléotides.

4) Taux de subtitution non constant entre les différentes séquences.

5) Absence d'équilibre de la composition en nucléotides de la séquence étudiée.

6) Hétérogénéité du taux de substitution suivant la portion de séquence étudiée.

Réseau de types et phylogénie

Ces problèmes méthodologiques nous ont donc amené à réexaminer les arbres phylogéniques concernant les types d'ADN-mt qui avaient déjà été proposés par différents auteurs (Johnson *et al.*, 1983; Wallace *et al.*, 1985; Bonné-Tamir *et al.*, 1986; Brega *et al.*, 1986a) et de proposer une phylogénie des types se basant à la fois sur la différenciation moléculaire des morphes, mais aussi sur les réseaux phylogéniques possibles entre les types trouvés dans une même population.

TABLE 4.18: Nombre de substitutions minimum entre types d'ADN-mt

Types	*Distances*
2	*2*
3	*4 4*
4	*5 5 1*
5	*5 3 1 2*
6	*1 3 5 6 6*
7	*1 1 3 4 4 2*
8	*1 3 5 6 6 2 2*
9	*2 4 6 7 7 1 3 1*
10	*2 2 2 3 3 3 1 3 4*
11	*3 5 5 6 6 4 4 4 5 3*
12	*1 3 5 6 6 2 2 2 3 3 4*
13	*1 3 5 6 6 2 2 2 3 3 4 2*
14	*6 4 2 3 1 7 5 7 8 4 7 7 7*
15	*1 3 5 6 6 2 2 2 3 3 4 2 2 7*
16	*2 4 6 7 7 1 3 3 2 4 5 3 3 8 3*
17	*1 3 5 6 6 2 2 2 3 3 4 2 2 7 2 3*
18	*2 4 6 7 7 3 3 3 4 4 5 3 3 8 3 4 3*
19	*3 3 7 8 6 4 4 4 5 5 6 4 4 7 4 5 4 1*
20	*3 5 7 8 8 4 4 4 5 5 6 4 4 9 4 5 4 1 2*
21	*1 3 3 4 4 2 2 2 3 1 2 2 2 5 2 3 2 3 4 4*
22	*2 2 4 5 3 3 3 3 4 2 3 3 3 4 3 4 3 4 3 5 1*
23	*1 3 5 6 6 2 2 2 3 3 2 2 2 7 2 3 2 3 4 4 2 3*
24	*2 4 4 5 5 3 3 3 4 2 3 3 3 6 3 4 3 2 3 3 1 2 3*
25	*1 3 5 6 6 2 2 2 3 3 4 2 2 7 2 3 2 3 4 4 2 3 2 3*
26	*1 3 5 6 6 2 2 2 3 3 4 2 2 7 2 3 2 3 4 4 2 3 2 3 2*
27	*2 4 6 7 7 3 3 3 4 4 5 3 3 6 3 4 3 4 5 5 3 4 3 4 3 3*
28	*1 3 5 6 6 2 2 2 3 3 4 2 2 7 2 3 2 1 2 2 2 3 2 1 2 2 3*
29	*4 6 8 9 9 3 5 3 2 6 7 5 5 8 5 4 5 6 7 7 5 6 5 6 5 5 4 5*
30	*3 3 3 4 4 4 2 4 5 1 4 4 4 3 4 5 4 5 6 6 2 3 4 3 4 4 3 4 5*
31	*3 1 3 4 2 4 2 4 5 1 4 4 4 3 4 5 4 5 4 6 2 1 4 3 4 4 5 4 7 2*
32	*4 4 2 3 3 5 3 5 6 4 7 5 5 4 5 6 5 6 7 7 5 6 5 6 5 5 6 5 8 5 5*
33	*4 2 2 3 1 5 3 5 6 4 7 5 5 2 5 6 5 6 5 7 5 4 5 6 5 5 6 5 8 5 3 2*
34	*3 1 5 6 4 2 2 4 3 3 6 4 4 5 4 3 4 5 4 6 4 3 4 5 4 4 5 4 5 4 2 5 3*
35	*3 1 5 6 4 4 2 4 5 3 6 4 4 5 4 5 2 5 4 6 4 3 4 5 4 4 5 4 7 4 2 5 3 2*
36	*3 3 7 8 6 4 4 4 5 5 6 4 4 7 4 5 2 5 4 6 4 3 4 5 4 4 5 4 7 6 4 7 5 4 2*
37	*3 3 5 6 4 4 4 4 5 3 4 4 4 5 4 5 4 5 4 6 2 1 4 3 4 4 5 4 7 4 2 7 5 4 4 4*
38	*1 3 5 6 6 2 2 2 3 3 4 2 2 7 2 3 2 3 4 4 2 3 2 3 2 2 3 2 5 4 4 5 5 4 4 4*
39	*1 3 5 6 6 2 2 2 3 3 4 2 2 7 2 3 2 3 4 4 2 3 2 3 2 2 1 2 5 4 4 5 5 4 4 4 2*
40	*1 3 3 4 4 2 2 2 3 3 4 2 2 5 2 3 2 3 4 4 2 3 2 3 2 2 3 2 5 4 4 3 3 4 4 4 2 2*
41	*1 3 5 6 6 2 2 2 3 3 4 2 2 7 2 3 2 3 4 4 2 3 2 3 2 2 3 2 5 4 4 5 5 4 4 4 4 2 2 2*
42	*1 3 5 6 6 2 2 2 3 3 4 2 2 7 2 3 2 3 4 4 2 3 2 3 2 2 3 2 5 4 4 3 5 4 4 4 4 2 2 2 2*
43	*2 2 4 5 5 3 1 3 4 2 5 3 3 6 3 4 3 4 5 5 3 4 3 4 3 3 4 3 6 3 3 4 4 3 3 3 5 3 3 3 3*
44	*2 2 6 7 5 3 3 3 4 4 5 3 3 6 3 4 3 2 1 3 3 2 3 2 3 3 4 1 6 5 3 6 4 3 3 3 3 3 3 3 3 4*
45	*1 3 5 6 6 2 2 2 3 3 4 2 2 7 2 3 2 3 4 4 2 3 2 3 2 2 3 2 5 4 4 5 5 4 4 4 4 2 2 2 2 2 3 3*
46	*1 3 5 6 6 2 2 2 3 3 2 2 2 7 2 3 2 3 4 4 2 3 2 3 2 2 3 2 5 4 4 5 5 4 4 4 4 2 2 2 2 2 3 3 2*
47	*1 1 5 6 4 2 2 2 3 3 4 2 2 5 2 3 2 3 2 4 2 1 2 3 2 2 3 2 5 4 2 5 3 2 2 2 2 2 2 2 2 3 1 2 2*
48	*1 3 5 6 6 2 2 2 3 3 4 2 2 7 2 3 2 3 4 4 2 3 2 3 2 2 3 2 5 4 4 5 5 4 4 4 4 2 2 2 2 2 3 3 2 2*
49	*1 3 5 6 6 2 2 2 3 3 4 2 2 7 2 3 2 3 4 4 2 3 2 3 2 2 3 2 5 4 4 5 5 4 4 4 4 2 2 2 2 2 3 3 2 2 2*
50	*1 3 5 6 6 2 2 2 3 3 4 2 2 7 2 3 2 3 4 4 2 3 2 3 2 2 3 2 5 4 4 5 5 4 4 4 4 2 2 2 2 2 3 3 2 2 2 2*
51	*2 4 6 7 7 3 3 3 4 4 5 3 3 8 3 4 3 2 3 3 3 4 3 2 3 3 4 1 6 5 5 6 6 5 5 5 5 3 3 3 3 3 4 2 3 3 3 3 3*
52	*1 3 5 6 6 2 2 2 3 3 4 2 2 7 2 3 2 3 4 4 2 3 2 3 2 2 3 2 5 4 4 5 5 4 4 4 4 2 2 2 2 2 3 3 2 2 2 2 2 3*
53	*2 4 6 7 7 3 3 3 4 4 5 3 3 8 3 4 3 2 3 3 3 4 3 2 3 3 4 1 6 5 5 6 6 5 5 5 5 3 3 3 3 3 4 2 3 3 3 3 3 2 1*
54	*2 4 6 7 7 3 3 3 4 4 5 3 3 8 3 4 3 4 5 5 3 4 3 4 3 3 4 3 6 5 5 6 6 5 5 5 5 3 3 3 3 3 4 4 3 3 3 1 3 3 4 3 4*
55 ≡ 47	
56	*1 3 5 6 6 2 2 2 3 3 4 2 2 5 2 3 2 3 4 4 2 3 2 3 2 2 1 2 3 2 4 5 5 4 4 4 4 2 2 2 2 2 3 3 2 2 2 2 2 3 2 3 3 2*
57	*3 5 7 8 8 4 4 4 5 5 6 4 4 9 4 5 4 1 2 2 4 5 4 3 4 4 5 2 7 6 6 7 7 6 6 4 6 4 4 4 4 4 3 3 4 4 4 4 4 4 3 4 3 5 4 4*
58	*4 6 6 7 7 5 5 5 6 4 1 5 5 8 5 6 5 6 7 7 3 4 3 4 5 5 6 5 8 5 5 8 8 7 7 7 5 5 5 5 5 5 6 6 5 3 5 5 5 5 6 5 6 6 5 5 7*
59	*2 4 6 7 7 3 3 3 4 4 5 3 3 6 3 4 3 4 5 5 3 4 3 4 3 3 2 3 4 3 5 6 6 5 5 5 5 1 3 3 3 3 4 4 3 3 3 3 3 3 4 3 4 4 3 1 5 6*
60	*2 4 4 5 5 3 3 3 4 2 3 3 3 6 3 4 3 4 5 5 1 2 3 2 3 3 4 3 6 3 3 6 6 5 5 5 3 1 3 3 3 3 4 4 3 3 3 3 3 3 4 3 4 4 3 3 5 4 2*
61	*4 6 6 7 7 5 5 5 6 4 5 5 5 8 5 6 5 2 3 3 3 4 5 2 5 5 6 3 8 5 5 8 8 7 7 5 5 5 5 5 5 5 4 4 5 5 5 5 5 5 4 5 4 6 4 5 1 6 6 4*
62	*1 3 5 6 6 2 2 2 3 3 4 2 2 7 2 3 2 3 4 4 2 3 2 3 2 2 3 2 5 4 4 5 5 4 4 4 4 2 2 2 2 2 3 3 2 2 2 2 2 2 3 2 3 3 2 2 4 5 3 3 5*

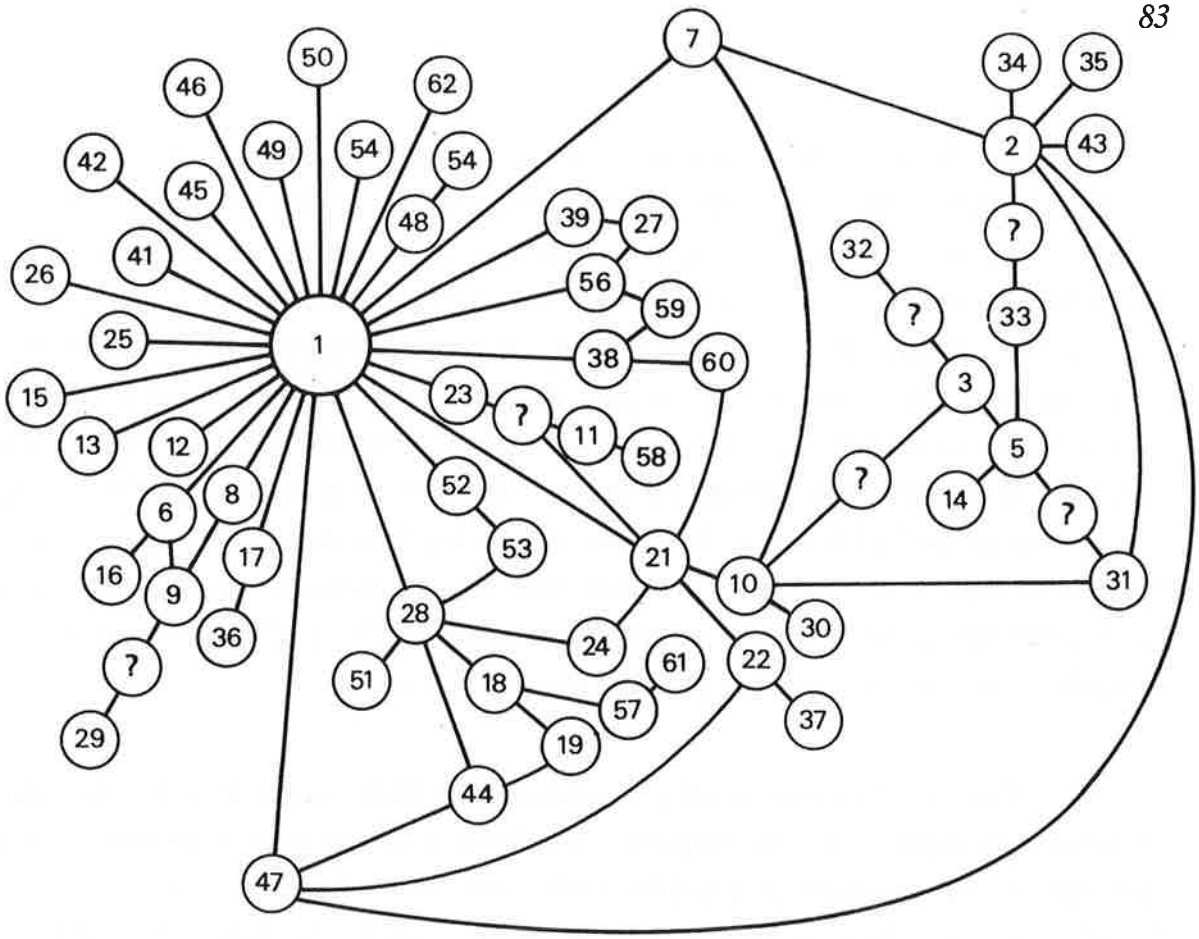

FIGURE 4.10 : Réseau de différenciation des 61 types d'ADN-mt définis dans 10 populations (voir texte). Les types sont reliés si ils diffèrent par une simple substitution (liaison d'ordre 1). Toutes les liaisons d'ordre 1 possibles ont été représentées. Les types indiqué par un ? représentent des types non identifiés dans les échantillons et nécessaires pour relier certains types ou groupes de types au réseau principal.

Le réexamen des liens entre les morphes de chaque enzyme nous a permis de constater que certaines liaisons entre types, jusqu'alors acceptées et reprises par d'autres auteurs, devaient être exclues et que d'autres qui n'avaient pas été perçues pouvaient bel et bien exister (Excoffier et Langaney, 1988). Les principales modifications des liens entre types, par rapport aux phylogénies déjà publiées, peuvent s'énoncer comme suit. Le passage entre les types 18 et 28 est possible par l'apparente double mutation $55^{-}/56^{+}$, le type 36 peut être directement rattaché au type 17, une simple liaison existe entre les types 1 et 23, le type 33 se branche sur le type 5 (et non sur le 2), ainsi que le type 43 sur le 2 (et non sur le 7). Du fait de la double mutation effective nécessaire pour passer du type 10 au type 3, ces 2 type ne peuvent être liés directement, un type intermédiaire reste à trouver. Il nous a fallu, enfin, placer les types 56 à 62 qui n'avaient pas été intégrés à un réseau de ce genre par Brega *et al.* (1986b).

Ces modifications n'ont pas fondamentalement bouleversé la structure des liens entre les types, et ont même parfois conduit à la rendre plus cohérente. Il est ainsi remarquable qu'il ne manque que 4 intermédiaires pour relier les 61 types. Cela dénote la présence d'une forte structure dans la différenciation des types. Les populations analysées conservent donc la trace de la quasi-totalité des types formés à partir des types ancestraux, que nous tenterons de définir ultérieurement. Les types intermédiaires restant à découvrir plaident en fait pour la constitution d'échantillons plus importants, qui auraient ainsi plus de chance de contenir des types rares, et également pour l'échantillonnage de populations où ils pourraient résider. On pense notamment au type manquant pour relier une bonne partie des types "africains" décrits au reste du réseau qui pourrait très bien être présent dans d'autres régions d'Afrique. Il en va de même pour le type 11 (lié au type 58), dont on trouve la trace dans plusieurs populations d'origine européenne ou moyen-orientale, ce qui montre que la diversité de ce groupe n'a pas été entièrement explorée.

Il convient maintenant de définir des critères qui vont nous permettre de déterminer, dans certains cas ambigûs, quelles seraient les substitutions les plus à même d'être à la source de la diversité des types actuels. Tout d'abord, on s'attend à ce qu'un type rare (<5%) soit issu d'un type plus fréquent (>5%), de manière analogue au gradient de fréquence défini dans le cas des morphes *Ava II*. Le deuxième critère sera la préférence des liens entre des types se trouvant dans les mêmes populations, ou du moins dans le même groupe continental. Finalement, des substitutions bien définies au niveau moléculaire seront préférées à des substitutions qui ne le sont pas. Nous pensons ici aux enzymes qui possèdent des séquences de reconnaissance multiples et dont la

perte ou le gain d'un site peuvent être dus à plusieurs évènements. La discussion qui va suivre va donc consister à justifier le choix des sites impliqués dans les passages entre types de la phylogénie présentée dans la Figure 4.11.

En premier lieu, nous allons examiner à quel type rattacher le groupe des types 3, 4, 5, 14 et 33. Ce groupe de types pourrait en effet être lié à 3 autres types (2, 10 ou 31). Le type 31 ne semble pas être à la source de ce groupe car il est rare et lui-même issu du type 2 ou du 10. Le type 2 étant très fréquent dans la population Bantou ou le 31 est également rencontré, nous privilégierons donc une liaison 2-31 aux dépens de la liaison 10-31. Si le type 2 était à l'origine du groupe 3-4-5-14-33, cela impliquerait que le type 33 ait produit les 4 autre types, ce qui est improbable étant donné sa faible fréquence dans le seul échantillon (Bantou) où il est rencontré. Par contre, le rattachement du type 3 au type 10 par un intermédiaire est plus logique, le type 3 étant clairement à la source des types 4, 5 et aussi très probablement du type 32, qui est détaché (voir la différenciation des morphes *Msp I*).

Le type 10 peut lui-même être issu de 2 types (le 7 ou le 21). La présence des types 10 et 21 n'ayant jamais été observée dans un même échantillon, alors que c'est le cas des types 7 et 10, nous privilégierons l'occurrence d'une liaison 7-10. Selon cette phylogénie, la presque totalité des types africains serait issue du type 7, que l'on retrouve également chez des populations méditerranéennes.

Toujours selon le même principe de la présence de types parents dans une même population, nous favoriserons l'attachement du type 60 au type 21 (et non au type 38), et du type 24 au type 21 également (et non au type 28). Le type 22 pourrait être obtenu à partir du type 21 et non du type 47, car en cas de présence simultanée dans le même échantillon, le type 21 est plus fréquent que les 2 autres (voir Table 4.17).

Le type 59 serait issu du type 38 plutôt que du type 56, selon leurs fréquences dans l'échantillon Romain. Le type 27, trouvé chez les Orientaux, se rattache au type 39 que l'on retrouve dans une autre population d'origine asiatique, plutôt qu'au type 56, trouvé pour l'instant uniquement chez les Romains.

Les types 8 et 9 partagent le morphe rare *Hpa I* 1, qui pourrait être très ancien, puisqu'on le retrouve chez un autre hominien. De ce fait, le type 9 découlerait du type 8 et non du type 6.

Au vu du schéma de différenciation des morphes *Ava II*, le type 44 ne peut pas être lié au type 19, mais plus vraisemblablement aux types 28 ou 47. Cette situation est exemplaire des cas où il est impossible de trancher avec une certitude raisonnable sur une phylogénie donnée, et c'est là que la notion de réseau prend tout son sens. Un problème similaire est rencontré pour l'origine du type 53, qui peut aussi bien être dérivé du type 52 que du type 28. Ces 2 situations concernent toutefois un ensemble de types peu fréquents, qui ne sont eux-mêmes vraisemblablement pas à l'origine d'autres types. Cette incertitude n'a pas de conséquences importantes pour la topologie globale du réseau phylogénique.

Un dernier cas litigieux se présente pour le rattachement des types 11 et 58. Les types 11 et 23 partagent une mutation bien caractérisée au niveau moléculaire (site 68/69), alors que les types 11 et 21 partageraient une mutation d'un site (37) dont nous avons vu que la définition moléculaire était très imprécise (plusieurs sites potentiels sont possibles et plusieurs évènements moléculaires peuvent être à la source du gain commun de ce site). Aussi, nous privilégierons la liaison du type 11 au type 23 par un intermédiaire non identifié pour l'instant.

Détermination de la racine hypothétique de la phylogénie

Afin de devenir une véritable phylogénie, notre réseau a besoin d'être polarisé en lui trouvant une racine. Il faut donc définir un type possédant les κ sites de restriction polymorphes dans leur état originel. Il est tout d'abord très vraisemblable de pouvoir considérer les sites monomorphes comme étant dans leur état ancestral. Par simple extension de cette hypothèse, nous pouvons postuler que l'état ancestral d'un site polymorphe est simplement l'état le plus fréquent parmi les M différents types trouvés, sans se préoccuper de leurs fréquences. Formellement, on considérera un type comme un vecteur booléen de 1 et de 0 pour chaque site, ,dépendant si celui-ci est présent ou non. L'état ancestral du site i sera déterminé après l'addition de l'état de chaque site i chez les M types. Nous obtiendrons un nombre S_i compris entre 1 et M-1. L'état ancestral sera considéré comme

0 si $S_i < M/2$

1 si $S_i > M/2$

indéterminé si $S_i = M/2$

Un type ancestral hypothétique est obtenu en répétant le processus pour les κ sites polymorphes. De cette manière nous trouvons que le type ancestral hypothétique reconstruit est équivalent au type 1

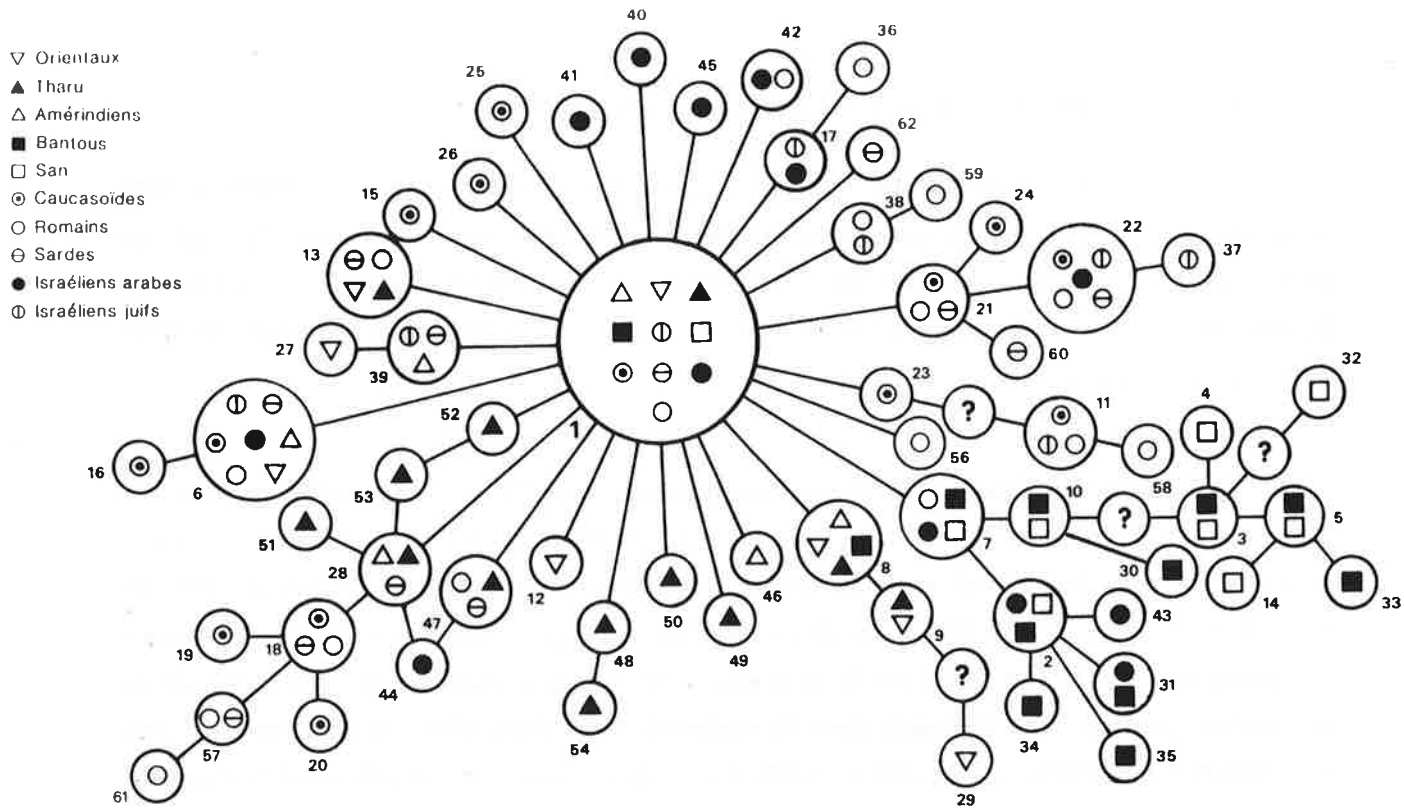

FIGURE 4.11 : Réseau phylogénique probable des 61 types d'ADN-mt définis dans 10 populations. La taille des cercles représentant les types est proportionnelle au nombre de populations où on les retrouve. Les liaisons des types 44 et 53 ne sont pas établies avec certitude, d'où la notion de réseau phylogénique au lieu d'une simple phylogénie. Ce réseau n'est pas définitif, car l'étude de nouvelles populations pourrait remettre en cause certaines liaisons actuelles entre types.

Rapports entre les groupes continentaux sur la base des types ancestraux

Nous définirons des types comme "ancestraux", si l'on a des raisons de penser qu'ils étaient présents dans des populations à partir desquelles auraient divergé les grands groupes continentaux africains, orientaux et occidentaux. Le nombre limité des échantillons ne nous permet malheureusement pas de distinguer des rapports existants entre des sous-groupes plus précis.

Hormis le type 1 que nous considérerons comme primitif (sensu stricto), les meilleurs candidats pour être des types ancestraux sont ceux qui sont retrouvés dans des populations actuelles, et qui proviennent de groupes continentaux différents, comme les types 2, 6, 7, 8, 13, 28, 39 et 47. La détermination du type 1 comme ancestral et central est confortée par le fait qu'il est le seul que l'on retrouve dans toutes les populations recensées, qu'il est très fréquent dans la majorité des échantillons et qu'il semble être directement à l'origine de 26 autres types. Les autres types potentiellement ancestraux sont d'ailleurs tous situés à une substitution près du type 1 (à l'exception des types 2 et 31). La perception d'une radiation de tous les types à partir du type 1, déjà sensible au niveau de la Figure 4.10, a été accentuée dans la construction phylogénique de la Figure 4.11. On peut donc supposer que certaines populations anciennes étaient déjà polymorphes et comprenaient des types proches du type 1.

La centralité du type 1 pour les populations africaines n'est pas soutenue au niveau des fréquences des types, mais est confortée par la présence, dans l'échantillon Bantou, du type potentiellement ancestral 8 qui dérive également du type 1. Selon cette vue centrifuge, la différenciation de la majorité des types du groupe continental africain se serait faite ensuite à partir du type 7. Quelques considérations sont nécessaires à ce propos. Le groupe africain est le seul qui présente une structure hiérarchisée aussi claire entre les types et une différenciation aussi poussée à partir du type 1. Ce phénomène n'est pas perceptible pour les groupes orientaux et occidentaux où certains types ont parfois comme ancêtre probable des types provenant d'autres groupes continentaux (liaison 18-28 ou 27-39), et où la majorité des types s'est différenciée autour du type 1, à quelques exceptions près qu'il sera intéressant de discuter.

La nette séparation apparente des types africains des autres branches de la phylogénie est quelque peu tempérée si l'on songe à la nature des échantillons représentant le stock génétique africain. Les populations Bantoues sont connues pour

avoir migré relativement récemment (1000 BP) à partir de la frontière Nigéria-Cameroun (Philipson, 1980; Pellegrini, 1987) et elles se seraient différenciées progressivement, tant du point de vue linguistique que du point de vue génétique, des populations d'Afrique de l'Ouest (Excoffier *et al.*, 1988). L'origine des San serait plus ancienne et ils auraient été l'objet d'un effet fondateur précoce à partir d'un stock génétique africain encore peu différencié (Excoffier *et al.*, 1987). Ils auraient ensuite développé leurs propres particularités. Ces 2 populations ont donc été rapidement isolées de tout contact avec d'autres groupes continentaux et ne sont pas représentatives de l'ensemble du groupe africain. La branche des types "africains" pourrait être coupée en deux groupes, l'un comportant la présence des sites *Msp I* 35 et 36 et l'autre où ils sont absents. Ce dernier groupe serait retrouvé principalement chez les San, comme nous l'avons déjà suggéré lors de la discussion des morphes *Msp I*, et correspondrait au groupe 3-4-5-14-32-33, qui est le plus différencié à partir du type 1, et qui serait donc issu du groupe où ces sites sont présents et qui est évolutivement plus proche du type 1. Deux autres arguments vont également dans le sens d'une différenciation des types "San" à partir des autres types africains. Le premier est la constatation déjà formulée que des individus d'origine africaine, mais résidant aux Etats-Unis, possèdent effectivement ces 2 sites, et seraient donc succeptibles de correspondre aux types "africains" proches du type 1 ou même d'être principalement du type 1. Le deuxième argument est la présence des types "africains" 2, 7 et 31 dans l'échantillon d'Arabes Israéliens, suggérant ainsi un lien ancien entre les groupes occidentaux et africains à proximité de cette région du monde, qui était en fait un passage obligé pour entrer ou sortir d'Afrique. On peut donc s'attendre à ce que d'autres populations africaines possèdent des types proches des types 1, 2, 7 et 10 plutôt que des type 3 et 5. De nouvelles études sont nécessaires pour aller plus avant dans la connaissance de la diversité africaine.

Le groupe oriental, constitué ici des échantillons Orientaux, Tharu et Amérindiens est fortement centré autour du type 1 qui y dépasse partout 50% et atteint même plus de 90% chez les Amérindiens. Aucun type ne diffère d'un autre par plus de 4 substitutions, et ne diffère du type 1 de plus de 2 substitutions, à l'exception du type 29 qui est à rattacher au type 9 par un intermédiaire absent. La branche 8-9-?-29 est aussi singulière par le fait que le type 8 est également retrouvé dans l'échantillon Bantou et qu'elle comprend le morphe *Hpa I* 1 reconnu chez l'Orang-Outan (Denaro *et al.*, 1981). Les autres types potentiellement ancestraux (6, 13, 28, 39 et 47) sont tous partagés avec des populations occidentales, ce qui est indicateur de liens anciens et étroits entre ces 2 groupes de populations.

Le groupe occidental, composé des échantillons Caucasoïdes, Romains, Sardes, Israéliens Juifs et Arabes semble être celui dont la diversité a été la mieux étudiée. Comme dans le cas du groupe oriental, la diversité de ses types y est centrée autour du type 1 qui dépasse souvent des fréquences de l'ordre de 50 à 60 %. Par contre, certains types sont présents sur des branches progressivement bien différenciées à partir du type 1 ou de types ancestraux communs avec le groupe oriental, comme les branches 28-18-19-20-57-61, 21-24-60-22-37, 23-?-11-58 et même la branche 7-2-31-43. Ainsi, certains types sont distants les uns des autres par au moins 9 substitutions. Donc, si le type 1 est toujours central au groupe occidental, une portion non négligeable de la diversité des types s'est apparemment développée à partir de types périphériques.

Cette plus grande diversité des types peut évidemment être une conséquence du meilleur échantillonnage du groupe occidental, mais elle pourrait aussi refléter une plus grande ancienneté de ce groupe par rapport aux 2 autres. Ce dernier point est soutenu par l'observation que tous les types potentiellement ancestraux sont présents dans le groupe occidental, à l'exception du type 8. D'autre part, certains types retrouvés uniquement dans ce groupe sont également potentiellement anciens, comme le type 22 qui atteint environ 10% au Moyen-Orient et le type 18 qui atteint des fréquences du même ordre en Italie et qui se trouve être à la source directe ou indirecte de 4 autres types.

Le groupe occidental présente une diversité telle qu'elle pourrait bien être à l'origine de celle des 2 autres groupes et notamment du groupe africain dont certains types importants sont retrouvés au Moyen-Orient.

A ce stade de nos connaissances et de l'échantillonnage d'un nombre limité de populations, il est tentant de proposer une constitution génétique hypothétique pour une population primitive de laquelle les populations actuelles seraient dérivées. Une telle population aurait pu comprendre les types ancestraux et anciens que nous avons définis (1, 2, 6, 7, 8, 13, 18, 22, 28, 31, 39 et 47). Celle-ci serait proche des populations occidentales actuelles et notamment des populations du Moyen-Orient qui possèdent encore les types 1, 2, 6, 7, 22 et 39, mais également le type 44 qui ne peut être issu que des types ancestraux 28 ou 47 (voir Figure 4.11). Par contre, il semble peu probable au vu de la constitution des échantillons à disposition que ce soient les populations africaines qui aient pu engendrer la diversité des groupes occidentaux ou orientaux, comme cela a été proposé dans le cas de l'ADN-mt (Johnson *et al.*, 1983; Cann *et al.*, 1987) ou de l'ADN nucléaire (Wainscoat *et al.*, 1986).

TABLE 4.19: Nombre de substitutions entre types d'ADN-mt calculé à partir du réseau phylogénique de la Figure 4.11

Types	*Distances*
2	*2*
3	*44*
4	*551*
5	*5512*
6	*13566*
7	*113442*
8	*1356622*
9	*24677331*
10	*222333134*
11	*3578844455*
12	*135662223334*
13	*1356622233342*
14	*6623175784977*
15	*135662223342277*
16	*2467713344533883*
17	*13566222334227223*
18	*24677333445338343*
19	*357884445564494541*
20	*3578844455644945412*
21	*135662223342272323344*
22	*246773334453383434551*
23	*1356622233222723234423*
24	*24677333445338343455123*
25	*135662223322272323442323*
26	*1356622233222723234423232*
27	*24677333445338343455343533*
28	*135662223342272321222323223*
29	*4789955326755105656775656565565*
30	*333444245164454545664545445547*
31	*315664245364474545664545445474*
32	*662337578497747878997878778710 57*
33	*66231757849772787899787878710574*
34	*31566424536447454566454544547427 7*
35	*315664245364474545664545445474277 2*
36	*24677333446448341455343433436558855*
37	*357884445564494565662143445476699665*
38	*135662223342272323442323223225447744 34*
39	*1356622233422723234423232232544774434 2*
40	*13566222334227232344232322325447744342 2*
41	*135662223342272323442323223254477443422 2*
42	*1356622233422723234423232232544774434222 2*
43	*31566424536447454566454544547427722564444 4*
44	*24677333445338343233343433416558855453333335*
45	*135662223342272323442323223254477443422222 43*
46	*1356622233422723234423232232544774434222224 32*
47	*13566222334227232344232322325447744342222243 22*
48	*1356622233422723234423232232544774434222243222*
49	*13566222334227232344232322325447744342222243222 2*
50	*135662223342272323442323223254477443422222432222 2*
51	*24677333445338343233343433436558855453333352333333*
52	*1356622233422723234423232232544774434222224322222 23*
53	*246773334453383432333434334365588554533333523333321*
54	*24677333445338343455343433436558855433333356433313343 4*
55	*135662223342272323442123223254477443222222432222223233*
56	*1356622233422723234423232232544774434222224322222232332*
57	*3578844455644945412245454452766996656444446344444434354 4*
58	*468995556615510565677563655658771010776755555765555556566557*
59	*24677333445338343455343433436558855431333564333333434433356*
60	*246773334453383434551232344365588554333333343333334344335 64*
61	*4689955566755105652335656556387710107767555557444444434354418 48*
62	*13566222334227232344232322325445744342222243222222323322453 35*

Mesure de la diversité moléculaire des échantillons sur la base des sites de restriction

La dérivation d'un réseau phylogénique entre les types d'ADN-mt nous permet de calculer le nombre d'évènements mutationnels minimum qui sépare 2 types d'ADN-mt pris au hasard. Cette mesure peut parfois être très éloignée du nombre net de différences de sites de restriction entre 2 types, de la même manière que le nombre de substitutions de nucléotides entre 2 séquences d'ADN différait du nombre observé de nucléotides dissemblables (voir chapitre sur l'étude des séquences d'ADN-mt). Comme nous avons défini un réseau phylogénique et non un arbre, cet indice de dissimilarité entre 2 types pourrait sous-estimer le nombre réel de différences de sites de restriction entre 2 types i et j que l'on notera par v_{ij}. La matrice des v_{ij} estimés est reportée dans la Table 4.19.

Variation génétique intrapopulation

A partir de l'indice v_{ij}, Nei et Tajima (1981) ont derivé une mesure du nombre moyen de différences de sites de restriction à l'intérieur d'une population entre 2 types pris au hasard. Cette quantité (v) est définie par

$$v = \sum_i \sum_j p_i p_j v_{ij} \qquad \text{(Nei and Tajima, 1981) (4.11)}$$

où p_i et p_j sont les fréquences des types i et j dans la population. Un estimateur non-biaisé de v sur le plan de l'échantillonnage sera donc

$$\hat{v} = \frac{n}{n-1} \sum_i \sum_j x_i x_j v_{ij} \qquad \text{(Nei and Tajima, 1981) (4.12)}$$

et les mêmes auteurs ont déterminé sa variance due à l'échantillonnage par

$$\hat{V(v)} = \frac{4}{n(n-1)} [(6-4n) (\sum_{i<j} p_i p_j v_{ij})^2 + (n-2) \Sigma p_i p_j p_k v_{ij} v_{ik} + \sum_{i<j} p_i p_j v_{ij}^2] \quad (4.13)$$

où n représente le nombre de gènes dans l'échantillon et x_i la fréquence du type i dans l'échantillon. Selon des arguments développés par Nei et Li (1979) ainsi que par Nei et Tajima (1981), la moyenne et la variance de v peuvent être étudiés par le modèle des sites infinis (Watterson, 1975), ceci lorsque l'échantillon est grand et à la condition que les mutations parallèles soient négligeables et que la séquence étudiée soit neutre. Dans ces conditions,

$$E(v) = \theta,$$
$$V(v) = \theta + \theta^2 \qquad \text{(Watterson, 1975) (4.14)}$$

où $\theta = 4N_e u$ dans le cas de l'ADN nucléaire et Nu dans le cas de l'ADN-mt (N_e représente la taille effective de la population et u le taux de mutation de la séquence en question). Une autre estimation de θ peut être obtenue en utilisant le nombre de sites de restriction polymorphes (κ) dans un échantillon de n gènes, grâce au modèle des sites infinis développé par Watterson (1975) qui n'utilise pas de fréquences géniques, ce qui le rend donc moins sensible à leurs variations, aléatoires ou non. Ainsi, nous avons

$$\theta = \kappa / (\sum_{i=1}^{n-1} \frac{1}{i}), \qquad \text{(Watterson, 1975) (4.15)}$$

et en première approximation,

$$V(\theta) = \theta \log_e(n) \qquad \text{(Ewens, 1983 d'après Watterson, 1975) (4.16)}$$

La comparaison des valeurs de $E(v)$ et de θ obtenus par les formules (4.12) et (4.15) est intéressante, car elle peut permettre de situer la validité des hypothèses de neutralité et/ou d'absence d'évolution parallèle des sites. Cependant, aucun test statistique n'a pu être développé pour la comparaison de ces 2 valeurs, étant donné la complexité de la distribution de ces variables aléatoires. Un tel test se rapprocherait sans doute de celui développé par Watterson (1978) qui définit un intervalle de confiance pour l'homozygosité en fonction du nombre d'allèles et de la taille d'un échantillon (voir Annexe B).

Dans la Table 4.20, nous avons reporté les estimations de θ obtenues par les 2 méthodes. Les chiffres entre parenthèses de la Table 4.20 représentent les estimations de v obtenues par Johnson *et al.* (1983) au moyen de la formule (4.12) pour quelques populations. La différence entre ces chiffres et les nôtres provient du fait que notre phylogénie des types est différente de la leur, ce qui consiste parfois à diminuer la diversité moléculaire de l'échantillon Caucasoïde et à augmenter celle des échantillons africains.

TABLE 4.20 : Comparaison entre le nombre moyen de différences de sites de restriction et son espérance.

Populations	Nb gènes	Nb. sites polym.	$\hat{v}$ [1] (4.12)	θ (4.15)
Caucasoïdes	50	13	1,32 *(1,64)*	2,90
Romains	95	14	1,13	2,73
Sardes	134	10	0,78	1,83
Isr. J.	39	10	1,39	2,37
Isr. A.	39	10	1,46	2,37
Bantou	40	9	2,14 *(1,83)*	2,12
San	34	8	1,96 *(1,66)*	1,96
Orientaux	46	8	0,89 *(0,89)*	1,82
Tharu	91	11	0,77	2,16
Amérind.	74	4	0,19	0,82

[1] Ces estimations ont été calculées à partir des v_{ij} définis dans la Table 4.19, à l'exception des nombres entre parenthèses qui correspondent aux estimations de v_{ij} par Johnson *et al.* (1983) effectuées à partir de leur propre phylogénie des types.

D'une manière générale, l'estimation de θ obtenue par la formule (4.15) produit des valeurs très supérieures à celles de (4.12), à l'exception des populations Bantou et San où ces 2 valeurs sont très similaires. Ce dernier fait semble suggérer que notre réseau phylogénique concernant ces 2 échantillons serait correct. Les différences considérables -d'un facteur 2 environ- observées pour les autres populations indiquent soit que le réseau sous-estime fortement le nombre de mutations parallèles, soit que les fréquences géniques perturbent une estimation correcte de θ par (4.12).

Nous observons donc une nouvelle fois un clivage entre les populations africaines et les autres populations. Les causes de cette séparation qualitative des échantillons pourraient être les mêmes que précédemment: la fréquence très élevée du type 1 dans les échantillons non-africains tendrait à diminuer fortement la moyenne du nombre de différences de sites entre 2 gènes tirés au hasard, étant donné que ce tirage fournit souvent des gène identiques. Les échantillons africains, bien que de tailles restreintes, semblent se comporter conformément à la théorie du point de vue de leur diversité moléculaire, alors que les autres échantillons seraient biaisés dans le sens d'une trop grande homogénéité moléculaire.

Cette constatation diffère radicalement de la vue obtenue par l'étude des fréquences des types par les analyses multivariées où l'on avait l'impression que les populations San et Bantou constituaient des cas particuliers. Il semblerait plutôt que ce soient l'ensemble des autres populations qui serait sous-diversifiées sur le plan moléculaire. Les causes de cette anomalie restent à préciser.

Une hypothèse alternative serait que les formules (4.12) et (4.15) fournissent des estimateurs différents pour toutes les populations et que la concordance observée pour les Bantou et les San serait due au seul hasard. Cette hypothèse n'obtient pas notre aval, en raison, d'une part, du comportement effectivement différent des échantillons africains du point de vue des fréquences géniques et des types présents, et d'autre part, en vertu d'une trop bonne concordance entre v et E(v) pour ces 2 échantillons qui ne semble guère fortuite. Nous regrettons encore une fois l'absence de test statistique qui pourrait permettre de trancher plus clairement entre ces différentes hypothèses.

Diversité nucléotidique

La mesure du nombre moyen de différences de sites de restrictions va, bien évidemment, dépendre de la longueur de la séquence étudiée et du nombre d'enzymes employés, si bien que les valeurs générées pour différentes études ne seront pas directement comparables. Il est possible de contourner cette difficulté en ramenant l'étude de la diversité moléculaire au niveau du nucléotide en formulant plusieurs hypothèses: les fréquences des 4 nucléotides A, C, G et T sont supposées être en équilibre dans la séquence et au cours du temps; les nucléotides sont répartis aléatoirement dans la séquence étudiée; les changements de sites de restriction sont dus à des substitutions; les taux de substitutions sont identiques pour les 4 nucléotides; les substitutions multiples à un site donné sont rares. L'estimation de la diversité nucléotidique (π) qui représente en fait la probabilité d'hétérozygosité à une position de nucléotide a été donné par Nei et Tajima (1981) comme

$$\pi = n/(n\text{-}1)\ \Sigma_{i \neq j} x_i x_j \pi_{ij} \qquad (4.17)$$

où x_i et x_j sont les fréquences des types i et j dans l'échantillon et π_{ij} représente le nombre de différences de nucléotides entre les types i et j. On notera l'analogie avec (4.12). La variance de π due au processus d'échantillonnage est donnée par (4.13) en remplaçant v_{ij} par π_{ij}. D'une façon commode, π peut être estimée directement à partir de l'estimation de v en posant

$$\pi_1 = v / R \qquad \text{(Nei, 1987)} \quad (4.18)$$

où R est le nombre de nucléotides effectivement surveillées par les divers enzymes de restriction, et j est le nombre d'enzymes employés, m_i est le nombre de sites de restrictions et r_i le nombre de nucléotides dans la séquence de reconnaissance. En fait, cette quantité est fonction du nombre de sites détectés, mais également du nombre de sites potentiels différant d'une séquence de reconnaissance par un seul nucléotide (Johnson et al., 1983). Celle-ci tend à long terme vers,

$$R' = a_0 m_T r' + a_1 m_T/3 \qquad (4.19)$$

où r' représente le nombre effectif de nucléotides compris dans une séquence de reconnaissance (Nei and Tajima, 1981), m_T est le nombre de nucléotides de l'ADN-mt, a_0 est donné par l'équation (4.3) et a_1 représente le nombre de sites potentiels qui est donné par

$$a_1 = a_0 \sum_{i=1}^{r} (1-f_i)/f_i \qquad (4.20)$$

où f_i est donné par l'équation située après (4.3). Cette dernière formule a également été trouvée sous une autre forme par Nei et Tajima (1983, Appendix). Le facteur 1/3 de l'équation (4.19) provient du fait qu'une seule substitution sur 3 par site donnera un gain de site et sera donc perceptible. On peut donc reformuler la quantité R' pour plusieurs enzymes comme

$$R' = 2m_T \sum_{i=1}^{h} (a_{0i} r_i' + a_{1i}/3), \qquad (4.21)$$

où h représente le nombre d'enzymes utilisés. Il faut noter que plusieurs auteurs (Engels, 1981; Ewens, 1981; Hudson, 1982) ont proposés d'autres méthodes pour estimer la probabilité d'hététozygosité à un site quelconque. La principale différence entre ces 2 types d'estimation réside dans le fait que le premier (Nei and Tajima, 1981) mesure la variabilité génétique de l'échantillon à un moment précis, alors que les autres méthodes s'intéressent plutôt à le mesure de la variabilité génétique à long terme, sans souci de l'hétérozygosité de la population actuelle, qui est sujette à de multiples variations aléatoires (voir Primard (1985) pour une revue des différentes méthodes et surtout Ewens (1983) pour une discussion des différentes propriétés des estimateurs).

Cette remarque étant faite, π_1 peut donc être estimé par v/R'. Si l'on utilise le modèle des sites infinis de la formule (4.15) on peut obtenir une autre estimation de π par

$$\pi_2 = \kappa / [(\sum_{i=1}^{n-1} \frac{1}{i}) R']) = \theta/R' \qquad \text{(Nei, 1987) (4.22)}$$

qui donnera une estimation de π indépendante des fréquences des types dans l'échantillon qui sera représentative d'un processus à long terme et, donc, moins sujette aux conditions du moment. A partir de l'équation (3.3), on obtient le nombre attendu de substitutions de nucléotide par site comme

$$s = -3/4 \log_e(1-4\pi/3) \qquad \text{(Nei and Tajima, 1983) (4.23)}$$

Comme $s=2\lambda t$, il est aisé de déterminer le temps nécessaire pour créer une telle diversité si l'on a une estimation de s pour un temps de divergence connu. Brown *et al.* (1982) ont estimé s, en admettant que la séparation entre la lignée humaine et celle du chimpanzé date de 5 millions d'années, comme étant égal à 2 % par million d'année de divergence entre les lignages (espèces), ce qui correspond à 1% par million d'années par lignage. Il est important de mentionner que le temps défini ne correspond pas à un temps de divergence pour une certaine population, mais au temps nécessaire pour qu'une diversité nucléotidique π_1 ou π_2 se développe à partir d'une population monomorphe. Pour des populations ayant connu de sévères "bottlenecks", cette valeur peut approcher le temps de divergence à partir d'une population mère. Inversément, si une population s'est implantée quelque part en étant déjà polymorphe, le temps t sera supérieur au temps de divergence. Nous estimerons donc 2 temps de divergence t_1 et t_2 qui correspondent respectivement aux estimations de π_1 et π_2 des équations (4.17) et (4.22).

Nous avons reporté dans la Table 4.21 les résultats de nos estimations de t_1 et t_2 après avoir calculé une valeur de R' égale à 618,4. Comme les valeurs de t sont directement proportionnelles aux valeurs des différences moyennes de sites de restriction entre types reportés dans la Table 4.20, nous retrouvons ici des rapports entre t_1 et t_2 identiques à ceux que nous avions indiqué entre v et θ. Cependant, le fait d'avoir affaire à des estimations temporelles nous amène à de nouveaux commentaires. Les échantillons recensés semblant relativement homogènes, les estimations de t_1 et t_2 devraient avoir tendance à surestimer les temps de divergence entre ces populations et des populations ancestrales. D'une manière générale, les estimations de t_1 et t_2 dépassent souvent nettement les valeurs admises actuellement pour les datations

concernants l'apparition des premiers hommes modernes. Celles.ci se situeraient aux alentours de 150-100'000 ans dans une région regroupant l'Afrique et le Moyen-Orient (Delson, 1988; Stringer, 1988; Stringer and Andrews, 1988; Valladas *et al.*, 1988). Ceci peut être dû à au moins 2 phénomènes qui tendent à augmenter l'écart entre les temps de divergence réels et les temps de différenciation. D'une part, ces populations n'ont probablement pas été totalement isolées depuis leur divergence et ont échangé des gènes avec des populations voisines. D'autre part, l'âge des gènes peut largement excéder l'âge d'une population particulière où même d'une espèce (Takahata and Nei, 1985) et ceci à plus forte raison s'ils sont déjà polymorphes au moment du processus de divergence.

TABLE 4.21: Estimation du temps t (en années) nécessaire pour créer une diversité nucléotidique Π à partir d'une population monomorphe.

Population	Π_1 ($\times 10^4$) (4.18)	Π_2 ($\times 10^4$) (4.22)	t_1	t_2
Caucasoïdes	21,34	46,90	213'704	470'473
Romains	18,27	44,16	182'923	442'905
Sardes	12,61	29,60	126'206	296'586
Isr. J.	22,48	38,34	225'138	384'383
Isr. A.	23,62	38,34	236'573	383'383
Bantous	34,62	34,30	347'001	343'787
San	31,70	31,70	317'672	317'672
Orientaux	14,38	29,44	143'938	294'979
Tharu	12,46	34,94	124'704	350'216
Amér.	3,08	13,28	30'806	132'717

Les estimations de t_2 étant peu dépendantes de phénomènes perturbateurs des fréquences géniques, elles apparaissent plus indicatrice que t_1 de la diversité moléculaire réelle des populations actuelles. Elles montrent que plus de 400'000 ans auraient été nécessaires pour obtenir la diversité de certaines populations caucasoïdes. Cela suggère que certains types retrouvés actuellement existaient probablement avant l'apparition des premiers hommes modernes et que ces derniers ont hérités d'un patrimoine génétique déjà polymorphe.

Un autre aspect de ces divergences de populations à l'état polymorphe est retrouvé dans le cas des Amérindiens. Bien que l'on suppose qu'ils aient été sujets d'un fort effet fondateur lors du peuplement de l'Amérique que l'on situe à plus de 30'000 ans (Dillelay and Collins, 1988; Guidon and Delibrias, 1986), une grande partie de leur différenciation génétique pourrait remonter bien au delà. Ceci est confirmé lorsque l'on remarque que 4 types sur 5 sont retrouvés dans d'autres groupes continentaux (types 1,

6, 8 et 39), ce qui montre que le polymorphisme de la ou les population(s) fondatrice(s) s'est propagé jusqu'à aujourd'hui.

Il est intéressant de constater que les datations produites par t_2 sont en accord avec l'interprétation que nous avons tirée de la Figure 4.11, avec une diversification progressive à partir du type 1. Ceci confirmerait la validité de cet estimateur face à t_1 et nous incite à l'utiliser pour dresser des hypothèses sur l'ancienneté relative de certaines populations. Au vu de ces 2 sources d'information (phylogénie des types et datations), certaines étapes de l'histoire du peuplement humain semblent pouvoir être précisées. Les populations occidentales possèdent des temps de diversification moléculaire plus élevés que les autres groupes de populations, ainsi qu'une variété de types telle qu'elles pourraient être directement issues d'une population ancestrale d'hommes modernes qui se serait différenciée il y a environ 150-100'000 ans.

Les 2 populations africaines recensées ont développé leur diversité moléculaire actuelle dans un temps plus restreint, ce qui suggère qu'elles ont entamé un processus de différenciation postérieurement, à partir d'autres populations, africaines ou non, possédant certains types encore retrouvés chez les populations du Moyen-Orient. Le fait que plusieurs types soient partagés entre les populations africaines et celles du Moyen-Orient suggère que les premières n'ont pas évolué à partir de populations monomorphes et donc que le temps de divergence entre ces populations serait bien inférieur à 300'000 ans et daterait probablement d'après l'apparition des premières formes d'hommes modernes (un raisonnement identique s'applique aux temps de divergence entre les groupes qui possèdent des types communs). Un effet fondateur a également pu bouleverser considérablement les fréquences géniques des populations migrantes et éliminer certains gènes.

Les populations orientales du continent asiatique ont apparemment développé leur diversité moléculaire dans une période comparable aux populations africaines, mais d'une toute autre manière, c'est à dire principalement autour du type 1, ce qui semble indiquer que le processus de différenciation n'a pas débuté par un effet fondateur important. Les Amérindiens, par contre, semblent avoir perdu une bonne partie de leurs gènes tout en restant polymorphes lors du processus de peuplement de l'Amérique qui daterait d'environ 30'000 ans (Dillelay and Collins, 1988; Guidon and Delibrias, 1986).

Variation génétique inter-population

L'introduction de connaissances sur la nature exacte des différences moléculaires existant entre les types d'ADN-mt peut donc apporter une information sur la nature et la qualité des divergences entre des populations. Ceci serait plus délicat dans le cas d'un polymorphisme non défini au niveau moléculaire. La diversité moléculaire entre 2 populations peut également être estimée par une méthode comparable à l'étude de la diversité intrapopulation. Nei et Tajima (1979) ont ainsi essayé de mesurer la différence moyenne de sites de restrictions entre 2 populations X et Y par une quantité v_{XY} définie par

$$v_{XY} = \sum_{ij} p_i q_j v_{ij} \qquad (4.24)$$

où $v_{ii} = 0$, p_i et q_j sont respectivement les fréquences des types i et j dans les populations X et Y. Cette valeur peut être estimée par

$$\hat{v}_{XY} = \sum_i \sum_j x_i x_j v_{ij}, \qquad \text{(Nei and Tajima, 1981) (4.25)}$$

où x_i et x_j représentent les fréquences des types dans les échantillons cette fois-ci. Contrairement à la formule équivalente (4.12), qui introduisait une correction pour tenir compte de la variabilité de l'estimation des fréquences géniques dans les populations à partir des échantillons, cette formule n'en introduit aucune et pourrait donc être biaisée, surtout lorsque les tailles des échantillons sont faibles.

Afin de tenter de tenir compte de la variabilité moléculaire existant à l'intérieur de chaque population, Nei et Tajima (1981) ont proposé de calculer le nombre *net* (d_{XY}) de différences de sites de restriction entre 2 populations X et Y comme

$$\hat{d}_{XY} = \hat{v}_{XY} - (v_X + v_Y)/2 \qquad (4.26)$$

où v_X et v_Y sont donnés par (4.12). Il est clair que d_{XY} peut prendre des valeurs négatives lorsque les populations X et Y sont évolutivement proches et lorsque la taille des échantillons est faible. Il semble donc que cet estimateur ne soit guère approprié pour mesurer des distances génétiques lorsque des populations ont divergé depuis peu. Néanmoins, nous le présentons ici, d'une part afin de le critiquer, et d'autre part il est encore couramment employé pour estimer la divergence de populations sur la base des sites de restriction.

Dans la Table 4.22, nous avons reporté les différentes estimations de d_{XY} et v_{XY}, calculées au moyen des formules (4.25) et (4.26). Une analyse multivariée de ces matrices de d'indices de dissimilarité entre populations (résultats non présentés ici) fournit essentiellement le même type de représentation que la Figure 4.7, à savoir une nette différenciation des échantillons africains par rapport au reste des autres populations. Cependant, il est clair que ces estimations, fortement dépendantes des valeurs des v_X et des v_Y, sont certainement soumises au même type de biais. Pour ces raisons, il nous semble délicat de vouloir tirer des scenari de la divergence de ces populations à partir des d_{XY}, comme cela l'a été fait auparavant (Johnson et al., 1983), et à plus forte raison d'en tirer un arbre évolutif ou une estimation du temps de divergence entre ces populations.

TABLE 4.22 : Différences[1] moyennes de sites de restriction entre populations

				Populations					
Cauc.	Rom.	Sard.	Isr. J.	Isr. A.	Bant.	San	Orient.	Tharu	Amér.
	1,250	*1,110*	*1,413*	*1,460*	*2,475*	*4,876*	*1,170*	*1,136*	*0,787*
0,016		*0,996*	*1,396*	*1,386*	*2,381*	*4,777*	*1,088*	*1,042*	*0,712*
0,024	-0,002		1,259	1,230	2,230	4,631	0,925	0,889	0,547
0,055	0,126	0,137		*1,544*	*2,621*	*5,023*	*1,311*	*1,287*	*0,921*
0,068	0,082	0,073	0,115		*2,306*	*4,716*	*1,271*	*1,231*	*0,888*
0,742	0,736	0,733	0,852	0,502		*3,778*	*2,245*	*2,166*	*1,868*
3,233	3,222	3,224	3,343	3,003	1,723		*4,655*	*4,465*	*4,271*
0,063	0,069	0,054	0,168	0,094	0,726	3,227		*0,894*	*0,568*
0,089	0,083	0,077	0,204	0,113	0,707	3,096	0,061		*0,562*
0,033	0,046	0,029	0,131	0,064	0,703	3,196	0,028	0,082	

[1] Les chiffres en italique en dessus de la diagonale représentent les estimations de v_{XY}. Les estimations de d_{XY} sont représentées en dessous de la diagonale.

Analyse qualitative des substitutions ayant conduit à des gains de site

Lorsqu'une substitution provoque l'apparition d'un site de restriction à partir d'un site potentiel, il est en général possible de déterminer la nature de la mutation si l'on connait la séquence d'ADN-mt non-mutante (dans notre cas la séquence de Cambridge) et la séquence de reconnaissance de l'enzyme concerné. Parmi les 68 sites de restriction recensés (voir Table 4.3), 37 nucléotides distincts appartenant à 35 sites de restriction ont été l'objet d'une ou plusieurs mutation. Nous avons pu ainsi observer la perte de 11 sites et le gain de 26 autres par rapport à la séquence de Cambridge correspondant au type 1.

Gènes codant pour des protéines

Dans la Table 4.23 sont reportées les caractéristiques des substitutions ayant abouti à des gains de sites. La plupart de ces gains se situent dans des régions codant pour des protéines. Certains sites ont été localisés approximativement et plusieurs substitutions peuvent être à la source du gain observé. C'est notamment le cas des sites 39 et 51. Les 2 possibilités de substitution du site 39 ne provoquent aucun changement d'acide aminé dans le gène CO III. Pour le site 51, une transition A ⇒ G à la position 12819 est phénotypiquement silencieuse, alors que la transition C ⇒ G à la position 12817 provoquerait un changement Arg → Gly. Nous postulerons donc que la substitution s'est produite à la position 12819 et qu'elle n'a pas provoqué de modifications de la structure protéique.

Sur les 19 sites observés (en incluant les sites 33 et 64), 11 ne provoquent pas de changement phénotypique et 8 induisent une modification de la structure primaire de la protéine. Toutes les mutations se produisant à la 3ème position des codons sont silencieuses. En admettant que les mutations se produisent au hasard du point de vue du cadre de lecture, on devrait s'attendre à trouver des mutations sur les 3 positions du codon à une fréquence équivalente. Or, nous les trouvons selon les proportions respectives de 5:3:11. Cela est en accord avec les proportions trouvées pour d'autres protéines d'origine nucléaire chez les mammifères (voir la Table 4.8 dans Kimura, 1983; Nei *et al.*, 1984), où les substitutions sont les plus fréquentes sur la 3ème position du codon, suivie par la 1ère position et enfin par la 2ème. Tout en ayant conscience des problèmes d'échantillonnage des mutations, on peut estimer grossièrement que plus de la moitié des substitutions se produisant sur la 1ère ou la 2ème position des codons ne sont

pas retrouvées au niveau des types d'ADN-mt. Il est probable que les types portant ces mutations ont été éliminés ou maintenus à des fréquences très faibles. Ceci confirme aussi la prépondérance des substitutions silencieuses dans les gènes mitochondriaux, déjà observée par Miyata *et al.* (1982) qui mesuraient un taux de substitutions silencieuses 6 fois plus élevé pour l'ADN-mt que pour l'ADN nucléaire.

TABLE 4.23 : Répercussion phénotypique des mutations ayant conduit à un gain de site.

N° du site	Enzyme	Position de la mutation[1]	Région[2]	Cadre de lecture	Substitution	Modif. a. a.	Effet sur phénotype	Nombre d'occurrence[3]
4	*Hpa I*	1118	ARN 12 S	-	A ⇒ C	-	oui	1
6	*Hpa I*	2157	ARN 16 S	-	T ⇒ G	-	oui	1
8 ou 64	*Hae II*	2540	ARN 16 S	-	C ⇒ G	-	oui	1
12 ou 33	*Ava II*	3135	ARN 16 S	-	A ⇒ C	-	oui	1
15	*Hpa I*	3594	NAD 1	3	C ⇒ T	Val → Val	silence	2
16	*Ava II*	3881	NAD 1	2	A ⇒ G	Glu → Gly	oui	1
17	*Ava II*	4311	ARN-t Ile	-	G ⇒ C	-	oui	2
20	*Ava II*	4811	NAD 2	3	A ⇒ G	Trp → Trp	silence	1
21	*Hae II*	4833	NAD 2	1	A ⇒ G	Thr → Ala	oui	1
24	*Ava II*	5262	NAD 2	1	G ⇒ A/T	Ala → Thr ou Ser	oui	1
33 ou 12	*Ava II*	7856	CO II	1	A ⇒ C	Asn → His	oui	1
34	*Msp I*	7975	CO II	3	A ⇒ G	Pro → Pro	silence	1
37	*Ava II*	8270	N.C.	-	C ⇒ G	-	silence	7
39	*Hae II*	9266	CO III	3	G ⇒ C	Gly → Gly	silence	1
		9269	CO III	3	C ⇒ G	Ala → Ala	silence	-
41	*Hae II*	9692	CO III	3	A ⇒ G	Ala → Ala	silence	1
43	*Hae II*	11002	NAD 4	3	A ⇒ G	Gln → Gln	silence	1
44	*Msp I*	11457	NAD 4	2	C ⇒ G	Ala → Gly	oui	1
46	*Hpa I*	12026	NAD 4	1	A ⇒ G	Ile → Val	oui	1
48	*Ava II*	12191	ARN-t His	-	C ⇒ G	-	probable	1
51	*Msp I*	12817	NAD 5	1	C ⇒ G	Arg → Gly	oui	-
		12819	NAD 5	3	A ⇒ G	Arg → Arg	silence	1
52	*Msp I*	13101	NAD 5	3	A ⇒ C	Ala → Ala	silence	1
55/56	*Ava II/Bam HI*	13368	NAD 5	3	G ⇒ A	Gly → Gly	silence	1
59	*Msp I*	14206	NAD 5	1	A ⇒ G	Ser → Gly	oui	1
63	*Ava II*	15487	Cyt B	3	A ⇒ C	Pro → Pro	silence	1
64 ou 8	*Hae II*	15501	Cyt B	2	A ⇒ C	Asp → Ala	oui	1
65	*Msp I*	15505	Cyt B	3	A ⇒ G	Pro → Pro	silence	1
66	*Ava II*	15887	N.C.	-	T ⇒ G	-	silence	6
68/69	*Ava II/Bam HI*	16391	N.C.	-	G ⇒ A	-	silence	1

[1] Par rapport à la séquence de Cambridge.
[2] N.C. : Région non codante.
[3] Selon l'arbre phylogénique des types de la Figure 4.11.

Les différents changements d'acides aminés (a.a.) n'ont pas tous la même importance pour la fonction de la protéine. Grantham (1974) a calculé une distance chimique entre a.a. qui est fonction de leur composition, de leur polarité et de leur

volume. La distance moyenne entre les 190 couples d'a.a. a été étalonnée à 100 (Table 2 dans Grantham, 1974). Cette distance chimique est corrélée négativement (-0,72) avec les différences évolutives entre a.a. d'une même protéine (Grantham, 1974), indiquant par là que les a.a. sont plutôt remplacés par d'autres a.a. chimiquement proches au cours de l'évolution des protéines.

TABLE 4.24 : Distances chimiques entre acides aminés mutants et originaux (séquence de Cambridge).

Site	Changement d'a.a.	Types concernés	Distance[1]
16	Glu → Gly	16	98
21	Thr → Ala	13	58
24	Ala → Thr	52, 53	58
24	Ala → Ser	52, 53	99
33	Asn → His	58	68
44	Ala → Gly	4	60
46	Ile → Val	12	29
59	Ser → Gly	48, 54	56
64	Asp → Ala	62	126

[1] Distance chimique entre acides aminés (Grantham, 1974)

Les distances chimiques entre a.a. mutants et originaux ont été reportées dans la Table 4.24. Deux alternatives sont possibles pour le changement d'a.a. du site 23. On remarque que le changement Ala → Thr semble moins dommageable que le changement Ala → Ser, et donc plus probable. D'une manière générale, les distances chimiques sont toutes inférieures à la moyenne de tous les changements possibles, à l'exception du changement Asp → Ala au site 63. Dans la Figure 4.12, nous avons représenté le résultat d'une analyse en composantes principales des différences chimiques entre a.a. à partir des données de la Table 1 de Grantham (1974). Nous y avons également indiqué quelles mutations ont été observées. Le premier axe est bien corrélé avec les 3 caractères (composition, polarité, volume) utilisés dans l'analyse multivariée (Table 4.25) et représente 60 % de la variance totale. On constate que les mutations se font entre a.a. relativement proches et jamais entre a.a. possédant des coordonnées très différentes selon l'axe principal. Il est donc probable que ces modifications n'aient pas d'effets trop importants sur les protéines et que leurs fonctions soient équivalentes à celles de la protéine originale ou réduites, mais qu'elles ne sont pas totalement supprimées.

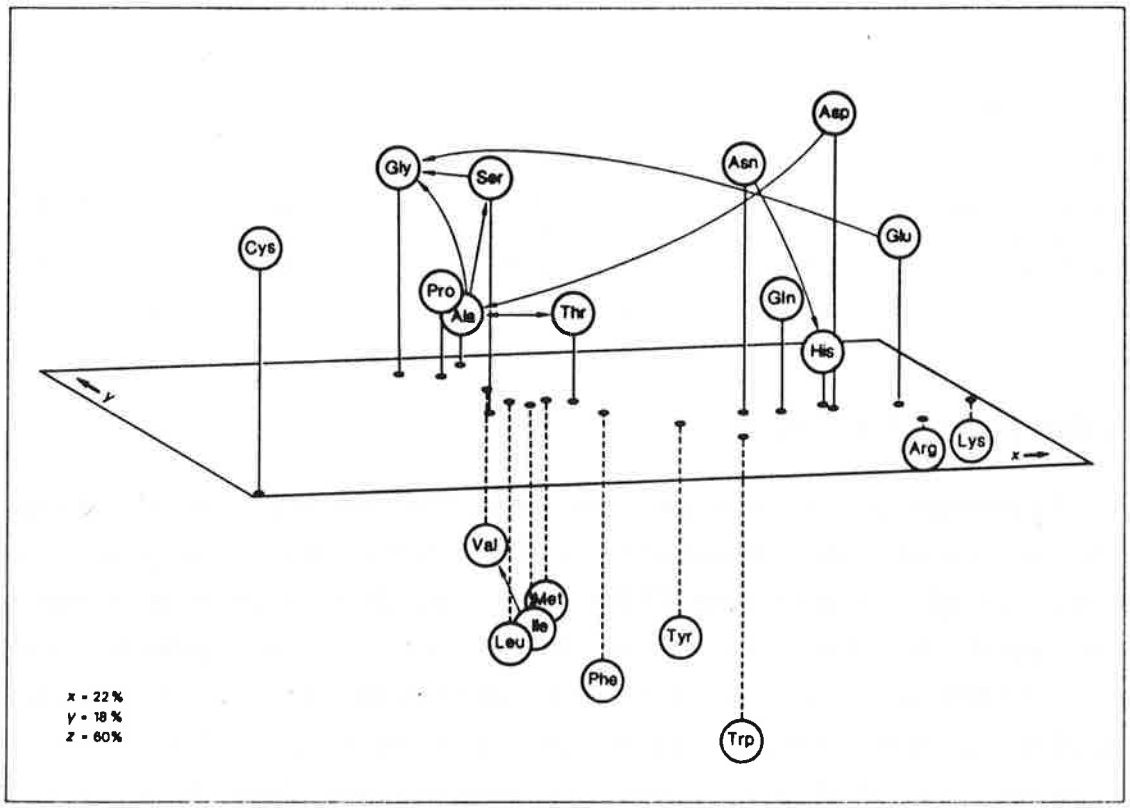

FIGURE 4.12 : Représentation tridimensionnelle d'une analyse en composantes principales représentant les distances chimiques entre les 20 acides aminés.

TABLE 4.25: Corrélations entre les axes principaux de l'analyse en composante principale de la Figure 4.12 et les 3 caractères utilisés.

	Axes		
Caractères	1	2	3
Composition	0,80	-0,25	0,54
Polarité	0,72	0,69	-0,07
Volume	-0,79	0,38	0,49

Gènes codant pour l'ARN ribosomique

La position de 2 sites de restriction n'a pu être définie précisément Il s'agit du site *Hae II* 8 ou 64 et du site *Ava II* 12 ou 33. Dans les 2 cas, le site peut soit se trouver dans un gène codant pour l'ARN 16 S, soit dans un gène de structure (respectivement Cyt B et Co II). Le gène de l'ARN 16 S est l'une des régions les moins variables de l'ADN-mt (Cann *et al.*, 1982). Il est donc soumis à de fortes contraintes fonctionnelles. Les substitutions survenues dans cette région de l'ADN-mt ont été étudiées du point de vue de leur impact sur la structure secondaire de l'ARN 16 S déterminée par Glotz *et al.* (1981) à partir de la séquence de Cambridge. La transversion C ⇒ G à la position 2540 de l'ADN-mt (site 8) provoquerait une mutation à la position 870 de l'ARN 16 S qui se situe dans une région fonctionnellement très importante, car conservée par rapport à l'ARN 23 S de *Escherichia coli* (Glotz *et al.*, 1981). La transversion A ⇒ C à la position 3135 (site 12) provoquerait une mutation à la position 1465 de l'ARN 16 S. Elle est localisée dans une hélice où l'on observe un appariement entre brins d'ARN homologues et qui est également conservée par rapport à *E. coli*. Ces 2 mutations, si elles se sont effectivement produites dans cette région auraient donc très probablement eu des répercussions phénotypiques. Dans le cas où ces mutations se seraient produites dans les gènes codant pour des protéines (sites 33 et 64), elles auraient provoqué des changements d'acide aminé (voir Table 4.23) qui auraient aussi très certainement affecté la fonctionnalité des protéines.

Une autre mutation s'est produite dans le gène de l'ARN 16 S. Il s'agit de la transversion T ⇒ G à la position 2157 de l'ADN-mt (site 5), correspondant à une mutation U ⇒ G à la position 487 de l'ARN 16 S qui se situe dans une région d'appariement entre brins. Elle conduit donc à une déstabilisation de la structure secondaire en l'absence d'une mutation compensatrice sur le brin homologue à la position 562 (position 1746 de l'ADN-mt). Le site 4 (position 1118 de l'ADN-mt) se

situant dans le gène de l'ARN 12 S a subi une mutation A ⇒ C qui entraîne une substitution à la position 471 de l'ARN 12 S. On peut penser que ce changement est phénotypiquement détectable car la base adénosine (A) est conservée chez les hominidés (*Gorilla gorilla, Pan paniscus, P. troglodytes* et *Pongo pygmaeus*) (Hixson and Brown, 1986) ainsi que chez la souris (*Mus musculus*) (Küntzel and Köchel, 1981). Elle se situe dans une boucle d'ARN simple brin de l'hélice 15 (Hixson and Brown, 1986). Ces boucles jouent plutôt un rôle dans la structure tertiaire et les interactions entre l'ARN-r et d'autres molécules.

Gènes codant pour des ARN-t

Les sites 16 et 47 se trouvent à l'intérieur de gènes codant respectivement pour les ARN-t^{Ile} et ARN-t^{His}. Les gènes de l'ARN-t de l'ADN-mt des primates évoluent apparemment 100 fois plus vite que ceux de l'ADN nucléaire (Brown *et al.*, 1982) qui sont extrêmement conservés, mais leur variabilité est quand même plus faible que celle des gènes structuraux ou celle des gènes de l'ARN-r codés sur l'ADN-mt (Cann, 1982; Cann *et al.*, 1982, 1984; Whittam *et al.*, 1986). Dans les 2 cas observés ici (Figures 4.13 et 4.14), la mutation s'est produite dans le bras T ψ C de l'ARN-t , région dont la structure diffère passablement de celle des ARN-t des gènes nucléaires (Anderson *et al.*, 1981). La mutation G ⇒ C à la position 4311 dans le gène de ARN-t^{Ile} se produit à un endroit particulièrement conservé où il y a généralement un appariement des bases G et C. Cet appariement n'étant plus possible, les structures secondaire et tertiaire peuvent en être fortement destabilisées, ce qui peut conduire à une fonctionalité réduite de l'ARN-t (Figure 4.13). La mutation C ⇒ G à la position 12191 dans le gène de l'ARN-t^{His} se situe dans la boucle simple brin du bras T ψ C (voir Figure 4.14) et pourrait avoir un impact sur la structure tertiaire de l'ARN-t, bien que cela soit moins clair que dans le cas précédent. Ces 2 mutations ont donc vraisemblablement eu des répercussions sur la fonctionalité des ARN-t.

Portions non-codantes de l'ADN-mt

Trois séries de gains de site se sont produites dans des portions non-codantes de l'ADN-mt. Il s'agit des mutations des sites 37, 66 et 68/69. La première citée est localisée dans une petite région non-codante d'environ 30 pb comprise entre les gènes de CO II et de l'ARN-t^{Lys}. La seconde se situe entre les gènes du Cyt B et de l'ARN-t^{Thr}. Il est intéressant de noter que l'espace entre ces 2 gènes n'est que d'une seule paire de base et que le gain de site a précisément été issu de la transversion de cette base. Enfin, le site 68/69 se trouve dans la région de la D-Loop.

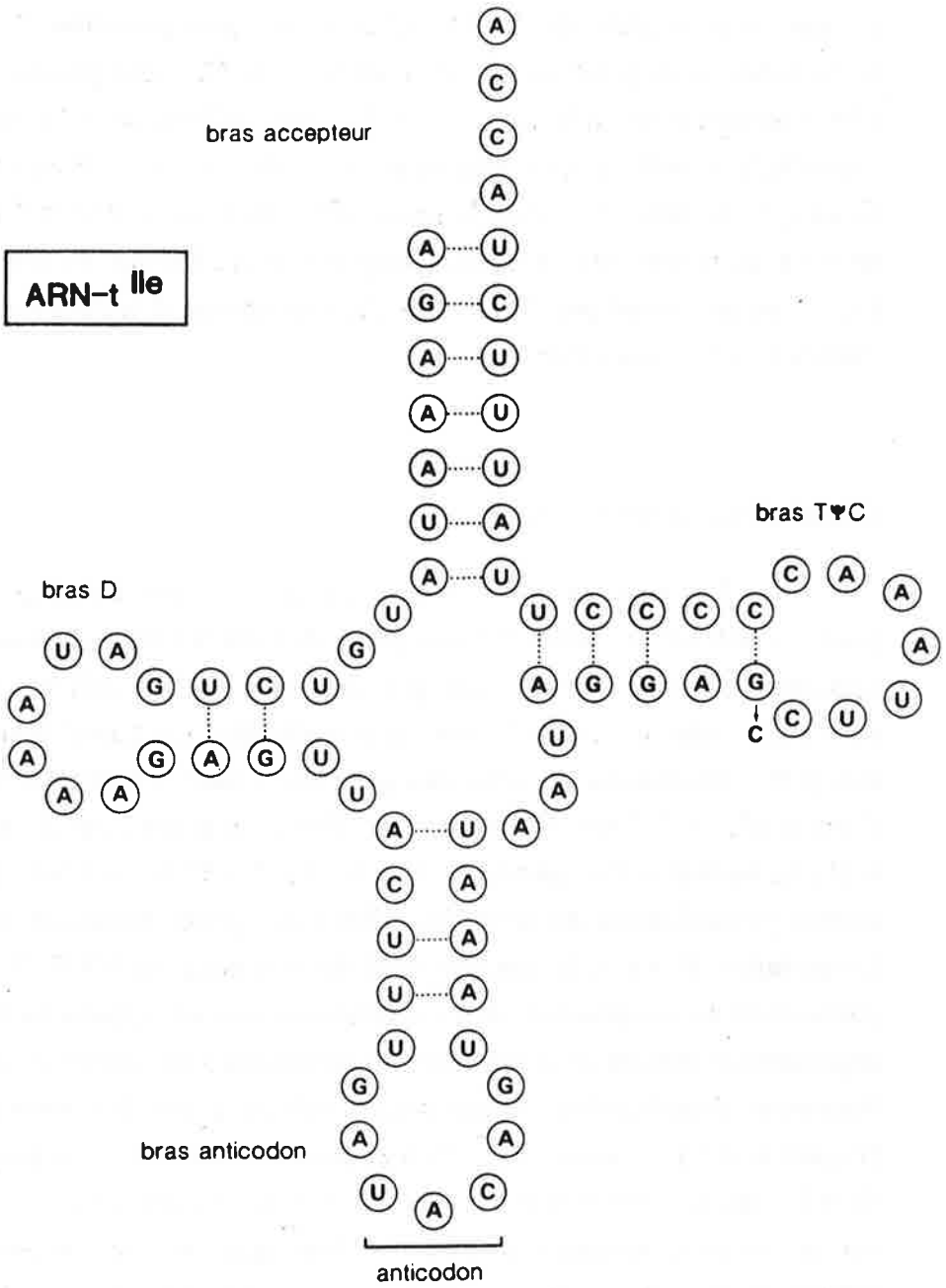

FIGURE 4.13 : Structure secondaire de l'ARN-t^{Ile}. Les bases sont représentées sous leur forme non modifiées et sont déterminées directement à partir de la séquence d'ADN.

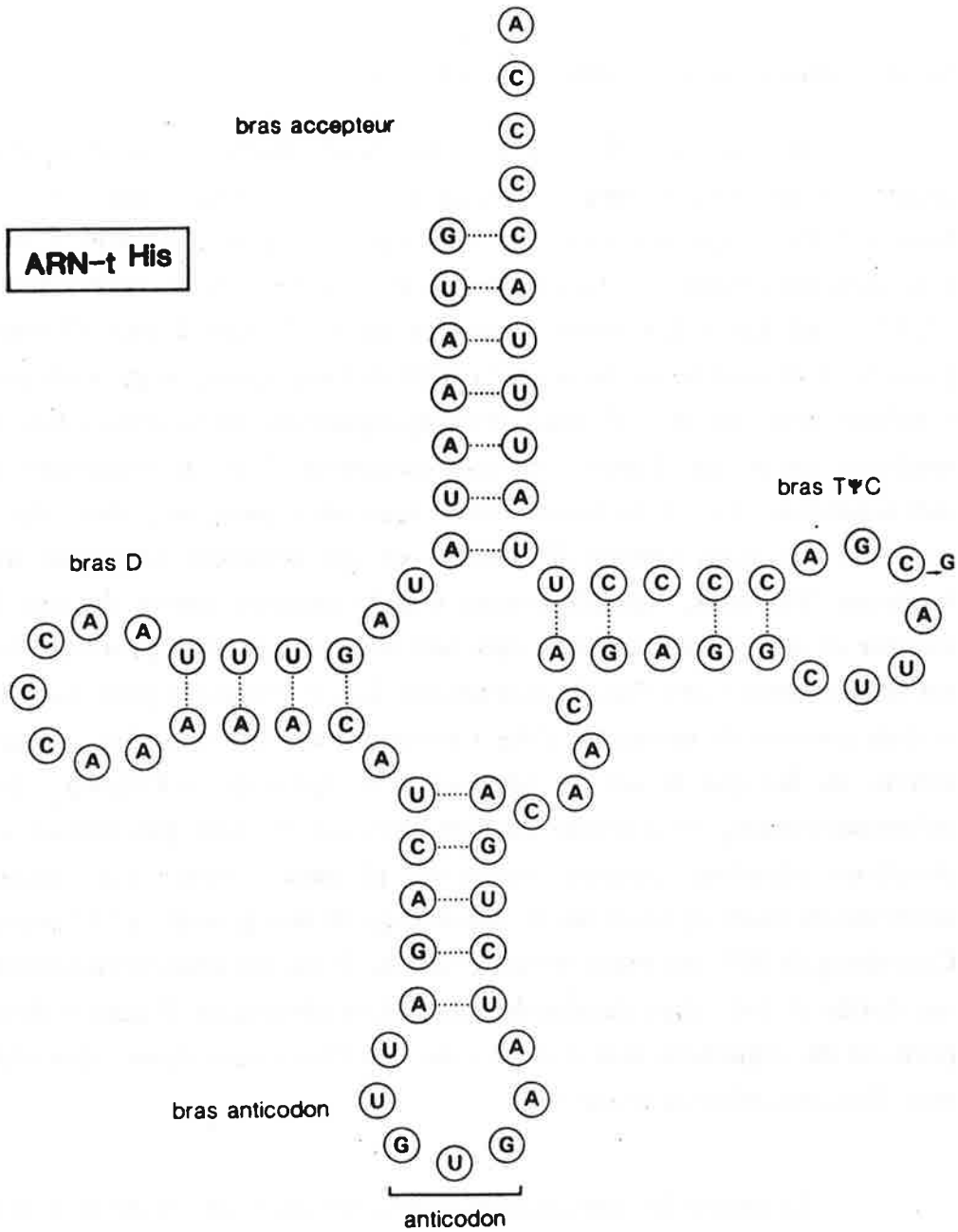

FIGURE 4.14 : Structure secondaire de l'ARN-t^{His}. Les bases sont représentées sous leur forme non modifiées et sont déterminées directement à partir de la séquence d'ADN.

Nombre d'occurrences et nature des substitutions

Dans la Table 4.23 ont également été reportés les nombres d'occurrences des différentes substitutions dues à des gains de site en s'appuyant sur la phylogénie des types d'ADN-mt (Figure 4.11). On s'aperçoit ainsi que la plupart de ces mutations ne sont apparues qu'une fois. Des mutations récurrentes sont observées dans 4 cas (sites 15, 17, 37 et 66). Les substitutions survenues au site 17 dans l'ARN-t^{Ile} sont les seules qui puissent avoir un effet sur la fonctionnalité de l'expression du gène où elles figurent. Les 2 substitutions du site 15 sont phénotypiquement silencieuses. Les sites 37 et 66 semblent avoir été l'objet de respectivement 7 et 6 mutations récurrentes et indépendantes. Les 2 positions concernées sont comprises dans des portions non-codantes que nous venons de décrire et qui semblent libres de toute contrainte évolutive. Toutefois, rappellons-nous que la position exacte du site 37 n'a pas été attestée et que plusieurs autres sites potentiels se situent à proximité (voir le chapitre sur la localisation des sites de restriction). Les 7 mutations pourraient donc cacher un certain nombre de mutations s'étant produites sur d'autres sites proches. Si l'on tient compte du fait que le site 69 (situé dans la région de la D-Loop a été perdu 5 fois indépendemment, on constate que les sites qui ne sont pas soumis à des pressions sélectives négatives peuvent connaître plusieurs évènements mutationnels. Ceci confirme les études portant sur le séquençage de la région de la D-Loop par Aquadro et Greenberg (1983) reportées précédemment. Il est toutefois remarquable qu'un même nucléotide ait été l'objet de plus de 5 mutations identiques. Il semble donc que certaines portions de séquences non codantes de l'ADN-mt constitutent des régions "hot-spot" pour l'accumulation de mutations.

La nature des substitutions dues à des gains de site est reportée dans la Table 4.26. Les parties codantes de l'ADN-mt ont été l'objet de 25 substitutions avec une nette prédominance des transitions A ↔ G, comme cela avait déjà été noté pour les substitutions dans la région de la D-Loop (voir Figure 3.1). En raison des incertitudes liées à la localisation de 2 sites (8 ou 33 et 12 ou 64) et à la position de la mutation du site 39 (2 possibilités), le nombre exact de substitutions peut parfois osciller entre 2 valeurs. Ainsi, le taux de transitions/transversions pour la partie codante de l'ADN-mt est comprise entre 1,1:1 et 1,3:1, ce qui est très inférieur au rapport 24:1 obtenu pour la région de la D-Loop (Greenberg *et al.*, 1983) ou au rapport 11:1 obtenu dans la comparaison d'une séquence de 896 pb d'une région codante entre 5 primates (Brown *et al.*, 1982). Par contre, il se rapproche du taux de 1,9:1 déduit de Cann *et al.* (1984) pour les gains de site d'une analyse de PLFR dans les régions codantes de l'ADN-mt humain,

ainsi que du taux de 1,75:1 obtenu par Horai *et al.* (1984) pour des gains de sites. Le très fort biais transitionnel des régions non-codantes n'est pas retrouvé pour les parties codantes de l'ADN-mt qui présentent toutefois un rapport transition/transversion au moins 2 fois plus élevé que celui correspondant à des taux de substitution identiques entre les 4 nucléotides (1 transition pour 2 transversions).

TABLE 4.26 : Nature des substitutions de nucléotides observées lors de gains de sites.

Type de substitutions	Nombre observé			
	Portion codante		Portion non-codante	Totalité ADN-mt
Transitions				
A → G	10	11-12	0	10
G → A	1-2		1	2-3
T → C	0	2		0
C → T	2			2
		13-14		14-15
Transversions				
A → T	0	0	0	0
T → A	0		0	0
C → G	1-3	5-6	7	8-10
G → C	3-4		0	3-4
A → C	4-5	4-5	0	4-5
C → A	0		0	0
T → G	1	1-2	6	7
G → T	0-1		0	0-1
		11-12		24-25

Répartition des différentes substitutions dans la phylogénie des types.

En intégrant l'information concernant l'impact probable de chaque substitution sur la fonctionnalité des types d'ADN-mt dans la phylogénie de la Figure 4.11, il est possible de déterminer quels types sont porteurs de mutations potentiellement défavorables. Dans la Figure 4.15, nous avons repris la phylogénie de la Figure 4.11 et y avons indiqué les passages entre types qui provoquaient un changement phénotypique. Dans le cas des 18 liaisons dues à des pertes de site, il n'est pas possible de déterminer quelle nucléotide a subi une mutation, et par la-même, de connaître son effet phénotypique. Enfin, nous avons également indiqué dans la Figure 4.15, pour quels passages entre sites une mutation survenue dans une région non-codante (vraisemblablement silencieuse) s'était produite.

L'occurence de 13 évènements mutationnels potentiellement défavorables va concerner 16 types (Nº 4, 12, 13, 16, 25, 29, 37, 38, 48, 52, 53, 54, 58, 59, 60 et 62) du fait que les types 53, 54 et 59 sont directement issus de types déjà porteurs de ces mutations (respectivement des types 52, 48 et 38). A ce propos, il apparaît que le type 53 est probablement issu du type 52, car il est plus probable que la mutation du site 23 ne soit apparue qu'une seule fois si elle a des effets négatifs sur le phénotype de son porteur.

Généralement, les types ayant acquis une mutation phénotypiquement exprimée ne sont donc pas à la source d'autres types (10 cas sur 13) et sont soit issus du type 1 sans être à la source d'autres types (4 cas), soit sur une branche terminale de la phylogénie (6 cas). Les 3 cas où un type défavorisé est à la source d'autres types ne concernent pas les types qui sont à l'origine de branches importantes de la phylogénie (comme les types 2, 7, 8, 10, 21, 23 ou 28). Si l'on s'intéresse aux types définis comme potentiellement ancestraux, c'est-à-dire retrouvés dans plusieurs groupes continentaux, on s'aperçoit que les types 2, 7, 31, 39 et 47 sont phénotypiquement identiques au type 1. Les types 6, 8 et 28 ont subi des pertes de sites, et seul le type 13 porte clairement une mutation potentiellement défavorable. Il est intéressant de constater que les types à la source de la diversité africaine (types 2, 7 et 10) sont tous porteurs de mutations silencieuses.

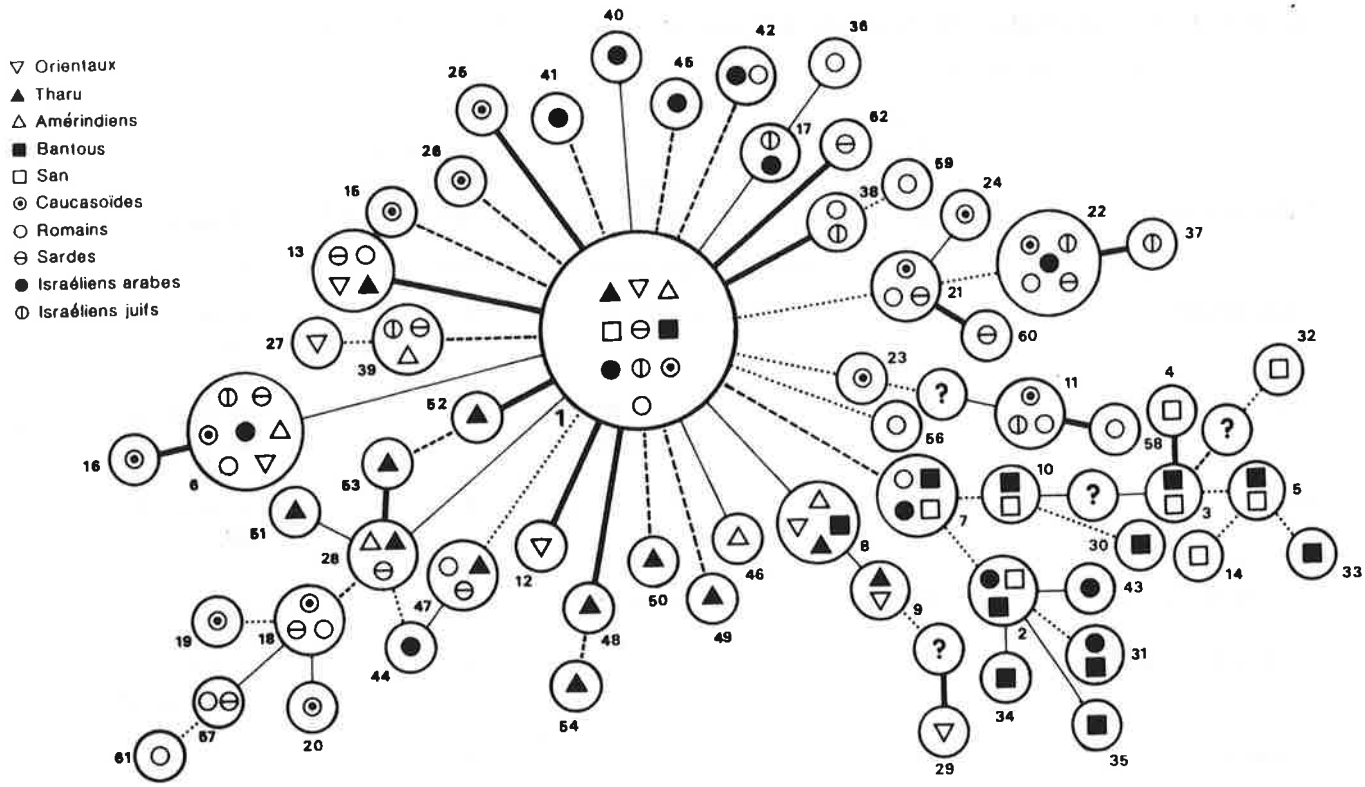

FIGURE 4.15 : Représentation des mutations qui peuvent provoquer un changement phénotypique des types d'ADN-mt. Notations : ..... : Mutation dans une région non-codante; ----- : Mutation phénotypiquement silencieuse; ——— : Mutation provoquant un changement phénotypique; ——— : Mutation due à une perte de site.

TABLE 4.27 : Répartition des types défavorisés ou non selon les populations.

Echantillons	Nombre de types potentiellement défavorisés[1]	Types et fréquences	Nombre de types non défavorisés[2]	Types
Caucasoïdes	2	16 *(2,0)*, 25 *(2,0)*	5	1, 15, 21, 22, 23
Romains	4	13 *(2,1)*, 38 *(2,1)*, 58 *(1,1)*, 59 *(1,1)*	7	1, 21, 22, 42, 47, 56
Sardes	3	13 *(1,5)*, 60 *(0,7)*, 62 *(1,5)*	5	1, 21, 22, 39, 47
Israéliens Juifs	2	37 *(2,6)*, 38 *(2,6)*	3	1, 22, 39
Israéliens Arabes	0	-	9	1, 2, 7, 22, 31, 40, 41, 42, 45
Bantous	0	-	6	1, 2, 7, 10, 30, 31
San	1	4 *(26,5)*	4	1, 2, 7, 10
Orientaux	1	13 *(4,3)*	2	1, 27
Tharu	5	13 *(25,3)*, 48 *(2,2)*, 52 *(2,2)*, 53 *(1,1)*, 54 *(1,1)*	4	1, 47, 49, 50
Amérindiens	0	-	2	1, 39

[1] Types ayant acquis des mutations exprimées au niveau phénotypique (changement d'acide aminé ou modification de la strucure secondaire ou tertiaire de l'ARN). Ce nombre constitue un minimum, car le comportement phénotypique des types caractérisés par des pertes de sites n'est pas définissable.

[2] Types portant des mutations qui sont phénotypiquement non exprimées (mutations silencieuses ou mutations dans des régions non-codantes). Le calcul du nombre de ces types ne tient pas compte des types ayant perdu des sites.

Dans la Table 4.27, nous avons reporté, pour chaque population, le nombre de types porteurs de mutations silencieuses et de mutations exprimées. Trois populations seulement ne possèdent aucun des types définis comme potentiellement défavorables (Arabes Israéliens, Bantous et Amérindiens). Les fréquences des types défavorisés sont généralement inférieures à 5 % à l'exception du type ancestral 13 chez les Tharu (25,3%) et du type 4 chez les San (25,3%). Les Arabes Israéliens possèdent

une nette majorité de types phénotypiquement identiques au type 1, suivis par les Romains et les Bantous.

Bien que les types issus de pertes de sites nous empêchent d'avoir un discours plus précis, nous pouvons faire quelques remarques générales sur la répartition des types défavorisés ou non. En premier lieu, il convient de remarquer que sur les 16 types phénotypiquement différents du type 1, un seul est retrouvé parmi les populations africaines. Les autres se répartissent à raison de 5 chez les populations d'origine orientales et de 10 parmi les populations caucasoïdes. Les types des populations africaines pourraient ainsi être, pour la plupart, phénotypiquement équivalents et les différences entre les fréquences géniques pourraient être principalement dues aux effets de la dérive génétique. Dans les autres groupes de populations, la présence de plusieurs types potentiellement défavorisés pourrait perturber le processus de dérive étant donné la présence de types sélectivement non-équivalents. L'explication de l'apparente absence de types fortement sélectionnés dans les populations africaines n'est pas triviale et nécessiterait une connaissance précise de la nature de tous les passages entre types. De plus, il est évident que certains types ont subi des mutations qui ne sont pas visibles dans ces études de PLFR menées avec un nombre restreint d'enzymes, mutations qui nous sont donc inconnues. Nous préférons donc remettre la discussion sur la sélection de certains types à la partie de notre travail traitant spécifiquement du problème de la sélection. Nous nous bornerons simplement à constater qu'il existe une bonne cohésion entre la neutralité apparente des populations africaines du point de vue de leur diversité moléculaire et le fait que l'on n'y trouve pas de types très défavorisés.

A) Autres études de PLFR portant sur des populations

En dehors des 10 populations étudiées d'une manière analogue par différents auteurs, très peu d'études ont été menées sur des échantillons tirés d'une population homogène et relativement bien définie sur le plan ethnique et géographique. De plus, ces quelques autres études ne sont pas comparables entre elles. Nous les analyserons donc ici avant tout par souci d'exhaustivité concernant ce système génétique, en espérant que des études ultérieures permettront d'utiliser ces résultats à des fins comparatives qui sont les seules qui puissent contribuer à l'étude de l'histoire du peuplement humain.

1) Etude d'une population japonaise

Un échantillon de 120 japonais a été analysé, d'une part, au moyen d'enzymes possédant une séquence de reconnaissance de 4 ou 5 pb (Horai and Matsunaga, 1986), et, d'autre part, avec des enzymes reconnaissant 6 pb (Horai *et al.*, 1984). Ces 2 études doivent être analysées séparément, bien que portant sur le même échantillon, car les types d'ADN-mt obtenus dans chacune des études n'ont pas été combinés par les auteurs, et aucune liste d'haplotypes pour l'ensemble des enzymes n'est disponible.

L'ADN-mt utilisé dans ces études a été extrait à partir de placentas humains. Cette méthode d'échantillonage exclut *a priori* que l'échantillon surveillé soit très homogène. Il s'ensuit donc que les résultats qui suivront concerneront une population de "japonais" au sens large, sans caractérisation ethnique ou géographique plus précise.

a) Analyse de la population au moyen d'enzymes reconnaissant 4 ou 5 pb

Localisation des sites de reconnaissance polymorphes

Contrairement aux études traitées précédemment, les emplacements des sites de reconnaissance polymorphes des différents enzymes ont été déterminés directement par les auteurs. Dans la Table 4.22, nous avons reporté et classé les sites polymorphes selon leur place sur la molécule d'ADN-mt.

TABLE 4.28: Sites de restriction polymorphes par substitution de nucléotides (Horai et Matsunaga, 1986)

Position	Site Nº	Polymorphisme[1]	Fréquence(%) (N=116)	Enzyme	Région
255	1	1	0,9	*Hha I*	ARN 7S
496	2	1	0,9	*Hpa II*	non codante
1307	3	0	1,7	*Rsa I*	ARN-r 12S
1315	4	1	1,7	*Rsa I*	"
1515	5	1	21,6	*Hae III*	"
2120	6	1	5,2	*Rsa I*	ARN-r 16S
3090	7	0	0,9	*Hae III*	"
3315	8	0	0,9	*Hae III*	NAD 1
3337	9	0	0,9	*Rsa I*	"
3391	10	1	0,9	*Hae III*	"
3427	11	0	0,9	*Hae III*	"
4360	12	0	1,7	*Hinf I*	ARN-t Gln
4621	13	1	1,7	*Rsa I*	NAD 2
4711	14	0	0,9	*Hpa II*	"
4794	15	1	5,2	*Hae III*	"
4831	16	1	7,8	*Hha I*	"
4848	17	0	2,6	*Hae III*	"
5009	18	1	14,7	*Rsa I*	"
5261	19	0	0,9	*Hae III*	"
5742	20	0	0,9	*Hpa II*	OL
6333	21	1	0,9	*Hae III*	CO I
6425	22	0	0,9	*Hae III*	"
7214	23	1	0,9	*Taq I*	"
7598	24	0	3,4	*Hha I*	CO II
7828	25	1	0,9	*Hha I*	"
7859	26	0	0,9	*Sau3A I*	"
7902	27	1	11,2	*Hinf I*	"
8022	28	1	18,1	*Taq I*	"
8150	29	0	0,9	*Hpa II*	"
8249/8250	30	1	4,3	*Hae III/Ava II*	"
8249	31	0	0,9	*Ava II*	"
8356	32	1	1,7	*Rsa I*	ARN-t Lys
8391	33	0	0,9	*Hae III*	URF A6L
8565	34	1	0,9	*Sau3A I*	"
8592	35	1	0,9	*Sau3A I*	ATPase 6
8750	36	1	13,8	*Hinf I*	"
8783	37	0	1,7	*Hinf I*	"
8994	38	0	0,9	*Hae III*	"
8998	39	0	0,9	*Rsa I*	"
9376	40	1	1,7	*Hinf I*	CO III
9400	41	0	0,9	*Hha I*	"
9438	42	0	0,9	*Hae III*	"
10180	43	0	1,7	*Taq I*	NAD 3
11421	44	0	0,9	*Taq I*	NAD 4
11454	45	1	0,9	*Hpa II*	"
12192	46	1	19,0	*Hinf I*	ARN-t His
12501	47	1	2,6	*Hha I*	NAD 5
12629	48	0	1,7	*Ava II*	"

Position	Site Nº	Polymorphisme[1]	Fréquence(%) (N=116)	Enzyme	Région
13116	49	1	0,9	*Sau3A I*	"
13595	50	0	0,9	*Hha I*	"
13605	51	1	2,6	*Hinf I*	"
13938/13939	52	1	1,7	*Hha I/Acc II*	"
14158	53	1	0,9	*Taq I*	URF 6
14279	54	1	7,8	*Hae III*	ARN-t Glu
14749	55	1	0,9	*Hae III*	Cyt b
14976	56	0	8,6	*Hinf I*	"
15047	57	0	0,9	*Hae III*	"
15152	58	0	0,9	*Hae III*	"
15172	59	1	0,9	*Hae III*	"
15234	60	0	0,9	*Hinf I*	"
15282	61	1	0,9	*Rsa I*	"
15431	62	1	2,6	*Hae III*	"
15723	63	0	0,9	*Hinf I*	"
15812	64	0	0,9	*Rsa I*	"
15925	65	0	2,6	*Hpa II*	ARN-t Thr
15949	66	0	1,7	*Rsa I*	"
16049	67	0	0,9	*Rsa I*	ARN-t Pro
16096	68	0	1,7	*Rsa I*	"
16208	69	0	0,9	*Rsa I*	D-Loop
16215	70	1	1,7	*Sau3A I*	"
16254	71	1	1,7	*Hae III*	"
16389/16390	72	1	4,3	*Hinf I/Ava II*	"
16398	73	1	0,9	*Hae III*	"
16503	74	1	0,9	*Ava II*	"
16517	75	0	44,0	*Hae III*	"
16534	76	1	3,4	*Hae III*	"

[1] La notation "1" correspond à un polymorphisme dû à un gain de site par rapport à la séquence de Cambridge, alors que la notation "0" correspond à la perte d'un site.

Certains sites sont chevauchant pour 2 enzymes (sites 30, 52 et 72), et conduisent à un double polymorphisme apparent. Dans les 3 cas concernés, la notation employée a consisté à indiquer les 2 sites de coupure, ainsi par exemple "8249/8250".

De cette façon, 76 sites polymorphes dus à des substitutions ont pu être mis en évidence en employant 9 enzymes différents sur 116 individus. Les sites monomorphes n'ont pas été reportés ici, car ils sont très nombreux et ils peuvent être aisément repéré sur la séquence de Cambridge. Globalement, les sites de restriction polymorphes sont répartis relativement uniformément sur l'ensemble de l'ADN-mt (voir Figure 4.12), avec une légère accumulation dans la région de la D-Loop (8 sites polymorphes pour une longueur de 330 pb).

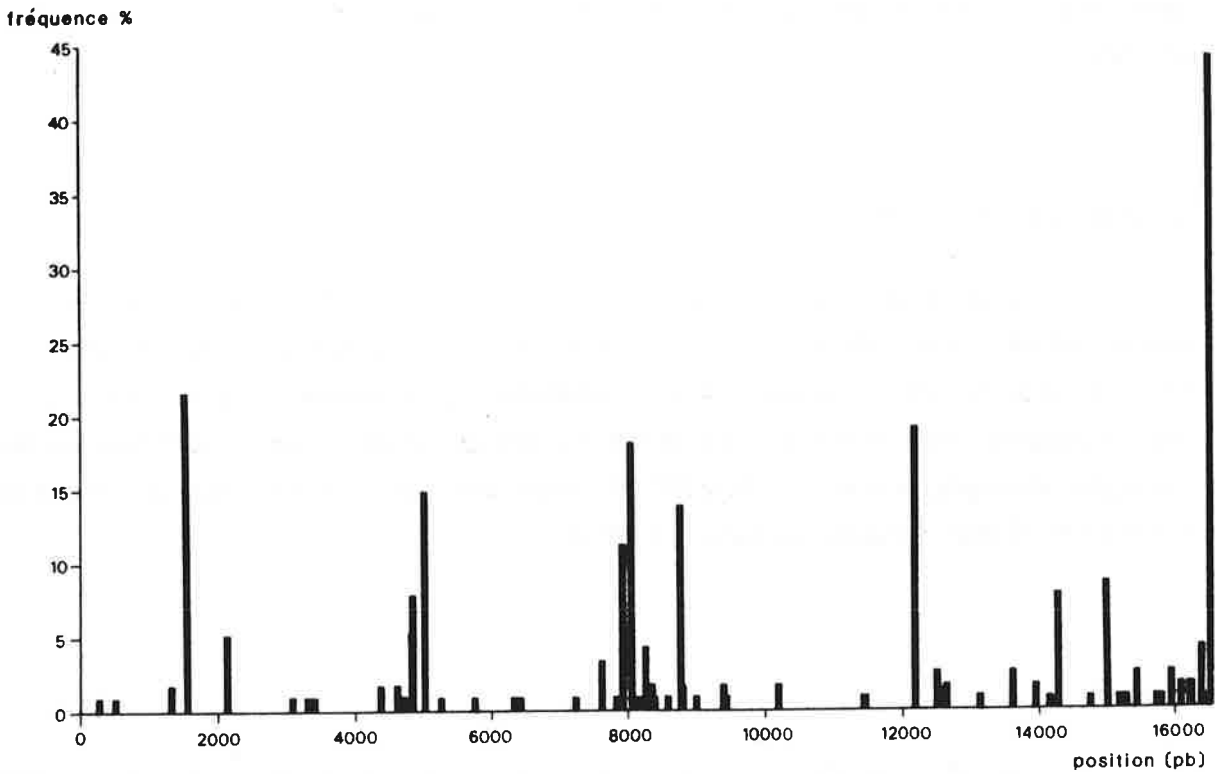

FIGURE 4.16 : Positions et fréquences des sites de restriction polymorphes définis par Horai et Matsunaga (1986).

Un certain nombre de délétions et d'insertions semblent également avoir été mises en évidence par Horai et Matsunaga (1986). Celles-ci n'ont pas été reportées ici, étant donné qu'une interrogation subsiste quant à leur nature. En effet, leur apparition pourrait résulter de simples substitutions ayant des effets sur la vitesse de migration dans les gels électrophorétiques à base de polyacrylamide (Horai *et al.*, 1987; Singh *et al.*, 1987).

Définition des morphes

Nous avons développé notre propre nomenclature concernant la numérotation des morphes, étant donné que Horai et Matsunaga n'ont pas défini les morphes comme une combinaison de présences ou absences de plusieurs sites de reconnaissance, mais plutôt comme le simple fait de couper à un certain site ou non. Ainsi, par exemple, le morphe *Hae III* Nº 8 représente une combinaison des "morphes" notés 16 et 35 par Horai et Matsunaga (1986).

Hae III

L'emploi de l'enzyme de restriction *Hae III* a produit une quantité considérable de morphes (Table 4.29 et Figure 4.17) très différenciés les uns des autres. Les morphes 1, 2 et 4 sont de loin les plus fréquents dans cette population de japonais. Ils correspondent également aux morphes à partir desquels d'autres morphes sont issus. Les morphes 1 et 2 semblent représenter 2 centres de différenciation clairement séparés. De ce point de vue, il est difficile de savoir si ils correspondent à une divergence ancienne de ces 2 morphes à partir desquels se seraient développées 2 voies évolutives à l'intérieur d'une même population, ou si tout cela résulte de la fusion, au niveau de l'échantillon, de 2 stocks génétiques ayant évolué séparément depuis longtemps. Etant donné le manque de précision quant à la constitution de cet échantillon, il ne nous sera pas possible de trancher à ce niveau de la discussion. On notera également qu'un troisième centre de différenciation, plus restreint, semble également s'articuler autour du morphe 4.

On notera cependant le haut degré de différenciation existant dans ce système, qui le rendrait particulièrement informatif si il était utilisé systématiquement dans l'étude de populations. Cela est bien entendu conditionné par une connaissance précise du schéma de divergence des morphes, comme celui que nous proposons dans la

Figure 4.17. En effet, comme nous l'avons vu dans le cas des morphes *Ava II* et *Msp I* définis sur 10 autres populations, un certains nombre de sites paraissent avoir subi des mutations parallèles. C'est notamment le cas des sites Nº 5, 8, 54, 62, 71, 75 et 76 qui ont tous été l'objet d'au moins 2 séries de substitutions indépendantes.

TABLE 4.29: Liste et fréquence des morphes *Hae III*

Morphe	Sites polymorphes[1]	Fréquence (%) (N = 116)
1	-	25,0
2	75	25,0
3	30	0,9
4	5	15,5
5	54	1,7
6	17	1,7
7	33	0,9
8	5,55	0,9
9	5,76	2,6
10	5,21,73	0,9
11	15	4,3
12	15,76	0,9
13	11,75	0,9
14	62,75	1,7
15	42,75	0,9
16	8,75	0,9
17	38,75	0,9
18	57,75	0,9
19	10,75	0,9
20	30,75	3,4
21	5,75	0,9
22	54,75	5,2
23	22,71,75	0,9
24	54,58,59,75	0,9
25	5,7,19,62,71,75	0,9
26	8,17	0,9

[1] Ces nombres indiquent les sites qui diffèrent de la séquence de Cambridge, leur numérotation correspond à celle de la Table 4.28

Ce système n'est cependant pas complétement défini, car 5 morphes intermédiaires sont absents de l'échantillon. Il s'agit en particulier de 4 morphes consécutifs qui sont nécessaire pour relier le morphe 4 au morphe 25 (voir la Figure 4.17), ce qui suggère qu'une partie de la diversité de ce système reste à explorer.

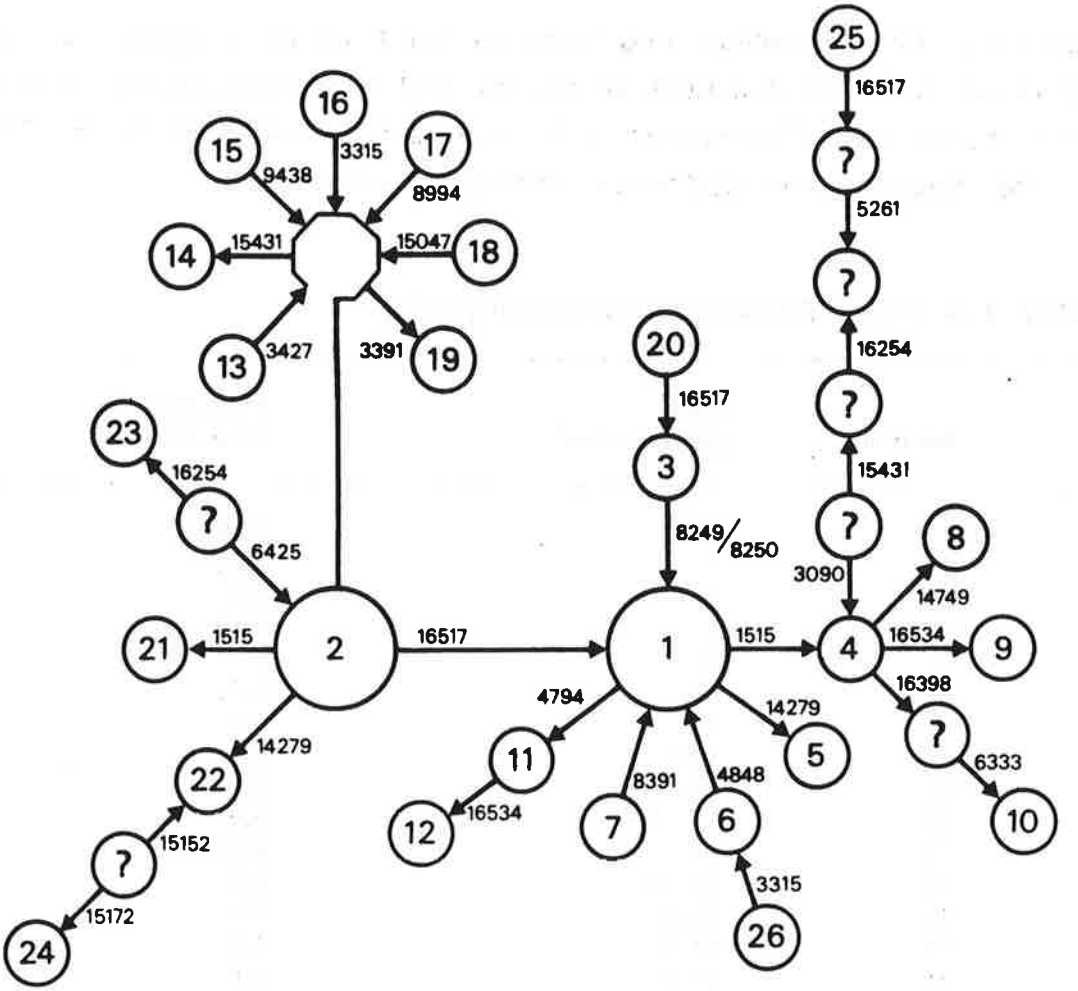

FIGURE 4.17 : Différenciation des morphes *Hae III*. Le sens des flèches indique un gain de site.

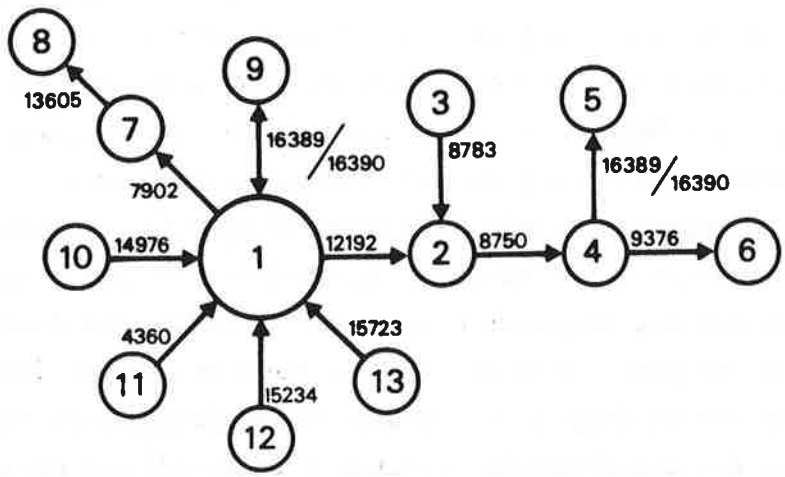

FIGURE 4.18 : Différenciation des morphes *Hinf I*. Le sens des flèches indique un gain de site.

Hinf I

Cet enzyme a permis de générer 13 morphes. Le morphe 1 est clairement central, puisque les 12 autres morphes en sont issus directement ou indirectement (voir Figure 4.18). Sa fréquence élevée (55 %) concorde également avec cette interprétation des faits (Table 4.30). Néanmoins, les morphes 4, 7 et 10 possèdent également des fréquences appréciables. Les morphes 7 et 10 dérivent directement du morphe 1 et le premier semble à l'origine du morphe 8. Le cas du morphe 4 est quelque peu particulier en ce sens qu'il se situe à 2 substitutions du morphe 1 et que le morphe 2 qui est intermédiaire possède une fréquence moindre. Le même morphe 4 est également à la source de 2 autres morphes (5 et 6), ce qui souligne son importance. Donc, une partie de la différenciation des morphes s'est déroulée indépendamment du morphe 1, ce qui pourrait évoquer des considérations similaires à celles qui ont été émises pour les origines des morphes *Hae III*.

TABLE 4.30: Liste et fréquence des morphes *Hinf I*

Morphe	Sites polymorphes[1]	Fréquence (%) (N = 116)
1	-	55,2
2	46	3,4
3	37,46	1,7
4	36,46	10,3
5	36,46,72	1,7
6	36,40,46	1,7
7	27	8,6
8	27,51	2,6
9	72	2,6
10	56	8,6
11	12	1,7
12	60	0,9
13	63	0,9

[1] Ces nombres indiquent les sites qui diffèrent de la séquence de Cambridge, leur numérotation correspond à celle de la Table 4.28

Il faut aussi noter la double mutation apparente *Hinf I/Ava II* du site 72, apparue indépendamment sur les morphes 5 et 9, qu'il faut confronter à la double mutation apparente *BamH I/Ava II* rencontrée précédemment et notée 68/69 dans la Table 4.3. En comparant ces 2 doubles mutations, on s'aperçoit qu'elles sont différentes

et concernent la substitution de 2 nucléotides distincts. Dans le cas de la mutation *Hinf I/Ava II*, c'est le nucléotide de la position 16390 qui a subi une transition G → A, alors que pour la double mutation *BamH I/Ava II*, il s'agit du nucléotide 16391 qui a subi une transition identique. Cela confirme l'hypothèse selon laquelle cette région de l'ADN-mt (D-Loop) serait sujette à des transitions fréquentes qui seraient conservées, vu la faible pression sélective pesant sur cette portion du génome.

Hha I

Le réseau de différenciation des 10 morphes *Hha I* est moins diversifié que celui de *Hinf I* ou de *Hae III*. Il s'articule nettement autour du morphe 1 qui possède une fréquence de 83,2 % (voir Table 4.31 et Figure 4.19). Tous les autres morphes sont distants du morphe 1 par une seule substitution à l'exception du morphe 3 qui peut théoriquement provenir des morphes 2 ou 4. Etant donné les fréquences de ces 3 morphes dans l'échantillon, et en admettant une différenciation primordiale à partir du morphe 1, une liaison 2-3 semble nettement plus vraisemblable.

TABLE 4.31: Liste et fréquence des morphes *Hha I*

Morphe	Sites polymorphes[1]	Fréquence (%) (N = 116)
1	-	83,6
2	16	5,2
3	16,24	2,6
4	24	0,9
5	50	0,9
6	25	0,9
7	52	1,7
8	47	2,6
9	41	0,9
10	1	0,9

[1] Ces nombres indiquent les sites qui diffèrent de la séquence de Cambridge, leur numérotation correspond à celle de la Table 4.28

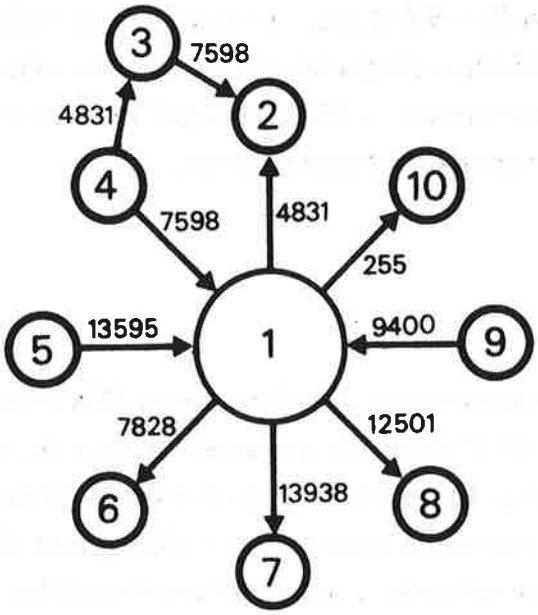

FIGURE 4.19 : Différenciation des morphes *Hha I*. Le sens des flèches indique un gain de site.

Rsa I

Le réseau de différenciation des 14 morphes définis par la digestion de l'ADN-mt avec *Rsa I* se développe essentiellement autour du morphe 1 (voir la Figure 4.20 et la Table 4.30). Le morphe 2 atteint presque une fréquence de 14 % et serait à l'origine du morphe 3.

TABLE 4.32: Liste et fréquence des morphes *Rsa I*

Morphe	Sites polymorphes[1]	Fréquence (%) (N = 116)
1	-	67,2
2	18	13,8
3	9,18	0,9
4	3	1,7
5	4	1,7
6	6	5,2
7	13	1,7
8	39	0,9
9	61	0,9
10	67	0,9
11	67,68	1,7
12	64	0,9
13	69	0,9
14	32,66	1,7

[1] Ces nombres indiquent les sites qui diffèrent de la séquence de Cambridge, leur numérotation correspond à celle de la Table 4.28

Le morphe 10 semble également avoir produit un autre morphe (le N°11), mais l'on constate curieusement que ce dernier est plus fréquent que le morphe duquel il serait issu. Néanmoins, les faibles fréquences de ces 2 morphes n'exclue pas que cela provienne du processus aléatoire de l'échantillonnage. En effet, il convient de mentionner que les morphes possédant une fréquence de 0,9 % dans cet échantillon n'ont été trouvé que chez un seul individu. Par conséquent, leur fréquence dans la population sera passablement mal estimée.

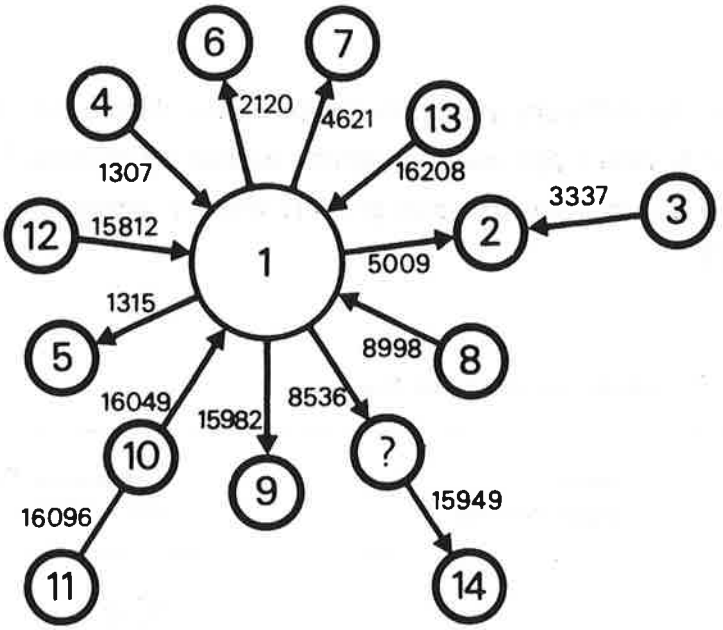

FIGURE 4.20 : Différenciation des morphes *Rsa I*. Le sens des flèches indique un gain de site.

Taq I

Seuls 6 morphes *Taq I* différents ont pu être mis en évidence. Le morphe 1, nettement prépondérant, s'avère central. On notera aussi la fréquence appréciable du morphe 2 (17,8%) (voir Table 4.33), à partir duquel a évolué le morphe 3 (voir Figure 4.17). Ce système est relativement simple et ne soulève pas de problème majeur.

TABLE 4.33: Liste et fréquence des morphes *Taq I*

Morphe	Sites polymorphes[1]	Fréquence (%) (N = 116)
1	-	78,4
2	28	17,2
3	28,53	0,9
4	23	0,9
5	43	1,7
6	44	0,9

[1] Ces nombres indiquent les sites qui diffèrent de la séquence de Cambridge, leur numérotation correspond à celle de la Table 4.28

Ava II

Cet enzyme avait déjà été étudié précédemment, et nous avions vu qu'il avait fourni une grande quantité de morphes différents dans les 10 populations où il avait été étudié. Cependant, cette vaste diversité provenait principalement des populations d'origine européenne, méditerranéenne et africaine. Les populations orientales présentaient un polymorphisme moindre avec une forte prépondérance du morphe 1 (voir Table 4.13). La diversité observée dans la population japonaise est cohérente avec cette interprétation. Les morphes 1, 2 et 3, déjà définis auparavant y sont retrouvés, alors que 3 nouveaux morphes (27, 28, 29) ont été définis. Il est à noter que le morphe 29 a subi une mutation provoquant simultanément la perte d'un site *Hae III* et le gain d'un site *Ava II*. Dans le cas du morphe 2, le gain de site se produit à un endroit différent et ne provoque pas de perte de site *Hae III* (voir les explications de Horai et Matsunaga , 1986, p.112)

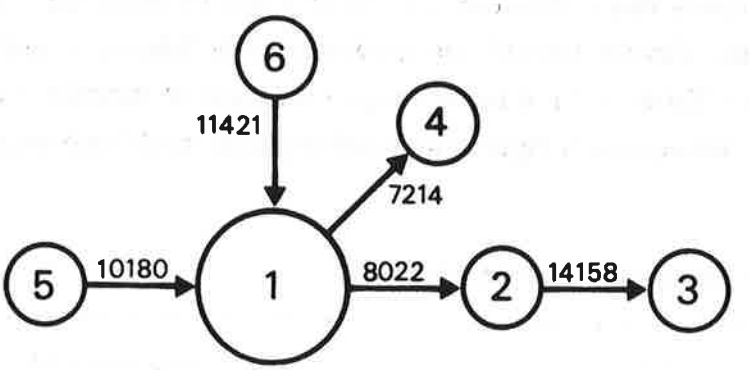

FIGURE 4.21 : Différenciation des morphes *Taq I*. Le sens des flèches indique un gain de site.

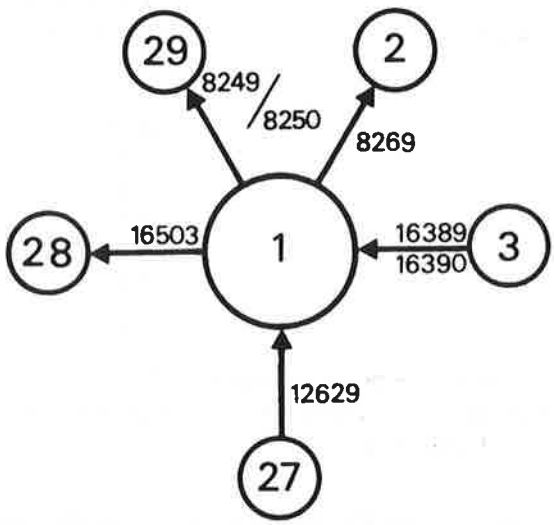

FIGURE 4.22 : Différenciation des morphes *Ava II*. Le sens des flèches indique un gain de site.

TABLE 4.34: Liste et fréquence des morphes *Ava II*

Morphes[1]	Sites polymorphes[2]	Fréquence (%) (N = 116)
1	-	87,9
2	31	0,9
3	72	4,3
27	48	1,7
28	74	0,9
29	30	4,3

[1] La numérotation adoptée ici correspond à celle de la Table 4.12.
[2] Ces nombres indiquent les sites qui diffèrent de la séquence de Cambridge, leur numérotation correspond à celle de la Table 4.28

Tous les morphes semblent être différenciés à partir du type 1 et aucun ne dépasse une fréquence de 5 %. Il est intéressant de constater que l'on retrouve le morphe 2 dans une population asiatique, alors que l'on pensait qu'il était restreint aux seules populations européennes et africaines. Cela confirme l'importance de la constitution de meilleurs échantillons du point de vue de l'effectif qui permettraient de détecter un nombre de morphes nettement plus élevé (voir Figure 2.1). Le morphe 3, déja retrouvé dans une population orientale (Tharu) semble être présent ici avec une fréquence non négligeable. Notons que la différenciation entre les morphes 3 et 22 n'est normalement possible qu'avec une double digestion *Ava II* et *BamH I*; cependant il semble que nous ayons bien affaire au morphe 3, car la double mutation *Ava II/Hinf I* diffère bien de la double mutation *Ava II/BamH I* du morphe 22 (voir la discussion pour les morphes *Hinf I*).

Hpa II

Hpa II est un isoschizomère de l'enzyme *Msp I*, c'est à dire qu'il reconnaît la même séquence sur l'ADN. La différence entre ces 2 enzymes réside dans l'incapacité de *Hpa II* de reconnaître la séquence CCGG si des nucléotides de cytosine sont méthylés. On rapprochera donc les morphes *Hpa II* des morphes *Msp I* définis dans les Tables 4.10 et 4.11.

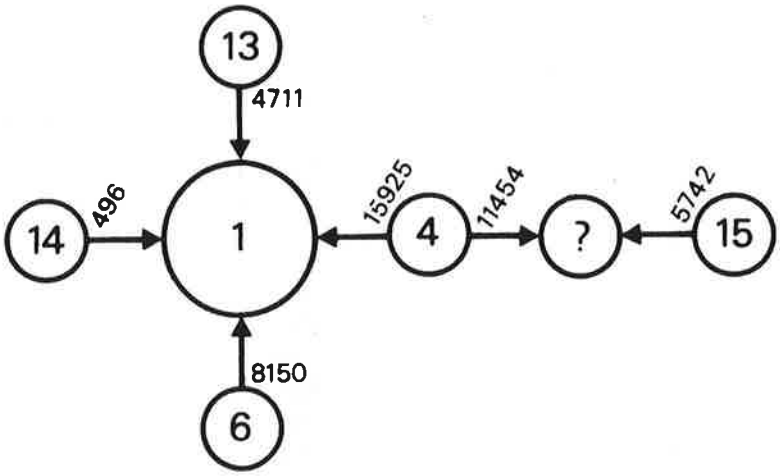

FIGURE 4.23 : Différenciation des morphes *Hpa II*. Le sens des flèches indique un gain de site.

TABLE 4.35: Liste et fréquence des morphes *Hpa II*

Morphes[1]	Sites polymorphes[2]	Fréquence (%) (N = 116)
1	-	94
4	65	2,6
6	29	0,9
13	14	0,9
14	2	0,9
15	20,45,65	0,9

[1] La numérotation adoptée pour les morphes 1, 4 et 6 correspond à celle de la Table 4.10 pour les morphes de l'enzyme *Msp I* qui est un isoschizomère de *Hpa II*.
[2] Ces nombres indiquent les sites qui diffèrent de la séquence de Cambridge, leur numérotation correspond à celle de la Table 4.28

Dans l'échantillon japonais, 6 morphes ont pu être définis, dont 3 nouveaux (Nº 13, 14 et 15) (voir Figure 4.23 et Table 4.35). Parmi les morphes déjà recensés, nous trouvons bien évidemment le morphe 1, à une fréquence comparable à celles des populations orientale (97,8 %) et Tharu (90,1 %). Le morphe 4 est également présent avec des fréquences comparables entre ces 3 populations. Enfin, le morphe 6 est également présent chez 1 individu (seulement). Ce morphe n'avait ét retrouvé, jusqu'ici, que parmi un échantillon d'Israéliens Arabes, mais il avait une importance considérable dans le réseau de différenciation des morphes *Msp I* (Figure 4.4), car il faisait la liaison entre le morphe 1 (central) et une série de morphes retrouvés uniquement dans les échantillons africains. Le fait de retrouver ce morphe dans une population japonaise tend à suggérer qu'il pourrait être très ancien et être apparu avant la séparation des populations occidentales, orientales et africaines. Il faut toutefois rester prudent quant à cette interprétation, étant donné la fréquence très faible de ce morphe dans les 2 échantillons où il a été retrouvé et la possibilité de l'occurrence de 2 mutations parallèles et indépendantes.

Sau3A I

La digestion de l'ADN-mt par cet enzyme a généré 6 morphes différents dans cet échantillon. Le morphe 1 atteint presque une fréquence de 95 % (Table 4.36). Les 5 autres morphes sont directement dérivés du morphe 1 (Figure 4.24).

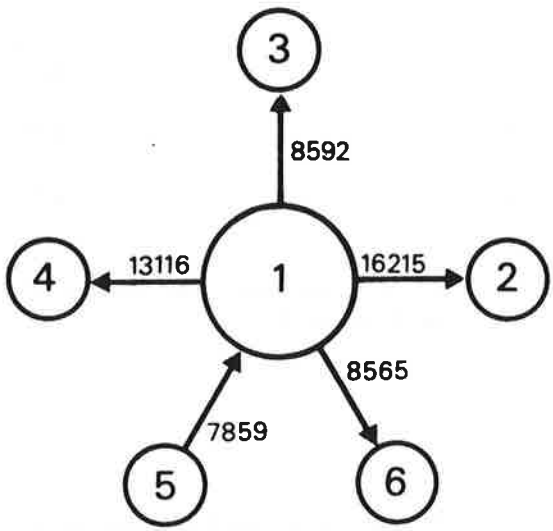

FIGURE 4.24 : Différenciation des morphes *Sau3 AI*. Le sens des flèches indique un gain de site.

TABLE 4.36: Liste et fréquence des morphes *Sau3A I*

Morphes	Sites polymorphes[1]	Fréquence (%) (N = 116)
1	-	94,8
2	70	1,7
3	35	0,9
4	49	0,9
5	26	0,9
6	34	0,9

[1] Ces nombres indiquent les sites qui diffèrent de la séquence de Cambridge, leur numérotation correspond à celle de la Table 4.28

Acc II

Seuls 2 morphes *Acc II* ont été identifiés (Figure 4.25). Ils diffèrent l'un de l'autre par une seule substitution. La fréquence du morphe 1 dépasse 98 % (Table 4.37). vu le faible degré de polymorphisme obtenu (qui devrait être confirmé par l'examen d'autres populations), l'utilité de cet enzyme semble limitée pour l'étude de la différenciation de populations.

TABLE 4.37: Liste et fréquence des morphes *Acc II*

Morphes	Sites polymorphes[1]	Fréquence (%) (N = 116)
1	-	98,3
2	52	1,7

[1] Ces nombres indiquent les sites qui diffèrent de la séquence de Cambridge, leur numérotation correspond à celle de la Table 4.28

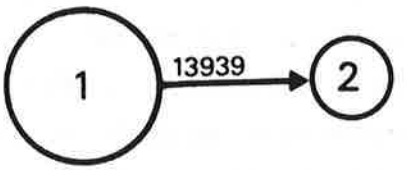

FIGURE 4.25 : Différenciation des morphes *Acc II*. Le sens des flèches indique un gain de site.

Définition et fréquence des types d'ADN-mt

La combinaison des morphes de 9 enzymes a permis à Horai et Matsunaga de définir 62 types d'ADN-mt parmi les 116 individus examinés. La nomenclature utilisée se base sur celle de Horai et Matsunaga (1986). Cependant, comme nous n'avons pas employé l'information provenant des délétions et insertions pour des raisons méthodologiques (voir plus haut), la constitution qualitative des types sera quelque peu différente de la leur. Ainsi, par exemple, le type 32 s'avère pour nous équivalent au type 30, car ils ne diffèrent que par une mutation de longueur, ce qui ramène le nombre de types à 61. Ces différences tendent toutes à diminuer le nombre de mutations entre les types, mais notre définition des types se base sur des mutations dont la nature est connue, contrairement à certaines mutations de longueurs.

Ce nombre de 61 types se rapproche fortuitement de celui trouvé pour l'ensemble des 10 populations étudiées précédemment avec 5 enzymes. Il est aussi élevé en raison du grand nombre d'enzymes employés. Bien que ce système de types s'avère potentiellement très informatif, il se situe à la limite de l'utilité pour l'étude d'une population, car pas moins de 47 types ne sont retrouvés qu'à un unique exemplaire. Les 15 autres types ne possèdent pas des fréquences très élevées (≈11 % pour les 2 plus fréquents).

Dans ces conditions, il semble clair que les fréquences des types rares sont mal estimées. L'augmentation de la taille de l'échantillonnage aurait aussi pour conséquence de faire apparaître de nouveaux types, vraisemblablement retrouvés chez un seul individu, ce qui ne contribuerait pas beaucoup à de meilleures estimations des types rares. Une telle situation montre qu'une étude quantitative des types s'avère dénuée de sens et plaide pour une approche qualitative de la constitution en types de cette population. Il semble presque plus judicieux de classer les types en 2 catégories -les types rares et les types plus fréquents-, et de ne comparer que les fréquences de ces derniers dans une comparaison éventuelle entre 2 populations, les types rares ne pouvant être confrontés qu'en terme de présence ou absence.

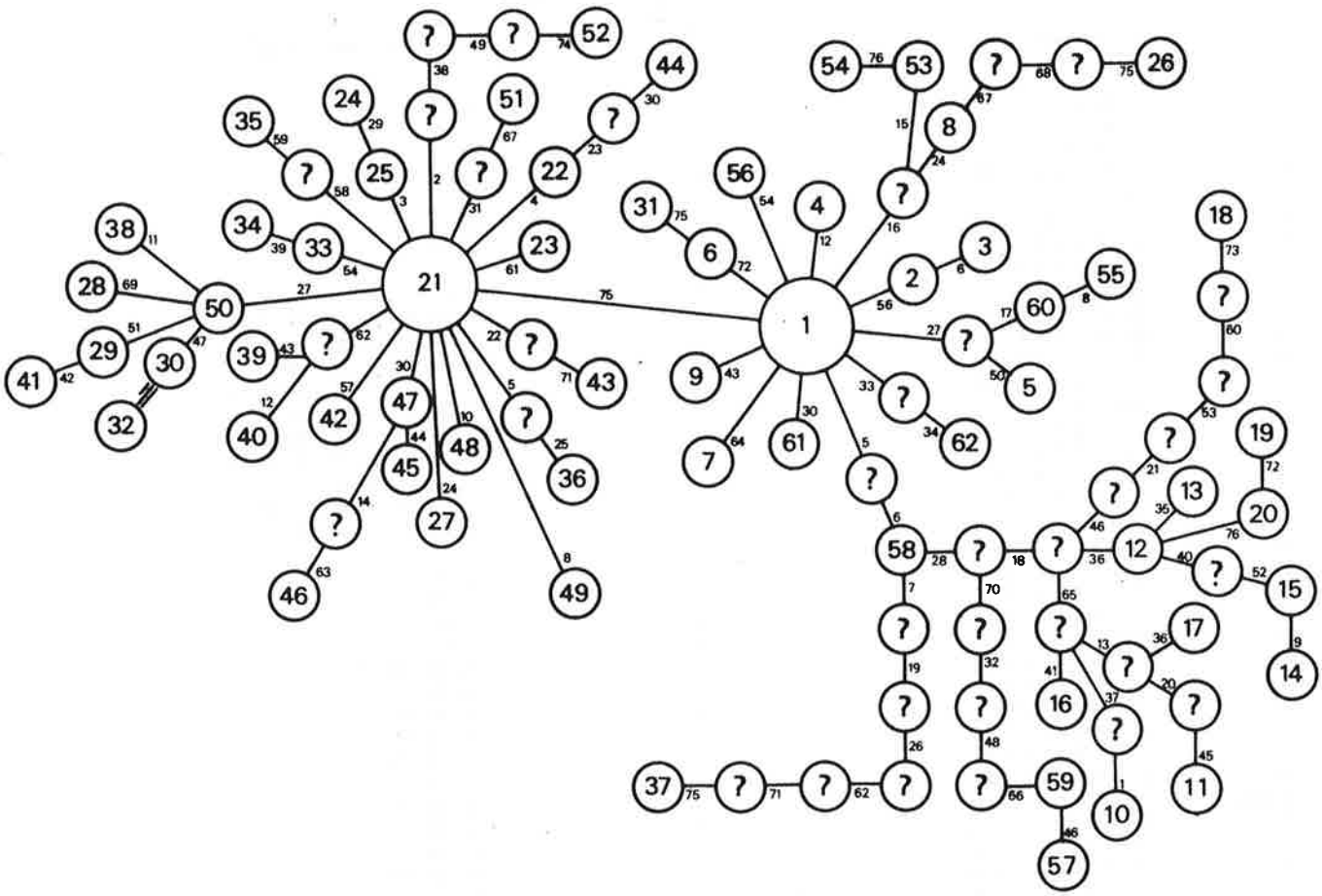

FIGURE 4.26 : Phylogénie des 62 types d'ADN-mt définis par Horai et Matsunaga (1986). Le type 21 est le type ancestral hypothétique (voir texte).

TABLE 4.38: Liste et fréquence des types d'ADN-mt définis dans Horai and Matsunaga (1986)

Type[1]	Morphes[2]	Nombre observé dans l'échantillon	Sites polymorphes[3]
1	(1 1 1 1 1 1 1 1 1)	12	
2	(1 10 1 1 1 1 1 1 1)	4	56
3	(1 10 1 6 1 1 1 1 1)	6	6; 56
4	(1 11 1 1 1 1 1 1 1)	1	12
5	(1 7 5 1 1 1 1 1 1)	1	27; 50
6	(1 9 1 1 1 3 1 1 1)	2	72
7	(1 1 1 12 1 1 1 1 1)	1	64
8	(1 1 3 1 1 1 1 1 1)	1	16; 24
9	(1 1 1 1 5 1 1 1 1)	1	43
10	(4 3 10 2 2 1 4 1 1)	1	1; 5; 18; 28; 37; 46; 65
11	(4 4 1 7 2 1 15 1 1)	1	5; 13; 20; 28; 36; 45; 46; 65
12	(4 4 1 2 2 1 1 1 1)	8	5; 18; 28; 36; 46
13	(4 4 1 2 2 1 1 3 1)	1	5; 18; 28; 35; 36; 46
14	(4 6 7 3 2 1 1 1 2)	1	5; 9; 18; 28; 36; 40; 46; 52
15	(4 5 7 2 2 1 1 1 2)	1	5; 18; 28; 36; 40; 46; 52
16	(4 3 9 2 2 1 4 1 1)	1	5; 18; 28; 36; 41; 46; 65
17	(8 4 1 7 2 1 4 1 1)	1	5; 13; 18; 28; 36; 46; 55; 65
18	(10 12 1 2 3 1 1 1 1)	1	5; 18; 21; 28; 53; 60; 73
19	(9 6 1 2 2 3 1 1 1)	2	5; 18; 28; 36; 46; 72; 76
20	(9 4 1 2 2 1 1 1 1)	1	5; 18; 28; 36; 46; 76
21	(2 1 1 1 1 1 1 1 1)	13	75
22	(2 1 1 5 1 1 1 1 1)	1	4; 75
23	(2 1 1 9 1 1 1 1 1)	1	61; 75
24	(2 1 1 4 1 1 6 1 1)	1	3; 29; 75
25	(2 1 1 4 1 1 1 1 1)	1	3; 75
26	(2 1 3 11 1 1 1 1 1)	2	16; 24; 67; 68; 75
27	(2 1 4 1 1 1 1 1 1)	1	24; 75
28	(2 7 1 13 1 1 1 1 1)	1	27; 69; 75
29	(2 8 1 1 1 1 1 1 1)	2	27; 51; 75
30	(2 7 8 1 1 1 1 1 1)	2	27; 47; 75
31	(2 9 1 1 1 3 1 1 1)	1	72; 75
32≡30	(2 7 8 1 1 1 1 1 1)	1	27; 47; 75
33	(22 1 1 1 1 1 1 1 1)	5	54; 75
34	(22 1 1 8 1 1 1 1 1)	1	39; 54; 75
35	(24 1 1 1 1 1 1 1 1)	1	58; 59; 75
36	(21 1 6 1 1 1 1 1 1)	1	5; 25; 75
37	(25 2 1 1 1 1 1 5 1)	1	5; 7; 19; 26; 46; 62; 71; 75
38	(13 7 1 1 1 1 1 1 1)	1	11; 27; 75
39	(14 1 1 1 5 1 1 1 1)	1	43; 62; 75
40	(14 11 1 1 1 1 1 1 1)	1	12; 62; 75
41	(15 8 1 1 1 1 1 1 1)	1	27; 42; 51; 75
42	(18 1 1 1 1 1 1 1 1)	1	57; 75
43	(23 1 1 1 1 1 1 1 1)	1	22; 71; 75
44	(20 1 1 5 4 29 1 1 1)	1	4; 23; 30; 75
45	(20 1 1 1 6 29 1 1 1)	1	30; 44; 75
46	(20 13 1 1 1 29 13 1 1)	1	14; 30; 63; 75
47	(20 1 1 1 1 29 1 1 1)	1	30; 75
48	(19 1 1 1 1 1 1 1 1)	1	10; 75
49	(16 1 1 1 1 1 1 1 1)	1	8; 75
50	(2 7 1 1 1 1 1 1 1)	1	27; 75
51	(2 1 1 10 1 2 1 1 1)	1	31; 67; 75
52	(17 1 1 1 1 28 14 4 1)	1	2; 38; 49; 74; 75
53	(11 1 2 1 1 1 1 1 1)	5	15; 16
54	(12 1 2 1 1 1 1 1 1)	1	15; 16; 76
55	(26 7 1 1 1 1 1 1 1)	1	8; 17; 27
56	(5 1 1 1 1 1 1 1 1)	2	54
57	(4 1 1 14 2 27 1 2 1)	1	5; 28; 32; 48; 66; 70
58	(4 2 1 1 1 1 1 1 1)	2	5; 46
59	(4 2 1 14 2 27 1 2 1)	1	5; 28; 32; 46; 48; 66; 70
60	(6 7 1 1 1 1 1 1 1)	2	17; 27
61	(3 1 1 1 1 29 1 1 1)	1	30
62	(7 1 1 1 1 1 1 6 1)	1	33; 34

[1] Cette nomenclature des types est conforme à celle qui a été établie par Horai and Matsunaga (1986)

[2] L'ordre des 9 enzymes est *Hae III*, *Hinf I*, *Hha I*, *Rsa I*, *Taq I*, *Ava II*, *Hpa II*, *Sau3A I*, *Acc II*

[3] Les sites mutants sont définis par rapport à la séquence de Cambridge qui est équivalente au type 1 et correspondent à la nomenclature de la Table 4.28.

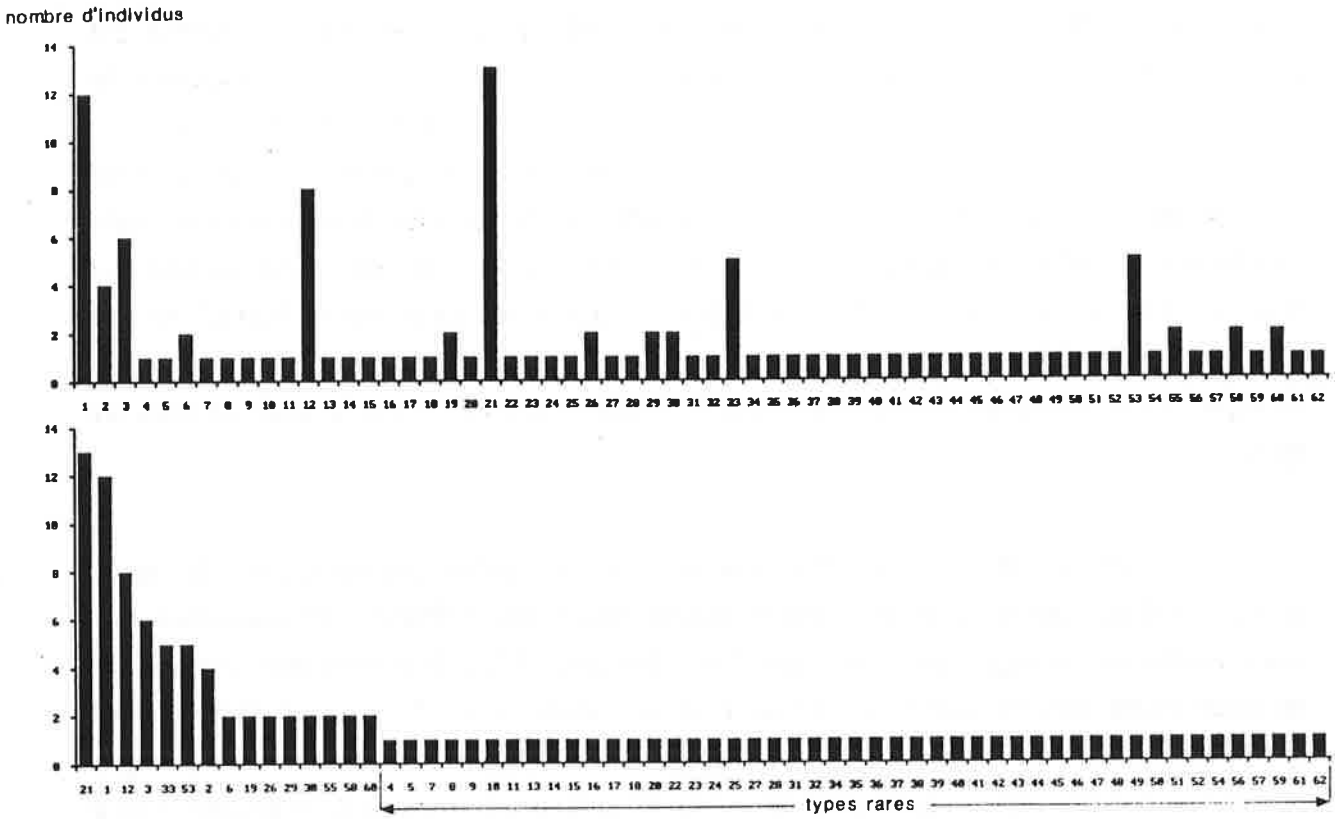

FIGURE 4.27 : Fréquences des 62 types d'ADN-mt définis par Horai et Matsunaga (1986).

La Figure 4.26 présente un réseau phylogénique entre les 61 types, tels que nous les avons définis, qui répond au critère de maximum de parcimonie. Ce réseau est composé de 3 subdivisions principales, à savoir un ensemble de types centré autour du type 1, un autre centré autour du type 21, et enfin un dernier ensemble articulé à partir du type 58 et regroupé autour du type 12, mais comprenant un nombre considérable de types manquants. Les types 1, 12 et 21 sont d'ailleurs les types les plus fréquents. Cette structure en 3 pôles est également perceptible dans le réseau des morphes *Hae III* (Figure 4.17), les types 1, 12 et 21 possédant respectivement les morphes *Hae III* Nº 1, 4 et 2 qui correspondent également aux 3 centres de différenciation des morphes *Hae III*. Le type 21 est identique au type ancestral hypothétique calculé sur la base de tous les types.

Horai et Matsunaga (1986) ont construit un "arbre phylogénique" des types au moyen d'une méthode de classement automatique dite UPGMA ("*Unweighted pair group arithmetic average clustering*", Sokal and Sneath, 1973). Il semble bon d'examiner les différences existant entre ce type de représentation et le nôtre. Tout d'abord, leur arbre ne distingue que 2 groupes qui correspondent, d'une part, à la réunion des groupes centrés autour des types 1 et 21 plus les types 37 et 58 et, d'autre part, le groupe centré autour du type 12. Ce type de représentation ne permet pas de voir les types manquants, de préciser des types centraux ou de déterminer une succession précise de mutations ayant conduit à un type donné (ex: 21 → 50 → 29 → 41). Cette représentation sépare des groupes surtout en fonction de leur différence moyenne mais peine à montrer les ressemblances et les liens phylogéniques d'ordre 1 (types séparés par une seule substitution). Ainsi, au vu de la Figure 4.26, on observe que les différences entre les types du groupe centré autour du type 12 et les autres types sont plus importantes que les différences calculées entre les types des groupes centrés autour du 1 et du 21, ce qui conduit à ignorer l'importance historique des types 1 et 21.

Horai et Matsunaga (1986) ont cependant associé une échelle de temps, dépendante de la fraction de sites de restriction partagés entre 2 types (déterminé au moyen de l'équation 4.6), aux niveaux de regroupement des types. Ils ont également déterminé une racine au niveau du dernier regroupement qui situerait la divergence à partir d'un ancêtre commun hypothétique à environ 125'000 ans. Sur la Figure 4.26, cela consisterait à placer la racine entre les types 12 et 58, ce qui serait très difficile à justifier. En principe, la détermination de l'emplacement d'une racine est conditionnée par la connaissance d'un ancêtre commun extérieur au groupe étudié, ce qui n'est pas le cas ici. Nous avons cependant pu déterminer que le type 21 était équivalent au type

ancestral hypothétique à partir duquel les autres types seraient dérivés. Cela suppose que l'on accepte l'hypothèse d'une différenciation des types s'étant effectuée à l'intérieur même du groupe considéré, ce qui exclut le cas, pourtant très envisageable, d'une différenciation séparée, suivie d'un regroupement dans un échantillon non homogène.

Nous avons déjà émis quelques réserves quant au processus d'échantillonnage, et le fait d'observer l'absence d'un grand nombre d'intermédiaires dans les branches ayant conduit aux types 10 à 20, ainsi qu'aux types 37 et 57 à 59, laisse supposer que ce processus n'a pas été homogène pour tous les types. Il semblerait que quelques types extérieurs ont été introduits à partir d'une autre population où un processus différent de diversification moléculaire des types aurait pu avoir lieu. La constitution d'un meilleur échantillon et l'étude de populations voisines des japonais pourrait permettre de vérifier ces hypothèses.

En consultant la Figure 4.22, on est forcé de constater l'importance des types 1 et 21 à partir desquels la majorité des autres types s'est différenciée. Le fait qu'ils soient les plus fréquents est compatible avec une grande ancienneté potentielle. Une population ancestrale composée principalement de ces 2 types aurait très bien pu se diversifier de la manière observée actuellement. Le principe du calcul d'un type ancestral ne comporte pas la possibilité d'une diversification d'une population ou d'un groupe de populations à partir de plusieurs types de fréquences plus ou moins équivalentes, et donc d'une population déjà polymorphe. C'est pourtant un événement fort vraisemblable dans le cas de cet échantillon de japonais, à moins que la population japonaise actuelle ait connu une origine mixte due à des vagues de peuplement successives.

Diversité moléculaire

Nous avons déterminé le nombre moyen de différences de sites de restriction entre 2 gènes sur l'ensemble des 61 types (ν) au moyen de l'équation (4.12). Le résultat a été reporté dans la Table 4.33. Nous avons également calculé l'estimation de $E(\nu)=\theta$ au moyen de (4.15).

TABLE 4.39 : Nombre moyen de différences de sites de restriction

Nombre de gènes	Nombre de types	Nombre de sites polymorphes	$\hat{\nu}$	θ
Ensemble des types				
116	61	76	4,59	14,27
Ensemble réduit[1]				
91	47	59	3,01	11,61

[1] Les types 10 à 20, 37, et 57 à 59 n'ont pas été considérés ici. Voir le texte pour les justifications.

Les 2 types d'estimations fournissent des valeurs très différentes (d'un facteur supérieur à 3), alors que sous l'hypothèse de neutralité sélective des mutations, elle devraient fournir des valeurs comparables. Une constitution non homogène de l'échantillon japonais pourrait également conduire à surestimer la valeur de θ donnée par (4.15). Afin de vérifier si la branche issue du type 58 perturbait nos estimations, nous avons recalculé les valeurs de ν et de θ en ne considérant que les types centrés autour des types 1 et 21 et en éliminant donc les types 10 à 20, 37 et 57 à 59. Les résultats sont également reportés dans la Table 4.39. Comme l'on pouvait s'y attendre, les valeurs des 2 estimateurs ont baissé, mais le rapport d'un facteur 3 a été conservé. Il semble donc bien que, comme dans le cas des populations orientales et occidentales déjà étudiées, le nombre de différences de sites de restriction soit inférieur à ce que l'on attendrait, au vu du nombre élevé de sites polymorphes, dans un un système génétique neutre.

TABLE 4.40: Estimation du temps t (en années) nécessaire pour créer une diversité nucléotidique Π à partir d'une population monomorphe.

Population	Π_1 ($\times 10^4$) (4.18)	Π_2 ($\times 10^4$) (4.22)	t_1	t_2
Totalité des types	14,46	44,96	144'740	450'973
Ensemble réduit[1]	9,48	36,58	94'860	366'694

[1] Les types 10 à 20, 37, et 57 à 59 n'ont pas été considérés ici. Voir le texte pour les justifications.

Les estimations du temps t nécessaire pour créer les diversités moléculaires estimées au moyen des équations (4.18) et (4.22) ont été reportées dans la Table 4.40.

La valeur de R' calculée par (4.21) est de 3'173,1. On retrouve, bien sûr, une valeur de t_2 plus élevé que celle de t_1 (voir les explications dans la discussion concernant la Table 4.21). Les chiffres de t_2 concernant l'ensemble de types réduit sont du même ordre de grandeur que ceux qui ont été trouvés pour d'autres populations asiatiques (Tharu népalais), alors que l'estimation concernant la totalité des types est plus élevée d'environ 100'000 ans. La phylogénie des types de la Figure 4.26 suggère que l'échantillon de japonais analysé par Horai et Matsunaga est hétérogène, ce qui impliquerait que les gènes mitochondriaux des japonais proviennent de centres de plusieurs foyers de différenciation, d'où un temps de diversification apparent très élevé.

b) Analyse de la population au moyen d'enzymes reconnaissant 6 pb

Horai *et al.* (1984) ont utilisé 15 enzymes possédant des séquences de reconnaissance de 6 pb. Parmi ceux-ci, 4 enzymes (*Xba I*, *Kpn I*, *Bam HI* et *Dra I*) n'ont produit qu'un seul morphe correspondant à la séquence de Cambridge et conduisent donc à des profils de digestion monomorphes. Les 11 autres enzymes (*Hinc II*, *Hae II*, *Eco RV*, *Pst I*, *Xho I*, *Hind III*, *Stu I*, *Sac I*, *Sca I*, *Eco RI* et *Pvu I*) ont produit au moins 2 morphes différents

Localisation des sites de reconnaissance polymorphes

Nous avons reporté les positions des 24 sites polymorphes par rapport à la séquence de Cambridge dans la Table 4.41. Tous les sites sont indépendants et ne conduisent pas à des doubles mutations apparentes pour 2 enzymes. Les sites 19 et 21 sont absents par rapport à la séquence de Cambridge, mais cette absence totale dans l'échantillon laisse penser que la séquence mutante est plutôt la séquence de Cambridge elle-même.

Les fréquences des autres sites polymorphes sont inférieures à 5 %, à l'exception des sites 6 et 24. Le site Hae II Nº 6 n'est d'ailleurs pas indépendant du site Hha I Nº 16 vu précédemment. Aucune mutation de longueur n'a pu être mise en évidence. Les sites polymorphes ne sont pas assez nombreux pour déterminer si ils se répartissent uniformément sur le génome mitochondrial.

TABLE 4.41: Sites de restriction polymorphes par substitution de nucléotides (Horai *et al.*, 1984)

Position	Site Nº	Polymorphisme[1]	Fréquence(%) (N=120)	Enzyme	Région
981	1	1	0,8	*Pvu II*	ARN-r 12 S
1006	2	0	0,8	*Hinc II*	"
4121	3	0	0,8	*Eco R I*	NAD 1
4582	4	1	1,7	*Hind III*	NAD 2
4740	5	0	1,7	*Sca I*	"
4830	6	1	8,3	*Hae II*	"
7196	7	1	2,5	*Stu I*	CO I
7364	8	1	1,7	*Pst I*	"
7855	9	0	2,5	*Hinc II*	CO II
9056	10	0	1,7	*Hae II*	ATPase 6
9643	11	0	0,8	*Sac I*	CO III
9736	12	1	0,8	*Hind III*	"
9745	13	0	0,8	*Sca I*	"
9758	14	1	0,8	*Xho I*	"
12026	15	1	2,5	*Hinc II*	NAD 4
12408	16	0	4,2	*Hinc II*	NAD 5
13259	17	0	2,5	*Hinc II*	"
13598	18	0	0,8	*Hae II*	"
13701	19	0	100	*Stu I*	"
14157	20	1	0,8	*Xho I*	URF 6
14201	21	0	100	*Hinc II*	"
14862	22	0	0,8	*Hae II*	Cyt b
15046	23	0	0,8	*Stu I*	"
16274	24	1	5,0	*Eco R V*	D-Loop

[1] La notation "1" correspond à un polymorphisme dû à un gain de site par rapport à la séquence de Cambridge, alors que la notation "0" correspond à la perte d'un site.

Définition des morphes

Nous avons repris la nomenclature adoptée par Horai *et al.* (1984) pour la numérotation des morphes des différents enzymes, à l'exception des morphes *Hae II* vus précédemment. D'une manière générale, les profils de digestion sont beaucoup moins polymorphes que dans les cas précédents, où les digestions étaient effectuées avec des enzymes reconnaissant 4 ou 5 pb. Cela est dû au fait que le nombre de sites potentiels pour les enzymes reconnaissant 6 pb est plus faible que les derniers cités (voir Table 4.1). Le nombre de sites étant moins élevé, le nombre total de pb surveillées le sera également, ce qui fait que, pour un même taux de mutation, la probabilité qu'une base ait muté à l'intérieur d'un site de 6 pb sera plus faible que dans un site de 4 ou 5 pb.

Hinc II

Les morphes *Hinc II* ont été décrit par Blanc *et al.* (1983) qui en ont recensé 7 dans un échantillon de 116 individus. Parmi ceux-ci, les morphes 1, 2, 3 et 5 ont été retrouvés parmi l'échantillon japonais. Trois nouveaux morphes ont été reconnu par Horai *et al.* (1984) et numérotés 9 à 11. (Figure 4.24). Notons que le morphe correspondant à la séquence de Cambridge n'a pas encore été retrouvé dans un échantillon. Cette séquence apparaît donc comme un mutant probablement rare pour l'enzyme *Hinc II*.

TABLE 4.42: Liste et fréquence des morphes *Hinc II*

Morphes	Sites polymorphes[1]	Fréquence (%) (N = 120)
1	16, 21	3,3
2	21	88,3
3	15, 21	2,5
5	9, 21	2,5
9	17, 21	1,7
10	2, 21	0,8
11	16, 17, 21	0,8

[1] Ces nombres indiquent les sites qui diffèrent de la séquence de Cambridge, leur numérotation correspond à celle de la Table 4.41

La fréquence élevée du morphe 2 (Table 4.42) est compatible avec le réseau de différenciation proposé dans la Figure 4.28 où il apparaît comme central. Le morphe 11 peut être issu des morphes 1 ou 9. Sans pouvoir préciser duquel il s'agit, il s'avère donc qu'un des sites 16 ou 17 a subi 2 mutations indépendantes. Il faut noter que le morphe N°6 (issu du morphe 5) défini par Blanc *et al.* (1983) avait déja connu la perte du site 17.

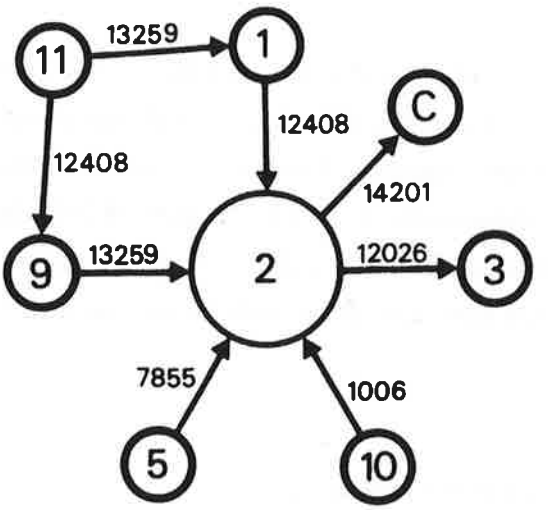

FIGURE 4.28 : Différenciation des morphes *Hinc II*. Le sens des flèches indique un gain de site.

Hae II

La diversification des morphes Hae II est relativement simple et seuls 5 morphes ont été trouvés dans cette population, dont 2 nouveaux (Nº 10 et 11) d'après notre propre numérotation. Ces derniers viennent s'ajouter aux 9 morphes déjà définis plus haut (voir Figure 4.3).

Le morphe 1 apparaît comme un centre de différenciation assez net puisque la majorité des autres morphes en sont directement issus. Les fréquences observées (Table 4.43) sont relativement proches de celles d'autres populations orientales où les morphes 1, 2 et 5 avaient également été observés (voir Table 4.9). Le morphe 11 peut être issu de 2 autres morphes (voir Figure 4.3), mais il semble plus probable qu'il soit dérivé du morphe 5 que l'on retrouve dans cet échantillon avec une fréquence de 7,5 %.

TABLE 4.43: Liste et fréquence des morphes *Hae II*

Morphes	Sites polymorphes[1]	Fréquence (%) (N = 120)
1	-	89,1
2	10	1,7
5	6	7,5
8	18	0,8
9	6, 22	0,8

[1] Ces nombres indiquent les sites qui diffèrent de la séquence de Cambridge, leur numérotation correspond à celle de la Table 4.41

Autres enzymes (*Eco RV*, *Pst I*, *Xho I*, *Hind III*, *Stu I*, *Sac I*, *Sca I*, *Eco RI* et *Pvu II*)

Les autres enzymes qui possèdent un profil de digestion polymorphe présentent des schémas de diversification extrêmement simples (Figure 4.29) qui s'articulent généralement autour d'un morphe très fréquent (> 95 %) qui occupe une position centrale (lorsque l'on trouve plus de 2 morphes).

Seuls les morphes Stu I méritent d'être examinés de plus près. En effet, le morphe 1 (96,7%) diffère de la séquence de Cambridge par l'absence du site 19. Cette

particularité s'explique de la même façon que la perte du site Hinc II Nº 21, c'est à dire par une mutation de la séquence de Cambridge ayant conduit à un gain de site à partir du morphe le plus fréquent qui serait analogue au 1.

TABLE 4.44: Liste et fréquence des morphes *Eco RV*, *Pst I*, *Xho I*, *Hind III*, *Stu I*, *Sac I*, *Sca I*, *Eco RI* et *Pvu II*.

	Morphes	Sites polymorphes[1]	Fréquence (%) (N = 120)
Eco RV			
	1	-	95,0
	2	24	5,0
Pst I			
	1	-	98,3
	2	8	1,7
Xho I			
	1	-	98,3
	2	20	0,8
	3	14	0,8
Hind III			
	1	-	97,5
	2	4	1,7
	3	12	0,8
Stu I			
	1	19	96,7
	2	7, 19	2,5
	3	7,19, 23	0,8
Sac I			
	1	-	99,2
	2	11	0,8
Sca I			
	1	-	97,5
	2	5	1,7
	3	13	0,8
Eco RI			
	1	-	99,2
	2	3	0,8
Pvu II			
	1	-	99,2
	2	1	0,8

[1] Ces nombres indiquent les sites qui diffèrent de la séquence de Cambridge, leur numérotation correspond à celle de la Table 4.41

Le morphe 2 semble issu du morphe 1 et a subi une mutation qui a donné naissance au morphe 3. Les autres enzymes n'amènent pas de commentaires particuliers et l'examen des Table 4.44 et Figure 4.29 suffisent à percevoir les mécanismes ayant permis de créer les différents morphes.

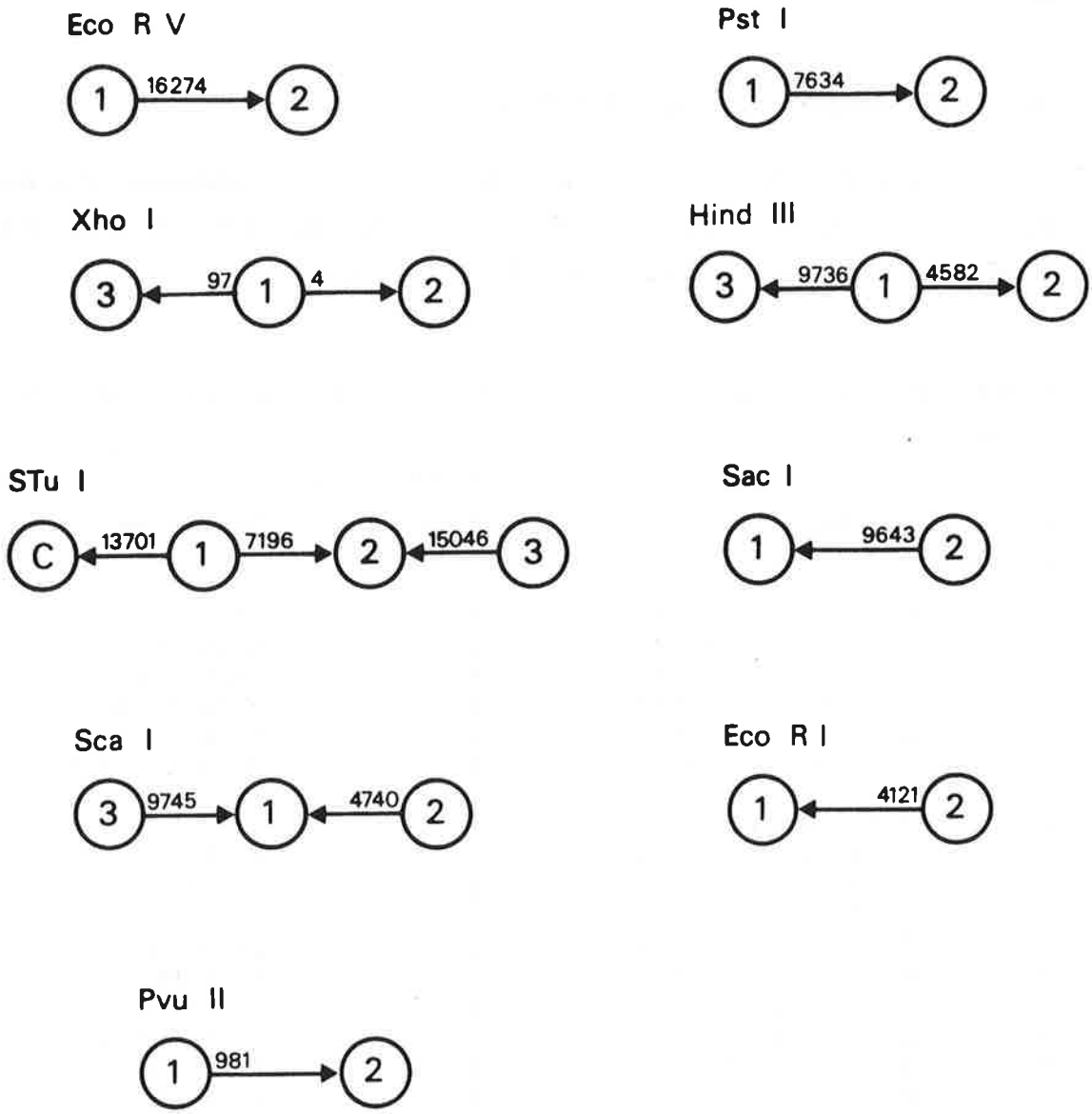

FIGURE 4.29 : Différenciation des morphes *Eco RV*, *Xho I*, *Stu I*, *Sca I*, *Pvu II*, *Pst I*, *Hind III*, *Sac I* et *Eco RI*. Le sens des flèches indique un gain de site.

Définition et fréquence des types d'ADN-mt

En combinant les morphes des 11 enzymes produisant des profils de digestion polymorphes, 22 types d'ADN-mt ont pu être mis en évidence (Table 4.45). La numérotation des types de Horai *et al.* (1984) a été respectée.

TABLE 4.45: Liste et fréquence des types d'ADN-mt définis dans Horai *et al.* (1984)

Type[1]	Morphes[2]	Nombre observé dans l'échantillon	Sites mutants[3]
1	(2 1 1 1 1 1 1 1 1 1 1)	86	19, 21
2	(2 5 1 1 1 1 1 1 1 1 1)	5	6, 19, 21
3	(2 1 2 1 1 1 1 1 1 1 1)	3	19, 21, 24
4	(2 5 2 1 1 1 1 1 1 1 1)	2	6, 19, 21, 24
5	(2 2 1 1 1 1 2 1 1 1 1)	1	7, 10, 19, 21
6	(2 11 2 1 1 1 1 1 1 1 1)	1	6, 19, 21, 22, 24
7	(2 10 1 1 1 1 1 1 1 1 1)	1	18, 19, 21
8	(2 1 1 2 1 1 1 1 1 1 1)	2	8, 19, 21
9	(2 1 1 1 1 2 1 1 1 1 1)	2	4, 19, 21
10	(2 1 1 1 1 3 1 1 1 1 1)	1	12, 19, 21
11	(2 1 1 1 1 1 1 2 1 1 1)	1	11, 19, 21
12	(2 1 1 1 3 1 1 1 1 1 1)	1	14, 19, 21
13	(1 1 1 1 1 1 1 1 2 1 1)	2	5, 16, 19, 21
14	(1 1 1 1 1 1 3 1 1 2 1)	1	3, 7, 16, 19, 21, 23
15	(1 2 1 1 1 1 1 1 1 1 1)	1	10, 16, 19, 21
16	(3 1 1 1 1 1 1 1 1 1 1)	3	15, 19, 21
17	(5 1 1 1 1 1 1 1 1 1 1)	2	9, 19, 21
18	(5 5 1 1 1 1 1 1 1 1 1)	1	6, 9, 19, 21
19	(9 5 1 1 2 1 2 1 3 1 1)	1	6, 7, 13, 17, 19, 20, 21
20	(9 1 1 1 1 1 2 1 1 1 1)	1	7, 17, 19, 21
21	(10 1 1 1 1 1 1 1 1 1 2)	1	1, 2, 19, 21
22	(11 1 1 1 1 1 1 1 1 1 1)	1	16, 17, 19, 21
		120	

[1] Cette nomenclature des types est conforme à celle qui a été établie par Horai et al. (1984)
[2] L'ordre des 11 enzymes est *Hinc II*, *Hae II*, *Eco RV*, *Pst I*, *Xho I*, *Hind III*, *Stu I*, *Sac I*, *Sca I*, *Eco RI* et *Pvu II*
[3] Les sites mutants sont définis par rapport à la séquence de Cambridge qui est équivalente au type 1 et correspondent à la nomenclature de la Table 4.41.

Le type 1 est nettement dominant (71,7 %) dans cet échantillon. Il correspond également au type duquel la majorité des autres types semble issue. Ceux-ci sont relativement rares et aucun n'atteint 5 % dans l'échantillon. Mentionnons néanmoins le type 2, retrouvé chez 5 individus, qui semble avoir donné naissance à 2 autres types (4 et 18) et qui pourrait constituer un centre de différenciation secondaire.

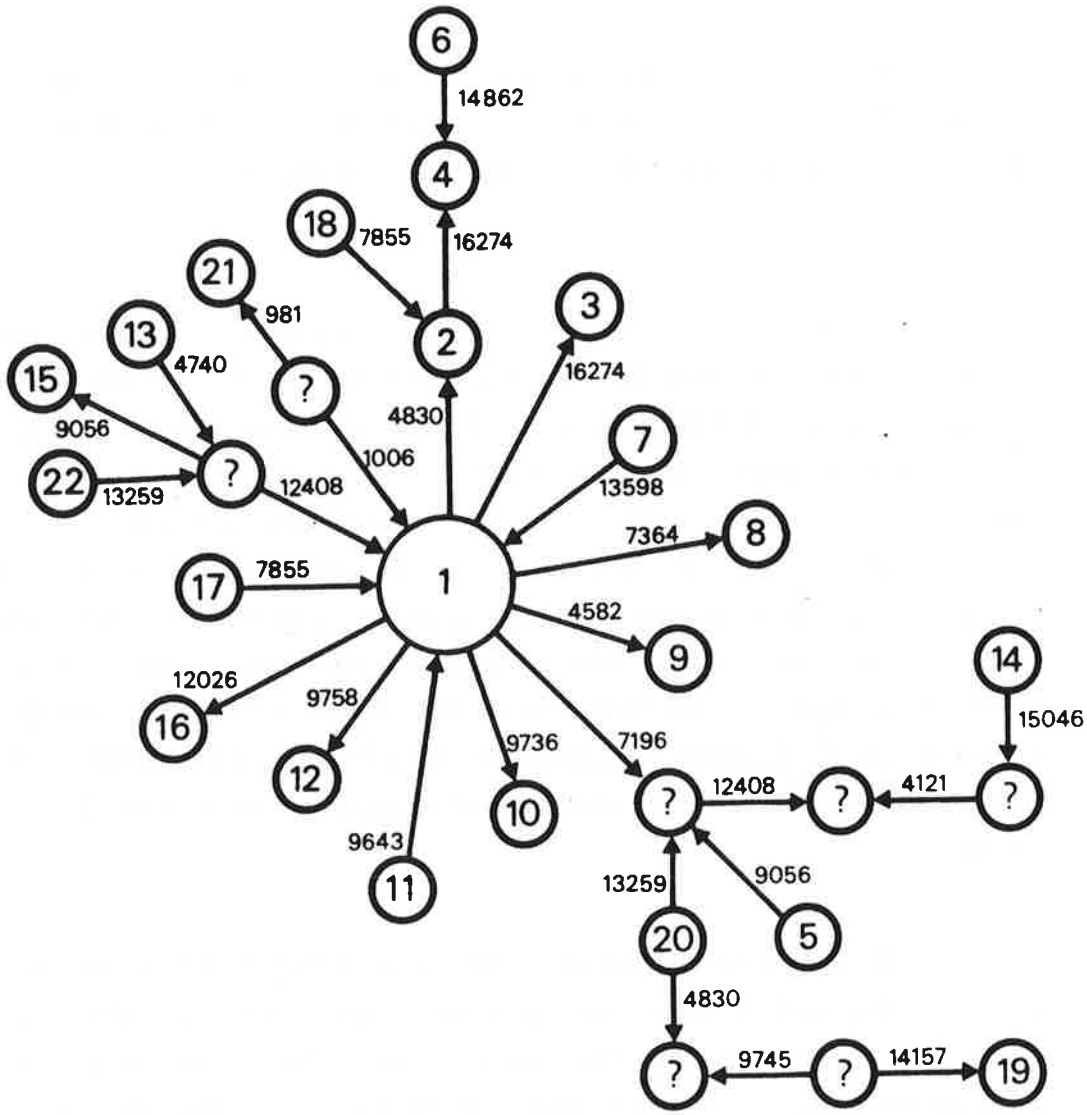

FIGURE 4.30 : Phylogénie des types d'ADN-mt définis par Horai *et al.* (1984). Le type 1 est le type ancestral hypothétique (voir texte).

Un certain nombre de sites ont apparemment été l'objet de mutations parallèles. Il s'agit des sites 6, 9, 16 et 17 (respectivement aux positions 4830, 7855, 12408 et 13259) (voir Figure 4.30) qui sont tous impliqués 2 fois dans des passages entre types.

Contrairement au schéma de différenciation des types provenant de l'analyse du même échantillon avec des enzymes reconnaissant 4 ou 5 pb, il n'existe pas de structure tripolaire bien marquée. La grande majorité des types est issue directement ou indirectement du type 1 qui est identique au type ancestral hypothétique calculé sur la base de tous les types. Cependant, il faut noter qu'une branche de diversification contenant les types 5, 14, 19 et 20 n'est pas directement rattachée au type 1. De nombreux types intermédiaires (5) sont nécessaires pour relier tout ces types au N°1. Cela doit être mis en rapport avec ce que l'on avait observé auparavant pour la branche contenant les types 37, 58 à 59 et 10 à 20 de la Figure 4.26. De ce point de vue, il aurait été intéressant de disposer des haplotypes complets pour les enzymes reconnaissant 4, 5 et 6 pb, afin de vérifier si certains des individus appartenant à ces 2 branches sont les mêmes ou non.

D'une manière générale, le réseau de la Figure 4.30 est moins diversifié que celui de la Figure 4.26, étant donné le nombre plus restreint de types trouvés. Bien que n'étant pas structuré de manière identique, ces 2 réseaux ne sont pas contradictoires. Leur confrontation semble confirmer l'hypothèse d'un apport extérieur de types très différenciés et d'une diversification relativement plus réduites à l'intérieur de la population japonaise.

Diversité moléculaire

Le nombre moyen de différences de sites de restriction entre 2 gènes tirés au hasard (ν) a été déterminé au moyen de l'équation (4.12) et reporté dans la Table 4.40.

On retrouve une différence très nette entre ν et $E(\nu) = \theta$ calculée par (4.15). Cela confirme le défaut de diversité moléculaire constaté précédemment pour cette population. Ce résultat est compatible avec l'existence d'une valeur sélective relative plus élevée du type 1 (le plus fréquent), qui conduirait à réduire considérablement la valeur de l'estimation de ν qui dépend fortement des fréquences des types dans la population.

TABLE 4.46 : Nombre moyen de différences de sites de restriction

Nombre de gènes	Nombre de types	Nombre de sites polymorphes	$\hat{v}$	θ
120	22	22	0,86	4,10

La Table 4.47 présente les estimations des temps de diversification des types. Le nombre de nucléotides effectivement surveillés dans cette étude est de 768,95. L'estimation d'environ 500'000 ans nécessaires pour aboutir à la diversité moléculaire observée est certainement surestimée du fait de la probable hétérogénéité de l'échantilnon japonais, déjà décrite précédemment. Nous constatons cependant qu'il existe un écart de 80'000 ans dans les estimations de t_2 provenant des études de Horai *et al.* (1984) et d'Horai et Matsunaga (1986), qui portent sur le même échantillon, mais pas sur la même portion de l'ADN-mt. Ceci nous conduit à constater l'importance de la variabilité de t_2. Le nombre de nucléotides effectivement surveillés étant 4 fois plus important dans l'étude précédente, la première estimation de t_2 nous semble plus fiable et aussi plus raisonnable. Un problème reste en suspens et concerne la source de l'hétérogénéité de l'échantillon de japonais. Celle-ci pourrait soit provenir d'une hétérogénéité propre à la population, soit du processus d'échantillonnage.

TABLE 4.47: Estimation du temps t (en années) nécessaire pour créer une diversité nucléotidique Π à partir d'une population monomorphe.

Population	Π_1 ($\times 10^4$) (4.18)	Π_2 ($\times 10^4$) (4.22)	t_1	t_2
Japonais	11,08	53,30	110'882	534'903

2) ETUDE D'UNE POPULATION JAPONAISE DE L'ILE DE HOKKAIDO

Harihara *et al.* (1986) ont étudié l'ADN-mt extrait de cellules du sang périphérique de 122 individus habitants l'île de Hokkaido, au nord du Japon. Cette étude avait, entre autres, pour objectif d'analyser d'éventuelles différences au niveau génétique entre un sous-échantillon de 48 Ainu (qui seraient les représentants d'un peuplement ancien des îles du Japon) et d'un autre sous-échantillon de 74 non-Ainu (japonais qui seraient issus d'une population arrivée plus tardivement du continent asiatique).

Ces 2 sous-échantillons ont été étudiés avec 5 enzymes (*Ava II, Hinc II, Hpa I, Bam HI* et *Pvu II*) déjà utilisés dans d'autres études de l'ADN-mt.

Localisation des sites polymorphes

La position et la fréquence des 10 sites polymorphes dans l'ensemble de l'échantillon ont été reportés dans la Table 4.48.

TABLE 4.48: Sites de restriction polymorphes par substitution de nucléotides (Harihara *et al.*, 1986)

Position	Nº	Polymorphisme[1]	Fréquence (%) (N=122)	Enzyme	Région
3881	1	1	2,5	*Ava II*	NAD I
6384	2	1	0,8	*Ava II*	CO I
7855	3	0	1,6	*Hinc II*	CO II
8249	4	1	0,8	*Ava II*	"
12026	5	1	3,3	*Hinc II/Hpa I*	NAD 4
12408	6	0	4,9	*Hinc II*	NAD 5
12629	7	0	0,8	*Ava II*	"
12753	8	1	0,8	*Pvu II*	"
14201	9	0	100	*Hinc II/Hpa I*	URF 6
16390	10	0	2,5	*Ava II*	D-Loop

[1] La notation "1" correspond à un polymorphisme dû à un gain de site par rapport à la séquence de Cambridge, alors que la notation "0" correspond à la perte d'un site.

Les sites 5 et 9 sont à la fois des sites pour l'enzymes *Hpa I* et *Hinc II*, car ce dernier possède une séquence de reconnaissance multiple (G T C/T A/G A C) qui inclut la séquence de reconnaissance de *Hpa I* (G T T A A C). Le site 9 est polymorphe

par rapport à la séquence de Cambridge, mais monomorphe dans l'échantillon. Seules les mutations des sites 1, 5, 6 et 10 ont été retrouvées chez plus de 2 individus, sans toutefois atteindre une fréquence de 5 % dans l'échantillon.

Définition des morphes

Tous les morphes recensés dans cette étude avaient déjà été définis auparavant, à l'exception d'un morphe *Pvu II* et d'un morphe *Ava II*. Nous avons modifié l'appellation de ce dernier (noté 14 par Harihara *et al.* (1986) et modifié en 30), pour tenir compte des 29 autres morphes définis pour cet enzyme. La digestion des 122 ADN-mt n'a pas produit de profils polymorphes pour *Bam HI* par rapport à la séquence de Cambridge

Ava II

La définition et les fréquences des 6 morphes trouvés pour l'enzyme *Ava II* sont reportés dans la Table 4.49. On observe que les Ainu sont monomorphes pour cet enzyme. Un nouveau morphe (noté N°30), différencié à partir du morphe 3 a été trouvé dans l'échantillon de japonais non-Ainu. Les 5 autres morphes *Ava II* avaient déjà été observés dans d'autres populations humaines. La fréquence élevée du morphe 1 est très semblable à celles qui ont été trouvées dans d'autres populations orientales (voir Tables 4.13 et 4.34). Le morphe 3 a également été retrouvé dans les populations Tharu et japonaises avec des fréquences du même ordre, quoiqu'un peu plus élevées. Le nouveau morphe (30) que l'on trouve ici semble issu du morphe 3 (Figure 4.31), ce qui souligne le fait que ce dernier serait bien ancré dans les populations orientales. Le morphe 27 avait été trouvé dans une autre population japonaise (Horai and Matsunaga, 1986), alors que le morphe 10 n'avait été localisé que dans une population de caucasoïdes (voir Table 4.13).

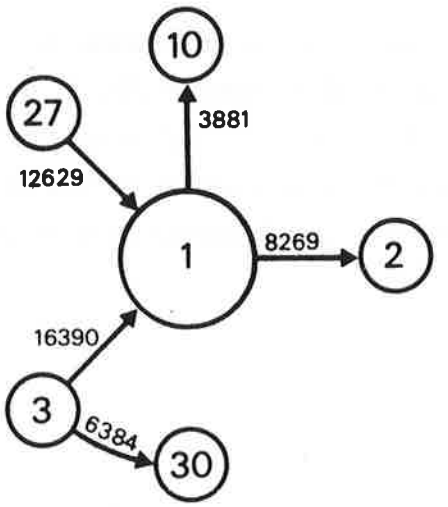

FIGURE 4.31 : Différenciation des morphes *Ava II*. Le sens des flèches indique un gain de site.

TABLE 4.49: Liste et fréquence des morphes *Ava II*

Morphes	Sites polymorphes[1]	Fréquence (%) Ainu (N = 48)	non-Ainu (N = 74)
1	-	100	89,2
2	4	0	1,4
3	10	0	2,7
10	1	0	4,1
27	7	0	1,4
30	2,10	0	1,4

[1] Ces nombres indiquent les sites qui diffèrent de la séquence de Cambridge, leur numérotation correspond à celle de la Table 4.48

Si l'on ne doit pas attacher trop d'importance aux fréquences des morphes rares, étant donné la faible taille des échantillons, on s'aperçoit, une fois de plus, qu'un morphe que l'on croyait restreint à une aire géographique est retrouvé avec une faible fréquence sur un autre continent. L'absence de mutations parallèles ne peut être certifiée, car nous ne possédons pas de types comparables entre différentes études, mais il est clair qu'il est impossible de prétendre à une absence totale d'un morphe donné dans une région quelconque.

La comparaison des populations Ainu et non-Ainu de cette partie du Japon montre que ces derniers semblent plus polymorphes. Etant donné la faiblesse numérique des échantillons en présence, le monomorphisme apparent des Ainu peut très bien être dû à un défaut d'échantillonnage. Par conséquent, nous nous abstiendrons de mettre en exergue des différences minimes de fréquences géniques probablement dues en grande partie à des fluctuations aléatoires.

Hinc II

Les morphes *Hinc II* 1, 2, 3 et 5 avaient déjà été retrouvés dans une population japonaise (Horai *et al.*, 1984) avec des fréquences du même ordre que celles de l'échantillon non-Ainu (voir Tables 4.42 et 4.50). Le morphe 2 apparaît bien comme central du point de vue de la diversification des autres morphes (voir Figure 4.28).

TABLE 4.50: Liste et fréquence des morphes *Hinc II*

Morphes	Sites polymorphes[1]	Fréquence (%) Ainu (N = 48)	non-Ainu (N = 74)
1	6, 9	2,1	6,8
2	9	95,8	86,5
3	5, 9	2,1	4,1
5	3, 9	0	2,7

[1] Ces nombres indiquent les sites qui diffèrent de la séquence de Cambridge, leur numérotation correspond à celle de la Table 4.48

Hpa I

Les morphes *Hpa I* 1, 2 et 4 ont tous été identifiés dans des populations orientales (Table 4.5), le morphe 2 étant le plus fréquent dans les populations orientales et occidentales et apparemment le centre de la différenciation des autres morphes (Figure 4.1).

TABLE 4.51: Liste et fréquence des morphes *Hpa I*

Morphes	Sites polymorphes[1]	Fréquence (%) Ainu (N = 48)	non-Ainu (N = 74)
1	6	2,1	6,8
2	-	95,8	89,2
4	5	2,1	4,1

[1] Ces nombres indiquent les sites qui diffèrent de la séquence de Cambridge, leur numérotation correspond à celle de la Table 4.48

La structure des populations Ainu et non-Ainu semble comparable (Table 4.51), tant du point de vue de la composition des morphes que de leur fréquence, au vu des variations aléatoires dues à l'échantillonnage.

Pvu II

Un nouveau morphe *Pvu II* (N°3) apparaît à un unique exemplaire dans l'échantillon non-Ainu (Table 4.52). Celui-ci est dérivé du morphe 1, nettement

dominant et défini auparavant dans la Figure 4.28, par une unique substitution à la position 12753.

TABLE 4.52: Liste et fréquence des morphes *Pvu II*

Morphes	Sites polymorphes[1]	Fréquence (%) Ainu (N = 48)	non-Ainu (N = 74)
1	-	100	98,6
3	8	0	1,4

[1] Ces nombres indiquent les sites qui diffèrent de la séquence de Cambridge, leur numérotation correspond à celle de la Table 4.48

Définition des types d'ADN-mt

Un total de 11 types d'ADN-mt ont été définis parmi les japonais non-Ainu en combinant les morphes des 3 enzymes *Ava II*, *Hinc II* et *Pvu II* (Table 4.47). Parmi ceux-ci, seuls les types 1, 2 et 4 sont retrouvés chez les Ainu. Les morphes *Hpa I* n'ont pas été reportés ici, car ils sont équivalents à certains morphes *Hinc II*.

TABLE 4.53: Liste et fréquence des types d'ADN-mt définis dans Horai and Matsunaga (1986)

Type[1]	Morphes[2]	Fréquence Ainu (N=48)	non-Ainu (N=74)	Sites mutants[3]
1	(1 2 1)	95,8	77,7	9
2	(1 1 1)	2,1	5,4	6, 9
3	(10 2 1)	0	4,1	1, 9
4	(1 3 1)	2,1	2,7	5, 9
5	(1 5 1)	0	2,7	3, 9
6	(2 2 1)	0	1,4	4, 9
7	(3 2 1)	0	1,4	9, 10
8	(3 3 1)	0	1,4	5, 9, 10
9	(27 1 1)	0	1,4	6, 7, 9
10	(30 2 1)	0	1,4	2, 9, 10
11	(1 2 3)	0	1,4	8, 9

[1] Cette nomenclature des types est conforme à celle qui a été établie par Harihara *et al.* (1986))
[2] L'ordre des 3 enzymes est *Ava II*, *Hinc II* et *Pvu II*
[3] Les sites mutants sont définis par rapport à la séquence de Cambridge et correspondent à la nomenclature de la Table 4.48.

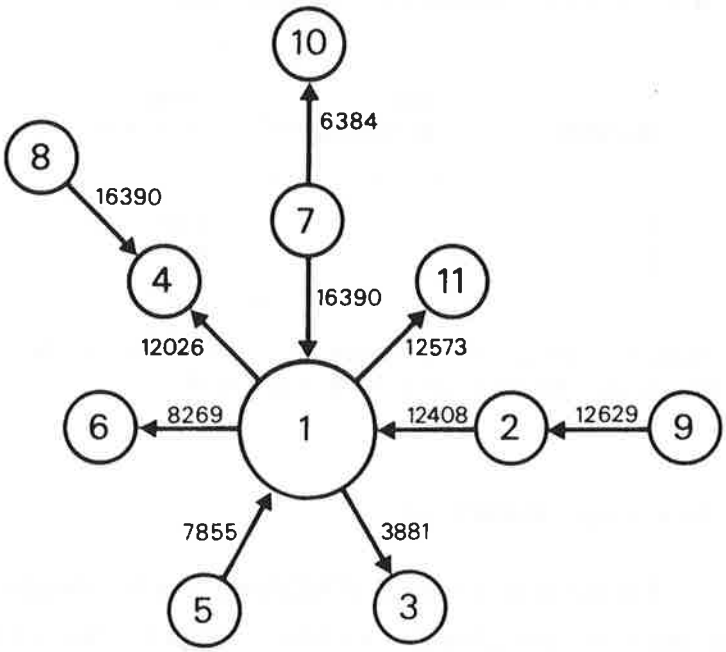

FIGURE 4.32 : Phylogénie des types d'ADN-mt définis par Harihara *et al.* (1986). Le type 1 est le type ancestral hypothétique (voir texte).

Le type 1 est le plus fréquent dans les 2 populations et correspond au type ancestral hypothétique du réseau de différenciation des types de la Figure 4.32. Le type 2 est le second type le plus fréquent et semble avoir conduit au type 9. Le type 8 peut théoriquement être issu des types 4 ou 7. A notre sens, la liaison 4 → 8 est la plus probable en raison de la fréquence plus élevée du type 4 par rapport au type 7 et du fait que le site 10 se trouve dans la région de la D-Loop. A ce propos, notons que ce site 10 semblerait avoir disparu 2 fois de façon indépendante dans la formation des types 7 et 8. Ce phénomène de mutation parallèle avait déjà été observé dans la Figure 4.5.

Les type 1, 2 et 4, communs aux 2 échantillons, sont tous à la source d'au moins un autre type, ce qui plaide pour leur ancienneté. Dans ce sens, les 2 populations Ainu et non-Ainu apparaissent très proche l'une de l'autre et semblent avoir évolué à partir d'un même stock génétique, ceci dans l'hypothèse où il n'y aurait pas eu d'échange génétique majeur ultérieurement, ce qui ne peut être totalement exclu. L'absence de nombreux types parmi les Ainu peut simplement résulter de la petite taille de cet échantillon.

Diversité moléculaire

Le nombre moyen de différences de sites de restriction entre 2 gènes tirés au hasard a été calculé au moyen de la formule (4.12) pour les 2 sous-échantillons de japonais et reportés dans la Table 4.54. Dans les 2 cas, ce nombre est nettement inférieur à la valeur attendue ($E(v)=\theta$) calculée en fonction du nombre de sites polymorphes dans l'échantillon par l'équation (4.15). Ceci confirme ce que nous avions déjà observé dans les autres populations orientales et occidentales.

TABLE 4.54 : Nombre moyen de différences de sites de restriction

	Nombre de gènes	Nombre de types	Nombre de sites polymorphes	$\hat{v}$	θ
Ainu	48	3	2	0,08	0,45
non-Ainu	74	11	9	0,53	1,85

L'utilisation des 3 enzymes *Ava II*, *Pvu II* et *Hinc II* ont permis de surveiller approximativement 342 nucléotides. Les temps correspondant à la diversification des types dans les échantillons Ainu et japonais ont été reporté dans la Table 4.55. Les estimations de t_2 donnent des valeurs d'environ 130'000 ans pour les Ainu et de plus de 500'000 ans pour les japonais, ce qui est proche de l'estimation de Horai *et al.* (1984). Ce dernier fait suggère que l'hétérogénéité constatée dans l'échantillon japonais serait plutôt due à une propriété de la population japonaise dans son ensemble. Plusieurs vagues de peuplement du Japon par différentes populations auraient pu créer une telle hétérogénéité observée actuellement et expliquer des temps de différenciation aussi élevés.

TABLE 4.55: Estimation du temps t (en années) nécessaire pour créer une diversité nucléotidique π à partir d'une population monomorphe.

Population	π_1 ($\times 10^4$) (4.18)	π_2 ($\times 10^4$) (4.22)	t_1	t_2
Ainu	2,32	13,16	23'204	131'716
non-Ainu	15,37	54,10	153'858	542'961

En ce qui concerne les Ainu, considérés comme provenant d'un peuplement ancien du Japon, ils présentent une moins grande diversité moléculaire que les japonais, bien que les types qu'ils possèdent soient également retrouvés chez ces derniers (voir Figure 4.32 et Table 4.53). Ces types communs montrent qu'ils proviennent d'un stock génétique proche de celui des japonais, mais moins différencié, ce qui suggère que les quelques 130'000 ans constituent une nette surestimation d'une éventuelle date de divergence à partir d'une population commune avec les japonais.

4) DONNÉES PORTANT SUR UN ENSEMBLE D'INDIVIDUS

Lors du chapitre précédent, nous avons passé en revue les études de l'ADN-mt humain ayant porté sur des populations. Dans tous les cas, nous avions affaire à un échantillon tiré d'une population relativement bien définie sur le plan géographique, ethnique ou religieux. Les raisons qui ont poussé des chercheurs à tenter d'amasser des échantillons homogènes sont liées à des considérations méthodologiques dérivant des théories de la génétique des populations (reposant maintenant sur des études précises et nombreuses: voir Malécot, 1966; Wright, 1969; Crow and Kimura, 1970; Nei, 1975, 1987; Ewens, 1979; Kimura, 1983; Hedrick, 1985) qui permettent de modéliser de façon probabiliste l'apparition, la transmission et l'évolution des fréquences de gènes mutants, et cela à l'intérieur de populations de taille finie.

La constitution d'un échantillon est donc censée fournir des estimations de certains paramètres inconnus qui concernent la population d'où il est tiré. Meilleur sera l'échantillon du point de vue quantitatif et qualitatif, et meilleures seront les estimation de ces paramètres, et donc plus précise sera la connaissance de la population. La comparaison de ces paramètres entre populations peut ensuite être utilisée pour en tirer des inférences concernant leur différenciation génétique.

Ce processus de comparaisons de populations est donc très différent de la confrontation 2 à 2 de tous les individus composant ces populations. Lewontin (1974) a montré depuis longtemps que la principale composante de la diversité entre 2 groupes de populations était la variabilité inter-individuelle et donc que la diversité intrapopulation était supérieure à la diversité interpopulation.

Il est donc nécessaire d'estimer cette variabilité intrapopulation avant de se lancer dans des comparaisons de populations. Cette démarche, qui a été entreprise pour des systèmes génétiques classiques (groupes sanguins ou tissulaires,) n'est pas encore pleinement assurée dans le cas de l'étude du polymorphisme de l'ADN qui débute à peine. Dans certains cas, cela peut conduire à des interprétations hasardeuses sur la divergence de populations ou groupes humains qui sont largement tributaires d'un manque de rigueur dans la constitution même des échantillons.

Pour ces raisons, nous avons tenu à séparer les études de l'ADN-mt effectuées sur des échantillons tirés de populations, d'autres études qui ont plus tendu à mettre en évidence les différences existant entre des individus issus de groupes continentaux sans le souci de constituer des échantillons homogènes. Ces dernières

études tiennent donc plus de l'analyse de différences entre individus avec une extension injustifiée à un discours sur les groupes d'où ils sont tirés, car cette démarche est forcément biaisée par l'absence de connaissance sur la représentativité de ces individus dans les groupes en question.

Cependant, bien que ces études ne remplissent pas les critères de qualité que nous nous sommes assignés, elles ne peuvent pas être négligées et ignorées. De plus, elles sont à la base de nouvelles théories largement médiatisées concernant l'origine de l'homme (ou de la femme) qui prônent l'apparition d'une Eve africaine il y a environ 200'000 ans. Il nous semble donc plus judicieux d'analyser ces études de façon détaillée, de les confronter aux données de l'ADN-mt récoltées sur des échantillons plus homogènes et de voir dans quelle mesure les biais méthodologiques peuvent influer sur les résultats obtenus.

Analyse d'un échantillon de 176 individus provenant de 5 groupes humains

Cann *et al.* (1987) ont entrepris d'analyser l'ADN-mt de 147 individus répartis sur divers continents. Il convient tout d'abord de préciser quelque peu la manière dont a été constitué l'échantillon. L'ADN-mt a été extrait à partir de 145 placentas humains de donneurs volontaires et de 2 lignées cellulaires. Les 147 individus ont été répartis selon 5 groupes de populations: 20 individus de souche africaine (dont 18 noirs américains, 1 San (!Kung) et 1 nigérian), 34 asiatiques (chinois américains, vietnamiens, laotiens, philippins, indonésiens et habitants des îles Tonga), 46 caucasoïdes (Nord américains, européens, africains du Nord et moyen-orientaux), 21 aborigènes australiens et 26 aborigènes de Nouvelle-Guinée. On peut d'emblée constater la vaste diversité de provenance des individus rassemblés pêle-mêle dans les échantillons africains, caucasoïdes et asiatiques, ainsi que la très faible taille des échantillons.

Une autre étude portant sur le même échantillon et utilisant les mêmes enzymes a été publiée (Stoneking *et al.*, 1986) avec 29 individus supplémentaires originaires de Nouvelle-Guinée. Malheureusement, les fréquences des types de ces nouveaux individus n'ont pas été donnés de façon détaillée. Par conséquent, il ne nous sera pas possible d'inclure ce nouvel échantillon dans les aspects de notre travail qui incluent des fréquences des types d'ADN-mt bien définis. Cependant, nous avons inclus cette étude pour la localisation des sites polymorphes, ainsi que pour la construction de la phylogénie des types.

L'ADN-mt a été analysé au moyen de 12 enzymes (*Hpa I*, *Ava II*, *Fnu DII*, *Hha I*, *Hpa II*, *Mbo I*, *Taq I*, *Rsa I*, *Hinf I*, *Hae III*, *Alu I* et *Dde I*) qui ont permis de mettre en évidence 204 sites de restriction polymorphes.

Localisation des sites de reconnaissance polymorphes

Nous avons reporté la position et la fréquence de 204 sites de restriction polymorphes indépendants dans la Table 4.56 et la Figure 4.33. Notons que 195 sites avaient été recensés dans la seule étude de Cann *et al.* (1987) et que 9 sites polymorphes supplémentaires ont été identifiés par Stoneking et ses collaborateurs chez 29 autres individus. La Figure 4.33 représente schématiquement la position des 195 sites polymorphes trouvés par Cann *et al.* (1987). La notation du polymorphisme des sites diffère quelque peu de celle qui a été utilisée précédemment. Ainsi, il n'est plus défini par rapport à la séquence de Cambridge, qui correspond en fait au type Nº 110, car ce dernier semble posséder des présences ou absences de sites nettement minoritaires dans l'échantillon qui suggèrent que cette séquence aurait accumulé plusieurs mutations peu fréquentes. Aussi, pour chaque site, la référence sera l'état majoritaire dans l'échantillon. Signalons qu'un seul type possède les 195 sites dans leur état majoritaire: il s'agit du type 69 qui peut donc être considéré comme le type ancestral hypothétique et constituera une référence pour cette étude.

Une grande partie des sites ne sont retrouvés qu'à des fréquences très faibles à l'exception des sites 4, 14, 16, 78, 124, 140, 148, 180, 199, 200, et 204 qui dépassent tous 10 %. Le fait que ces sites soient polymorphes chez un grand nombre d'individus peut être expliqué par 2 causes différentes : le polymorphisme est ancien et s'est répandu progressivement dans la (les) population(s) ou bien le site a été le sujet de substitutions indépendantes sur plusieurs types d'ADN-mt. Ces 2 hypothèses ne sont, bien sûr, pas en opposition ni exclusives pour le même site.

TABLE 4.50: Sites de restriction polymorphes par substitution de nucléotides (Cann *et al.*, 1987; Stoneking *et al.*, 1986)

Position	Site Nº	Polymorphisme[2]	Fréquence(%)[1] (N=147)	Enzyme	Région
8	1	1	0.07	*Mbo I*	D-Loop
64	2	1	0.01	*Hpa II*	"
134	3	1	0.01	*Taq I*	"
207	4	1	0.12	*Hpa I*	"
255	5	1	0.01	*Hha I*	"
259	6	1	0.01	*Alu I*	"
340	7	1	0.01	*Mbo I*	"
663	8	1	0.02	*Hae III*	ARN-r 12S
712	9	1	0.01	*Hpa I*	"
740	10	0	0.03	*Mbo I*	"
748	11	1	0.01	*Ava II*	"
1240	12	0	0.03	*Alu I*	"
1043	13	1	?	*Dde I*	"
1403	14	1	0.18	*Alu I*	"
1463	15	0	0.01	*Hae III*	"
1484	16	0	0.22	*Hae III*	"
1536	17	1	0.01	*Hha I*	"
1610	18	0	0.03	*Alu I*	ARN-t^{Val}
1637	19	0	0.02	*Dde I*	"
1667	20	0	0.01	*Dde I*	"
1715	21	0	0.03	*Dde I*	ARN-r 16S
1917	22	0	0.01	*Alu I*	"
2208	23	0	0.01	*Alu I*	"
2223	24	1	0.01	*Alu I*	"
2384	25	1	0.01	*Alu I*	"
2390	26	1	0.05	*Mbo I*	"
2734	27	0	0.03	*Alu I*	"
2578	28	0	0.05	*Rsa I*	"
2849	29	0	0.01	*Rsa I*	"
3123	30	0	0.01	*Rsa I*	"
3315	31	0	0.01	*Hae III*	"
3337	32	0	0.02	*Rsa I*	"
3391	33	1	0.02	*Hae III*	"
3537	34	0	0.01	*Alu I*	"
3592	35	1	0.08	*Hpa I*	"
3698	36	0	0.02	*Hha I*	"
3842	37	1	0.01	*Hae III*	"
3849	38	0	0.01	*Hae III*	"
3899	39	1	0.07	*Taq I*	"
3930	40	1	0.01	*Dde I*	"
3944	41	0	0.01	*Taq I*	"
4092	42	1	0.01	*Hinf I*	"
4411	43	0	0.03	*Alu I*	ARN-t^{Met}
4464	44	0	0.01	*Rsa I*	"
4481	45	1	0.01	*Ava II*	NAD 2
4631	46	0	0.01	*Alu I*	"
4643	47	1	0.01	*Rsa I*	"
4732	48	1	0.01	*Rsa I*	"
4769	49	1	0.01	*Alu I*	"

Position	Site Nº	Polymorphisme[1]	Fréquence(%)[2] (N=147	Enzyme	Région
4793	50	1	0.01	*Hae III*	NAD 2
5176	51	0	0.04	*Alu I*	"
5261	52	0	0.01	*Hae III*	"
5269	53	0	0.01	*Taq I*	"
5351	54	1	0.02	*Hha I*	"
5538	55	1	0.01	*Hha I*	ARN-t^{Trp}
5552	56	0	0.01	*Dde I*	"
5742	57	0	0.01	*Hpa II*	O_L
5754	58	1	0.01	*Hpa II*	"
5978	59	0	0.01	*Alu I*	CO I
5983/5984	60	1	0.01	*Hinf I/Ava II*	"
5983/5985	61	1	0.08	*Hinf I/Rsa I*	"
5996	62	0	0.01	*Alu I*	"
6022	63	0	0.01	*Alu I*	"
6166	64	1	0.01	*Hha I*	"
6211	65	0	0.03	*Hinf I*	"
6260	66	0	0.01	*Hae III*	"
6356	67	1	0.01	*Dde I*	"
6377	68	0	0.01	*Dde I*	"
6409	69	1	0.01	*Taq I*	"
6501	70	1	0.01	*Hpa II*	"
6610	71	1	0.01	*Hinf I*	"
6699	72	1	0.01	*Ava II*	"
6871	73	0	0.01	*Hinf I*	"
6904	74	0	?	*Mbo I*	"
6915	75	1	0.01	*Rsa I*	"
6931	76	0	0.01	*Hinf I*	"
6957	77	0	0.01	*Hae III*	"
7025	78	0	0.18	*Alu I*	"
7055	79	0	0.02	*Alu I*	"
7103	80	0	?	*Dde I*	"
7241	81	1	0.01	*Rsa I*	"
7335	82	0	0.03	*Taq I*	"
7347	83	1	0.01	*Hae III*	"
7461	84	0	0.01	*Taq I*	ARN-t^{Ser}
7474	85	0	0.01	*Alu I*	"
7598	86	0	?	*Hha I*	"
7617	87	1	0.01	*Hha I*	CO II
7750	88	0	0.07	*Dde I*	"
7859	89	0	0.01	*Mbo I*	"
7970	90	1	0.05	*Hinf I*	"
8074	91	0	0.01	*Alu I*	"
8112	92	0	0.01	*Hpa II*	"
8150	93	0	0.01	*Hpa II*	"
8165	94	1	0.01	*Hae III*	"
8249/8250	95	1	0.02	*Ava II/Hae III*	"
8299	96	1	0.01	*Rsa I*	ARN-t^{Lys}
8391	97	0	0.01	*Hae III*	ATPase 8
8466	98	1	?	*Alu I*	"
8515	99	0	0.01	*Dde I*	"
8592	100	0	0.01	*Mbo I*	ATPase 6
8729	101	0	?	*Hha I*	"
8783	102	0	0.01	*Hinf I*	"
8852	103	0	0.01	*Hha I*	"
8994	104	0	0.01	*Hae III*	"
9009	105	1	0.01	*Alu I*	"

Position	Site Nº	Polymorphisme[1]	Fréquence(%)[2] (N=147)	Enzyme	Région
9053	106	0	0.09	*Hha I*	ATPase 6
9070	107	1	0.02	*Taq I*	"
9150	108	1	0.01	*Mbo I*	"
9266	109	0	0.02	*Hae III*	CO III
9294	110	0	0.01	*Hae III*	"
9342	111	0	0.01	*Hae III*	"
9380	112	0	0.01	*Hha I*	"
9429	113	1	0.01	*Rsa I*	"
9553	114	0	0.01	*Hae III*	"
9714	115	1	0.01	*Hae III*	"
9746	116	0	0.01	*Rsa I*	"
9751	117	0	0.01	*Taq I*	"
9859	118	1	0.01	*Hinf I*	"
10028	119	1	0.01	*Alu I*	"
10066	120	1	0.01	*Hha I*	NAD 3
10084	121	1	0.01	*Taq I*	"
10352	122	0	0.01	*Alu I*	"
10364	123	0	0.01	*Hae III*	"
10394	124	1	0.39	*Dde I*	"
10413	125	1	0.03	*Alu I*	ARN-t^{Arg}
10644	126	1	0.01	*Rsa I*	NAD 4L
10689	127	0	0.01	*Hae III*	"
10694	128	1	0.01	*Alu I*	"
10725	129	1	0.02	*Hae III*	"
10806	130	1	0.07	*Hinf I*	NAD 4
10830	131	0	?	*Hinf I*	"
10893	132	1	0.01	*Taq I*	"
11146	133	0	0.01	*Dde I*	"
11161	134	1	0.01	*Hpa II*	"
11329	135	1	0.01	*Hae III*	"
11350	136	1	0.02	*Alu I*	"
11806	137	1	0.02	*Alu I*	"
11922	138	0	0.01	*Mbo I*	"
12026	139	1	0.01	*Hpa I*	"
12345	140	1	0.13	*Rsa I*	NAD 5
12406	141	0	0.02	*Hpa I*	"
12560	142	0	0.02	*Alu I*	"
12795	143	1	0.03	*Mbo I*	"
12810	144	1	0.02	*Rsa I*	"
12925	145	1	0.01	*Hinf I*	"
12990	146	1	0.01	*Alu I*	"
13004	147	1	0.01	*Mbo I*	"
13031	148	0	0.42	*Hinf I*	"
13051	149	0	0.03	*Hae III*	"
13065	150	0	?	Dde I	"
13068	151	1	0.01	*Alu I*	"
13096	152	1	0.01	*Rsa I*	"
13100	153	1	0.01	*Hpa II*	"
13103	154	0	0.01	*Hinf I*	"
13208	155	0	0.04	*Hha I*	"
13268	156	0	0.02	*Hinf I*	"
13367	157	0	0.03	*Ava II*	"
13404	158	0	0.02	*Hae III*	"

Position	Site Nº	Polymorphisme[1]	Fréquence(%)[2] (N=147)	Enzyme	Région
13635	159	1	0.02	*Taq I*	NAD 5
13702	160	0	0.03	*Hae III*	"
14015	161	0	0.01	*Alu I*	"
14050	162	1	0.01	*Taq I*	"
14279	163	1	0.01	*Hae III*	"
14279	164	1	0.01	*Mbo I*	URF 6
14322	165	1	0.01	*Alu I*	"
14385	166	1	0.01	*Dde I*	"
14509	167	1	0.03	*Fnu DII*	"
14567	168	1	0.01	*Hpa II*	"
14608	169	0	0.01	*Dde I*	"
14749	170	1	0.01	*Hae III*	Cyt b
14869	171	0	0.01	*Mbo I*	"
14956	172	0	0.01	*Taq I*	"
15005	173	1	0.01	*Hinf I*	"
15047	174	0	?	*Hae III*	"
15172	175	0	0.01	*Hae III*	"
15195	176	1	0.01	*Mbo I*	"
15234	177	0	0.01	*Hinf I*	"
15238	178	0	0.01	*Dde I*	"
15250	179	0	0.09	*Dde I*	"
15606	180	1	0.14	*Alu I*	"
15723	181	0	0.01	*Hinf I*	"
15790	182	1	0.01	*Mbo I*	"
15883	183	0	0.04	*Hae III*	"
15897	184	1	0.01	*Rsa I*	ARN-t^{Thr}
15907	185	1	0.01	*Rsa I*	"
15912	186	1	0.01	*Hpa II*	"
15925	187	0	0.02	*Hpa II*	"
15996/16000	188	0	0.01	*Dde I/Hinf* I	ARN-t^{Pro}
16049	189	0	0.03	*Rsa I*	D-Loop
16065	190	0	0.01	*Hinf I*	"
16089	191	1	0.01	*Rsa I*	"
16096	192	0	0.01	*Rsa I*	"
16125	193	0	0.09	*Rsa I*	"
16178	194	1	0.04	*Taq I*	"
16208	195	0	0.03	*Rsa I*	"
16217	196	1	0.01	*Taq I*	"
16246	197	1	0.01	*Hinf I*	"
16254	198	1	0.01	*Alu I*	"
16303	199	0	0.10	*Rsa I*	"
16310	200	0	0.33	*Rsa I*	"
16389/16390	201	1	0.07	*Hinf I/Ava II*	"
16398	202	1	0.05	*Hae III*	"
16490	203	1	0.01	*Hinf I*	"
16517	204	0	0.49	*Hae III*	"

[1] La notation "1" correspond à un polymorphisme dû à un gain de site, alors que la notation "0" correspond à la perte d'un site par rapport au type Nº 69 (voir Table 4.58).

[2] Les fréquences des haplotypes définis dans Stoneking *et al.* (1986) n'étant pas disponibles, les fréquences des sites ont été calculés sur la seule base des types définis dans Cann *et al.* (1987).

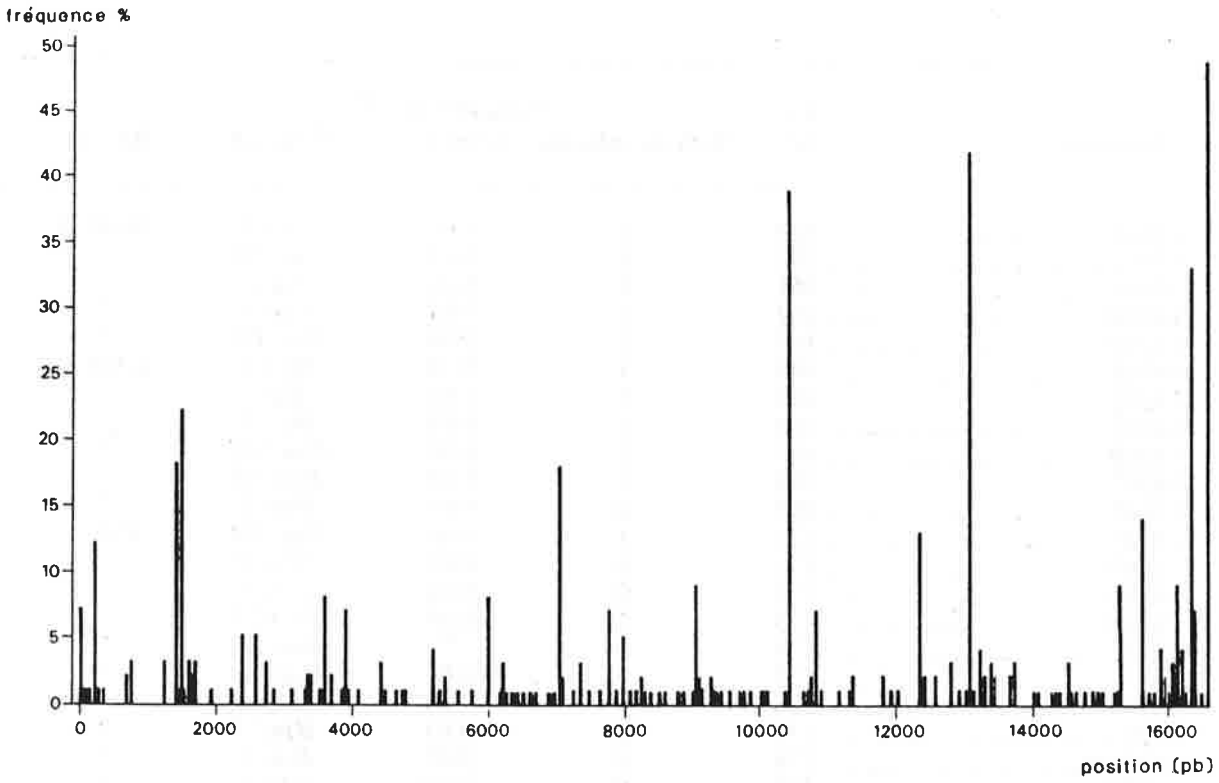

FIGURE 4.33 : Positions et fréquences des sites de restriction définis par Cann *et al.* (1987).

Le nombre très élevé de sites polymorphes nous offre une opportunité d'étudier leur répartion sur l'ADN-mt. Pour ce faire, nous avons scindé la molécule d'ADN-mt en 15 portions ou classes de longueur équivalente (≈ 1104 pb). La classe 15 comprend toute la portion non codante qui entoure l'origine de réplication du brin lourd de l'ADN-mt, dont la région de la D-Loop. Dans la Table 4.51, nous avons reporté le nombre de sites polymorphes trouvé dans chaque classe et nous l'avons confronté au nombre attendu dans l'hypothèse d'une répartition homogène des 195 sites polymorphes dans les 15 classes. Dans cette situation, la valeur de χ^2 calculée n'est pas significative. Toutefois, on remarque que les classes 2, 4 et 11 comprennent trop peu de sites polymorphes alors que la classe 15 (D-Loop) en contient nettement trop. Cependant, ces fluctuations peuvent être le fruit du seul hasard au niveau de confiance $\alpha = 0,05$.

TABLE 4.57: Test de l'uniformité de la répartition des sites polymorphes sur l'ADN-mt

Classes[1]	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Nb. de sites observés	13	7	11	12	13	17	16	16	13	11	6	18	11	17	23
Nb. de sites attendus	13,6	13,6	13,6	13,6	13,6	13,6	13,6	13,6	13,6	13,6	13,6	13,6	13,6	13,6	13,6

$\chi^2_{[14;0,05]} = 23,69$ $\chi^2_{calc} = 19,676$

Nb. de sites observés[2]		181	23
Nb. de sites attendus		190,4	13,6

$\chi^2_{[1;0,01]} = 6,64$ $\chi^2_{calc} = 6,96$

[1] Chaque classe correspond à une longueur approximative de 1104 pb, qui permet de fragmenter l'ADN-mt en 15 classes de taille équivalente. La classe 1 commence à la position 577 (début du gène codant pour l'ARN-t^{Phe}). La classe 15 comprend toute la partie non-codante qui entoure l'origine de réplication (O_H) du brin lourd de l'ADN-mt.
[2] Dans la seconde partie de la table, nous avons regroupé les 14 premières classes en une seule. La seconde classe est la même que la classe 15 décrite auparavant.

Nous avons ensuite regroupé les classes 1 à 14 afin de vérifier si la région de la D-Loop possédait statistiquement plus de sites que le reste de l'ADN-mt. Dans ce cas,

la différence de sites s'avère significative au niveau $\alpha = 0,01$ et confirme l'hypothèse d'une accumulation préférentielle de mutations dans cette région.

Définition des types

Les différents morphes enzymatiques ne sont pas disponibles dans Cann *et al.* (1987) ou dans Stoneking *et al.*(1986). On pourra néanmoins en retrouver dans Cann (1982), bien que seuls 113 individus y soient étudiés complétement. Aussi, nous ne définirons pas les types en fonction des morphes, mais directement en fonction du polymorphisme minoritaire des sites de restriction (Table 4.58).

A partir des 176 individus testés pour les 12 enzymes à disposition, 134 types on été définis par Cann *et al.* (1987) et 13 autres supplémentaires par Stoneking *et al.*(1986) (voir Table 4.58). Leur numérotation correspond à celle définie par Cann *et al.* (1987). Les 13 nouveaux types de Stoneking *et al.*(1986) ont été renumérotés de 135 à 147. Cette numérotation reprend simplement l'ordre de classement des types dans la Figure 3 de l'article de Cann *et al.* (1987), qui est censée représenter un "arbre généalogique" de 134 types.

En ce qui concerne l'étude de Cann *et al.* (1987), la très grande majorité des types (127 sur 134) ne sont retrouvés que chez un seul individu. Cependant, le type 134 est présent chez 6 individus, les types 29, 65 et 80 sont identifiés chez 3 individus et les types 59, 114 et 115 ont été trouvés chez 2 individus chacun. Notons aussi qu'en fonction de la liste des sites polymorphes fournie par Cann *et al.* (1987), les types 114 et 116 seraient identiques, ce qui réduirait le nombre de types différents à 133. Parmi les nouveaux types définis par Stoneking *et al.* (1986), le type 144 est également équivalent au type 114. Nous avons donc un total de 145 types différents pour ces 2 études.

On peut tout d'abord constater que la majorité des types possèdent un grand nombre de sites dans leur état minoritaire et semblent donc avoir subi une quantité de mutations par rapport à un type central hypothétique qui ne présenterait aucun site avec un état minoritaire dans l'échantillon. Le nombre moyen de mutations accumulées par type est donc condidérablement plus élevé que pour les types définis précédemment (voir Tables 4.14, 4.38, 4.45 et 4.53) dans d'autres études. Ceci est, bien sûr, lié à l'emploi d'un grand nombre d'enzymes sur un échantillon hétérogène. Le type central hypothétique est équivalent au type 69, et ceci aussi bien sur la base des 134 types de Cann *et al.* (1987) que sur les 147 types de Stoneking *et al.* (1986).

TABLE 4.58 : Liste des types d'ADN-mt définis par Cann *et al.* (1987) et Stoneking *et al.* (1986)

Type	Origine[1]	Sites polymorphes[2]
1	*Af*	*28 92 93 95 124 130 148 163 200*
2	*Af*	*1 16 18 26 28 35 46 124 130 135 148 193 200 204*
3	*Af*	*3 18 26 28 35 79 88 124 130 148 200*
4	*Af*	*16 28 35 88 124 130 135 148 155*
5	*Af*	*1 16 28 35 38 67 74 88 107 124 130 144 148 155 169 200 202*
6	*Af*	*28 35 79 82 88 102 107 124 130 134 144 148 200*
7	*Af*	*12 28 35 79 82 88 102 107 124 130 134 144 145 154 155 183*
8	*As*	*1 12 16 26 28 88 124 158 200*
9	*As*	*1 12 16 26 35 67 88 124 148 158 193 200*
10	*Au*	*16 56 90 94 124 148 166 175 200 203*
11	*As*	*14 16 31 124 148 195 200 201 202 204*
12	*Ng*	*14 41 71 124 148 200 204*
13	*Ng*	*14 124 148 200 204*
14	*Au*	*14 15 16 39 124 148 161 179 200*
15	*As*	*14 16 57 105 124 136 148 155 179 187 189 195 204*
16	*As*	*10 14 16 105 124 125 148 179 204*
17	*As*	*10 14 16 54 124 179 204*
18	*As*	*10 14 27 39 54 73 124 148 179 204*
19	*Eu*	*10 16 54 73 124 126 142 148 204*
20	*As*	*14 124 126 147 148 179 200 204*
21	*As*	*14 51 100 124 148 179 191 204*
22	*Au*	*14 21 27 39 124 129 148 189 204*
23	*Au*	*1 14 16 21 39 43 100 124 148 149 164 204*
24	*As*	*14 16 90 124 179 204*
25	*As*	*14 16 51 110 124 179*
26	*Ng*	*2 14 124 148 152 194 200 204*
27	*Ng*	*14 21 99 111 124 146 148 194 200 204*
28	*Ng*	*14 109 124 148 194 200 204*
29	*Ng*	*14 124 148 194 200 204*
30	*Eu*	*16 82 148 200*
31	*Au*	*17 97 148 200*
32	*Au*	*16 39 148*
33	*Au*	*16 148*
34	*Eu*	*14 16 132 148 202*
35	*Au*	*16 43 148 201*
36	*As*	*43 199 201*
37	*Af*	*12 24 35 65 88 114 136 148 201 202*
38	*Af*	*35 65 149 160 201*
39	*Af*	*5 35 65 124 148 149 159 160 201*
40	*Af*	*35 65 88 124 143 148 149 157 160 201*
41	*Af*	*16 33 35 65 88 124 149 160 201*
42	*Au*	*16 124 148*
43	*Eu*	*8 16 124 148 181*
44	*As*	*18 34 37 57 77 124 176*
45	*Af*	*16 18 77 120 121 124 148 183*
46	*Af*	*38 125 129 148 185 189 197*
47	*Eu*	*19 20 148*
48	*As*	*8 76 117 148 188*
49	*Ng*	*11 14 26 74 109 124 148*
50	*As*	*14 124 130 148*
51	*As*	*148 195 204*
52	*Au*	*16 59 106 142 148 195 204*
53	*Eu*	*8 16 148 204*
54	*As*	*16 27 51 148 179 204*
55	*As*	*14 16 51 87 148 195 204*
56	*As*	*14 16 51 113 125 148 173*
57	*Eu*	*14 16 43 51 113 148 173*
58	*Au*	*16 148 183 199 200 204*
59	*Au*	*148 183 199 204*
60	*Eu*	*104 148*
61	*Eu*	*16 95 104 133 148*
62	*As*	*116 148*
63	*Au*	*64 148*
64	*Au*	*148*
65	*Ng*	*137 156 180*
66	*Eu*	*202*
67	*Eu*	*109*
68	*Au*	*68 69 91 167 168*
69	*Au*	-
70	*Af*	*22*
71	*As*	*39 53 84 90*
72	*Af*	*81 90*
73	*As*	*27 87 90 200*
74	*As*	*63 90 189 193*
75	*Eu*	*78 155*
76	*As*	*50 78*
77	*As*	*78 90*
78	*Eu*	*70 72 78 193 200*
79	*Eu*	*26 78 158*
80	*Eu*	*78*
81	*Af*	*1 58 78 179 186 187 189*
82	*Af*	*1 78 148 182 199 200 201 204*
83	*Eu*	*1 78 130 148*
84	*Eu*	*78 124 178*
85	*Eu*	*1 50 78 124 125*
86	*Eu*	*1 78 118 124 171*
87	*Eu*	*23 47 60 62 124 170*
88	*Eu*	*124 140 155 157 193*
89	*As*	*124 193 199 201*
90	*As*	*9 112 143 157 179 187 193 199*
91	*As*	*106 141 193 199*
92	*As*	*39 82 106 141 148 165 199*
93	*As*	*26 78 106 141 142 148 193 199*
94	*Eu*	*32 78 124 200*
95	*Ng*	*66 106 124 196 200*
96	*As*	*106 123 124 127 129 138 200*
97	*Eu*	*106 124 184 200 204*
98	*Eu*	*88 106 115 124 125 143*
99	*Eu*	*106 124 143 157 200*
100	*Eu*	*47 106 124 178 200*
101	*Eu*	*27 106 200*
102	*Eu*	*19 39 106 200*

TABLE 4.58 : Liste des types d'ADN-mt définis par Cann *et al.* (1987) et Stoneking *et al.* (1986)

Type	Origine	Sites polymorphes	Type	Origine	Sites polymorphes
103	*Eu*	*78 108*	126	*Au*	*93 96 119 167 171 200 204*
104	*Eu*	*18 78 108 128*	127	*Ng*	*140 180 198 204*
105	*Eu*	*33 78 193 204*	128	*Ng*	*4 140 151 180 204*
106	*Eu*	*6 25 78 204*	129	*Ng*	*4 44 140 180 184 204*
107	*Eu*	*16 78 153 204*	130	*Ng*	*4 36 55 140 180 204*
108	*Eu*	*19 78 204*	131	*Ng*	*4 36 140 180 199 204*
109	*Eu*	*78 103 204*	132	*Ng*	*4 36 95 140 180 198 199 204*
110	*Eu*	*12 39 40 49 78 103 122 133 136 204*	133	*Ng*	*4 32 61 140 180 200 204*
111	*Af*	*45 78 124 179 199 204*	134	*Ng*	*4 61 140 180 200 204*
112	*Eu*	*7 29 42 45 78 106 183 199 204*	135	*Ng*	*14 124 150 200*
113	*Af*	*177 204*	136	*Ng*	*14 124 195 200 204*
114	*Eu*	*204*	137	*Ng*	*14 26 86 98 101 124 189 190 192 201 204*
115	*Eu*	*48 159 172 204*	138	*Ng*	*14 86 124 131 201*
116	*Eu*	*204*	139	*Ng*	*13 14 86 124 131 201*
117	*As*	*114 204*	140	*Ng*	*200*
118	*As*	*1 32 52 202 204*	141	*Ng*	*75*
119	*Eu*	*30 192 193 202 204*	142	*Ng*	*75 141*
120	*As*	*39 82 89 97 124 193 200 204*	143	*Ng*	*106 114 199 200*
121	*Eu*	*85 124 190 204*	144	*Ng*	*204*
122	*Eu*	*33 124 190 193 204*	145	*Ng*	*4 21 80 174 180 204*
123	*Au*	*21 124 167 204*	146	*Ng*	*4 140 180 195 199 204*
124	*Au*	*83 130 139 167 204*	147	*Ng*	*4 60 140 180 204*
125	*Au*	*162 167 200 204*			

[1] Les abbrévation d'origine pour les individus portant les différents types sont les suivantes : *Af*: Afrique ; *As* : Asie ; *Au* : Australie ; *Ng* : Nouvelle-Guinée ; *Eu* : Europe.

[2] On définira ici que les sites sont polymorphes par rapport au type 69 qui comprend les 204 sites dans leur état majoritaire dans l'échantillon.

Cette extraordinaire diversité des types nous montre la richesse potentielle de l'information apportée par l'emploi d'un nombre élevé d'enzymes qui permet presque d'individualiser complétement chaque personne testée. Nul doute que l'emploi d'autres enzymes permettrait d'affecter un type unique à chaque individu. Cette diversité nous montre aussi que l'on ne peut plus estimer de fréquences de types sur de tels échantillons quand le fait de choisir Monsieur X ou Monsieur Y modifie presque à coup sûr la composition de notre échantillon de taille N ou chaque type possède une fréquence avoisinant 1/N.

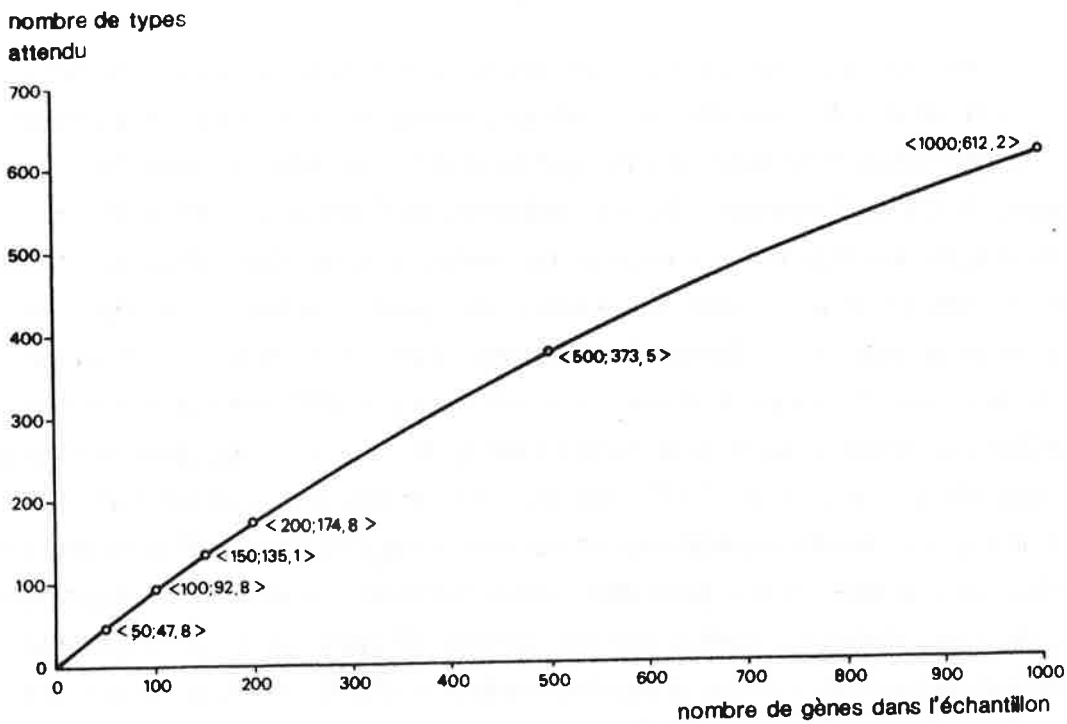

FIGURE 4.34 : Nombre attendu de types d'ADN en fonction de la taille d'un échantillon de n gènes tirés d'une population stationnaire. Les points de la courbe ont été calculés au moyen de l'équation (A.12) en fonction d'une valeur de θ égale à 670,6 établie par itérations successives de (A.12) en considérant arbitrairement que les 133 types trouvés par Cann *et al.* (1987) dans l'échantillon de 147 gènes provenaient d'une seule population.

Extrapolation du nombre de types définissables pour ce système

Cette conséquence est d'ailleurs prévue par le modèle des allèles infinis (voir Annexe A). Il est en effet possible d'estimer un paramètre de mutation θ qui dépend du nombre de gènes dans l'échantillon ainsi que du nombre de types trouvés. Si l'on calcule θ (équation A.12) sur l'ensemble des 147 individus de Cann *et al.*(1987), on trouve une valeur de θ égale à 670,6. En reportant ce paramètre et dans l'hypothèse selon laquelle nous avons affaire à une seule population, on peut estimer le nombre de types observables pour une taille d'échantillon donnée. Ainsi, pour de petits échantillons (en regard de la valeur élevée de θ) d'une taille inférieure à 1000, nous avons une relation presque linéaire entre la taille et le nombre de types (Figure 4.30), pour arriver à 612 types détectables à partir de 1000 individus. Cette linéarité disparaît ensuite et le nombre de types va tendre asymptotiquement vers $\theta \log_e(N)$ lorsque N tend vers l'infini. Le modèle des allèles infinis peut donc nous permettre d'estimer grossièrement le nombre de types différents définis par ces mêmes 12 enzymes pour l'ensemble des 5 milliards d'êtres humains comme étant environ égal à 10'000 (10'147,6), si l'on considère que la taille effective de la population humaine est égale au nombre de femmes ($2{,}5\ 10^9$). Ce chiffre important est cependant beaucoup plus faible que le nombre de type possibles pour 195 sites, soit $2^{195} = 5.10^{58}$. On peut penser que ce nombre est surestimé, car il est clair que nous n'avons pas affaire à un échantillon tiré d'une seule population homogène, mais à une collection d'individus provenant de populations et de groupes continentaux différents, ce qui a sans doute pour effet de gonfler artificiellement la diversité des types Ceci est confirmé par l'étude de Stoneking *et al.* (1986) qui n'a découvert "que" 12 types effectivement nouveaux parmi 29 individus supplémentaires, alors que l'on aurait dû en observer 23 selon la fonction de la Figure 4.34. La valeur de θ calculée sur les 176 individus et les 145 types effectivement différents donne une valeur de 383,6, ce qui conduit à un nombre d'allèles attendus de 6019 pour la population humaine. Néanmoins, l'ordre de grandeur étant donné, il est évident qu'il serait vain de prétendre étudier correctement la diversité des différentes populations humaines dans un tel système à partir d'échantillons de quelques centaines d'individus.

PHYLOGÉNIE DES TYPES

Une autre conséquence de l'estimation de ce nombre considérable de types est qu'il semble vain de vouloir définir une phylogénie de ces types avec nos moyens informatiques actuels. En effet, le nombre d'arbres phylogéniques binaires possibles est égal à

$$(2n-3)! / [2^{n-2} (n-2)!] \qquad \text{(Cavalli-Sforza and Edwards, 1967)} \quad (4.18)$$

où n est le nombre de types. Ce chiffre atteint déjà la valeur respectable de 34'459'425 lorsque n=10 et devient vite incalculable. De plus, un arbre trouvé pour ces 145 types serait certainement très différent d'un autre arbre défini sur d'autres types provenant d'un nouvel échantillon d'une taille similaire. En admettant qu'il soit possible de déterminer un arbre phylogénétique correct pour ces données, celui-ci ne pourrait conserver une structure stable étant donné le nombre considérable de types intermédiaires manquants pour relier tous les types par de simples mutations. Il ne serait donc pas représentatif de la différenciation historique des types d'ADN-mt.

La seule structure stable définissable à partir des données de cet échantillon correspond à une série de liens d'ordre 1 entre certains types, c'est à dire à un réseau de différenciation directe des types par des mutations uniques. Il est ainsi possible de reconstruire des lambeaux de l'arbre phylogénique global, que l'on peut ensuite comparer à la phylogénie de Cann *et al.* (1987) pour en vérifier la validité.

Ainsi, dans la Figure 4.35, nous avons reporté des liens d'ordre 1 non-ambigûs existant entre 30 types d'ADN-mt trouvés parmi les 5 "populations". Les 115 autres types ne peuvent être rattachés directement à ce réseau, et , mis à part une liaison simple entre les types 130 et 131, et une autre entre les type 133 et 134, il n'existe pas d'autres liens directs entre ces types. Le type 69, qui, rappelons-le, ne présente aucun site dans un état minoritaire pour cet échantillon, possède une place centrale dans ce réseau partiel. C'est également le type qui est lié au plus grand nombre d'autres types (8) par une simple mutation. Il est notamment lié aux type 80 et 116 qui ont été tout deux trouvés à 3 exemplaires parmi l'échantillon d'occidentaux, qui n'est d'ailleurs guère homogène.

Il est aussi frappant de constater que ce sont les types qui ont accumulé le moins de mutations qui sont associés en réseau. Ceci est compatible avec l'hypothèse selon laquelle ils seraient les représentants de types anciens peu différenciés, qui ont pu être présents dans les diverses populations à des fréquences plus élevées, et à partir

desquels les autres types se seraient progressivement diversifiés. Le fait que les types de ce petit noyau sont retrouvés dans des continents différents semble également attester d'une grande ancienneté potentielle datant d'avant la divergence des grands groupes humains. Selon cette vue, les populations occidentales et australiennes possèdent les types les moins différenciés, suggérant qu'ils possèdent une constitution génétique proche des populations primitives (au sens historique). Bien évidemment, étant donné la pauvreté des échantillons, il n'est pas possible de prétendre que ces types sont absents des autres continents. D'ailleurs, les groupes "africain" et asiatique possèdent des types directement dérivés de ces derniers, ce qui plaide en faveur de leur présence dans ces populations. Nous nous trouvons donc en présence d'un petit nombre de types (64, 69, 80 et 116) peu différenciés, probablement présents dans 4 groupes continentaux, qui seraient à la source des autres types. Il faut aussi constater que l'échantillon étendu de Nouvelle-Guinée comprend le type 116 qui est également retrouvé en Europe. Ce type est le seul qui soit retrouvé dans 2 groupes continentaux différents et confirme notre hypothèse d'ancienneté de ce noyau central de types. Par ailleurs, l'échantillon étendu de Nouvelle-Guinée nous montre qu'on y trouve des types proches des types ancestraux, alors que l'ancien échantillon (Cann *et al.*, 1987) n'avait permis de recenser que des types relativement éloigné du noyau central (12, 13, 26, 28 et 29) sur une branche bien différenciée à partir d'un type trouvé en Australie (voir Figure 4.35). Donc, les 5 sous-échantillons possèdent des types de base communs ou très proches qui laissent penser à une origine commune des populations correspondantes.

Les autres types, nettement plus différenciés, sont probablement la conséquence d'évènements mutationnels postérieurs à la séparation des grands groupes. Le fait qu'ils ne soient pas connectables entre eux, ni au réseau partiel gravitant autour du type 69, résulterait d'un processus d'échantillonnage insuffisant par rapport au nombre important de mutations accumulées. Le nombre de ces mutations semble être fortement sous-estimé, car une grande quantité de mutations parallèles a vraisemblablement eu lieu au cours de la diversification des types. Ceci est d'ailleurs une des raisons qui nous empêche de reconstruire un réseau phylogénique complet. Prenons par exemple le cas simple des types 108 et 109 qui possèdent tout deux 4 mutations par rapport au type 69 (voir Table 4.58), dont 2 mutations communes aux sites 78 et 204. Ils peuvent donc être rattachés par un seul intermédiaire soit au type 80, soit au type 116, et, dans les 2 cas, une mutation parallèle aura eu lieu. Le type 51 présente, lui, 3 sites polymorphes et peut être rattaché soit au type 64 soit au type 116 par une seule mutation du site 204, ce qui provoquerait aussi une mutation parallèle du site 148 ou 204.

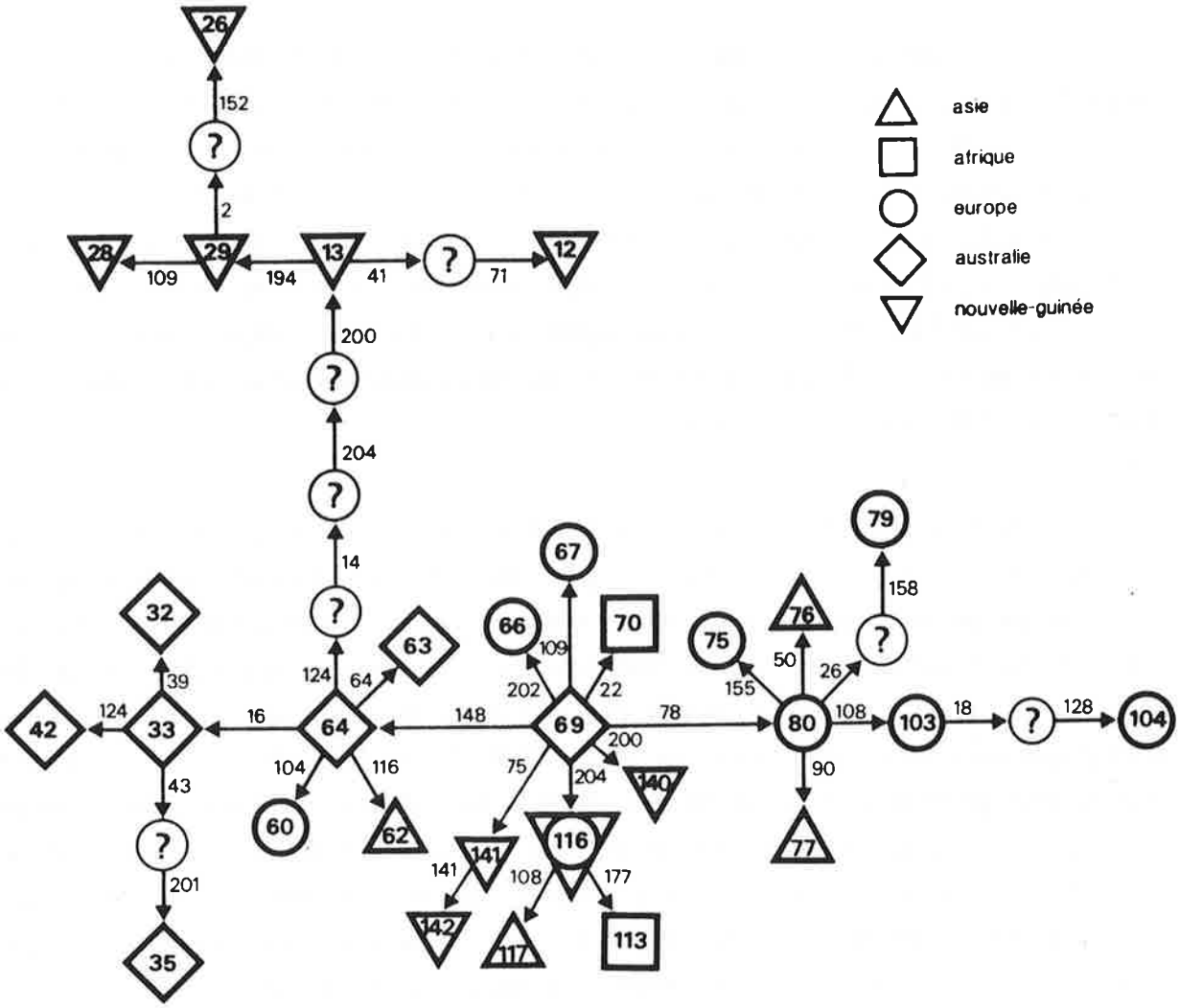

FIGURE 4.35 : Phylogénie partielle des types d'ADN-mt définis par Cann *et al.* (1987) et Stoneking *et al.* (1986). Le sens des flèches indique un changement de l'état d'un site par rapport au type 69 qui est le type ancestral hypothétique (voir texte).

On pourrait ainsi multiplier des exemples, surtout dans les cas plus complexes où des types ont accumulé un grand nombre de mutations, et où des apparentements entre types ne sont concevables qu'avec des mutations indépendantes de sites identiques, ou bien par des gains-pertes de sites sur une même branche. Les sites potentiellement impliqués dans ces mutations parallèles sont souvent les sites qui possèdent les plus hautes fréquences de polymorphisme, à savoir les sites 14, 16, 124, 148 et 204 (voir Table 4.56). Cela laisse supposer que ces fortes fréquences ne sont pas forcément dues à une plus grande ancienneté du polymorphisme de ces sites, mais plutôt à une série de mutations récurrentes.

Si l'on compare maintenant notre réseau partiel à "l'arbre phylogénique" de la Figure 3 de Cann *et al.* (1987), nous pouvons noter de profondes différences de structure qui peuvent entraîner des interprétations fort divergentes. On peut, en effet, attendre d'une méthode de classification basée sur le principe de parcimonie qu'elle groupe les types qui ne diffèrent que par une seule mutation apparente. Ceci n'est réalisé que pour les couples des types 28-29, 75-80, 76-80, 113-116, 116-117 et 133-134 dans la phylogénie de Cann *et al.*. Beaucoup de liaisons d'ordre 1 n'ont pas été mises en évidence. Ainsi, Cann et ses collaborateurs regroupent le type 69 aux types 65 à 68, le type 80 aux types 78 et 79, le type 103 au 102 ou le type 42 aux types 43 à 45. Donc, dans toute une série de points précis, leur topologie des types est tout simplement fausse. De plus, elle ne met pas du tout en évidence le réseau centré autour des types 64, 69, 80 et 116. Leur méthode de reconstruction phylogénique, déjà critiquée par Darlu et Tassy (1987b), semble avoir été perturbée par l'incorporation de types très différenciés et l'absence d'un grand nombre de types intermédiaires. Les méthodes de classification automatique obéissant au principe de parcimonie maximale minimisent le nombre total de mutations nécessaires pour relier tous les types, et privilégient les branchements de types très différenciés aux dépens des types très proches. Dans ce cas précis, la recherche d'une vision globale de l'apparentement de tous les types a complétement faussé l'ordre de branchement de types proches, qu'il était possible de distinguer par un simple examen visuel des données (Table 4.58). Le principe même de parcimonie, dans un système où les mutations récurrentes et parallèles sont nombreuses, est aussi à remettre en cause, car c'est le nombre de mutations visibles qui sera minimisé et non pas le nombre de mutations s'étant réellement produites.

Il en résulte que la phylogénie des types proposée par Cann *et al.* peut sérieusement être remise en cause. C'est également le cas du positionnement de la racine de leur phylogénie à mi-chemin entre un groupe de types "africains" ayant

accumulé un grand nombre de mutations et le reste des autres types. Il est bien clair que ce ne sont pas ces types très différenciés qui sont à l'origine des autres types, mais plutôt le groupe des types dérivés du type 69. En fait, les données récoltées par Cann et ses collaborateurs sont compatibles avec les données de l'ADN-mt analysées précédemment et portant sur 10 populations. Dans les 2 cas, nous avons un groupe de types centraux ayant accumulé peu de mutations et présents dans plusieurs populations. Le fait que l'on ne trouve ici qu'un seul type commun entre plusieurs populations peut être dû à la petitesse des échantillons et à la caractérisation très précise des types (étudiés avec 12 enzymes et sur 195 sites au lieu de 5 enzymes sur 71 sites précédemment). La deuxième caractéristique commune est une diversification poussée de certains types africains, comme les types 2, 5, 6 et 7 qui ont respectivement accumulés un minimum de 14, 17, 13 et 16 mutations. Pour pouvoir interpréter clairement cette dernière caractéristique, il est nécessaire d'étudier la diversité moléculaire des 5 échantillons.

Diversité moléculaire

Nous avons calculé les nombres moyens de différences de sites de restriction intra et interpopulation et les avons reportés dans les Tables 4.59 et 4.61. Avant d'analyser ces chiffres, il nous semble important de préciser qu'ils ne concernent que la diversité moléculaire des échantillons et qu'ils sont basés sur les différences observées de sites de restriction entre les types. Ainsi, le fait que les échantillons soient hétérogènes et non représentatifs d'une population donnée doit être clair lorsque nous extrapolerons nos conclusions à des populations ou à des groupes continentaux.

L'examen de la Table 4.59 nous apprend que le nombre théorique de différences de sites de restriction (θ) est nettement supérieur à ce que l'on observe (v) pour tous les échantillons, y compris les africains. La meilleure concordance entre les valeurs théoriques et les valeurs observées est trouvée en Nouvelle-Guinée.

Cette diversité moléculaire réduite peut être due au fait que les mutations parallèles n'ont pas été prises en compte pour mesurer le nombre réel de mutations entre 2 types quelconques (v_{ij}). D'autre part, l'hétérogénéité des échantillons peut avoir deux influences: la première est de gonfler artificiellement le nombre de sites polymorphes observés et, par la même, de biaiser vers le haut la valeur de θ; la seconde est de hausser également la valeur observée des v_{ij}. Quoiqu'il en soit, ces facteurs perturbateurs nous empêchent d'interpréter correctement ces différences. Nous noterons néanmoins que la valeur observée de v est quand même la plus forte pour

l'échantillon africain qui est presque un échantilon inter-population à lui tout seul étant donné sa constitution. Il est donc très concevable que sa pseudo diversité intra-population soit très importante.

TABLE 4.59: Nombre moyen de différences de sites de restriction intrapopulation

Echantillons	Nb. gènes	Nb. sites polymorphes	v (4.12)	θ (4.15)
"Européens"	46	81	7,16	18,34
"Africains"	20	68	13,50	19,17
"Asiatiques"	34	78	10,18	19,08
Australiens	21	48	7,67	13,34
Nv.-Guinée[1]	26	36	7,18	9,43

[1] Echantillon de Cann *et al.* (1987) uniquement.

Il est difficile de savoir si des mécanismes de réduction de la diversité moléculaire sont en jeu pour ces échantillons sur la base des seules comparaisons de v et de θ. Un indice paraissant aller dans ce sens provient des travaux de Whittam *et al.* (1986) qui ont étudié, pour ce même échantillon, la distribution des génotypes de 28 loci de l'ADN-mt. Ces auteurs ont en effet trouvé un excès d'allèles fréquents et d'allèles singletons par rapport à un modèle neutraliste. Ceci a été interprété comme un effet possible d'une expansion récente des populations humaines et de l'action d'une sélection éliminant des allèles défavorisés.

Les temps nécessaires pour créer la diversité moléculaire observée dans les 5 échantillons ont été reportés dans la Table 4.60. Le nombre de nucléotides effectivement surveillées par cette étude est de 4'158,2 pour les 195 sites, soit près d'un quart du génome mitochondrial, ce qui explique que l'on observe des types si différenciés les uns des autres pour chaque individu. Les estimations fournies par t_1 indiquent une plus grande ancienneté de l'échantillon africain alors que les estimations de t_2 montrent que les échantillons d'"européens" d'"asiatiques" et d'"africains" ont développé leur diversité en un laps de temps comparable de l'ordre de 450'000 ans. Etant donné la grande hétérogénéité des échantillons, ces temps surestiment passablement les temps de divergence entre groupes. Les divers degrés d'hétérogénéité des échantillons pourraient suffire à expliquer les différences obtenues dans les estimations de t_2, et nous empêchent de les interpréter correctement. Cependant, l'ordre de grandeur de 450'000 ans semble bien constituer une limite vers laquelle tendent les

estimations maximums de t_2 jusqu'ici. Cann *et al.* (1987) ont obtenu une estimation d'environ 140'000 à 290'000 ans pour ce processus de différenciation en considérant un regroupement de tous les types. Pour cela, ils ont utilisé une valeur de 2 à 4% de différence entre des séquences ayant divergé depuis un million d'années, ce qui correspond à un pourcentage de divergence entre espèces, alors que nous avons affaire ici à une seule espèce. Il en résulte que leurs estimations devraient être doublées. A nos yeux, ce regroupement de populations n'a pas de justification théorique et peut conduire à surestimer le temps nécessaire pour établir la diversité génétique de l'espèce humaine.

TABLE 4.54: Estimation du temps t (en années) nécessaire pour créer une diversité nucléotidique π à partir d'une population monomorphe.

Population	π_1 (x10^4) (4.18)	π_2 (x10^4) (4.22)	t_1	t_2
"Européens"	17,22	44,10	173'398	442'302
"Africains"	32.46	46,10	325'304	462'423
"Asiatiques"	24,48	45,88	245'200	460'209
Australiens	18,44	32,08	184'627	321'488
Nv.-Guinée	17,26	22,68	172'799	227'144

Les estimations de t_2 pour les échantillons d'Australie et de Nouvelle-Guinée semblent aussi plus élevées que les temps de divergence probables de ces 2 groupes de population, c'est-à-dire plus de 40'000 ans pour le continent Australien qui comprenait également la Nouvelle-Guinée à cette époque. Notons que la séparation physique de l'Australie et de la Nouvelle-Guinée n'aurait eu lieu que depuis 10'000 ans environ (Chappell and Thom, 1977) ce qui suggère que des échanges génétiques ont eu lieu avant cette date entre les populations de ces 2 sous-continents. Donc, il semble probable que les premières populations migrantes possédaient déjà un certain degré de polymorphisme. De plus, il a été suggéré que le peuplement de ces îles aurait été effectué en plusieurs vagues successives (Jones, 1979), ce qui conduit également à augmenter les temps de diversification apparents.

Si l'on mesure ensuite le nombre net de différences de sites de restriction entre 2 populations (d_{XY} obtenu par l'équation (4.18)) (voir Table 4.61), on s'aperçoit que c'est l'échantillon de Nouvelle-Guinée qui présente les valeurs de d_{XY} les plus élevées. Ces résultats concordent avec ceux de la Table 1 de Cann *et al.* (1987) où sont reportés les pourcentages de divergence de séquence entre les 5 échantillons.

TABLE 4.61: Nombre net (d_{XY}) de différences de sites de restriction interpopulations

Echantillons	Eur.	Afr.	Asie	Aus.
Afrique	*1,31*	-		
Asie	*0,45*	*0,96*	-	
Australie	*0,58*	*1,10*	*0,24*	-
Nv.-Guinée	*1,64*	*2,64*	*1,56*	*1,53*

Si les échantillons étaient représentatifs et s'il n'existait pas de mutations parallèles, ce résultat pourrait être interprété comme le fait que l'échantillon de Nouvelle-Guinée est celui qui est le plus différencié, bien que la plus grande partie de la diversité moléculaire soit présente à l'intérieur des populations. Ensuite, n'importe quelle méthode de classification automatique appliquée à ces indices de dissimilarités conduirait alors à isoler cet échantillon par rapport aux 4 autres (voir Darlu and Tassy, 1987c). Nous ne retiendrons bien évidemment pas cette interprétation, mais nous nous étonnons que Cann et ses collaborateurs ne l'aient pas mentionné, alors qu'elle était aisément perceptible à travers leurs données dont ils n'ont pas mis en doute la qualité. D'autre part, ils se sont basés sur un argument similaire au niveau des types pour proposer que l'échantillon africain soit considéré comme le plus ancien à partir de leur "généalogie" des types.

TEST DE LA NEUTRALITÉ SÉLECTIVE DE L'ADN-MT

La neutralité sélective d'un locus quelconque doit être vérifiée pour que celui-ci puisse être utilisé afin de dresser un réseau de relations génétiques entre populations à partir de fréquences géniques. En effet, l'interprétation de la plupart des distances génétiques assume, entre autres conditions, que les fréquences géniques varient selon un processus de pure mutation-dérive génétique et que tous les gènes présents sont sélectivement équivalents. L'usage extensif des données provenant du polymorphisme de l'ADN-mt à des fins anthropologiques nous a conduit à vérifier l'absence de mécanismes sélectifs différentiels pour les gènes de cette molécule. La présence de gènes avantagés ou désavantagés pourraient fausser certaines conclusions basées sur la diversité moléculaire dont le calcul utilise précisément les fréquences des types d'ADN-mt dans les échantillons.

Choix d'un test

Watterson (1977, 1978, 1986) a développé une procédure de test dite "test de l'homozygosité", basé sur la comparaison d'une valeur de F observée, équivalente à l'homozygosité d'un échantillon, avec une valeur de F attendue calculée selon le modèle des allèles infinis (Crow and Kimura, 1964; Ewens, 1972) et dépendant uniquement de la taille de l'échantillon et du nombre d'allèles observés (voir l'Appendice B pour les détails du test). Ce test n'est pas fait pour vérifier la validité globale de la théorie neutraliste, comme il lui a souvent été reproché (Kimura, 1983), mais permet de vérifier si un échantillon donné possède une distribution de fréquences alléliques en accord avec la neutralité sélective du locus considéré. Il peut donc permettre de mettre en évidence des comportements différents du point de vue sélectif entre plusieurs échantillons étudiés à un locus précis. Il évite ainsi de mettre toutes les populations étudiés dans un même sac. Il a donc l'avantage de distinguer les populations qui peuvent faire l'objet d'une analyse de distances génétiques et celles qui introduiraient des biais dans une telle analyse.

D'autres tests de neutralité sélective ont été proposés. Tout d'abord, un test portant sur les données électrophorétiques a été établi par Weir *et al.* (1976) en se basant sur le modèle "Charge-State" de Ohta et Kimura (1973). Un test basé sur le modèle des sites infinis (Kimura, 1969) a été développé par Ewens (1979). Il étudie la

fréquence des présences-absences de chaque site de restriction polymorphe. Ainsi, un test par site polymorphe peut être exécuté, mais des problèmes se posent lorsque l'on veut étendre le test à un ensemble de sites (un type) et lorsque des mutations récurrentes se produisent. Une série de différents autres tests de neutralité, dépendant de subdivisions géographiques (Lewontin and Krakauer, 1973) ou portant sur des procédures informelles (Ayala *et al.*, 1972; Johnson and Feldman, 1973; Milton and Koehn, 1973), ou encore des test basés sur des arbres phylogéniques (Langley and Fitch, 1973), sont décrits dans Ewens (1977). Plus récemment, Hudson (1983) a étendu le test de Langley et Fitch et l'a appliqué à des phylogénies établies à partir de séquences protéiques.

Une autre classe de test consistera à tester simultanément la neutralité de 2 ou plusieurs loci. Fuerst *et al.* (1977) ont ainsi testé la distribution de la variance de l'hétérozygosité d'une série de loci selon le modèles des allèles infinis. Chakraborty *et al.* (1979) ont également étudié la distribution regroupée des fréquences alléliques d'une vingtaine de loci chez 4 espèces, et ont conclu à un bon accord entre les fréquences attendues selon le modèle des allèles infinis et les fréquences observées. Il est à noter que ces 2 dernières études avaient pour ambition de tester la validité globale de la théorie neutraliste sans s'attacher à un système particulier. Enfin, plus récemment, Hedrick et Thomson (1985) ont développé un test pour vérifier la neutralité de 2 loci liés et Hudson *et al.* (1987) ont élaboré une procédure de test permettant de comparer le polymorphisme de 2 régions du génome chez 2 espèces différentes sous l'hypothèse d'une évolution moléculaire neutre.

Il existe également un moyen simple de vérifier la conformité d'une distribution de fréquences alléliques par un test de "*goodness of fit*" entre les fréquences observées et attendues, comme le G-test (Sokal and Rohlf, 1969) ou un test de Kolmogorov-Smirnov également employé par Fuerst *et al.* (1977). Ces tests sont indépendants du modèle utilisé pour calculer les fréquences géniques théoriques. Ils permettent surtout de quantifier l'impression retirée par une simple comparaison visuelle des fréquences observées et attendues.

Il existe donc à l'heure actuelle toute une série de tests de la neutralité sélective d'un locus ou d'un ensemble de loci, selon le modèle que l'on a considéré et la nature des données à disposition (PLFR, séquences de nucléotides, séquences de protéines, fréquences alléliques, etc...). Pour notre part, nous avons choisi d'employer le test de l'homozygosité de Watterson (1978), reposant sur le modèle des allèles infinis, qui décrit bien le comportement des types déterminés par des PLFR, et qui jouit d'une

formulation mathématique bien établie (voir l'Annexe B). Il permet des analyses relativement fines pour une population donnée à un locus précis. Son comportement face à certains phénomènes de sélection ou de changements de tailles de populations est connu. Enfin, il s'agit aussi du test qui a été le plus couramment utilisé sur des sytèmes génétiques très polymorphes (Hedrick and Thomson, 1983; Johnson *et al.*, 1983; Hedrick *et al.*, 1986; Klitz *et al.*, 1986; Whittam *et al.*, 1986; Clark, 1987).

Simulations de fréquences géniques et test de neutralité

Un total de 1000 simulations de fréquences alléliques ont été effectuées selon la procédure décrite dans l'annexe B pour chaque échantillon de types d'ADN-mt à notre disposition. Ces simulations ont permis de définir la distribution des fréquences des k allèles d'un échantillon de r gènes selon le modèle des allèles infinis. La valeur attendue de la quantité F en a été tirée et les limites d'intervalles de confiance pour une probabilité donnée ont été déterminées empiriquement. Il est évident que la valeur F n'est ici pas égale à l'homozygosité de l'échantillon, car cette notion n'est définissable que pour une population génétiquement diploïde, alors que les gènes mitochondriaux se comportent comme des unités haploïdes du point de vue de leur transmission. F ne représente ici qu'une quantité synthétisant la distribution des fréquences alléliques.

Dans notre procédure de test, nous assumons que l'hypothèse nulle (H_0) est la neutralité sélective de l'ADN-mt. L'hypothèse alternative est définie comme la présence d'allèles désavantagés, ce qui a pour conséquence une augmentation de la valeur observée de F. Deux indices nous ont incités à choisir cette hypothèse alternative. Le premier est le fait que la majeure partie des échantillons étudiés présente une diversité moléculaire réduite par rapport aux prédictions du modèle des sites infinis, équivalent au modèle des allèles infinis en l'absence de recombinaison (Ewens, 1979, p. 277). Le second est l'étude de la caractérisation moléculaire des mutations ayant conduit à des gains de site qui montre l'absence d'une partie des substitutions survenues en 1ère et en 2ème position des codons dans les régions codant pour des protéines. D'autre part, plusieurs types sont phénotypiquement différents du type ancestral et donc potentiellement désavantagés (voir Table 4.23 et Figure 4.15). Nous déterminerons donc une valeur de F (F_{sup}) qu'il est peu probable (<5%) de dépasser en cas de neutralité sélective de tous les allèles.

Nous avons ainsi effectué des tests unilatéraux sur la statistique F et avons déterminés la valeur de F pour laquelle 95% des simulations ont fourni des valeurs

inférieures (F_{sup} = borne supérieure de l'intervalle). Nous trouvons un intervalle de confiance [1/k ; F_{sup}] dont la borne inférieure (F_{inf} = 1/k) est trouvée, sachant k, de telle manière que tous les allèles possèdent une fréquence identique égale à 1/k et donc que F défini par (B.3) est lui-même équivalent à 1/k. Le test de neutralité sélective, dans le cas de l'ADN-mt, consistera donc à vérifier que la valeur observée de F tombe bien dans cet intervalle.

TABLE 5.1: Valeurs observées et atttendues de la statistique F

Echantillons	Nombre d'allèles	Nombre de gènes	F Attendu	F Observé		F_{sup} [1]	G^2	(d.l.)
Caucasoïdes[a]	15	50	0,139	0,395	>	0,210*	26,84*	(6)
Romains[b]	15	95	0,174	0,413	>	0,284*	44,59*	(7)
Sardes[b]	13	134	0,221	0,578	>	0,391*	88,25*	(7)
Isr. Juifs[c]	9	39	0,233	0,291	<	0,370	8,39	(4)
Isr. Arabes[c]	12	39	0,165	0,350	>	0,254*	15,35*	(4)
Bantous[a]	12	40	0,165	0,193	<	0,244	2,34	(4)
San[a]	9	34	0,224	0,196	<	0,344	1,86	(3)
Orientaux[a]	9	46	0,244	0,499	>	0,399*	19,21*	(4)
Tharu[d]	13	91	0,201	0,345	>	0,327*	28,11*	(6)
Amérindiens[e]	5	74	0,465	0,823	>	0,755*	35,70*	(3)
Caucasoïdes[f]	42	46	0,026	0,029	>	0,027*	2,50§	(0)
Nouvelle-Guinée[f]	17	26	0,081	0,101	>	0,098*	0,84	(2)
Nouvelle-Guinée[g]	30	55	0,053	0,066	=	0,066	4,80	(8)
Japonais[h]	22	120	0,118	0,519	>	0,183*	132,64*	(10)
Japonais[i]	62	116	0,027	0,042	>	0,032*	15,33*	(14)
Japonais[j]	11	74	0,227	0,600	>	0,361*	51,45*	(5)
Ainou[j]	3	48	0,629	0,919	=	0,919§	18,74*	(2)

[1] Valeur de F en-dessous de laquelle sont réparties 95% des simulations. Cela correspond à la borne supérieure empirique d'un intervalle de confiance au niveau 5%.

[2] Valeur du test de G provenant de la comparaison de la distribution des fréquences géniques observées et attendues. Le nombre de degrés de liberté est indiqué entre parenthèses. Cette valeur de G est à comparer avec une valeur de χ^2 de même nombre de degrés de liberté.

* Significatif au niveau 5%.

§ Test impossible.

Sources des échantillons: a: Johnson *et al.* 1983; b: Brega *et al.* 1986b; c: Bonné-Tamir *et al.* 1986; d: Brega *et al.* 1986a; e: Wallace *et al.* 1985; Cann *et al.* 1987; g: Stoneking *et al.*, 1986; h: Horai *et al.*, 1984; i: Horai and Matsunaga, 1986; j: Harihara *et al.*, 1986.

Nous avons reporté dans la Table 5.1 les résultats des simulations d'échantillons concernant les valeurs de F. Certaines simulations portant sur des échantillons de Cann *et al.* (1987) n'ont pas été reportés dans la Table 5.1. Il s'agit des échantillons d'africains, d'australiens et d'asiatiques qui présentaient autant de types que de gènes dans l'échantillon. Dans ces conditions, une fois fixé k et r, il n'existait aucune variabilité dans les simulations des distributions alléliques et tout test aurait été vide de sens sur ces données. Seule une procédure d'échantillonnage portant sur un plus grand nombre de gènes issus de chaque population naturelle aurait sans doute permis cette étude. Cependant, l'absence de ces 3 échantillons n'est pas déterminante pour pouvoir tirer des conclusions sur la neutralité sélective de ce système.

Mise en évidence d'une sélection différentielle entre les populations

D'une manière générale, toutes les valeurs de F_{obs} sont supérieures aux valeurs attendues, à l'exception de l'échantillon San. Au niveau 5%, seuls les distributions haplotypiques des échantillons d'Israéliens Juifs, de Bantous, de San et de Nouvelle-Guinée apparaissent compatibles avec l'hypothèse de neutralité sélective de l'ADN-mt. Cela est confirmé par le test de G, où les distributions alléliques attendues et observées ne diffèrent pas significativement au niveau 5%. Ces résultats sont clairement observables sur les Figures 5.1 à 5.17 où nous avons reportés les fréquences simulées à côté des fréquences observées des allèles classés par ordre de fréquences décroissantes.

Concernant les 10 premiers échantillons analysés, rappellons-le, avec les mêmes enzymes de restrictions, l'examen des Figures 5.1 à 5.10 nous permet de constater que le type 1 est effectivement surabondant, ce qui semble constituer le principal facteur de rejet de l'hypothèse neutraliste. La fréquence élevée du type 1 conduit aussi à une sous représentation des types de fréquences intermédiaires (non-singletons). On observe que les distributions haplotypiques attendues et observées sont en très bon accord pour les échantillons San et Bantous (Figures 5.7 et 5.8) et en relativement bon accord pour les Israéliens Juifs où l'on note la fréquence élevé du type 6.

Pour les 7 autres échantillons analysés avec des enzymes différents, seuls les échantillons de Nouvelle-Guinée présentent une distribution allélique pouvant prétendre être compatible avec la théorie de neutralité sélective. Rappellons toutefois que l'échantillon de Nouvelle-Guinée de Stoneking *et al.* (1986) constitue une extension de l'échantillon de Cann *et al.* (1987). Les 7 échantillons ont aussi pour caractéristique la

surabondance d'un ou plusieurs types par rapport à la distribution attendue (Figures 5.11 à 5.17), pouvant entraîner un déficit des types de fréquences intermédiaires et un nombre trop élevé de types singletons (Figures 5.12 à 5.17).

Donc, bien que les types ne soient pas comparables entre toutes les études, il apparaît que la présence de types trop fréquents constitue un trait particulier du polymorphisme de l'ADN-mt. Ceci a souvent pour conséquence (13 échantillons sur 17) de faire diverger la distribution des fréquences haplotypiques d'une distribution neutre. Il apparaît ainsi que la neutralité sélective de la molécule d'ADN-mt ne peut être supportée par les présents résultats et qu'une partie des types d'ADN-mt pourraient être sélectivement désavantagés. Cette observation confirme les résultats de Johnson *et al.* (1983) qui avaient effectué un test de l'homozygosité sur les fréquences alléliques d'un regroupement, discutable, de 5 échantillons provenant de 4 continents. Elle est aussi partiellement en accord avec les résultats de Whittam *et al.* (1986) qui avaient testé la neutralité de 28 loci du génome mitochondrial sur un échantillon regroupant 145 individus définis dans l'étude de Cann *et al.* (1987). Ces auteurs avaient trouvé que 10 tests de l'homozygosité sur 35 conduisaient à un rejet de l'hypothèse neutraliste, mais avaient conclu à une acceptation globale de la neutralité de la molécule d'ADN-mt.

Le fait de tester, pour la première fois, des échantillons individuels au lieu de regroupements arbitraires d'échantillons hétérogènes, permet de mettre en évidence des comportements sélectifs différentiels entre populations. Il est à noter que ces conclusions vont dans le même sens et éclairent les résultats trouvés lors de l'étude de la diversité moléculaire des échantillons. Il semble maintenant clair que certaines populations (plus particulièrement africaines) ont connu une évolution conforme à la thèse neutraliste des modèles des allèles et des sites infinis. Les types d'ADN-mt de ces populations sont bien différenciés les uns des autres et leurs fréquences observées sont compatibles avec un modèle reposant sur un équilibre entre dérive génétique et mutations. Pour les autres populations, il apparaît que les types qu'on y trouve ne sont moléculairement pas assez différenciés les uns des autres, ce qui suppose l'existence de mécanismes d'élimination de certains types vraisemblablement désavantagés. Les fréquences alléliques de ces populations reflètent ces différences sélectives entre types.

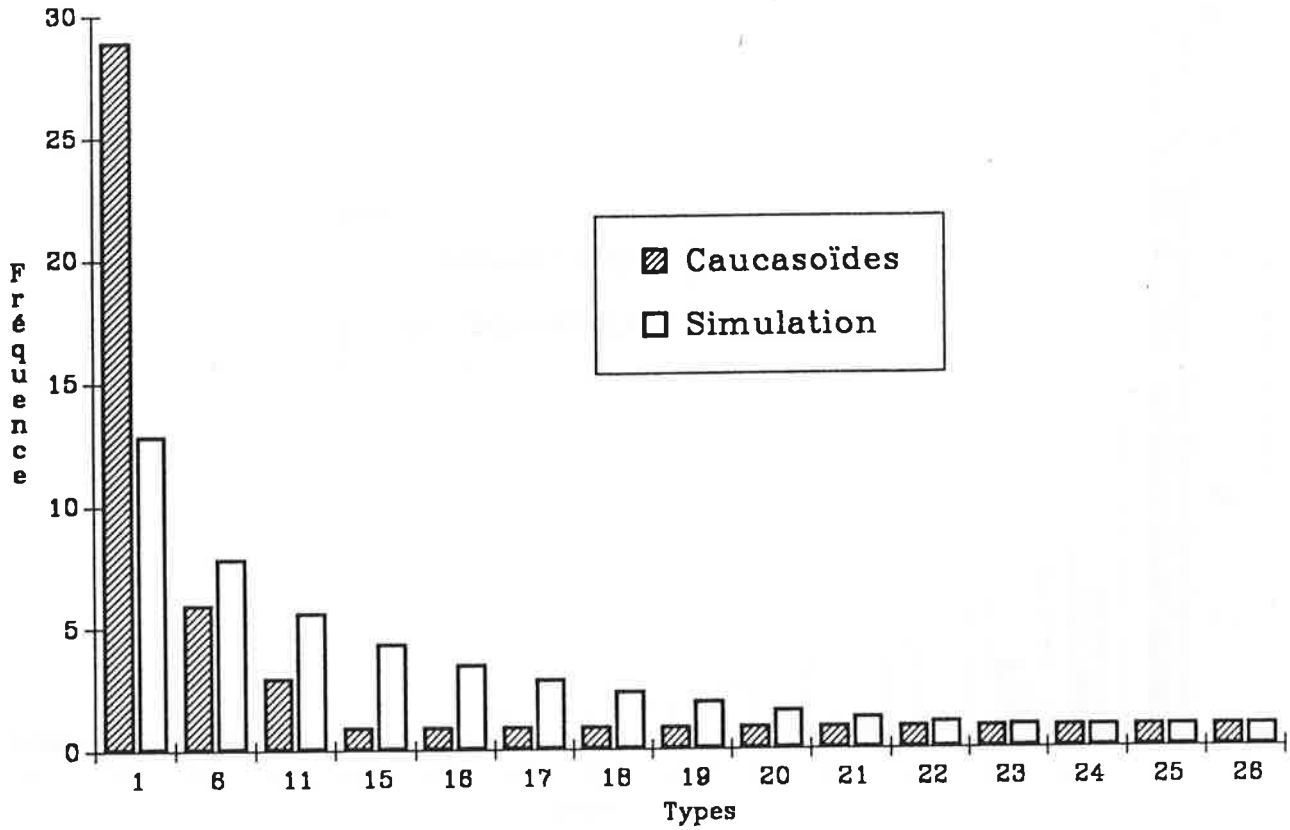

FIGURE 5.1 : Fréquences alléliques simulées et observées dans un échantillon de Caucasoïdes.

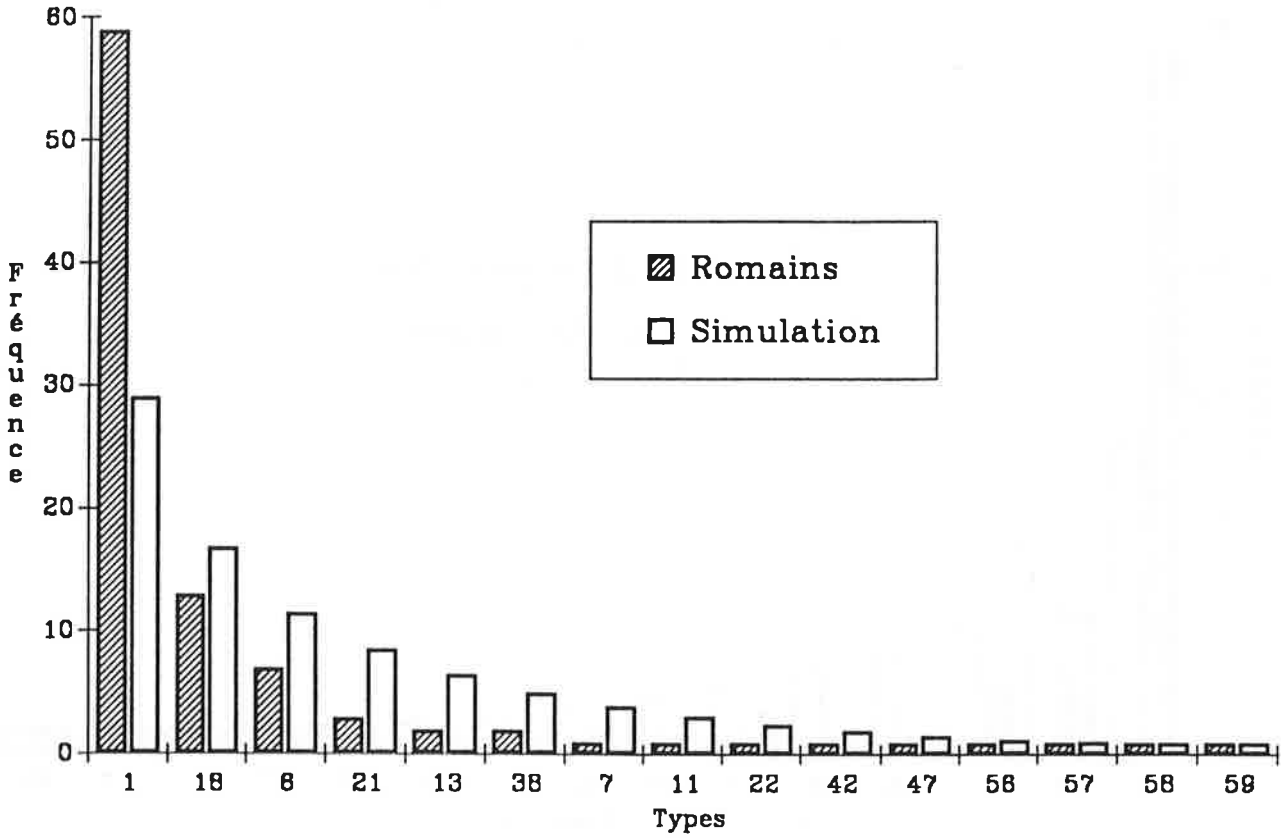

FIGURE 5.2 : Fréquences alléliques simulées et observées dans un échantillon de Romains.

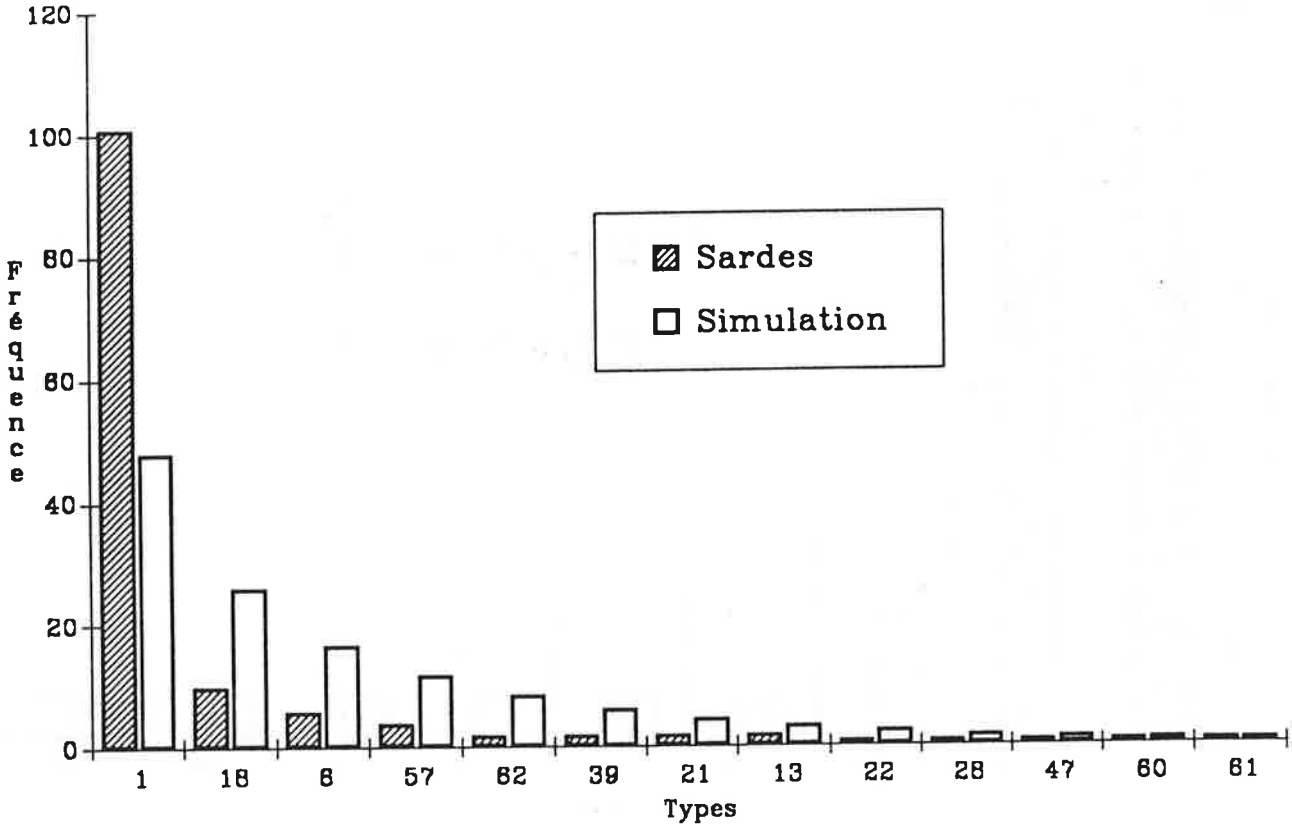

FIGURE 5.3 : Fréquences alléliques simulées et observées dans un échantillon de Sardes.

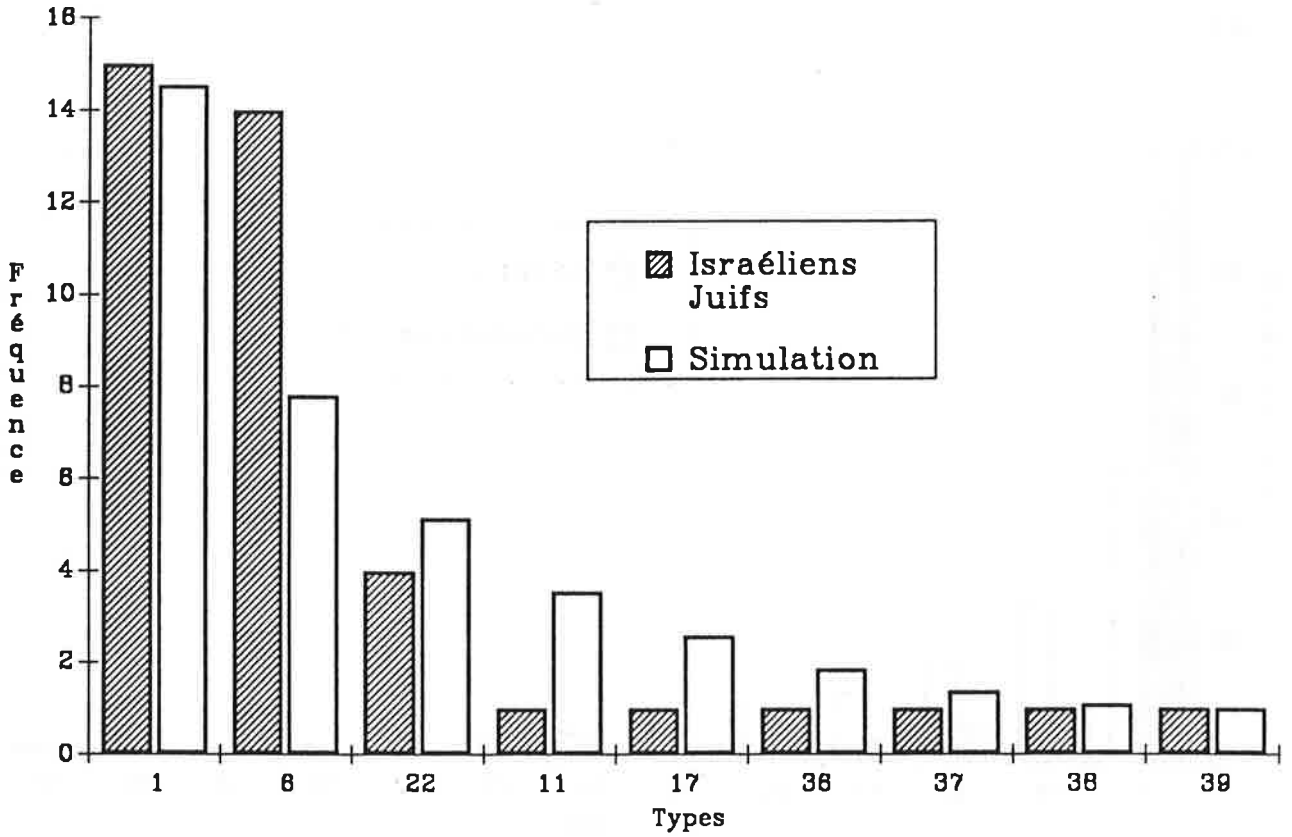

FIGURE 5.4 : Fréquences alléliques simulées et observées dans un échantillon d'Israéliens Juifs.

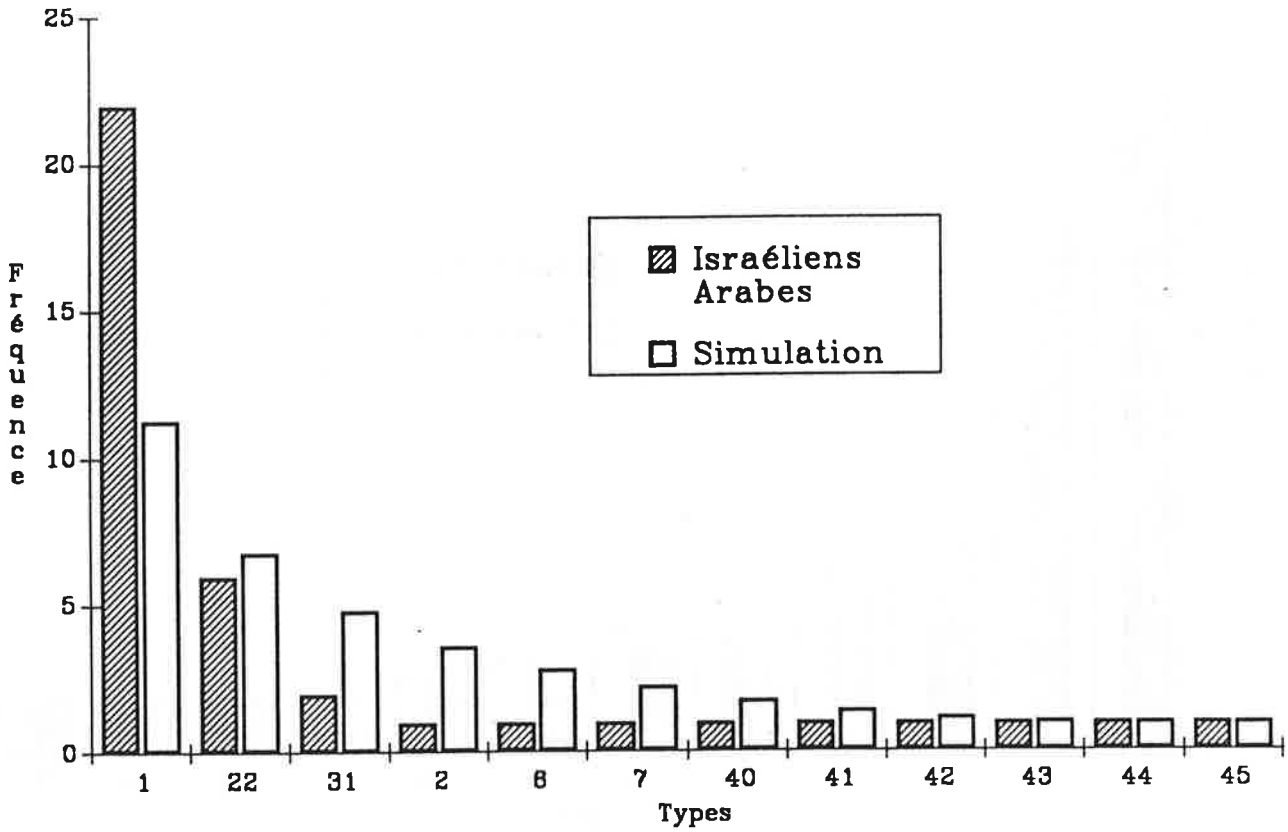

FIGURE 5.5 : Fréquences alléliques simulées et observées dans un échantillon d'Israéliens Arabes.

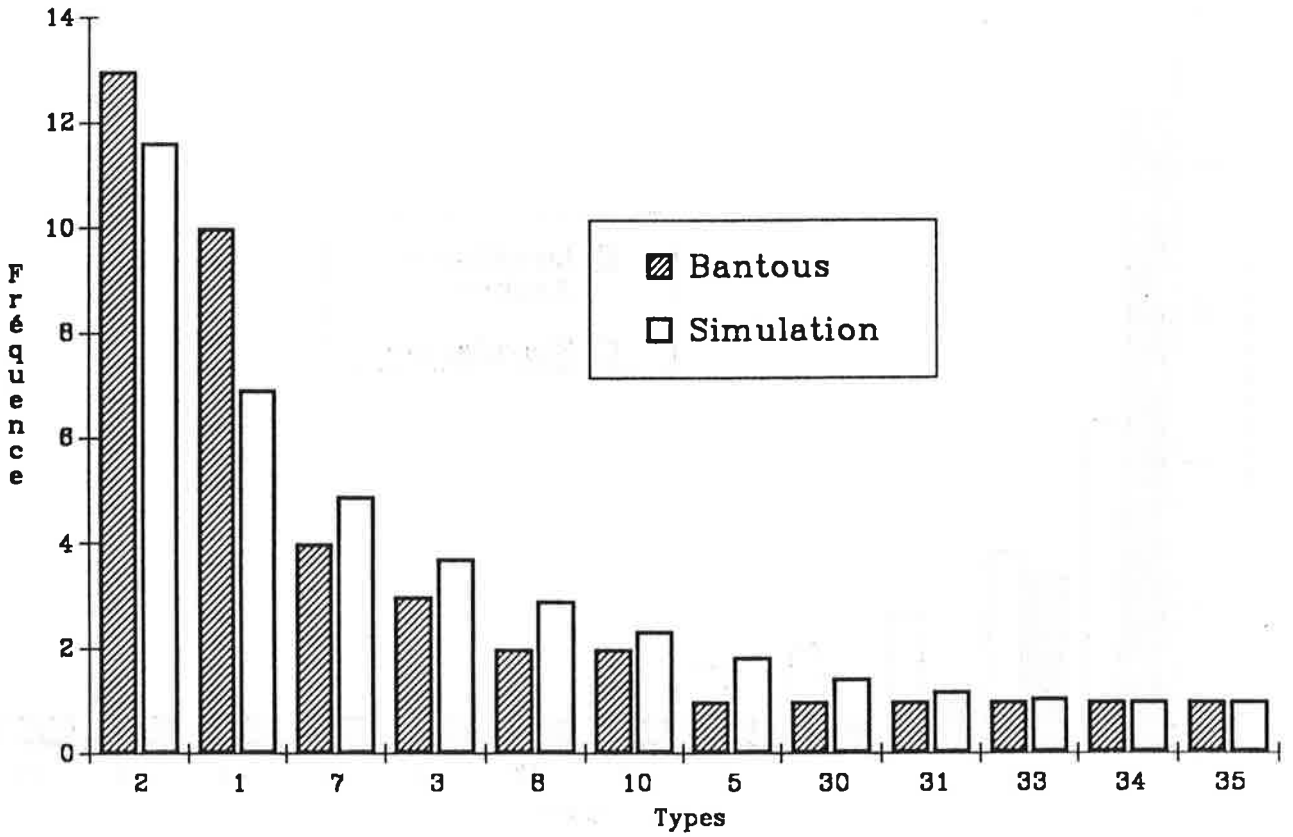

FIGURE 5.6 : Fréquences alléliques simulées et observées dans un échantillon de Bantous.

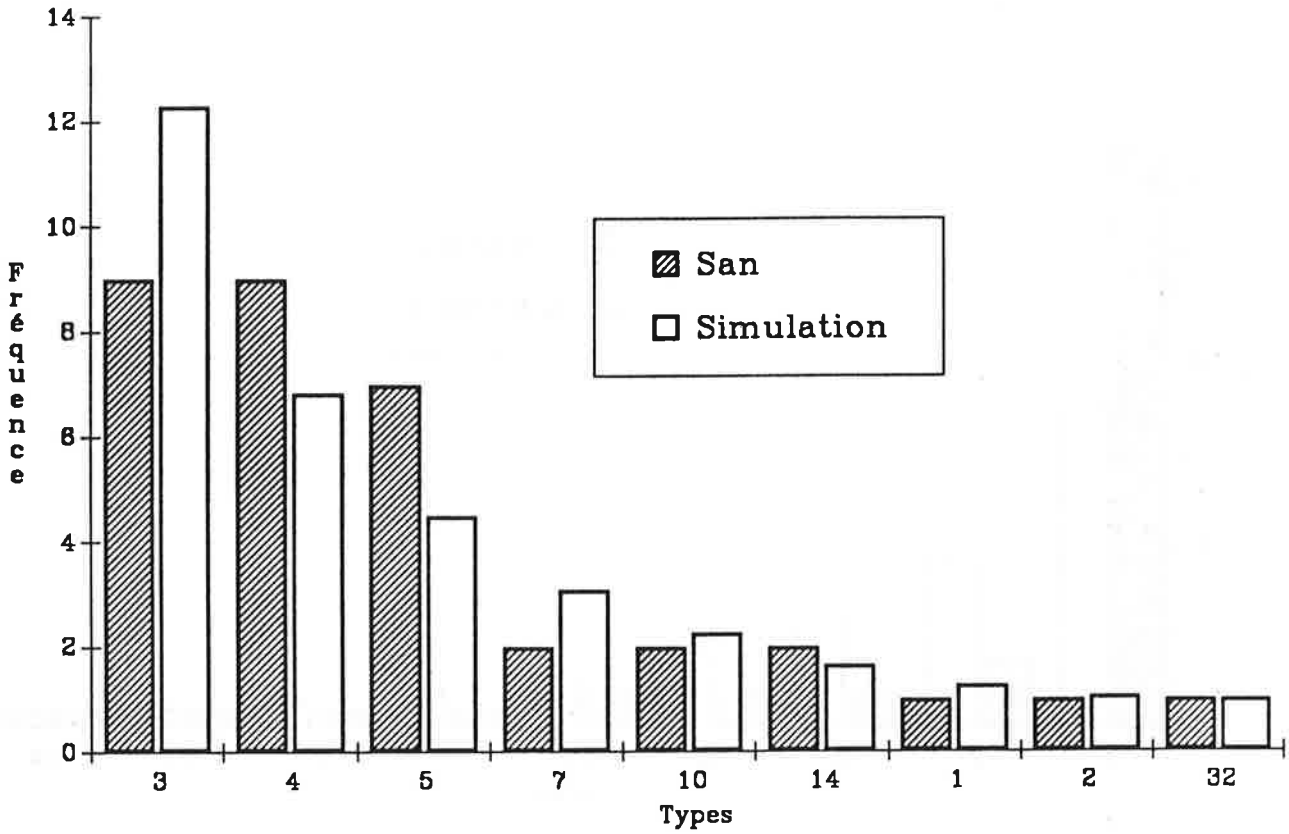

FIGURE 5.7 : Fréquences alléliques simulées et observées dans un échantillon de San.

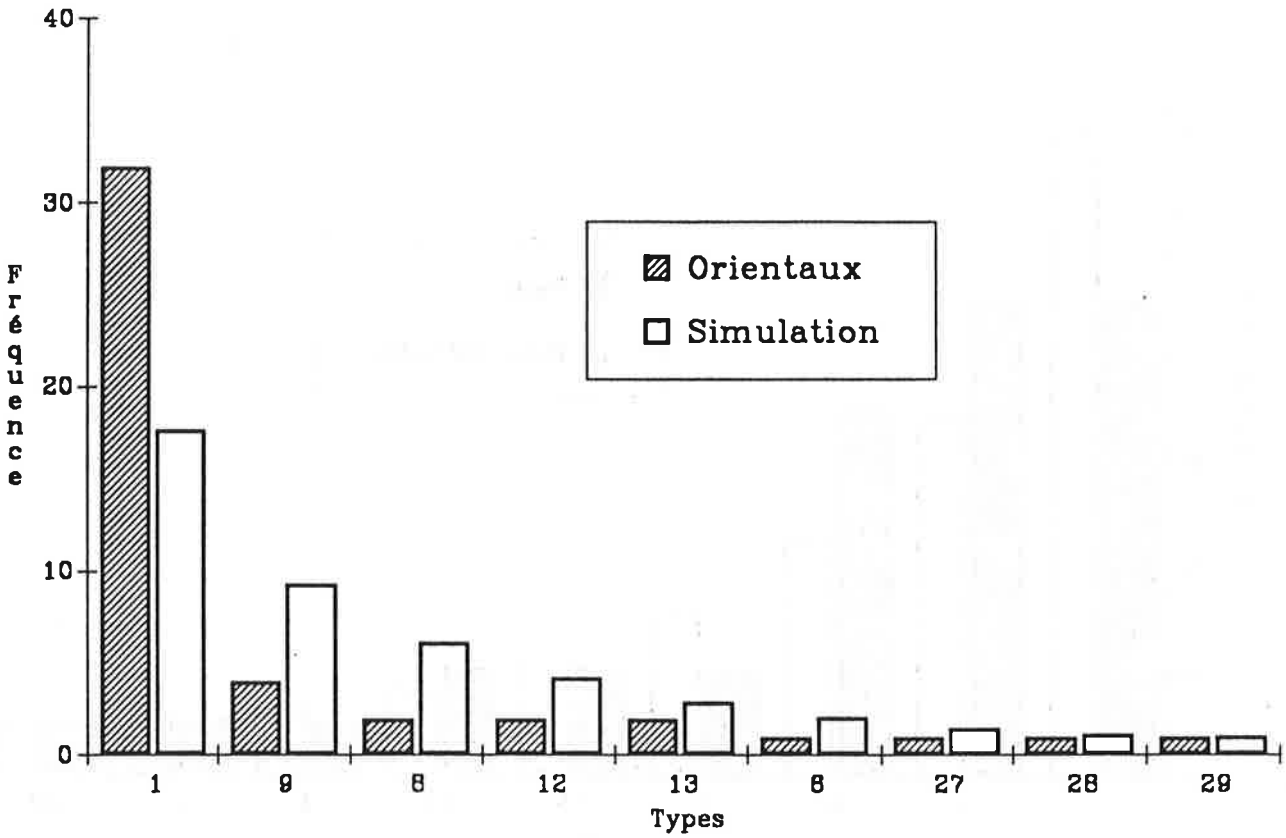

FIGURE 5.8: Fréquences alléliques simulées et observées dans un échantillon d'Orientaux.

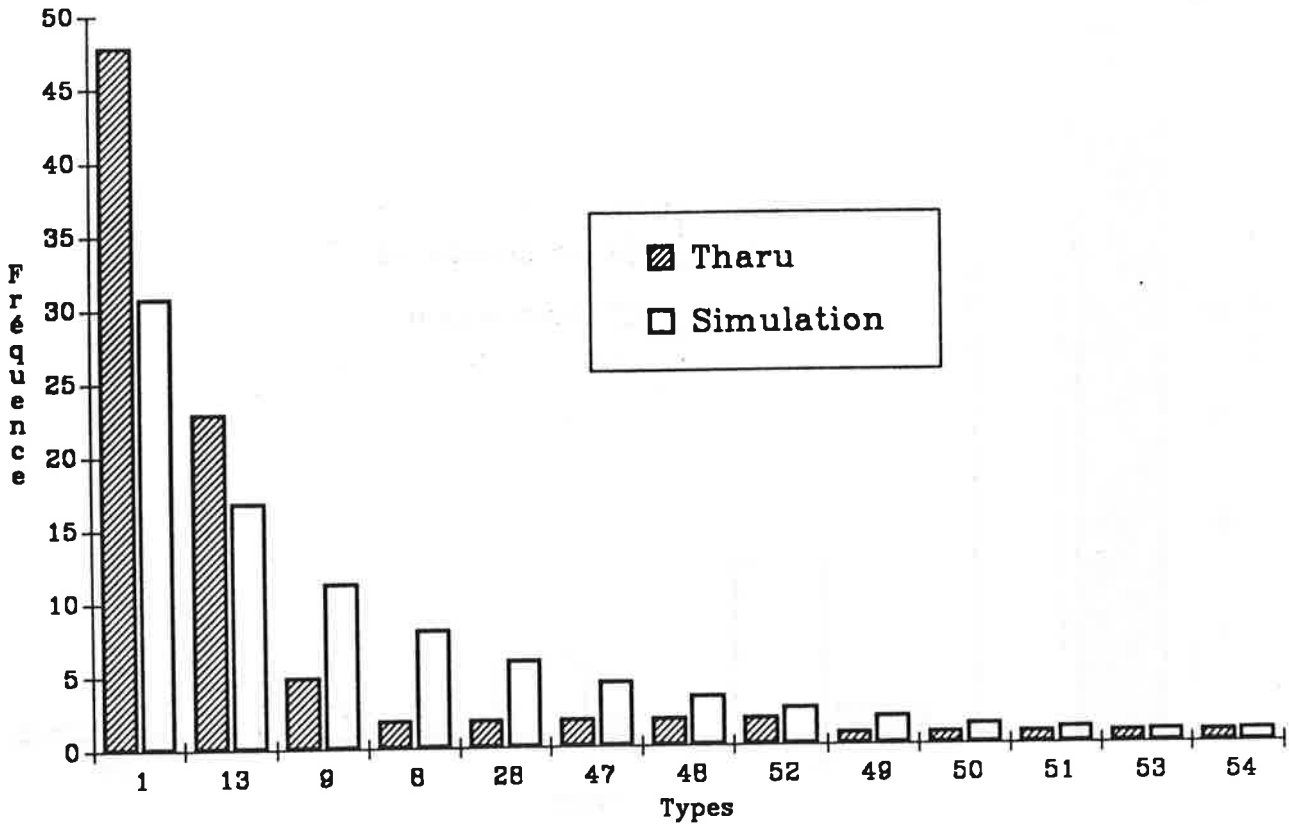

FIGURE 5.9 : Fréquences alléliques simulées et observées dans un échantillon de Tharu.

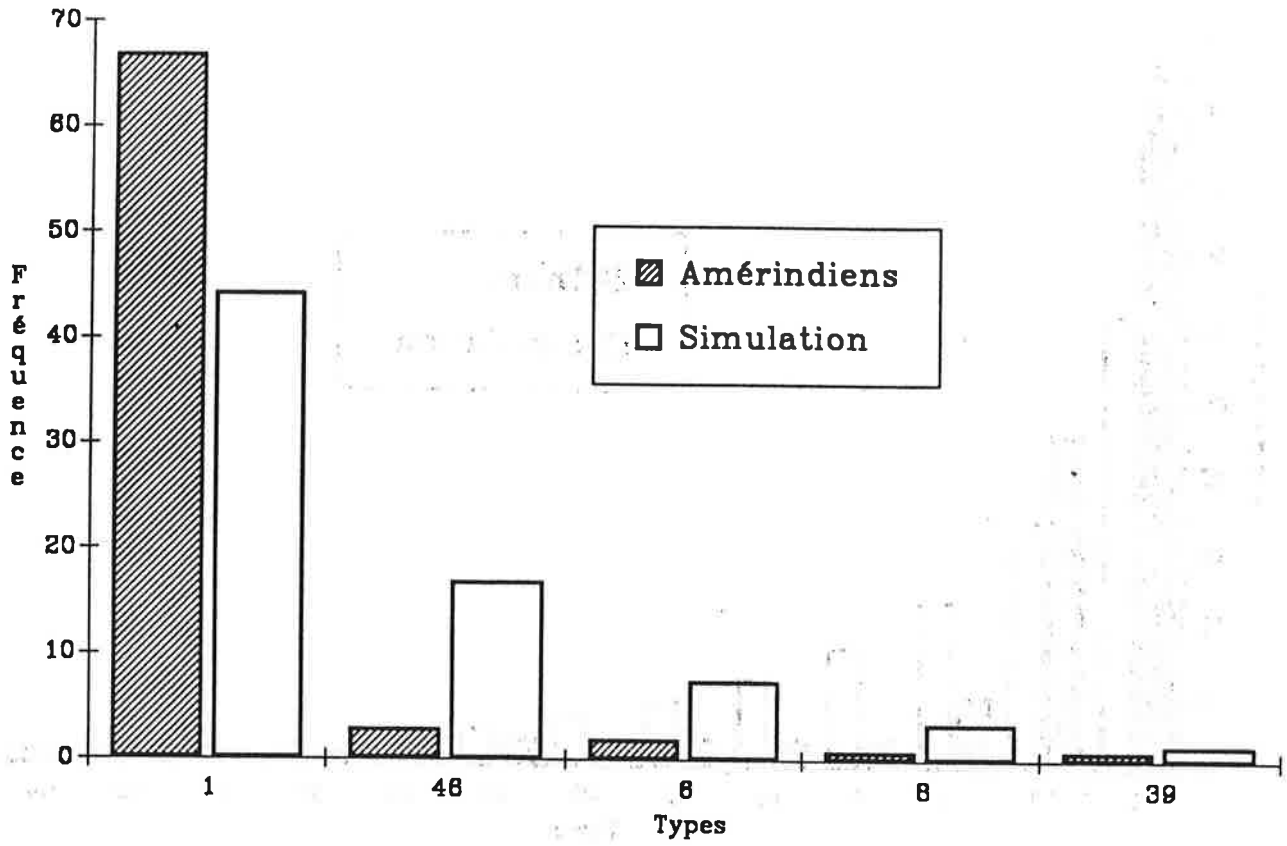

FIGURE 5.10 : Fréquences alléliques simulées et observées dans un échantillon d'Amérindiens.

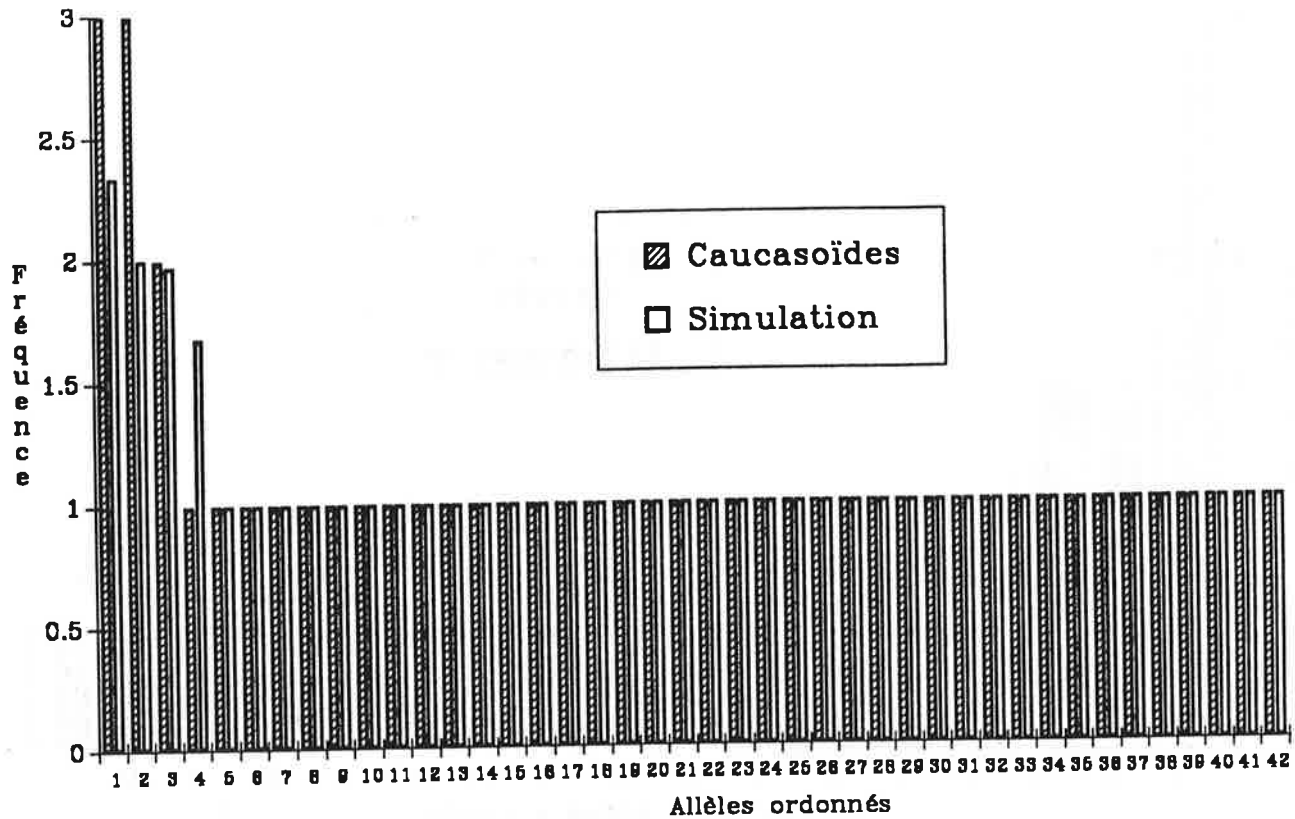

FIGURE 5.11 : Fréquences alléliques simulées et observées dans un échantillon de Caucasoïdes.

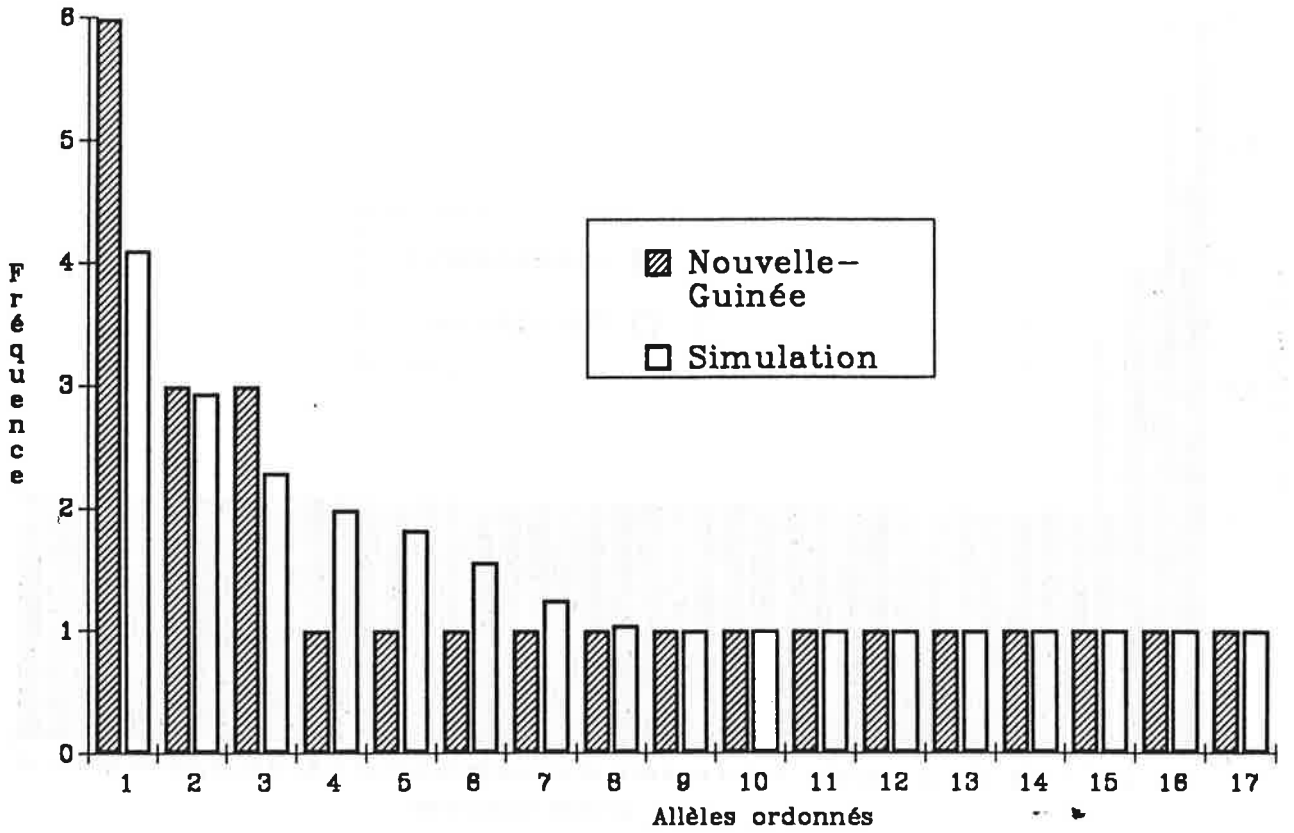

FIGURE 5.12 : Fréquences alléliques simulées et observées dans un échantillon de Nouvelle-Guinée.

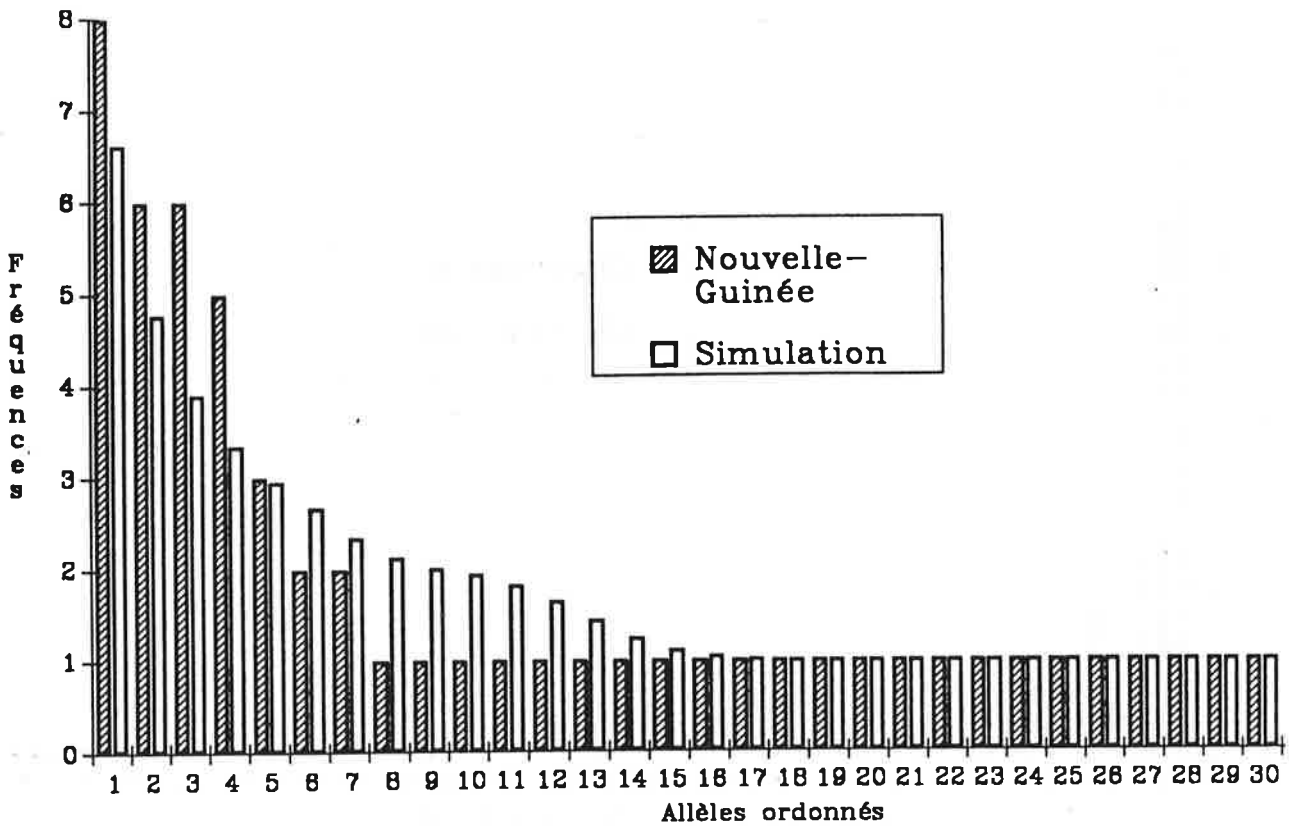

FIGURE 5.13 : Fréquences alléliques simulées et observées dans un échantillon de Nouvelle-Guinée.

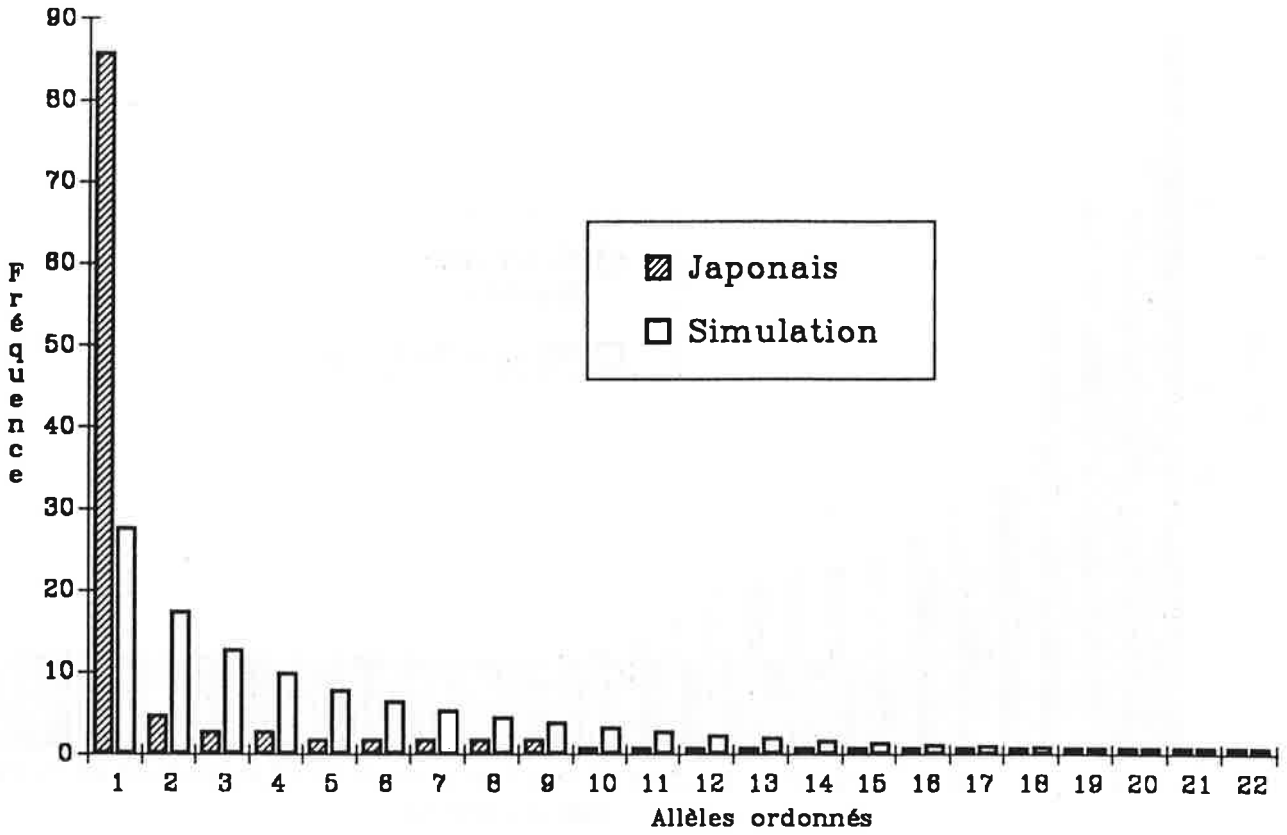

FIGURE 5.14 : Fréquences alléliques simulées et observées dans un échantillon de Japonais.

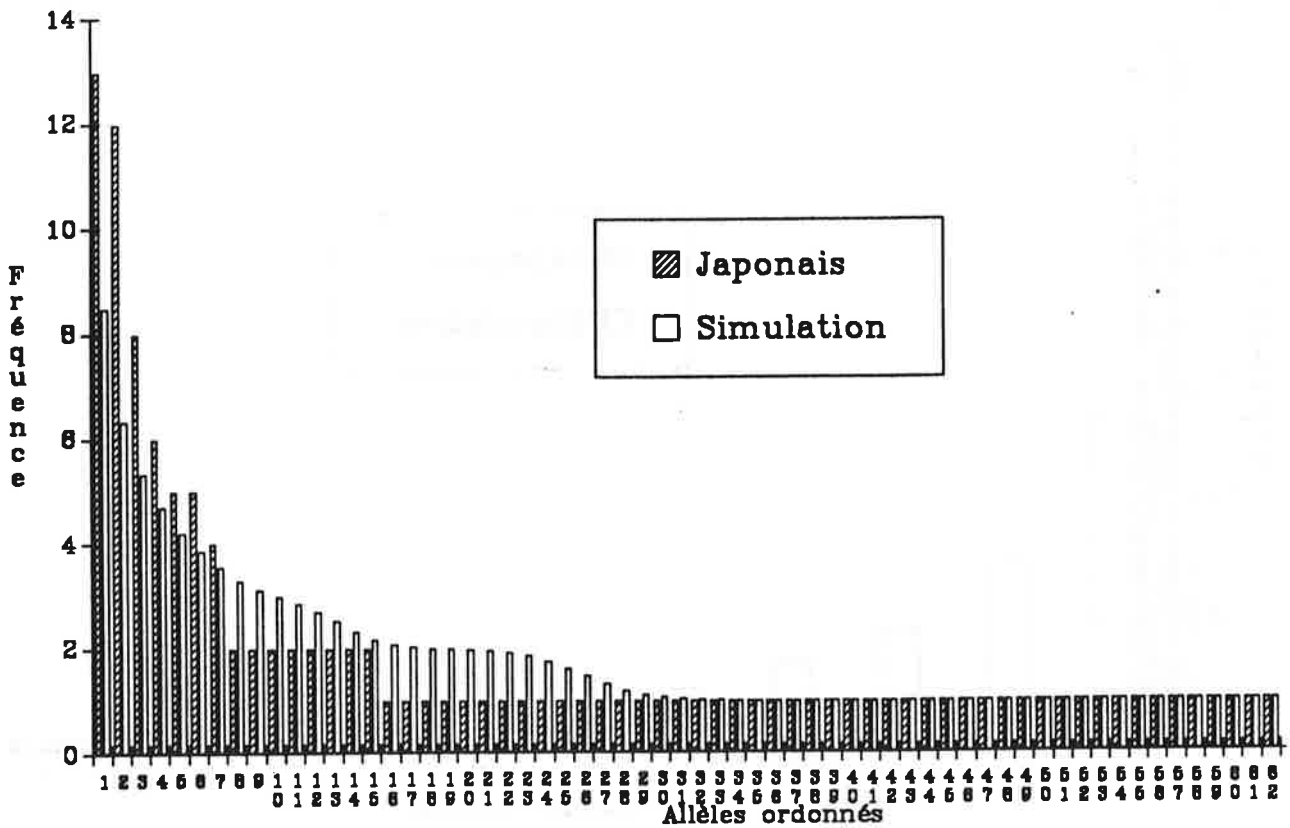

FIGURE 5.15 : Fréquences alléliques simulées et observées dans un échantillon de Japonais.

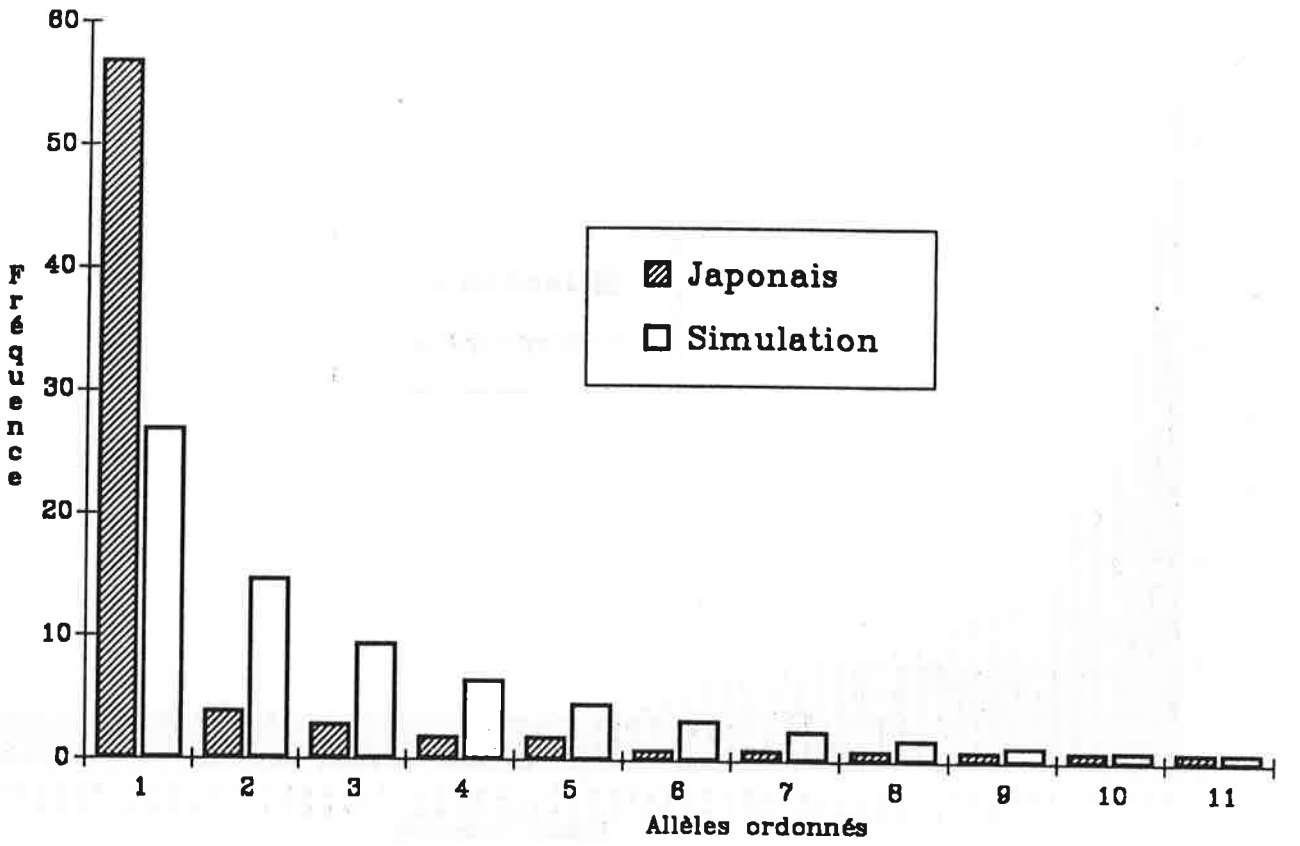

FIGURE 5.16 : Fréquences alléliques simulées et observées dans un échantillon de Japonais.

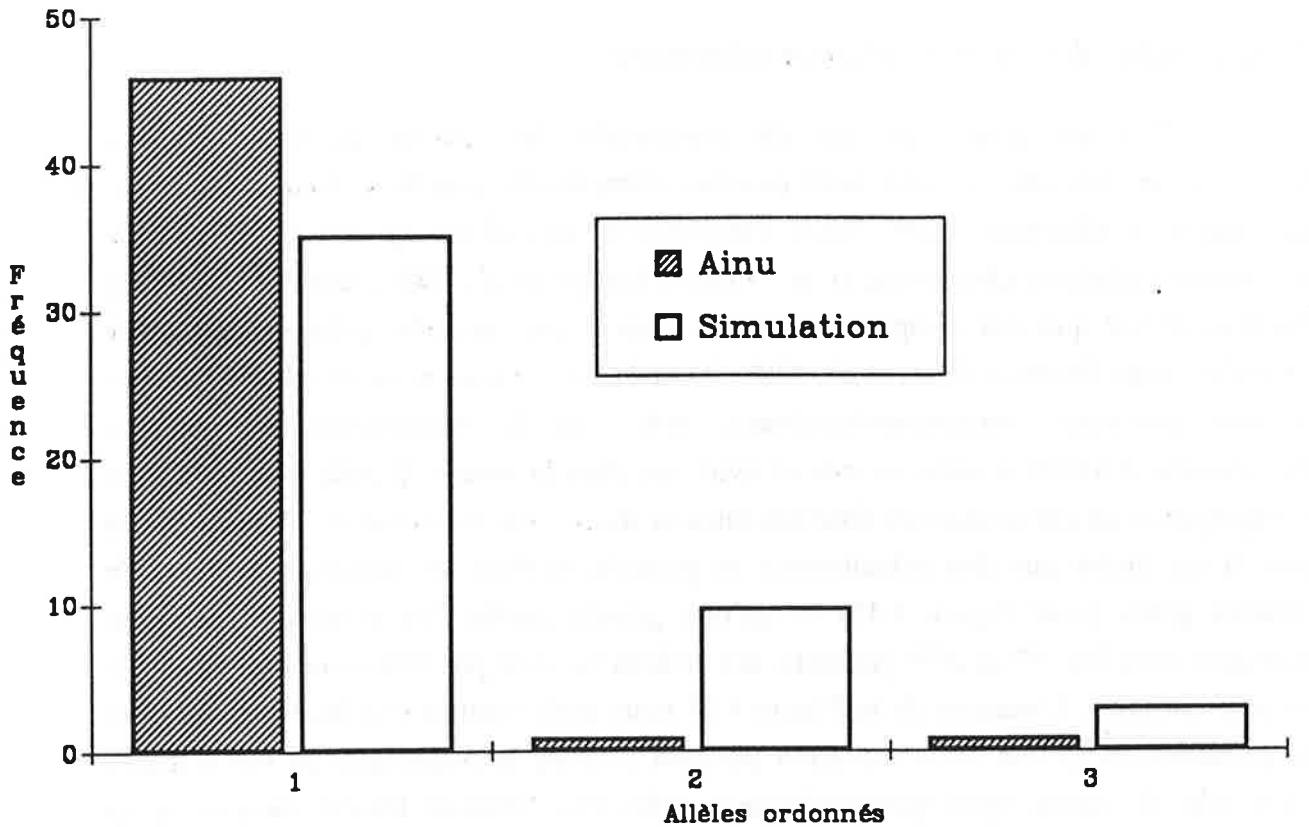

FIGURE 5.17 : Fréquences alléliques simulées et observées dans un échantillon de Ainou.

Causes possibles de l'apparente sélection différentielle

Il nous reste à essayer de comprendre les raisons de ces évolutions dissemblables. En premier lieu, nous pouvons comprendre que la molécule d'ADN-mt soit sujette à sélection, étant donné l'importance des gènes qu'elle porte pour la respiration cellulaire (Anderson *et al.*, 1981; Chomyn *et al.*, 1985; Spinner and King, 1986) et le fait que des myopathies ont été associé avec certains polymorphismes de l'ADN-mt chez l'homme (Holt *et al.*, 1988). L'ADN-mt pourrait aussi être impliqué dans certains processus extra-mitochondriaux, tels que la modulation des niveaux d'expression d'ARN's nucléaires mis en évidence chez la levure (Parikh *et al.*, 1987) ou la ségrégation de chromosomes observés chez la souris (Beermann *et al.*, 1988). D'autre part, il est établi que des substitutions se produisent dans les portions codantes de certains gènes (voir Figure 4.15) et qu'une grande partie des substitutions s'étant produites dans les 1ère et 2ème positions des codons ne sont pas retrouvées sur les types les plus courants. L'examen de la Figure 4.15 nous avait montré que les types africains ne présentaient qu'une seule mutation pouvant affecter le phénotype de son porteur, alors que 15 autres types potentiellement défavorisés étaient identifiables dans les autres groupes continentaux. Cela confirme le fait que la plupart des types africains sont vraisemblablement neutres, alors que des types sélectionnés sont présents en plus grand nombre dans les autres groupes continentaux. En ce qui concerne l'échantillon d'Israéliens Juifs, on constate une contradiction entre les résultats du test de l'homozygosité et ceux de l'analyse de la diversité moléculaire. Nous sommes tentés d'expliquer cette discordance par le fait que cet échantillon est probablement hétérogène du fait qu'il comporte des juifs ashkénazes et sépharades, originaires de régions différentes. Néanmoins, cet échantillon est le seul, à l'exception des africains, à présenter une fréquence du type 1 compatible avec les prédictions théoriques.

Si certains types présentent un désavantage sélectif par rapport à d'autres, il semble a priori étrange que leur effet se fasse sentir dans quelques populations et pas dans d'autres ou que des types désavantagés n'apparaissent que dans certaines populations. En fait, l'absence de sélection apparente dans une population peut très bien être expliquée théoriquement, si l'on réalise que le facteur de sélection n'est pas uniquement le coefficient de sélection s d'un certain allèle (ou d'un type), mais le produit $N_e s$ où intervient également la taille effective de la population (N_e). Il faut rappeller que Kimura (1968) avait montré que si $|N_e s| \leq 1$, un allèle pouvait être considéré comme neutre. Ainsi, le facteur $2N_e s$ apparaissant dans l'équation (B.4) sera le principal modulateur de $E(F|k)$ en cas de sélection. Si l'on suppose qu'un allèle

donné subit les mêmes contraintes sélectives dans toutes les populations humaines, il ressort que la taille effective de la population où il se trouve sera responsable du fait qu'il se comporte comme un allèle sélectionné ou non. Un mutant sélectionné dans une grande population pourra sembler neutre dans une petite population (Nei, 1987, p.411).

Un récent accroissement de population peut également entraîner une augmentation artificielle de la valeur de E(F|k), mais d'après les simulations de Watterson (1986), cet effet est très sensible juste après le bottleneck et E(F|k) tend ensuite vers sa valeur d'équilibre. De plus, il n'est pas clair que la non-stationarité des populations a un effet majeur sur les périphéries de la distribution de E(F|k), où les niveaux de signification sont déterminés (Watterson, 1986, p.903). Donc, bien qu'une récente augmentation de la taille des populations humaines ait eu lieu, il nous semble difficile de penser que la totalité de l'accroissement de E(F|k) soit dû à ce seul facteur. Cette impression est corroborée par les résultats de Avise *et al.* (1988) qui montrent que la diversité moléculaire de l'ADN-mt est trop réduite pour 3 autres vertébrés (*Anguilla rostrata*, *Arius felis* et *Agelaius phoenicius*). Ceci pourrait être dû également à un accroissement récent des populations où à un désavantage sélectif de certains allèles. Bien qu'il soit possible que des populations d'anguilles, de poissons-chats ou de merles aient vu leurs effectifs augmenter récemment, il semble plus probable que la molécule d'ADN-mt soit sujette à un même type de sélection dans la plupart des espèces.

Il faut aussi rappeller que la taille effective de la population n'est pas la même lorsque l'on étudie des gènes mitochondriaux ou nucléaires. D'une part, la taille effective de la population (N_e) d'ADN-mt est égale en principe au seul nombre de femelles. D'autre part, une population sera effectivement subdivisée pour les gènes mitochondriaux à des taux de migrations pour lesquels les gènes nucléaires se conduisent comme tirés d'une population panmictique (Birky *et al.*, 1983). La différence est d'un facteur 4, si les mâles et les femelles connaissent un taux de migration équivalent, mais s'accentue encore dans le cas d'une migration préférentielle des mâles. Il semble donc que la taille effective des populations actuelles ou passées constitue un facteur important dans le comportement différentiel de certaines populations vis-à-vis du test de neutralité sélective et que celle-ci agisse au niveau de la sélection ou qu'elle conduise à une augmentation de l'homozygosité apparente, ou encore que ces 2 facteurs agissent de façon synergique. Il est ainsi possible que des conditions démographiques et sociales aient été telles, pour les San, les Bantous et les Papous de Nouvelle-Guinée, qu'elles aient permis le développement d'un polymorphisme de l'ADN-mt apparemment neutre dans ces populations.

Conséquences de la sélection

La mise en évidence de mécanismes évolutifs différentiels selon les populations entraîne une révision de l'interprétation des données provenant du polymorphisme de l'ADN-mt et concernant l'histoire du peuplement humain. Tout d'abord, il apparaît que les conclusions basées sur des distances génétiques entre populations calculées à partir des fréquences géniques (Johnson *et al.*, 1983) doivent être réexaminées. Elles concluaient notamment à une divergence très précoce des San, puis ensuite des Bantous et enfin des populations caucasoïdes et orientales. Nous avons retrouvé cette nette divergence des populations africaines dans les Figures 4.7 et 4.8 où nous avions déjà montré qu'elle était due à de faibles fréquences du type 1 dans ces populations. La prépondérance du type 1 peut donc être liée à la présence d'allèles désavantagés qui seraient rapidement éliminés ou maintenus à de faibles fréquences (excès de types singletons dans les échantillons). La valeur sélective du type 1 et des types proches serait donc supérieure à la moyenne de celle des autres allèles mutants, ce qui contribuerait à le maintenir à de fortes fréquences dans certaines populations.

Les distances génétiques basées sur la diversité nucléotidique (Cann *et al.*, 1987) sont également sujette à caution, car la diversité sera moindre dans les populations où agirait la sélection, ce qui conduirait à une forte différenciation des populations neutres. C'est effectivement ce qui est observé dans la Table 4.55 ou dans la Figure 2 de l'étude de Darlu et Tassy (1987c) reprenant les données de Cann *et al.*, où l'échantillon de Nouvelle-Guinée apparaît comme le plus divergent, alors que nous avons trouvé que sa distribution de fréquences géniques était compatible avec la théorie neutraliste.

Il s'ensuit que l'étude du polymorphisme de l'ADN-mt s'est basée sur des données biaisées. L'utilisation des fréquences des types d'ADN-mt à des fins anthropologiques est sérieusement limitée et peut causer des erreurs d'appréciation des apparentements génétiques réels entre populations. Il s'avère notamment que l'apparente ancienneté des populations africaines serait un artefact dû à la sélection dans les autres populations. Il semble donc préférable de raisonner sur les arbres phylogéniques des types, leurs présences ou leurs absences, ou encore sur des indices qui ne dépendent pas directement des fréquences alléliques (nombres de sites polymorphes). Sans oublier que de telles études nécessitent de bons échantillons qui restent à constituer.

CONCLUSIONS

Tout au long de ce travail, nous avons cherché à évaluer les problèmes méthodologiques liés à l'analyse des données du polymorphisme de l'ADN-mt (Excoffier et Langaney, 1988). Parmi ceux-ci, le problème de la constitution des échantillons s'est révélé intervenir à plusieurs niveaux. Tout d'abord, il faut mentionner que la taille de tous les échantillons utilisés pour étudier les polymorphismes de longueur des fragments de restriction est nettement insuffisante. Ceci induit de médiocres estimations des fréquences des types d'ADN-mt ainsi que l'absence dans les échantillons d'une quantité importante de types pourtant présents dans les populations (Figure 2.1). Par conséquent, il n'est jamais possible d'être certain de l'absence totale de types particuliers dans les populations étudiées. Les échantillons se sont aussi souvent avérés très hétérogènes dans leur constitution ethnique ou géographique (Table 1.1 et 4.2). Cette hétérogénéité, particulièrement sensible dans l'étude de Cann *et al.* (1987) contribue à rassembler des types très différenciés dans les échantillons dont il est difficile de saisir les liens généalogiques lors de tentatives de reconstructions phylogéniques (Figure 4.35). Elle a aussi pour conséquence une augmentation apparente de la diversité moléculaire à l'intérieur de l'échantillon. Le nombre des échantillons analysés constitue également un facteur important pour une bonne interprétation des données. En effet, l'ADN-mt a été étudié sur un nombre encore restreint de populations ou même de regroupements de populations, ce qui introduit des discontinuités géographiques et ethniques entre les groupes. Celles-ci sont, bien sûr, répercutées au niveau génétique. De plus, le choix arbitraire et souvent conditionné d'une population ou d'une autre pour représenter un continent peut entraîner des erreurs d'interprétation sur les liens entre groupes continentaux. Nous avons ainsi constaté que le choix d'une population San comme indicateur de la diversité africaine n'était pas forcément très judicieux. (Excoffier *et al.*, 1987), et pouvait contribuer à une marginalisation de l'ensemble des populations africaines. Le nombre restreint de populations typées pour l'ADN-mt est encore réduit du fait de la non-standardisation du jeu d'enzymes de restriction employés pour les différentes études. L'emploi d'un nombre varié d'enzymes différents abouti ainsi à la définition de types non comparables entre les études. Il semblerait bon, qu'à l'avenir, un même ensemble d'enzymes soit utilisé sur toute une série de populations afin d'obtenir des résultats que l'on puisse grouper et comparer.

Une bonne partie de notre travail a aussi consisté à remettre de l'ordre dans la nomenclature des morphes et des types qui avaient parfois été mal définis et qui étaient repris tels quels d'études en études. Ceci a eu des répercussions sur la mesure de la diversité moléculaire des échantillons, de même que sur les topologies des phylogénies. Nous avons, entre autre, pu confirmer la présence de substitutions

multiples et indépendantes sur plusieurs sites de restriction, comme cela avait déjà été constaté au niveau des séquences de l'ADN-mt (Aquadro and Greenberg, 1983). Nous nous sommes aperçus que les substitutions multiples étaient mal prises en compte dans diverses mesures de la diversité moléculaire (Table 3.2) et qu'elles pouvaient fausser le calcul du nombre moyen de différences de sites de restriction inter et intra-population. Elles sont également la source de véritables casse-têtes dans l'établissement des phylogénies des morphes (Figure 4.5 par exemple) et des types (Figure 4.10).

Malgré ces difficultés, nous avons néanmoins pu mettre en évidence certaines propriétés du polymorphisme de l'ADN-mt dans la majorité des échantillons. Tout d'abord, il faut mentionner la formidable ampleur du polymorphisme qui est trouvé au sein des populations. Il semble même parfois exister un certain polymorphisme intra-individuel (Greenberg *et al.*, 1983; Monnat and Loeb, 1985; Monnat and Raey, 1986). Ce haut degré de polymorphisme conduit à une grande diversité des types (Figures 4.11, 4.26, 4.34 et 4.35).

Nous avons proposé une nouvelle méthode pour la détermination d'une racine des phylogénies qui passe par l'identification d'un type ancestral hypothétique. Une fois calculé, ce dernier s'est avéré présenter certaines propriétés générales intéressantes dans la majorité des phylogénies. Il est souvent très fréquent dans les populations (Figures 4.6 et 4.11, Tables 4.15, 4.38, 4.45 et 4.53). Il est toujours à la source d'un grand nombre d'autres types (Figures 4.11, 4.26, 4.30, 4.32 et 4.35). Quand plusieurs échantillons sont analysés simultanément, il se trouve être présent dans toutes les populations (Figure 4.11) ou être à l'origine directe de types trouvés dans des populations appartenant à des groupes continentaux différents (Figure 4.35).

Le temps nécessaire à la création du polymorphisme observé dans les populations à partir d'une population ancestrale monomorphe a été estimé par une quantité ne dépendant pas de fréquences géniques. Celle-ci dépasse très souvent (Excoffier and Langaney, 1989) le temps écoulé depuis l'apparition des premiers hommes modernes (150-100'000 ans en Afrique et au Moyen-Orient) (Stringer and Andrews, 1988; Valladas *et al.*, 1988). Ceci peut être dû au fait que l'âge des gènes peut très souvent excéder l'âge des espèces (Takahata and Nei, 1985) et peut aussi être causé par des échanges de gènes entre les populations au cours de leur histoire. Il faut donc comprendre que ces estimations temporelles ne constituent pas des temps de divergence à partir d'une population mère, mais leur sont habituellement très supérieurs.

La structure radiante des phylogénies et le noyau central de types trouvés dans des groupes continentaux différents suggèrent que les populations humaines actuelles sont les descendantes d'une population déjà polymorphe au moment de l'apparition des premiers hommes modernes il y a plus de 100'000 ans. Cela montre également que les populations actuelles comprennent encore une partie du patrimoine génétique des premiers hommes modernes. La constitution génétique du groupe des populations caucasoïdes semble être la plus proche de celle d'un groupe primitif (*sensu stricto*) au vu de la phylogénie de la Figure 4.11. En effet, ces populations possèdent 9 des 10 types trouvés dans des groupes continentaux différents, contre 7 pour les populations orientales et 5 seulement pour les populations africaines. La sructure radiante de cette phylogénie suggère également que la plus grande partie de la diversification des échantillons africains s'est effectuée à partir des types 2 et 7. Selon les échantillons actuels, il semble donc plus vraisemblable que les populations africaines aient connu un goulot ("*bottleneck*") relativement ancien qui les aurait isolées d'autres groupes et qu'elles se soient ensuite très différenciées. Ceci est aussi perceptible dans les temps de différenciation où ceux des populations caucasoïdes et orientales sont généralement plus élevés que ceux des populations africaines. Les populations caucasoïdes n'ont vraisemblablement pas subi de goulots et se sont diversifiées progressivement à partir des types ancestraux de populations polymorphes génétiquement extra-africaines. En suggérant que les populations caucasoïdes sont les plus proches d'une population ancestrale polymorphe, nous ne prétendons pas que les premiers hommes modernes étaient Caucasoïdes, ni qu'ils sont apparus dans une région occupée actuellement par les Caucasoïdes, car ces 2 hypothèses ne semblent pas testables avec les données actuelles. Nos résultats ne peuvent cependant pas supporter les affirmations de Cann *et al.* (1987) concernant l'existence d'une Eve africaine ayant vécu il y a 200'000 ans, du fait que la généalogie des types sur laquelle est principalement bâtie cette hypothèse s'est avérée contenir de multiples erreurs topologiques (voir la Figure 4.35).

Les conclusions concernant les rapports entre groupes continentaux sont principalement basées sur des phylogénies de types et leurs présences communes dans certaines populations. Pour ce qui est de l'analyse des apparentements entre populations, celle-ci semblait très prometteuse avec l'utilisation des données du polymorphisme de l'ADN-mt du fait de son taux d'évolution 5 à 10 fois plus rapide que celui de l'ADN nucléaire (Brown *et al.*, 1979, 1982). Nous avons cependant pu montrer que des rapports entre populations basés sur des distances génétiques utilisant les fréquences des types souffraient de plusieurs biais, et ne pouvaient être établis de manière fiable. Hormis les estimations imprécises des fréquences des types dans de

petits échantillons, il apparaît que la plupart des populations présentent une distribution de fréquences géniques non conformes avec les prédictions d'un modèle neutre de populations à l'équilibre. Les simulations des fréquences géniques attendues dans les divers échantillons ont montré qu'un ou plusieurs types étaient trop fréquents, que de trop nombreux autres types n'étaient retrouvés qu'à un seul exemplaire dans l'échantillon et que les types de fréquences intermédiaires étaient sous-représentés. Ces résultats sont en accord avec l'hypothèse de la présence d'allèles désavantagés dans certaines populations. Ceci a pour conséquence de perturber les fréquences des types dans les populations où le mécanisme de sélection agirait. L'ADN-mt ne semble pas être sélectionné dans toutes les populations, car nous constatons que 4 échantillons (San, Bantous, Israéliens Juifs et Papous de Nouvelle-Guinée), sur 17 testés, présentent une distribution de fréquences des types en accord avec le modèle neutraliste des allèles infinis (Table 5.1). Cet accord avec une hypothèse de neutralité sélective a également été constaté pour 3 de ces 4 échantillons lors du calcul de la diversité nucléotidique au moyen du modèle des sites infinis (Tables 4.20 et 4.59) (Excoffier and Langaney, 1989). La sélection possible de l'ADN-mt est également supportée par une analyse des mutations ayant conduit à des gains de sites de restriction. Celle-ci montre que plusieurs types sont phénotypiquement différents du type ancestral hypothétique et présentent des gènes dont la fonctionnalité peut être atteinte. D'autre part, plus de la moitié des substitutions s'étant produites sur les 2 premières positions des codons des gènes de structure ne sont pas retrouvées sur des types présents dans les échantillons. Cela suggère que ces types ont été éliminés des populations ou maintenus à des fréquences très faibles.

S'il semble probable qu'une fraction des types soient sélectivement désavantagés, une autre hypothèse a été envisagée pour rendre compte de l'apparente sélection de l'ADN-mt. En effet, une récente et forte augmentation de la taille des populations peut conduire à une augmentation de l'indice statistique utilisé pour tester l'hypothèse de neutralité des populations supposées à l'équilibre. Dans ce cas, ce serait l'équilibre de certaines populations qui serait perturbé et non la neutralité de l'ADN-mt. Cependant, il nous semble que les 2 facteurs ont pu intervenir simultanément, étant donné qu'un seul paramètre, lié à la taille effective des populations, est à la source des 2 déséquilibres. Ainsi, un mutant se comportera de manière neutre dans une petite population, alors qu'il sera sélectionné dans une grande (Nei, 1987, p.411). Des facteurs démographiques différentiels pourraient donc être à la source de fluctuations de fréquences géniques non-aléatoires entre les populations. Ceci a pour conséquence de rendre hasardeux l'emploi de distances génétiques basées sur ces fréquences afin de comparer des populations. Il n'est donc pas étonnant que l'utilisation de telles distances

montre que les populations africaines neutres et/ou en équilibre soient très divergentes (Johnson *et al.*, 1983) d'autres populations dont la diversité moléculaire est artificiellement réduite et où certains types se comportent comme s'ils étaient avantagés. Il s'ensuit que l'utilisation des fréquences du polymorphisme de l'ADN-mt à des fins anthropologiques doit être sérieusement révisée, que l'ADN-mt soit sélectionné ou non, étant donné le comportement différent et *a priori* imprévisible de certaines populations. De nouveaux indices de distance entre populations basés sur les présences-absences des allèles, les allèles communs, et la connaissance des processus généalogiques de la transmission (phylogénies) de ces allèles devraient être développés pour l'étude de systèmes très informatifs, comme l'ADN-mt. Certains travaux récents allant dans ce sens méritent d'être cités (Griffiths, 1979a; Padmadisastra, 1987; Watterson, 1985; Tavaré, 1984) avant de clore ce travail qui reste préliminaire et qui aura contribué, je l'espère, à une meilleure compréhension de quelques mécanismes en jeu dans l'origine, le maintien et la répartition du polymorphisme de l'ADN-mt.

ANNEXE A

THÉORIE DE L'ÉCHANTILLONNAGE DES ALLÈLES (HAPLOTYPES) NEUTRES DANS UNE POPULATION FINIE.

Comme la plupart des études génétiques sur l'histoire du peuplement se basent sur des fréquences géniques estimées à partir d'échantillons, il nous a semblé judicieux de présenter certaines théories, développées en génétique des populations, qui concernent l'inférence de paramètres de la population à partir des seuls échantillons. Ces théories s'appliquent relativement bien à l'échantillonage d'haplotypes de restriction définis par les études de polymorphisme de longueur des fragments de restriction (PLFR) dont il est question tout au long de ce travail.

Nous allons tout d'abord définir des notations pour certaines quantités qui seront fréquemment utilisées dans cette Annexe.

N_e : Taille effective de la population (normalement inconnue)

u : Taux de mutation vers de nouveaux allèles (normalement inconnu)

n : Nombre d'individus dans l'échantillon (pris à la génération t)

K : Nombre d'allèles dans la population à la génération t (une variable aléatoire inconnue)

k : Nombre d'allèles différents observés dans l'échantillon

$\theta = 4 N_e u$ ($\theta = N_e u$ dans le cas des haplotypes de restriction des mitochondries (Birky, Maruyama and Fuerst, 1983))

P(i) : Probabilité que le nombre d'allèles observé dans l'échantillon soit i ($i = 1, 2, 3, \ldots, 2n$ (n pour l'ADN mitochondrial))

$x_{(j)} = x(x+1)(x+2) \ldots (x+j-1)$

$x_{[j]} = x(x-1)(x-2) \ldots (x-j+1)$

$A_{(i)}$: Types alléliques

z : Vecteur des fréquences $z_{(i)}$ des types alléliques $A_{(i)}$

POPULATION STATIONNAIRE

Le modèle développé par Ewens (1972) se base sur le modèle de neutralité stricte de Kimura et Crow (1964) qui généralise le modèle de Wright-Fisher. Il considère une population de N individus diploïdes (2N gènes) avec une infinité d'allèles (A_i) pouvant être créés par des mutations se produisant à un taux constant u au locus A en question. Le type mutant est supposé être d'un type entièrement nouveau, qui n'a jamais existé dans la population et il n'existe pas de différences sélectives entre les allèles. Le processus est censé avoir duré suffisamment longtemps ($t \to \infty$) pour que la population soit en état stationnaire (équilibre mutation-dérive). Le modèle devient réaliste quand les allèles sont définis avec suffisamment de précision au niveau moléculaire, comme les haplotypes de restriction par exemple.

On assume donc qu'à la génération t, il y a X_i gènes d'un certain type allélique A_i. La probabilité qu'à la génération t+1, il y ait Y_i gènes de type allélique A_i et Y_0 nouveaux mutants est

$$P(Y_0, Y_1, Y_2, \ldots \mid X_1, X_2, \ldots) = (2N)!\, \Pi\, P(i)^{Yi} / (\Pi\, (Y_i)!) \qquad (A.1)$$

où $P(0) = u$ et $P(i) = X_i(1-u)/(2N)$

Comme tous les allèles vont disparaître de la population tôt ou tard, il n'y a pas de distribution stationnaire simple pour les fréquences des allèles, étant donné le nombre très élevé de configurations alléliques possibles. La démarche mathématique exacte peut être trouvée dans Ewens (1979). Nous ne développerons ici que quelques concepts simples utiles à la suite de notre discours.

Intéressons nous d'abord à la probabilité (H_2) que deux gènes tirés au hasard soient du même type allélique. Ainsi, ils doivent, soit provenir du même gène parental, soit de différents gènes parentaux du même type allélique. Il ne peut s'agir ici d'un gène ancien et d'un nouveau mutant, puisque par convention, tous les nouveaux mutants n'ont encore jamais existé. Nous avons donc :

$$H_2^{(t+1)} = (1-u)^2 [(2N)^{-1} + (1-(2N)^{-1})H_2^{(t)}] \qquad (A.2)$$

à l'équilibre, $H_2^{(t+1)} = H_2^{(t)} = H_2$, d'où

$$H_2 = [1 - 2N + 2N(1-u)^2]^{-1} \quad (A.3)$$

$$H_2 \approx (1+\theta)^{-1} \quad (A.4)$$

On notera que H_2 est aussi la probabilité que deux gènes soient identiques par ascendance ou encore la probabilité d'homozygosité. D'une manière similaire, on trouve la probabilité que 3 gènes tirés au hasard soient identiques, soit :

$$H_3^{(t+1)} = (1-u)^3(2N)^{-2}[1+3(2N-1)H_2^{(t)} + (2N-1)(2N-2)H_3^{(t)}] \quad (A.5)$$

A l'équilibre,

$$H_3 \approx 2(2+\theta)^{-1}H_2 \approx 2!/[(1+\theta)(2+\theta)] \quad (A.6)$$

En généralisant pour de faibles valeurs de i, Watterson (1979) montre que,

$$H_i \approx (i-1)!/[(1+\theta)(2+\theta) \ldots (i-1+\theta)] \quad (A.7)$$

H_i représente la probabilité qu'un échantillon de i gènes ne contienne qu'un type allélique. C'est donc la probabilité de la configuration allélique {i}. Le même raisonnement conduit à trouver la probabilité d'observer une configuration {i-1, 1} comme étant égale à la probabilité que dans un échantillon de i gènes, les premiers i-1 gènes soient d'un type allélique et le dernier d'un autre type, multiplié par le nombre de permutations possibles d'un gènes parmi i, soit pour $i \geq 3$,

$$P(i-1, 1) = (H_{i-1} - H_i).i \approx i(i-2)!\theta/[(1+\theta)(2+\theta) \ldots (i-1+\theta)] \quad (A.8)$$

De la même manière, on peut définir les probabilités d'obtenir toutes les configurations de r gènes dans un échantillon. Karlin et McGregor (1972), s'inspirant des travaux de Ewens (1972), ont trouvé, sous ce modèle, la distribution de probabilités exacte de n'importe quelle configuration de k allèles dans un échantillon de r gènes ($r << N$) :

$$H(r_1, r_2, \ldots, r_k; k) = \frac{r!\,\theta^k}{1^{\alpha_1}.2^{\alpha_2}.\ldots\ldots r^{\alpha_r}.\alpha_1!\alpha_2!\ldots\ldots\alpha_r!\,S_r(\theta)} \quad (A.9)$$

où $r_1+r_2+\ldots.+r_k = r$, α_i représente le nombre de valeurs égales à i dans l'ensemble $(r_1, r_2, \ldots, r_k)$, et

$$S_r(\theta) = \prod_{i=0}^{r-1} (\theta + i)$$

$H(r_1, r_2, ..., r_k; k)$ représente donc la probabilité qu'un échantillon contienne k types alléliques différents avec r_1 gènes d'un certain type allélique, r_2 gènes d'un autre type, etc...

Par une sommation appropriée de (A.9), on peut obtenir la distribution de probabilité de la variable aléatoire k comme,

$$P(\text{k types alléliques dans l'échantillon}) = | S_r^k | . \theta^k / S_r(\theta) \qquad (A.10)$$

où $| S_r^k |$ représente le coefficient de θ^k dans $S_r(\theta)$ qui est défini comme un nombre de Stirling du premier genre (voir Abramovitz and Stegun, 1970).

En combinant (A.9) et (A.10), on obtient la probabilité d'une certaine configuration allélique en sachant qu'il existe k allèles dans un échantillon,

$$P(r_1, r_2, \ldots, r_k \mid k) = \frac{r!}{| S_r^k | \, 1^{\alpha_1}.2^{\alpha_2} \ldots\ldots r^{\alpha_r}.\alpha_1!\alpha_2! \ldots\ldots \alpha_r!} \qquad (A.11)$$

Cette probabilité a la remarquable propriété d'être indépendante de facteurs inconnus tels que la taille de la population d'où est tiré l'échantillon ainsi que le taux de mutation du gène considéré.

De (A.10), Ewens obtient l'espérance et la variance du nombre d'allèles dans un échantillon de r gènes comme

$$E_s(k;r) = \theta \sum_{i=0}^{r-1} 1/(\theta + i) \qquad (A.12)$$

$$V_s(k;r) = E_s(k;r) - \theta^2 \sum_{i=0}^{r-1} 1/(\theta + i)^2 \qquad (A.13)$$

Par extension, la valeur attendue du nombre d'allèles dans la population s'obtient en remplaçant r par le nombre de gènes de la population (2N pour les gènes nucléaires ou N pour les gènes mitochondriaux), si celui ci est connu ou estimé. On

notera également que $E_s(k;r)$ tend asymptotiquement vers $\theta . \log_e(r)$ lorsque la taille de l'échantillon $r \rightarrow \infty$.

La formule (A.12) nous permet de déduire le paramètre de mutation θ de la population. Ewens (1972) a en effet démontré que l'estimateur du maximum de vraisemblance de θ était la valeur pour laquelle le nombre attendu d'allèles égalait la valeur observée du nombre d'allèles dans l'échantillon. Il a aussi montré que θ était le seul paramètre concernant la population pouvant être estimé à partir de l'échantillon et que k était un *paramètre suffisant* pour estimer θ. Cette dernière affirmation revient à dire que la distribution des nombres $r_1, r_2, \ldots, r_k$ est *indépendante* de θ et que leur utilisation dans l'estimation de θ biaiserait la valeur de l'estimateur.

POPULATION NON-STATIONNAIRE

Contrairement au modèle discret de Ewens (1972), dérivé pour des populations stationnaires, Griffiths (1979a) a développé un modèle de diffusion permettant d'estimer les valeurs du nombre d'allèles dans des échantillons tirés de populations non-stationnaires. Ce nombre d'allèles dépendra donc ici des valeurs des fréquences alléliques de la population initiale.

Il montre que dans une population suivant un modèle avec une infinité d'allèles, un paramètre de mutation θ et des fréquences alléliques initiales $z_{(1)}, z_{(2)}, \ldots$, le nombre attendu de types alléliques k dans un échantillon de r gènes est

$$E(k \mid \mathbf{z}) = E_s(k;r) + \sum_{n=2}^{r} \rho_n(t)\, r_{[n]}[(\theta+r)_{(n)}]^{-1}(\theta+2n-1)(n!)^{-1}C_n(\mathbf{z}), \qquad (A.14)$$

où,

$$C_n(\mathbf{z}) = \sum_{m=2}^{n} (-1)^{n-m} C_m^n (\theta+m)_{(n-1)}[k_m(\mathbf{z})-E_s(k;m)],$$

$$\rho_n(t) = \exp\{-\tfrac{1}{2}n(n-1)t-\tfrac{1}{2}\theta nt\}$$

$E_s(k;m)$ est le nombre de types alléliques attendu dans un échantillon de taille m tiré d'une population stationnaire et $k_m(\mathbf{z}) = \Sigma_i [1-(1-z_{(i)})^m]$.

La distribution des fréquences de la population au temps t_0 sont rarement connues, mais il est clair qu'elles doivent être comprises entre deux situations extrêmes qui sont, d'une part, la fixation d'un seul allèle dans la population, et d'autre part une fréquence égale pour tous les allèles. Pour toutes valeurs de θ, la valeur de $E_s(k;r)$ sera comprise entre les valeurs limites de $E(k \mid \mathbf{z})$ fixées par les deux distributions extrêmes (voir Table A.1). En utilisant (A.14), il a été possible de calculer précisément $E(k \mid \mathbf{z})$ pour les deux distributions de fréquences dans la population initiale que nous venons de décrire, et ceci pour différentes valeurs de θ et de t. On s'aperçoit que la non-stationarité de l'échantillon a peu d'influence sur la valeur du nombre de types attendu dès que $t > 1$ et lorsque $\theta > 1$ (ce qui est souvent le cas pour les polymorphismes moléculaires). Si l'on admet que les distributions de fréquences sont rarement extrêmes (fixation ou fréquences uniformes), et que les échantillons étudiés ont un effectif raisonnable ($\approx$ 100 gènes), il semble justifié d'utiliser l'équation (A.12) pour estimer θ. Lorsque la taille de l'échantillon $r \to \infty$, $E(k \mid \mathbf{z}) \approx \theta \log_e(r)$ pour n'importe quelle valeur de t et de $\mathbf{z}$. Il est à noter que $E(k \mid \mathbf{z})$ tend asymptotiquement vers la même valeur que $E_s(k;r)$. Les grandes populations vont donc vraisemblablement adopter rapidement un comportement stationnaire du point de vue du nombre d'allèles présents.

TABLE A.1 : Nombre attendu d'allèles k dans un échantillon tiré d'une population non-stationnaire au temps *t* (en unités de 2N générations)[1]

		t = 0,10		0,20		0,50		1,00		1,50		2,00		5,00		∞
Teta	**Taille**[2]															
0,10	10	1,06	6,92	1,08	5,34	1,13	3,29	1,19	2,18	1,22	1,75	1,24	1,54	1,24	1,54	1,27
	50	1,19	14,69	1,21	8,80	1,27	4,27	1,34	2,58	1,38	2,02	1,40	1,75	1,40	1,75	1,43
	100	1,25	17,13	1,28	9,62	1,34	4,48	1,41	2,68	1,45	2,10	1,47	1,83	1,47	1,83	1,50
	150	1,29	18,15	1,32	9,94	1,38	4,57	1,45	2,73	1,49	2,15	1,51	1,88	1,51	1,88	1,54
	200	1,32	18,71	1,34	10,12	1,41	4,62	1,48	2,77	1,52	2,18	1,54	1,91	1,54	1,91	1,57
	250	1,34	19,06	1,37	10,23	1,43	4,66	1,50	2,79	1,54	2,20	1,56	1,93	1,56	1,93	1,59
	300	1,36	19,31	1,38	10,31	1,45	4,68	1,52	2,82	1,56	2,22	1,58	1,95	1,58	1,95	1,61
	350	1,37	19,50	1,40	10,37	1,46	4,71	1,53	2,83	1,57	2,24	1,60	1,96	1,60	1,96	1,63
	400	1,39	19,64	1,41	10,42	1,47	4,73	1,54	2,85	1,59	2,25	1,61	1,98	1,61	1,98	1,64
	450	1,40	19,75	1,42	10,45	1,49	4,74	1,56	2,86	1,60	2,26	1,62	1,99	1,62	1,99	1,65
	500	1,41	19,84	1,43	10,49	1,50	4,76	1,57	2,87	1,61	2,27	1,63	2,00	1,63	2,00	1,66
0,50	10	1,20	6,97	1,35	5,46	1,64	3,59	1,89	2,68	2,02	2,37	2,08	2,24	2,08	2,24	2,13
	50	1,63	15,01	1,90	9,32	2,31	5,10	2,64	3,66	2,79	3,24	2,87	3,08	2,87	3,08	2,94
	100	1,90	17,65	2,20	10,38	2,64	5,57	2,98	4,03	3,14	3,60	3,21	3,43	3,21	3,43	3,28
	150	2,07	18,80	2,39	10,85	2,84	5,82	3,18	4,25	3,34	3,81	3,42	3,63	3,42	3,63	3,49
	200	2,20	19,47	2,53	11,13	2,98	5,98	3,32	4,39	3,48	3,95	3,56	3,78	3,56	3,78	3,63
	250	2,31	19,91	2,63	11,33	3,09	6,11	3,43	4,51	3,59	4,06	3,67	3,89	3,67	3,89	3,74
	300	2,39	20,22	2,72	11,48	3,18	6,21	3,52	4,60	3,68	4,16	3,76	3,98	3,76	3,98	3,83
	350	2,46	20,46	2,80	11,60	3,25	6,29	3,60	4,68	3,76	4,23	3,84	4,06	3,84	4,06	3,91
	400	2,53	20,65	2,86	11,70	3,32	6,36	3,66	4,75	3,83	4,30	3,91	4,12	3,91	4,12	3,98
	450	2,58	20,81	2,92	11,78	3,38	6,43	3,72	4,81	3,88	4,36	3,96	4,18	3,96	4,18	4,04
	500	2,63	20,94	2,97	11,86	3,43	6,48	3,77	4,86	3,94	4,41	4,02	4,24	4,02	4,24	4,09
1,00	10	1,40	7,03	1,69	5,60	2,23	3,95	2,66	3,23	2,83	3,03	2,89	2,97	2,89	2,97	2,93
	50	2,25	15,40	2,78	9,96	3,57	6,10	4,14	4,93	4,36	4,64	4,45	4,55	4,45	4,55	4,50
	100	2,78	18,30	3,39	11,31	4,22	6,89	4,81	5,63	5,05	5,34	5,13	5,24	5,13	5,24	5,19
	150	3,13	19,62	3,76	11,97	4,61	7,34	5,21	6,04	5,45	5,74	5,54	5,65	5,54	5,65	5,59
	200	3,39	20,41	4,03	12,39	4,89	7,64	5,49	6,33	5,73	6,03	5,82	5,93	5,82	5,93	5,88
	250	3,59	20,95	4,24	12,69	5,11	7,88	5,71	6,56	5,96	6,25	6,05	6,16	6,05	6,16	6,10
	300	3,77	21,35	4,42	12,93	5,29	8,07	5,90	6,74	6,14	6,44	6,23	6,34	6,23	6,34	6,28
	350	3,91	21,66	4,57	13,12	5,44	8,23	6,05	6,90	6,29	6,59	6,38	6,49	6,38	6,49	6,44
	400	4,04	21,92	4,70	13,28	5,57	8,36	6,18	7,03	6,42	6,72	6,52	6,62	6,52	6,62	6,57
	450	4,15	22,13	4,81	13,43	5,69	8,48	6,30	7,15	6,54	6,84	6,63	6,74	6,63	6,74	6,69
	500	4,25	22,31	4,92	13,55	5,79	8,59	6,40	7,26	6,65	6,95	6,74	6,85	6,74	6,85	6,79
1,50	10	1,60	7,08	2,01	5,74	2,77	4,27	3,31	3,72	3,47	3,59	3,52	3,56	3,52	3,56	3,54
	50	2,86	15,78	3,64	10,57	4,76	7,06	5,50	6,10	5,74	5,91	5,81	5,86	5,81	5,86	5,84
	100	3,66	18,93	4,54	12,22	5,74	8,17	6,51	7,14	6,76	6,94	6,84	6,89	6,84	6,89	6,87
	150	4,19	20,42	5,10	13,06	6,32	8,81	7,11	7,75	7,36	7,54	7,44	7,49	7,44	7,49	7,47
	200	4,57	21,34	5,51	13,62	6,74	9,25	7,54	8,18	7,79	7,97	7,87	7,92	7,87	7,92	7,90
	250	4,88	21,98	5,83	14,03	7,07	9,59	7,87	8,52	8,13	8,31	8,20	8,25	8,20	8,25	8,23
	300	5,13	22,46	6,09	14,35	7,34	9,87	8,14	8,79	8,40	8,58	8,47	8,53	8,47	8,53	8,51
	350	5,35	22,85	6,31	14,62	7,57	10,11	8,37	9,02	8,63	8,81	8,71	8,76	8,71	8,76	8,74
	400	5,54	23,17	6,51	14,85	7,77	10,31	8,57	9,22	8,83	9,01	8,90	8,96	8,90	8,96	8,94
	450	5,71	23,44	6,68	15,05	7,94	10,49	8,74	9,40	9,00	9,19	9,08	9,13	9,08	9,13	9,11
	500	5,86	23,67	6,84	15,22	8,10	10,65	8,90	9,56	9,16	9,35	9,24	9,29	9,24	9,29	9,27

		t = 0,10		t = 0,20		t = 0,50		t = 1,00		t = 1,50		t = 2,00		t = 5,00		t = ∞
2,00	10	1,79	7,14	2,33	5,88	3,27	4,57	3,85	4,14	4,00	4,06	4,03	4,04	4,03	4,04	4,04
	50	3,46	16,16	4,48	11,18	5,90	7,97	6,75	7,20	6,97	7,07	7,02	7,04	7,02	7,04	7,04
	100	4,53	19,56	5,68	13,12	7,20	9,40	8,09	8,56	8,33	8,43	8,38	8,40	8,38	8,40	8,39
	150	5,22	21,22	6,42	14,14	7,98	10,23	8,89	9,37	9,13	9,23	9,18	9,20	9,18	9,20	9,20
	200	5,74	22,26	6,96	14,83	8,54	10,82	9,46	9,94	9,69	9,80	9,75	9,77	9,75	9,77	9,77
	250	6,14	23,00	7,39	15,35	8,98	11,27	9,90	10,39	10,14	10,25	10,19	10,22	10,19	10,22	10,21
	300	6,48	23,57	7,74	15,76	9,34	11,64	10,26	10,75	10,50	10,61	10,56	10,58	10,56	10,58	10,57
	350	6,77	24,03	8,04	16,10	9,64	11,95	10,57	11,06	10,81	10,92	10,86	10,89	10,86	10,89	10,88
	400	7,02	24,41	8,30	16,40	9,90	12,22	10,83	11,32	11,07	11,18	11,13	11,15	11,13	11,15	11,14
	450	7,25	24,74	8,53	16,65	10,14	12,45	11,07	11,56	11,31	11,42	11,36	11,39	11,36	11,39	11,38
	500	7,45	25,02	8,73	16,88	10,34	12,67	11,28	11,77	11,52	11,63	11,57	11,60	11,57	11,60	11,59
2,50	10	1,98	7,19	2,63	6,00	3,72	4,85	4,32	4,51	4,43	4,46	4,45	4,46	4,45	4,46	4,46
	50	4,06	16,53	5,30	11,77	6,99	8,84	7,90	8,22	8,08	8,14	8,11	8,12	8,11	8,12	8,12
	100	5,39	20,18	6,79	14,00	8,61	10,59	9,57	9,91	9,76	9,82	9,80	9,81	9,80	9,81	9,80
	150	6,25	22,01	7,72	15,20	9,58	11,61	10,56	10,91	10,76	10,82	10,79	10,80	10,79	10,80	10,80
	200	6,89	23,18	8,40	16,02	10,28	12,34	11,27	11,62	11,47	11,53	11,50	11,52	11,50	11,52	11,51
	250	7,40	24,02	8,93	16,65	10,82	12,90	11,82	12,18	12,02	12,08	12,06	12,07	12,06	12,07	12,07
	300	7,83	24,67	9,37	17,15	11,27	13,35	12,27	12,63	12,47	12,54	12,51	12,52	12,51	12,52	12,52
	350	8,19	25,20	9,74	17,56	11,65	13,74	12,65	13,01	12,86	12,92	12,89	12,90	12,89	12,90	12,90
	400	8,50	25,64	10,06	17,92	11,98	14,07	12,98	13,35	13,19	13,25	13,23	13,24	13,23	13,24	13,23
	450	8,79	26,03	10,35	18,23	12,27	14,37	13,27	13,64	13,48	13,54	13,52	13,53	13,52	13,53	13,53
	500	9,04	26,36	10,61	18,51	12,53	14,63	13,54	13,90	13,74	13,81	13,78	13,79	13,78	13,79	13,79
3,00	10	2,16	7,25	2,92	6,13	4,14	5,10	4,71	4,84	4,80	4,81	4,81	4,81	4,81	4,81	4,81
	50	4,64	16,89	6,10	12,34	8,03	9,68	8,94	9,18	9,09	9,12	9,11	9,12	9,11	9,12	9,11
	100	6,23	20,79	7,89	14,86	9,96	11,74	10,94	11,19	11,10	11,13	11,12	11,12	11,12	11,12	11,12
	150	7,27	22,79	9,00	16,25	11,12	12,95	12,12	12,38	12,29	12,32	12,31	12,31	12,31	12,31	12,31
	200	8,04	24,08	9,81	17,20	11,96	13,81	12,97	13,23	13,14	13,17	13,16	13,17	13,16	13,17	13,16
	250	8,65	25,03	10,45	17,93	12,62	14,48	13,63	13,90	13,80	13,83	13,82	13,83	13,82	13,83	13,83
	300	9,16	25,76	10,97	18,51	13,15	15,03	14,17	14,44	14,34	14,38	14,36	14,37	14,36	14,37	14,37
	350	9,59	26,36	11,42	19,01	13,61	15,49	14,63	14,90	14,80	14,84	14,82	14,83	14,82	14,83	14,83
	400	9,97	26,87	11,81	19,43	14,00	15,89	15,03	15,30	15,20	15,23	15,22	15,23	15,22	15,23	15,22
	450	10,31	27,31	12,15	19,80	14,35	16,24	15,38	15,65	15,55	15,59	15,57	15,58	15,57	15,58	15,58
	500	10,61	27,69	12,46	20,13	14,66	16,56	15,70	15,96	15,86	15,90	15,89	15,89	15,89	15,89	15,89
3,50	10	2,35	7,30	3,20	6,24	4,52	5,34	5,05	5,14	5,11	5,12	5,12	5,12	5,12	5,12	5,12
	50	5,22	17,25	6,88	12,90	9,01	10,47	9,91	10,07	10,02	10,04	10,03	10,04	10,03	10,04	10,04
	100	7,07	21,40	8,96	15,70	11,26	12,85	12,22	12,40	12,35	12,36	12,36	12,36	12,36	12,36	12,36
	150	8,28	23,56	10,25	17,27	12,62	14,25	13,60	13,79	13,73	13,75	13,74	13,75	13,74	13,75	13,75
	200	9,17	24,98	11,20	18,36	13,59	15,25	14,59	14,78	14,72	14,74	14,73	14,74	14,73	14,74	14,74
	250	9,89	26,02	11,94	19,19	14,35	16,02	15,36	15,55	15,49	15,51	15,50	15,51	15,50	15,51	15,51
	300	10,48	26,84	12,55	19,87	14,98	16,66	15,99	16,18	16,12	16,14	16,14	16,14	16,14	16,14	16,14
	350	10,98	27,51	13,07	20,43	15,51	17,20	16,53	16,72	16,66	16,68	16,67	16,67	16,67	16,67	16,67
	400	11,42	28,08	13,53	20,92	15,97	17,66	16,99	17,18	17,12	17,14	17,13	17,14	17,13	17,14	17,14
	450	11,82	28,58	13,93	21,34	16,38	18,07	17,40	17,59	17,53	17,55	17,54	17,55	17,54	17,55	17,54
	500	12,17	29,01	14,29	21,73	16,74	18,44	17,76	17,96	17,90	17,92	17,91	17,91	17,91	17,91	17,91
4,00	10	2,52	7,35	3,47	6,35	4,86	5,55	5,34	5,40	5,38	5,39	5,39	5,39	5,39	5,39	5,39
	50	5,78	17,60	7,64	13,45	9,95	11,24	10,80	10,92	10,89	10,90	10,89	10,89	10,89	10,89	10,89
	100	7,90	21,99	10,01	16,53	12,51	13,92	13,43	13,56	13,53	13,54	13,53	13,53	13,53	13,53	13,53
	150	9,28	24,32	11,49	18,28	14,06	15,51	15,01	15,14	15,10	15,11	15,11	15,11	15,11	15,11	15,11
	200	10,30	25,87	12,56	19,50	15,17	16,64	16,13	16,27	16,23	16,24	16,24	16,24	16,24	16,24	16,24
	250	11,11	27,01	13,41	20,44	16,04	17,53	17,01	17,14	17,11	17,12	17,12	17,12	17,12	17,12	17,12
	300	11,78	27,91	14,11	21,20	16,75	18,25	17,73	17,87	17,83	17,84	17,84	17,84	17,84	17,84	17,84
	350	12,36	28,66	14,70	21,84	17,36	18,86	18,34	18,48	18,44	18,45	18,45	18,45	18,45	18,45	18,45
	400	12,87	29,29	15,22	22,39	17,88	19,39	18,87	19,00	18,97	18,98	18,98	18,98	18,98	18,98	18,98
	450	13,31	29,84	15,68	22,87	18,35	19,86	19,34	19,47	19,43	19,45	19,44	19,44	19,44	19,44	19,44
	500	13,72	30,32	16,09	23,30	18,77	20,28	19,75	19,89	19,85	19,86	19,86	19,86	19,86	19,86	19,86

		t													
		0,10		0,20		0,50		1,00		1,50		2,00		5,00	∞
4,50	10	2,69	7,40	3,73	6,46	5,17	5,75	5,59	5,63	5,62	5,63	5,63	5,63	5,63	5,63
	50	6,34	17,95	8,38	13,98	10,84	11,97	11,64	11,72	11,70	11,70	11,70	11,70	11,70	11,70
	100	8,72	22,58	11,04	17,34	13,71	14,95	14,58	14,67	14,65	14,65	14,65	14,65	14,65	14,65
	150	10,26	25,08	12,70	19,27	15,45	16,73	16,34	16,43	16,41	16,42	16,42	16,42	16,42	16,42
	200	11,41	26,75	13,91	20,63	16,70	18,01	17,61	17,70	17,68	17,68	17,68	17,68	17,68	17,68
	250	12,32	27,99	14,86	21,67	17,67	19,00	18,59	18,69	18,66	18,67	18,67	18,67	18,67	18,67
	300	13,08	28,98	15,65	22,51	18,48	19,81	19,40	19,49	19,47	19,48	19,48	19,48	19,48	19,48
	350	13,73	29,79	16,31	23,22	19,16	20,49	20,08	20,18	20,16	20,16	20,16	20,16	20,16	20,16
	400	14,30	30,48	16,90	23,84	19,75	21,09	20,68	20,77	20,75	20,76	20,76	20,76	20,76	20,76
	450	14,80	31,09	17,41	24,38	20,27	21,62	21,20	21,30	21,28	21,28	21,28	21,28	21,28	21,28
	500	15,25	31,62	17,87	24,86	20,74	22,09	21,67	21,77	21,75	21,75	21,75	21,75	21,75	21,75
5,00	10	2,86	7,44	3,98	6,56	5,45	5,94	5,82	5,85	5,84	5,84	5,84	5,84	5,84	5,84
	50	6,90	18,29	9,11	14,50	11,69	12,67	12,42	12,47	12,46	12,46	12,46	12,46	12,46	12,46
	100	9,52	23,17	12,05	18,14	14,86	15,96	15,66	15,73	15,71	15,72	15,72	15,72	15,72	15,72
	150	11,24	25,82	13,89	20,25	16,79	17,92	17,62	17,68	17,67	17,67	17,67	17,67	17,67	17,67
	200	12,51	27,62	15,23	21,73	18,17	19,33	19,02	19,08	19,07	19,07	19,07	19,07	19,07	19,07
	250	13,52	28,96	16,29	22,88	19,26	20,43	20,11	20,18	20,16	20,17	20,17	20,17	20,17	20,17
	300	14,37	30,03	17,16	23,81	20,15	21,33	21,01	21,07	21,06	21,06	21,06	21,06	21,06	21,06
	350	15,09	30,92	17,90	24,60	20,91	22,09	21,77	21,83	21,82	21,82	21,82	21,82	21,82	21,82
	400	15,72	31,67	18,55	25,27	21,56	22,75	22,43	22,49	22,48	22,48	22,48	22,48	22,48	22,48
	450	16,28	32,33	19,12	25,87	22,14	23,33	23,01	23,08	23,06	23,07	23,07	23,07	23,07	23,07
	500	16,78	32,92	19,63	26,41	22,66	23,86	23,53	23,60	23,58	23,59	23,59	23,59	23,59	23,59
5,50	10	3,03	7,49	4,22	6,66	5,70	6,11	6,02	6,04	6,03	6,03	6,03	6,03	6,03	6,03
	50	7,44	18,63	9,82	15,01	12,49	13,35	13,15	13,19	13,18	13,18	13,18	13,18	13,18	13,18
	100	10,32	23,74	13,04	18,92	15,97	16,93	16,70	16,74	16,73	16,74	16,74	16,74	16,74	16,74
	150	12,20	26,56	15,06	21,21	18,08	19,08	18,84	18,88	18,88	18,88	18,88	18,88	18,88	18,88
	200	13,60	28,48	16,53	22,82	19,61	20,62	20,38	20,42	20,41	20,42	20,42	20,42	20,42	20,42
	250	14,71	29,92	17,70	24,07	20,80	21,82	21,58	21,62	21,61	21,62	21,62	21,62	21,62	21,62
	300	15,64	31,07	18,65	25,09	21,78	22,81	22,56	22,61	22,60	22,60	22,60	22,60	22,60	22,60
	350	16,43	32,03	19,47	25,95	22,61	23,65	23,40	23,44	23,43	23,44	23,44	23,44	23,44	23,44
	400	17,13	32,85	20,18	26,69	23,33	24,37	24,12	24,17	24,16	24,16	24,16	24,16	24,16	24,16
	450	17,74	33,57	20,81	27,35	23,97	25,01	24,76	24,81	24,80	24,80	24,80	24,80	24,80	24,80
	500	18,30	34,20	21,37	27,93	24,54	25,59	25,34	25,38	25,37	25,37	25,37	25,37	25,37	25,37
6,00	10	3,19	7,54	4,45	6,76	5,93	6,26	6,20	6,21	6,21	6,21	6,21	6,21	6,21	6,21
	50	7,97	18,96	10,51	15,50	13,25	13,99	13,84	13,87	13,86	13,86	13,86	13,86	13,86	13,86
	100	11,11	24,31	14,01	19,69	17,03	17,87	17,69	17,72	17,71	17,72	17,72	17,72	17,72	17,72
	150	13,16	27,30	16,21	22,15	19,33	20,20	20,02	20,05	20,04	20,04	20,04	20,04	20,04	20,04
	200	14,68	29,33	17,82	23,90	20,99	21,88	21,69	21,72	21,72	21,72	21,72	21,72	21,72	21,72
	250	15,90	30,87	19,08	25,25	22,29	23,19	23,00	23,03	23,02	23,02	23,02	23,02	23,02	23,02
	300	16,91	32,11	20,13	26,36	23,36	24,26	24,07	24,10	24,09	24,10	24,09	24,09	24,09	24,09
	350	17,77	33,14	21,02	27,29	24,26	25,17	24,98	25,01	25,00	25,00	25,00	25,00	25,00	25,00
	400	18,53	34,02	21,79	28,09	25,05	25,96	25,77	25,80	25,79	25,79	25,79	25,79	25,79	25,79
	450	19,20	34,79	22,48	28,81	25,74	26,66	26,46	26,50	26,49	26,49	26,49	26,49	26,49	26,49
	500	19,80	35,47	23,09	29,44	26,37	27,29	27,09	27,12	27,12	27,12	27,12	27,12	27,12	27,12

1 Pour chaque valeur de t, la première colonne représente les valeurs de k dans le cas où la population initiale ne comprend qu'un seul allèle, Pour les valeurs de la seconde colonne, on suppose que tous les allèles présents initialement ont la même fréquence

2 La taille s'exprime en nombre de gènes dans l'échantillon

TEST DE LA NEUTRALITÉ SÉLECTIVE D'UN LOCUS : TEST DE L'HOMOZYGOSITÉ

La théorie neutraliste peut être envisagée sous deux aspects que Ewens (1979a) a défini comme une neutralité *stricte* et une neutralité *généralisée*. L'hypothèse de neutralité généralisée admet l'existence d'allèles désavantageux qui sont créés par mutation en nombre plus important que les allèles neutres (environ 10 fois plus selon une estimation de Kimura, 1983, pp. 209-210). Du fait de leur désavantage sélectif, ces allèles ne se répandraient pas dans la population et leur fréquence dépasserait rarement 10 %. Ainsi, ils contribueraient peu au polymorphisme des populations et encore moins à leur hétérozygosité. L'hypothèse de la neutralité stricte est, quant à elle, conforme au modèle des allèles infinis (Kimura, 1964) évoqué plus haut, où tous les allèles sont censés être sélectivement équivalents. Dans le cadre de cette hypothèse, Watterson (1978) a développé une procédure de test permettant de définir si l'échantillonnage des gènes d'un locus donné était compatible avec l'hypothèse selon laquelle tous les mutants sont sélectivement neutres. Sous cette condition, les gènes d'un échantillon aléatoire devraient posséder des fréquences suivant une distribution de probabilité définie par l'équation (A.9). Si l'on accepte comme hypothèse H_0 la neutralité sélective stricte, il est nécessaire de définir des hypothèses alternatives. H_0 a été opposée à l'hypothèse d'un avantage sélectif aux hétérozygotes, et à celle de la présence d'allèles désavantagés dans la population. Ainsi, pour ces deux schémas, Watterson (1977) a montré que pour des valeurs sélectives faibles, la distribution de probabilité de l'échantillont $\{k; r_1, r_2, ..., r_k\}$ était donné par

$$P(k;r_1, r_2, \ldots, r_k) =$$

$$P(k;r_1, r_2, \ldots, r_k \mid \text{neutralité}) [\, 1 + A\beta + 0(\beta^2) \,], \qquad \text{(B.1)}$$

où

$$A = r\{(1+\theta)^{-1} - r\hat{F}(r+\theta)^{-1}\}\{r+1+\theta\}^{-1} \qquad \text{(B.2)}$$

on remplace β par $2N_e s$ dans le schéma de sélection hétérotique, et par $-2(2N_e s)^2\gamma(1-\gamma)$, ainsi que $0(\beta^2)$ par $0(\beta^3)$ dans le schéma incorporant des allèles délétères (γ représente ici la fraction des allèles désavantagés). Le fait que la quantité F, équivalente à l'hétérozygosité de l'échantillon, définie comme

$$\hat{F} = \sum_{i=1}^{k} (r_i/r)^2, \qquad (B.3)$$

intervienne dans le rapport des probabilités (B.1) et (A.9) a motivé le fait qu'elle soit employée comme statistique permettant de déceler des écarts à la neutralité pour un locus donné (Watterson, 1978). Ici, F n'est pas simplement considérée comme la probabilité qu'un individu soit homozygote, mais plutôt utilisée pour décrire la distribution des fréquences alléliques. Le test va donc porter sur les valeurs de F qui, pour un échantillon donné, vont conduire à rejetter l'hypothèse de neutralité stricte de tous les allèles. F aura tendance à être plus faible que sa valeur sous hypothèse de neutralité dans le schéma de l'avantage des hétérozygotes, et à être plus élevé lors de présence d'allèles délétères. Ceci est perceptible dans la formule suivante de Watterson (1978)

$$E(F|k) = E(F|k, \text{neutralité}) + \beta[2\theta/\{(1+\theta)^2(2+\theta)(3+\theta)\}] + 0(\beta^2, r\text{-}1), \qquad (B.4)$$

valable dans le cas de différences sélectives faibles et où $\beta = -2Ns$ pour le modèle hétérotique, $\beta = 2(2Ns)^2\gamma(1-\gamma)$ et $0(\beta^2)$ est remplacé par $0(\beta^3)$ pour le modèle des allèles délétères.

Afin de déterminer un intervalle de confiance pour F, sa distribution sous l'hypothèse de neutralité peut être trouvée en utilisant (A.11), mais se révèle en pratique assez difficile à calculer. Il est plus adéquat de simuler les fréquences géniques d'une série d'échantillons aléatoires selon une procédure définie par Stewart (Appendice de Fuerst et al., 1977) qu'il convient de décrire ici brièvement.

Les équations (A.10) et (A.11) font usage de nombres de Stirling du premier genre qui sont relativement difficiles à calculer. Une autre quantité B(k, r) égale à

$$B(k, r) = \sum_{r_1, r_2, \ldots, r_k} (r_1 r_2 \ldots r_k)^{-1} \qquad (B.5)$$

où la sommation se fait sur toutes les valeurs possibles de $r_1, r_2, \ldots, r_k$ avec la restriction $r_i \geq 1$ et $\Sigma r_i = r$, est liée à S_r^k (nombre de Stirling du premier genre) par

$$B(k,r) = (-1)^{r-k} \frac{k!}{r!} S_r^k \qquad (B.6)$$

ANNEXE B

En utilisant cette quantité, l'équation (A.10) peut être reformulée comme

$$P(k) = \frac{r!\, B(k, r)}{k!\, S_r(\theta)} \tag{B.7}$$

et l'équation (A.11) peut être réécrite comme

$$P(r_1, r_2, ..., r_k \mid k) = [B(k, r)\, r_1 r_2 ... r_k]^{-1} \tag{B.8}$$

Un échantillon aléatoire de k allèles est obtenu en choisissant un ensemble de r_1, r_2, ..., r_k gènes correspondant respectivement aux allèles A_1, A_2, ..., A_k. La procédure de Stewart débute par la génération d'un nombre aléatoire à partir d'une distribution uniforme sur [0,1] et sa comparaison avec les probabilités cumulées de (B.8). Ceci est facilité lorsque Stewart montre que (B.8) peut être mis sous la forme de

$$P(r_1, r_2, ..., r_m \mid k) = \frac{B(k\text{-}m, r\text{-}r_1\text{-}...\text{-}r_m)}{B(k\text{-}m+1, r\text{-}r_1\text{-}...\text{-}r_{m-1}) r_m} \tag{B.9}$$

avec $m \leq k$. Donc, pour un seul allèle r_1,

$$P(r_1) = \frac{B(k\text{-}1, r\text{-}r_1)}{B(k, r)\, r_1} \tag{B.10}$$

Stewart poursuit en indiquant que

$$P(r_m \mid r_1, r_2, ..., r_{m-1}) = \frac{B(k\text{-}m, r\text{-}r_1\text{-}...\text{-}r_m)}{B(k\text{-}m+1, r\text{-}r_1\text{-}...\text{-}r_{m-1})\, r_m} \tag{B.11}$$

On commencera donc par trouver une valeur aléatoire de r_1 en générant un nombre aléatoire A sur [0,1] et en le comparant avec les probabilités cumulées de (B.10). P(1), P(2), ..., P(r) sont générés jusqu'à ce que $\Sigma^r_{i=1} P(i) \geq A$. Lorsque cette inégalité est vérifiée, r_1 prend la valeur i. Les valeurs de r_2, r_3, ..., r_k sont ensuite déterminées en recommençant la même procédure, mais en utilisant (B.11) à la place de (B.10). Les valeurs de B(i, j) sont calculées une fois pour toutes pour i = 1 à k et j = i à r en utilisant la formule de récurrence suivante

$$B(i, j+1) = [i\, B(i\text{-}1, j) + j\, B(i, j)]/(j + 1) \tag{B.12}$$

avec B(0, i) = 0 et B(i, i) = 1, pour $i \geq 1$.

Les programmes permettant le calcul des nombres B(i, j) et la simulation des fréquences géniques d'un échantillon aléatoire sont inclus à la fin de cet Annexe B. Il est

important de rappeller que l'équation (B.8) est indépendante du taux de mutation au locus considéré, ainsi que de la taille de la population dont est tiré l'échantillon. La procédure d'échantillonnage s'applique ainsi à tout locus neutre dans n'importe quelle population et elle est uniquement conditionnée par le nombre d'allèles recensés et la taille de l'échantillon. Le programme SIMFREQ développé dans le cadre de cette thèse permet de simuler 1000 échantillons aléatoires de k allèles et r gènes. Il calcule également les fréquences géniques moyennes des allèles A_1, ..., A_k classés par ordre de fréquence décroissante, ce qui permet de comparer visuellement ou par un test statistique les distributions des fréquences observées pour un échantillon quelconque et les fréquences attendues sous l'hypothèse neutraliste. Pour chaque simulation, une valeur de F est calculée selon (B.3). La valeur de E(F|k) est estimée en prenant la moyenne des F pour les 1000 simulations. Les niveaux de sigification des valeurs de F sont déterminées en calculant la probabilité d'observer un échantillon simulé possédant une valeur de F plus faible. L'intervalle de confiance empirique de E(F|k) pour un niveau de confiance α donné peut être déterminé pour un test bilatéral, après classement des 1000 valeurs de F, en trouvant les valeurs de F délimitant un intervalle d'où sont exclues les $(\alpha/2).1000$ valeurs les plus hautes et les $(\alpha/2).1000$ valeurs les plus basses obtenues par la procédure de simulation. Pratiquement, les procédures de test seront unilatérales en raison des hypothèses alternatives à la neutralité stricte.

TABLE B.1 : Valeurs de E(F|k) pour différentes tailles d'échantillons. Les valeurs en italique sont tirées de l'Appendice D de Ewens (1979b).

	k							
r	3	5	7	10	15	20	25	30
50	0,633	0,438	0,327	0,226	0,138	0,092	0,066	0,050
100	0,669 *0,671*	0,488 *0,490*	0,373 *0,376*	0,274 *0,271*	0,175 *0,176*	0,126 *0,125*	0,094 *0,094*	0,074 *0,073*
200	0,709 *0,705*	0,536 *0,532*	0,423 *0,421*	0,316 *0,313*	0,213 *0,212*	0,154 *0,156*	0,121 *0,120*	0,096 *0,096*
300	0,722 *0,722*	0,554 *0,554*	0,451 *0,444*	0,336 *0,336*	0,229 *0,232*	0,175 *0,173*	0,135 *0,135*	0,112 *0,110*
400	0,734 *0,732*	0,561 *0,568*	0,463 *0,459*	0,350 *0,351*	0,247 *0,245*	0,186 *0,185*	0,146 *0,146*	0,120 *0,119*
500	0,745 *0,740*	0,578 *0,579*	0,469 *0,470*	0,361 *0,362*	0,257 *0,255*	0,197 *0,193*	0,155 *0,153*	0,126 *0,126*
1000	0,769	0,610	0,497	0,392	0,283	0,222	0,177	0,145
2000	0,781	0,639	0,533	0,425	0,304	0,242	0,199	0,169

TABLE B.2 : Niveaux de signification de F déterminés empiriquement par simulation pour des valeurs données de k et de r.Les valeurs en italique sont tirées de l'Appendice C de Ewens (1979b).

	k							
r	3	5	7	10	15	20	25	30
50								
2,5%	0,36	0,25	0,19	0,14	0,09	0,07	0,05	0,04
5%	0,39	0,26	0,20	0,14	0,10	0,07	0,05	0,04
97,5%	N.S.	0,75	0,59	0,41	0,23	0,14	0,09	0,07
100								
2,5%	0,36	0,27	0,20	0,15	0,11	0,08	0,06	0,05
	0,36	*0,27*	*0,20*	*0,15*	*0,11*	*0,08*	*0,06*	*0,05*
5%	0,39	0,28	0,22	0,16	0,11	0,09	0,07	0,05
	0,40	*0,29*	*0,21*	*0,16*	*0,11*	*0,08*	*0,07*	*0,05*
97,5%	N.S.	0,83	0,68	0,52	0,31	0,22	0,15	0,12
	N.S.	*0,87*	*0,71*	*0,48*	*0,33*	*0,22*	*0,15*	*0,12*
200								
2,5%	0,38	0,28	0,22	0,17	0,12	0,09	0,08	0,06
	0,37	*0,28*	*0,22*	*0,17*	*0,12*	*0,09*	*0,08*	*0,06*
5%	0,42	0,30	0,23	0,18	0,12	0,10	0,08	0,06
	0,41	*0,30*	*0,23*	*0,18*	*0,13*	*0,10*	*0,08*	*0,07*
97,5%	N.S.	0,90	0,78	0,62	0,40	0,29	0,22	0,17
	N.S.	*0,89*	*0,78*	*0,63*	*0,41*	*0,29*	*0,23*	*0,17*
300								
2,5%	0,40	0,30	0,22	0,17	0,13	0,10	0,08	0,07
	0,38	*0,29*	*0,23*	*0,17*	*0,12*	*0,10*	*0,08*	*0,07*
5%	0,43	0,32	0,24	0,19	0,13	0,10	0,09	0,07
	0,43	*0,31*	*0,24*	*0,19*	*0,13*	*0,11*	*0,08*	*0,07*
97,5%	N.S.	0,92	0,82	0,65	0,47	0,32	0,24	0,21
	N.S.	*0,93*	*0,83*	*0,68*	*0,48*	*0,34*	*0,26*	*0,20*
400								
2,5%	0,41	0,30	0,22	0,18	0,13	0,11	0,08	0,07
	0,41	*0,29*	*0,23*	*0,17*	*0,13*	*0,10*	*0,08*	*0,07*
5%	0,45	0,32	0,25	0,20	0,14	0,11	0,09	0,08
	0,45	*0,31*	*0,25*	*0,19*	*0,14*	*0,11*	*0,09*	*0,08*
97,5%	0,99	0,93	0,86	0,72	0,52	0,36	0,28	0,21
	0,99	*0,93*	*0,86*	*0,71*	*0,51*	*0,35*	*0,28*	*0,21*
500								
2,5%	0,38	0,29	0,24	0,18	0,13	0,11	0,09	0,07
	0,40	*0,28*	*0,24*	*0,18*	*0,13*	*0,11*	*0,09*	*0,07*
5%	0,43	0,32	0,26	0,19	0,14	0,11	0,09	0,08
	0,45	*0,31*	*0,25*	*0,20*	*0,15*	*0,11*	*0,09*	*0,08*
97,5%	0,99	0,95	0,86	0,68	0,53	0,42	0,30	0,23
	0,99	*0,93*	*0,86*	*0,74*	*0,52*	*0,41*	*0,31*	*0,24*
1000								
2,5%	0,44	0,31	0,25	0,20	0,14	0,12	0,09	0,08
5%	0,47	0,33	0,26	0,21	0,15	0,13	0,10	0,09
97,5%	0,99	0,96	0,89	0,76	0,54	0,43	0,35	0,29
2000								
2,5%	0,43	0,33	0,26	0,20	0,15	0,13	0,10	0,09
5%	0,47	0,35	0,28	0,23	0,16	0,14	0,11	0,10
97,5%	1,00	0,97	0,92	0,79	0,63	0,49	0,43	0,33

Les Tables A.2 et A.3 présentent respectivement les valeurs de E(F|k) et des niveaux de signification de F obtenues par des simulations effectuées pour différentes valeurs de k et de r. Afin de vérifier la validité des simulations, celles-ci sont confrontées aux résultats obtenus par Anderson (1978) et repris dans les Appendice C et D de Ewens (1979b).

Généralement, il existe un bon accord entre les résultats de nos simulations et ceux d'Anderson (1978) (un centième d'écart au maximum), ce qui montre la validité de notre programme SIMFREQ, utilisé pour calculer E(F|k) pour des valeurs précises de k et de r. Les Tables A.2 et A.3 pourront cependant être utilisées pour trouver des valeurs intermédiaires en approximant E(F|k) et les pourcentiles par des interpolations linéaires.

Ce test de neutralité sélective n'a cependant pas fait l'unanimité et il a connu certaines critiques. Celles-ci sont notamment de 3 ordres. Tout d'abord, il lui a été reproché de ne s'appliquer qu'à un seul locus (Kimura, 1983, p. 212) et de ne pas pouvoir remettre en cause la globalité de la théorie neutraliste. Ceux qui prônent le rassemblement (*"pooling"*) de plusieurs loci, doivent se rendre compte qu'ils perdent la faculté de distinguer entre des loci neutres et sélectionnés et qu'ils admettent implicitement que tous les loci se comportent de la même manière. Or, il va sans dire que l'étude de la neutralité d'un seul locus peut être instructive. Le fait de rassembler plusieurs loci peut aussi entraîner des difficultés dans le calcul des intervalles de confiance de la statistique utilisée, lorsqu'il existe des effets de linkage entre les loci rassemblés.

Ensuite, Kimura (1983, p. 272) a poursuivi en arguant que la présence d'allèles délétères, compatibles avec la théorie neutraliste généralisée, pourrait conduire à observer une augmentation de l'homozygosité. Ewens (1979 a et b) a montré que le test de l'homozygosité s'appliquait également dans le cadre de la neutralité généralisée si la pression sélective ($\beta = 2N_e s$) était faible (peu d'écart avec la théorie neutraliste stricte) ou forte (un allèle fortement délétère a peu de chances d'être observé dans un échantillon de quelques centaines d'individus). Pour des valeurs de β intermédiaires (10-200), un écart non négligeable à la théorie neutraliste stricte peut être observé, mais, dans le cas de l'alternative de l'avantage des hétérozygotes, le test de neutralité stricte semble être conservatif par rapport à la neutralité généralisée.

Enfin, ce test a été critiqué sur la base du fait qu'il assume que les populations testées sont en équilibre de Hardy-Weinberg, alors qu'un récent goulot ("*bottleneck*") peut être la cause d'un déficit ou d'une augmentation de l'homozygosité (Griffiths, 1979b; Maruyama and Fuerst, 1985; Nei, Maruyama, and Chakraborty, 1975; Perlow, 1979). Il a notamment été suggéré qu'une récente augmentation de la taille d'une population pourrait conduire à observer une valeur de E(F|k) plus élevée que celle attendue sous l'hypothèse neutraliste stricte (Nei, 1987; Whittam *et al.*, 1986). Watterson (1986) a étudié l'influence de changements de taille de populations sur le test de l'homozygosité. Il a confirmé qu'une élévation de la taille d'une population conduisait à une augmentation temporaire de E(F|k), tout comme la présence d'allèles délétères. Cet effet est particulièrement sensible juste après l'acroissement de taille, mais E(F|k) tend par la suite à retrouver sa valeur initiale (qui est, rappelons le, indépendante de la taille de la population d'où l'échantillon est tiré) au même rythme que la population tend à retrouver sa stationnarité (voir la Figure 2 de Watterson, 1986). Toutefois, les cas étudiés par Watterson (1986) concernaient des loci possédant un paramètre de mutation faible (θ = 0,01-1) et connaissaient brusquement une augmentation d'effectif d'un facteur 100. La consultation de la Table A.1 nous a montré que les populations non-stationnaires avaient d'autant plus tendance à se comporter comme des populations stationnaires que les valeurs de θ étaient élevées et que les tailles d'échantillon étaient grandes. Il semble donc que le test de l'homozygosité portant sur de grands échantillons et/ou des loci possédant un paramètre de mutation élevé pourra être considéré comme valide, même si la population a passé récemment par un goulot.

L'utilisation du test de l'homozygosité afin d'éprouver l'hypothèse de la neutralité stricte d'un locus semble être unanimement admise. Le concept de neutralité généralisée, s'il est plus tolérant, est beaucoup plus vague que le concept de neutralité stricte (Watterson, 1986). En effet, la proportion des allèles délétères à un locus donné et la limite du désavantage sélectif qu'ils possèdent n'ont jamais été assez clairement précisés pour que l'on puisse encore qualifier un locus de neutre. La neutralité d'un *allèle* est assumée si $|N_e s| \leq 1$ (Kimura, 1968), sans que cette condition définisse pour autant la neutralité du *locus* auquel cet allèle appartient.

PROGRAMME GÉNÉRATEUR DE NOMBRES B(I,J), SELON LA PROCÉDURE DE L'ANNEXE B

{!!!!!! PROGRAMME EN TURBO PASCAL 4.0 !!!!!!}

```
program GENERATEUR_DE_NOMBRES_B;

uses crt,dos;

const   nball   = 20;                    {NB. MAXIMUM D'ALLELES POUVANT ETRE SIMULÉS}
        nbgene  = 1000;                  {NB. MAXIMUM DE GENES POUVANT ETRE SIMULÉS}

                                         {DÉFINITION DES VARIABLES}

var
   f        : file of real;
   g        : text;
   i,j,j1   : word;
   x,y      : word;
   B        : array [0..nball,0..nbgene] of real;
   count    : word;

begin
     ClrScr;
     Assign(g,'B.par');          { FICHIER CONTENANT LES PARAMETRES DU FICHIER
                                   "B.NB" (NB. DE GENES ET D'ALLELES MAXIMUMS) }
     Rewrite(g);
     Write(g,nball:10,nbgene-1:10);
     Close(g);
     Assign(f,'B.nb');                   { FICHIER DE STOCKAGE, A LIRE DANS LE
                                           PROGRAMME SIMFREQ}
     Rewrite(f);
     GotoXY(1,10);
     write('Nombres calculés : ');
     x:=Wherex;
     y:=Wherey;
     B[0,0]:=0;    { VALEURS INITIALES }
     B[0,1]:=0;
     for i:=1 to nball do B[i,i]:=1;
     for i:=1 to nbgene do B[0,i]:=0;
     count:=0;
     for i:=1 to nball do
     for j:=i to nbgene-1 do
     begin
          count:=count+1;
          j1:=j+1;
          B[i,j1]:=(i*B[i-1,j]+j*B[i,j])/j1;
          write(f,B[i,j]);
          if count mod 50 = 0 then
          begin
               Gotoxy(x,y);
               write(count:10)
          end;
     end;
     writeln;
     writeln('Nb. d''allèles : ',nball,' Nb. de gènes : ',nbgene-1,' Taille : ',count,'
enregistrements');
     Close(f);
end.
```

PROGRAMME DE CALCUL DE FRÉQUENCES GÉNIQUES DANS UN ÉCHANTILLON SELON LA PROCÉDURE DE L'ANNEXE B

{!!!!!! PROGRAMME EN TURBO PASCAL 4.0 !!!!!!}

```
program SIMFREQ;

uses dos,graph,crt;

                                {DÉFINITION DES VARIABLES}

const nbiter        = 1000;                  {NB. DE SIMULATIONS PAR POPULATION}

type fichier_reel       = file of real;
     vecword            = array[1..150] of word;
     vecreel            = array[1..nbiter] of real;
     recfreq            = record
                                Sx,Sx2  : real;
                          end;
     moyvartyp          = array [1..150] of real;
     freqtyp            = array [1..150] of recfreq;

var f                           : fichier_reel;
    ktabl,ntabl                 : array [1..150] of word;
    g                           : text;
    i,j,nb_de_gene,nb_d_allele: word;
    nvec                        : vecword;
    homvec                      : vecreel;
    nbgenemax,nballelemax       : word;
    totfreq                     : word;
    iteration                   : word;
    A                           : real;
    esphom,hom_inf,hom_sup      : real;
    freqvec                     : freqtyp;
    moyvec,varvec               : moyvartyp;

          {-------------------------------------------------------------------------}
                        {PROCÉDURE DE TRI POUR NOMBRES ENTIERS}

procedure quicksort_entier(var a: vecword; Lo,Hi: integer);

procedure sort(l,r: integer);
var
  i,j,x,y: integer;
begin
  i:=l; j:=r; x:=a[(l+r) DIV 2];
  repeat
    while a[i]<x do i:=i+1;
    while x<a[j] do j:=j-1;
    if i<=j then
    begin
      y:=a[i]; a[i]:=a[j]; a[j]:=y;
      i:=i+1; j:=j-1;
    end;
  until i>j;
  if l<j then sort(l,j);
  if i<r then sort(i,r);
end;
```

```
begin {quicksort};
  sort(Lo,Hi);
end;

            {-------------------------------------------------------------------------}
                          {PROCÉDURE DE TRI POUR NOMBRES RÉELS}

procedure quicksort_reel(var a: vecreel; Lo,Hi: integer);

procedure sort(l,r: integer);
var
  i,j    : integer;
  x,y    : real;
begin
  i:=l; j:=r; x:=a[(l+r) div 2];
  repeat
    while a[i]<x do i:=i+1;
    while x<a[j] do j:=j-1;
    if i<=j then
    begin
      y:=a[i]; a[i]:=a[j]; a[j]:=y;
      i:=i+1; j:=j-1;
    end;
  until i>j;
  if l<j then sort(l,j);
  if i<r then sort(i,r);
end;

begin {quicksort};
  sort(Lo,Hi);
end;

            {-------------------------------------------------------------------------}
                          {CALCUL DES FRÉQUENCES ALLÉLIQUES MOYENNES}

procedure calcul_moyenne_variance(var moyennes,variances: moyvartyp; frequences: freqtyp);

var  i    : word;

begin
     for i:=1 to nb_d_allele do
     with frequences[i] do
     begin
          moyennes[i]:=Sx/nbiter;
          variances[i]:=(Sx2-sqr(Sx)/nbiter)/(nbiter-1)
     end
end;
```

```
{-------------------------------------------------------------------------}
```

{CALCUL DE L'HOMOZYGOSITÉ}

```
procedure homozygosite(k: word; nvec: vecword; var hom: real);

var       i : word;

begin
     hom:=0;
     for i:=1 to nb_d_allele do hom:=hom+sqr(nvec[i]/nb_de_gene);
end;

{-------------------------------------------------------------------------}
```

{CALCUL DE L'HOMOZYGOSITÉ MOYENNE POUR TOUTES LES SIMULATIONS}

```
procedure homozygosite_moyenne(nbiter: word; homvec: vecreel; var esphom: real);

var       i: word;

begin
     esphom:=0;
     for i:=1 to nbiter do esphom:=esphom + homvec[i];
     esphom:=esphom/nbiter
end;

{-------------------------------------------------------------------------}
```

{CALCUL DES BORNES F_{inf} ET F_{sup} D'UN INTERVALLE POUR UN CERTAIN NIVEAU DE CONFIANCE}

```
procedure intervalle_confiance_homozygosite(nbiter: word; homvec: vecreel;
                                    alpha: real; var hom_inf,hom_sup: real);
var i,borne_inf,borne_sup: word;

begin   { Tri des homozygosités estimées }
     quicksort_reel(homvec,1,nbiter);

     borne_inf:=trunc(alpha/2*nbiter);
     borne_sup:=trunc((1-alpha/2)*nbiter);
     if borne_inf=0 then borne_inf:=1;
     hom_inf:=homvec[borne_inf];
     hom_sup:=homvec[borne_sup]
end;

{-------------------------------------------------------------------------}
```

{LECTURE DU NOMBRE D'ALLELES ET DU NOMBRE DE GENES MAXIMUMS QUE CONTIENT LE FICHIER "B.NB"}

```
procedure lecture_parametre_bernouli_nb;

begin
     Assign(g,'B.par');
     Reset(g);
     read(g,nballelemax,nbgenemax);
     Close(g);
end;
```

```
            {-------------------------------------------------------------------------}
      {PROCEDURE PERMETTANT DE TROUVER LES VALEURS DE B(I,J) DANS LE FICHIER "B.NB"}

function B(allele,gene: word): real;

var i        : integer;
    position : word;
    lu       : real;

begin
     position:=(allele-1)*nbgenemax-((allele-1)*(allele-2) div 2)+1+(gene-allele);
     Seek(f,position-1);                          {CAR LA PREMIERE POSITION DU FICHIER EST 0}
     Read(f,lu);
     B:=lu;
end;

            {-------------------------------------------------------------------------}
                        {CALCUL DE LA PREMIERE FRÉQUENCE SIMULÉE}

procedure calcul_premiere_frequence(var n1: word);

var imax         : word;
    i            : integer;
    prob         : real;
    denominateur: real;

function P(r:word): real;

begin
     P:=B(nb_d_allele-1,nb_de_gene-r)/(denominateur*r)
end;

begin
     A:= Random;
     prob:=0;
     denominateur:=B(nb_d_allele,nb_de_gene);
     imax:=nb_de_gene-nb_d_allele+1; { FRÉQUENCE MAXI POUR CET ALLELE, QUI DEPEND DU NOMBRE
                                       D'ALLELES ET DU NOMBRE DE GENES DANS L'ÉCHANTILLON }
     i:=0;
     repeat
          i:=i+1;
          prob:=prob+P(i)
     until (prob >= A) or (i>=imax);
     n1:=i;
end;

            {-------------------------------------------------------------------------}
 {CALCUL DES FRÉQUENCES SIMULÉES CONDITIONNÉES PAR CELLES QUI ONT DEJA ETE CALCULÉES}

procedure Calcul_frequences(var ni: word; ki: word; total: word);

var imax         : word;
    i            : integer;
    prob         : real;
    denominateur: real;
```

```
function P2(r:word): real;

var  l        : word;

begin
     l:=ki+1;
     P2:=B(nb_d_allele-l,nb_de_gene-total-r)/(denominateur*r)
end;

begin
     A:= Random;
     prob:=0;
     denominateur:=B(nb_d_allele-ki,nb_de_gene-total);
     imax:=nb_de_gene-total-nb_d_allele+ki+1;
     if imax>1 then
     begin
          i:=0;
          repeat
               i:=i+1;
               prob:=prob+P2(i);
          until (prob >= A) or (i>=imax);
          ni:=i;
     end else ni:=1;
end;

          {-------------------------------------------------------------------------}
           {CALCULS PRINCIPAUX ET INSCRIPTION DES RÉSULTATS DANS UN FICHIER}

procedure calcule_et_met_dans_fichier(all,gen: word);

var
   kstr,nstr : string[4];
   ff        : text;
   i         : word;

begin
   Randomize;
   Str(all,kstr);
   Str(gen,nstr);
   Assign(ff,'s'+kstr+nstr+'.res');                    {DÉFINITION D'UN NOM DE FICHIER DE
                                                        SORTIE POUR LES RÉSULTATS}
   Rewrite(ff);
   writeln;
   writeln('s'+kstr+nstr+'.res');
   writeln(ff,'Nombre de genes dans l''échantillon : ',gen:3,('sim'+kstr+nstr+'.res'):30);
   writeln(ff,'Nombre d''alleles différents        : ',all:3);
   nb_d_allele:=all;          {VALEURS GLOBALES, NON RESTREINTES A CETTE PROCEDURE}
   nb_de_gene:=gen;           {                IDEM                                }
   for i:=1 to nb_d_allele do
   with freqvec[i] do
   begin
        Sx:=0;
        Sx2:=0;                          { INITIALISATION }
   end;
   for iteration:=1 to nbiter do
   begin
     write(#13,iteration);
     Calcul_premiere_frequence(nvec[1]);
```

```
      totfreq:=nvec[1];
      for i:=2 to nb_d_allele-1 do
      begin
           Calcul_frequences(nvec[i],i-1,totfreq);
           totfreq:=totfreq+nvec[i];
      end;
      nvec[nb_d_allele]:=nb_de_gene-totfreq;

      quicksort_entier(nvec,1,nb_d_allele);          { TRI DES FRÉQUENCES }

      for i:=1 to nb_d_allele do                     { SAUVEGARDE DES FRÉQUENCES DANS UN }
      with freqvec[i] do                             { VECTEUR POUR CALCULS DES FRÉQUENCES }
      begin                                          { MOYENNES SUR TOUTES LES SIMULATIONS }
         Sx:=Sx+nvec[i];
         Sx2:=Sx2+sqr(nvec[i]);
      end;

      homozygosite(nb_d_allele,nvec,homvec[iteration]);  { Calcul de l'homogozité }

   end;

   Calcul_moyenne_variance(moyvec,varvec,freqvec);

   homozygosite_moyenne(nbiter,homvec,esphom);

                        {ECRITURE DES RÉSULTATS DANS UN FICHIER}

   intervalle_confiance_homozygosite(nbiter,homvec,0.05,hom_inf,hom_sup);
   writeln(ff,'Homozygosité moyenne         :  ',esphom:6:4);
   writeln(ff,'Intervalle de confiance 5 %  : [',hom_inf:6:4,' ; ',hom_sup:6:4,' ]');
   writeln;
   writeln('Homozygosité moyenne         :  ',esphom:6:4);
   writeln('Intervalle de confiance 5 %  : [',hom_inf:6:4,' ; ',hom_sup:6:4,' ]');

   intervalle_confiance_homozygosite(nbiter,homvec,0.1,hom_inf,hom_sup);

   writeln(ff,'Intervalle de confiance 10 % : [',hom_inf:6:4,' ; ',hom_sup:6:4,' ]');
   writeln(ff);
   writeln(ff,'  Freq.attendue', 'Variance':16,'Ecart-type':17);
   writeln('Intervalle de confiance 10 % : [',hom_inf:6:4,' ; ',hom_sup:6:4,' ]');
   writeln;
   writeln('  Freq.attendue', 'Variance':16,'Ecart-type':17);

   for i:=nb_d_allele downto 1 do
   begin
        writeln(moyvec[i]:10:3,varvec[i]:21:3,sqrt(varvec[i]):15:3);
        writeln(ff,moyvec[i]:10:3,varvec[i]:21:3,sqrt(varvec[i]):15:3);
   end;

   writeln(ff);
   writeln(ff);
   writeln;
   writeln;
   Close(ff);
end;
```

```
{-------------------------------------------------------------------------}
                          {PROGRAMME PRINCIPAL}

Begin
     ClrScr;
     Assign(f,'B.nb');
     Reset(f);
     lecture_parametre_bernouli_nb;

     ktabl[1]:=10;                    {NOMBRE D'ALLELES DANS L'ÉCHANTILLON}
     ntabl[1]:=100;                   {NOMBRE DE GENES DANS L'ÉCHANTILLON}

     Calcule_et_met_dans_fichier(ktabl[1],ntabl[1]);
End.
```

REFERENCES BIBLIOGRAPHIQUES

Abramovitz, M, and Stegun, IA (1970) Handbook of Mathematical Functions. New York : Dover Publ., Inc.

Adams, J, and Rothman, ED (1982) Estimation of phylogenetic relationships from restriction patterns and selection of endonuclease cleavage sites. Proc. Natl. Acad. Sci. USA. *79*:3560-3564.

Anderson, R (1978) M. Sci. Thesis. Monash University.

Anderson, S, Bankier, AT, Barrel, BG, de Bruijn, MHL, Coulson, AR, Drouin, J, Eperon, IC, Nierlich, DP, Roe, BA, Sanger, F, Schreier, PH, Smith, AJH, Staden, R, and Young, IG (1981) Sequence and organization of the human mitochondrial genome. Nature *290*:457-465.

Aquadro, CF, and Greenberg, BD (1983) Human mitochondrial DNA variation and evolution : Analysis of nuclotide sequences from seven individuals. Genetics *103*:287-312.

Aquadro, CF, Kaplan, N, and Risko KJ (1984) An analysis of the dynamics of mammalian mitochondrial DNA sequence evolution. Mol. Biol. Evol. *1*:423-434.

Avise, JC, Ball, RM, and Arnold J (1988) Current versus historical population sizes in vertebrate species with high gene flow: a comparison based on mitochondrial DNA lineages and inbreeding theory for neutral mutations. Mol. Biol. Evol. *5*:331-344.

Ayala, F, Powell, JR, Tracey, ML, Mouras, CA, and Perez-Salas, S (1972) Enzyme variability in the *Drosophila willistoni* group, IV. Genetic variation in natural populations of *Drosophila willistoni*. Genetics *70*:113-139.

Backer, JM, and Weinstein, IB (1980) Mitochondrial DNA is a major cellular target for a dihydrodiol-epoxide derivative of benzo[a]pyrene. Science *209*:297-299.

Beermann, F, Hummler, E, Franke, U, and Hansmann, I (1988) Maternal modulation of the inheritable meiosis I error Dipl I in mouse oocytes is associated with the type of mitochondrial DNA. Hum. Genet. *79*:338-340.

Birky, CW, Maruyama, T, and Fuerst, P (1983) An approach to population and evolutionary genetic theory for genes in mitochondria and chloroplasts, and some results. Genetics *103*:513-527.

Blanc, H, Chen, KH, D'Amore, MA, and Wallace. DC (1983) Amino acid change associated with the major polymorphic Hinc II site of Oriental and Caucasian mitochondrial DNAs. Am. J. Hum. Genet. *35*:167-176.

Bogenhagen, D, and Clayton, DA (1974) The number of mitochondrial deoxyribonucleic acid genomes in mouse L and HeLa cells. Quantitative isolation of mitochondrial deoxyribonucleic acid. J. Biol. Chem. *249*:7991-7995.

Bonné-Tamir, B, Johnson, MJ, Natali, A, Wallace, DC, and Cavalli-Sforza, LL (1986) Human mitochondrial DNA types in two Israeli populations - A comparative study at the DNA level. Am. J. Hum. Genet. *38*:341-351.

Brega, A, Gardella, R, Semino, O, Morpurgo, G, Astaldi Ricotti, GB, Wallace, DC, and Santachiara Berenecetti, AS (1986a) Genetic studies on the Tharu Population of Nepal: Restriction endonuclease polymorphisms of mitochondrial DNA. Am. J. Hum. Genet. *39*:502-512.

Brega, A, Scozzari, R, Maccioni, L, Iodice, C, Wallace, DC, Bianco, I, Cao, A, and Santachiara Berenecetti, AS (1986b) Mitochondrial DNA polymorphisms in Italy IPopulation data from Sardinia and Rome. Ann. Hum. Genet. *50*:327-338.

Brown, GG, and DesRosiers, LJ (1983) Rat mitochondrial DNA polymorphism: Sequence analysis of a hypervariable site for insertions/deletions. Nucl. Acids Res. *11*:6699-6708.

Brown, WM (1980) Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis. Proc. Natl. Acad. Sci. USA. *77*:3605-3609.

Brown, WM (1983) Evolution of animal mitochondrial DNA. In "Evolution of Genes and Proteins" (Nei, M, and Koehn, RK eds.) Sunderland, Massachussets : Sinauer Associates, Inc. pp. 62-88.

Brown, WM (1985) The mitochondrial genome of animals. In "Molecular Evolutionary Genetics". (MacIntyre, RJ ed.) New York, London : Plenum Press pp. 95-130.

Brown, WM, George, MJr, and Wilson, AC (1979) Rapid evolution of animal mitochondrial DNA. Proc. Natl. Acad. Sci. USA. *76*:1967-1971.

Brown, WM, Prager, EM, Wang, A, and Wilson, AC (1982) Mitochondrial DNA sequences of primates : Tempo and mode of evolution. J. Mol. Evol. *18*:225-239.

Cann, RL (1982) The Evolution of Human Mitochondrial DNA. Ph. D. Thesis. University of California, Berkeley.

Cann, RL, and Wilson, AC (1983) Length mutations in human mitochondrial DNA. Genetics *104*:669-711.

Cann, RL, Brown, WM, and Wilson, AC (1984) Polymorphic sites and the mechanism of evolution in human mitochondrial DNA. Genetics *106*:479-499.

Cann, RL, Stoneking, M, and Wilson, AC (1987) Mitochondrial DNA and human evolution. Nature *325*:31-36.

Castora, FJ, Arnheim, N, and Simpson, MV (1980) Mitochondrial DNA polymorphism: evidence that variants detected by restriction enzymes differ in nucleotide sequence rather than in methylation. Proc. Natl. Acad. Sci. USA. *77*:6415-6419.

Cavalli-Sforza, LL, and Edwards, AWF (1964) Analysis of human evolution. In "Genetics Today (Proc. XI Int. Congr. Genet., The Hague)" Oxford : Pergamon Press. pp. 923-933.

Cavalli-Sforza, LL, and Edwards, AWF (1967) Phylogenetic analysis: Models and estimation procedures. Amer. J. Hum. Genet. *19*:233-257.

Chappell, J, and Thom, BS (1977) Sea-levels and coasts. In "Sunda and Sahul: Prehistoric Studies in Southeast Asia, Melanesia and Australia" (Allen, J, Golson, J, and Jones, R eds.) London: Academic Press. pp. 275-291.

Chakraborty, R, Fuerst, PA, and Nei, M (1980) Statistical studies on protein polymorphism in natural populations. III. Distribution of allele frequencies and the number of alleles per locus. Genetics *94*:1039-1063.

Chomyn, A, Mariottini, P, Cleeter, MWJ, Ragan, IC, Matsuno-Yagi, A, Hatefi, Y, Doolittle, RF, and Attardi, G (1985) Six unidentified reading frames of human

mitochondrial DNA encode components of the respiratory chain NADH dehydrogenase. Nature *314*:592-597

Clark, AG (1987) Neutrality tests of highly polymorphic restriction-fragment-length polymorphisms. Am. J. Hum. Genet. *41*:948-956.

Clayton, DA (1982) Replication of animal mitochondrial DNA. Cell *28*:693-705.

Clayton, DA, Doda, JN, and Friedberg, EC (1974) The absence of a pyrimidine dimer repair mechanism in mammalian mitochondria. Proc. Natl. Acad. Sci. USA. *71*:2777-2781.

Crow, JF, and Kimura, M (1970) An Introduction to Population Genetics Theory. New York, Evanston, London : Harper & Row.

Darlu, P, and Tassy, P (1987a) Disputed African origin of human populations. Nature *329*:111.

Darlu, P, et Tassy, P (1987b) L'ADN, l'Afrique et l'homme. La Recherche *18*:979-981

Darlu, P, and Tassy, P (1987c) Roots (a comment on the evolution of human mitochondrial DNA and the origins of modern humans) Hum. Evol. *2*:407-412.

Dawid, IB (1972) Evolution of mitochondrial DNA sequences in *Xenopus*. Devel. Biol. *29*:139-151.

Delson, E (1988) One source not many. Nature *332*:206.

Denaro, M, Blanc, H, Johnson, MJ, *et al.* (1981) Ethnic variation in *Hpa I* endonuclease cleavage patterns of human mitochondrial DNA. Proc. Natl. Acad. Sci. USA. *78*:5768-5772.

Densmore, LD, Brown, WM, and Wright, JW (1985) Length variation and heteroplasmy are frequent in mitochondrial DNA from parthenogenetic and bisexual lizards (genus *Cnemidophorus*). Genetics *110*:689-707.

Dillehay, TD, and Collins, MB (1988) Early cultural evidence from Monte Verde in Chile. Nature *332*:150-152.

Eckardt, RB. (1987) Evolution east of Eden. Nature *326*:749.

Engels, WR (1981) Estimating genetic divergence and genetic variability with restriction endonucleases. Proc. Natl. Acad. Sci. USA. *78*:6329-6333.

Ewens, WJ (1972) The sampling theory of selectively neutral alleles. Theor. Popul. Biol. *3*:87-112.

Ewens, WJ (1977) Population genetics theory in relation to the neutralist-selectionist controversy. In "Advances in Human Genetics". Harris, H and Hirschhorn, K (eds.). New York & London : Plenum Press. pp. 67-134.

Ewens, WJ (1979a) Testing the generalized neutrality hypothesis. Theor. Pop. Biol. *15*:205-216.

Ewens, WJ (1979b) Mathematical Population Genetics. Biomathematics. Vol. 9. Berlin, Heidelberg, New York : Springer-Verlag.

Ewens, WJ, Spielman, RS, and Harris, H (1981) Estimation of genetic variation at the DNA level from restriction endonuclease data. Proc. Natl. Acad. Sci. USA. *78*:3748-3750.

Ewens, WJ (1983) The role of models in analysis of molecular genetic data with particular reference to restriction fragment data. In "Statistical Analysis of DNA Sequence Data" (Weir, BS ed.) New York and Basel: Marcel Dekker, Inc., pp. 45-73.

Excoffier, L, et Langaney, A (1988) Phylogénie des types d'ADN-mt mitochondriaux humains. Problèmes méthodologiques et principaux résultats. C. R. Acad. Sci. Paris *307*:541-546

Excoffier, L, and Langaney, A (1989) Origin and differentiation of human mitochondrial DNA. Am. J. Hum. Genet. (In Press)

Excoffier, L, Pellegrini, B, Sanchez-Mazas, A, Simon, C, and Langaney, A (1987) Genetics and history of sub-saharan Africa. Y. Phys. Anthrop. *30*:151-194.

Excoffier, L, Pellegrini, B et Sanchez-Mazas, A (1988) Peuplement de l'Afrique: Hypothèses génétiques. 2ème Congrès de démographie historique. (In Press).

Ferris, SD, Brown, WM, Davidson, WS, and Wilson, AC (1981) Extensive polymorphism in the mitochondrial DNA of apes. Proc. Natl. Acad. Sci. USA. *78*:6319-6323.

Fisher, RA (1930) The Genetical Theory of Natural Selection. Oxford : Clarendon Press.

Fitch, WM (1977) On the problem of discovering the most parsimonious tree. Amer. Natur. *3*:223-257.

Fuerst, PA, Chakraborti, R, and Nei, M (1977) Statistical studies on protein polymorphism in natural populations I. Distribution of single locus heterozygosity. Genetics *86*: 455-483.

Giles, RE, Blanc, H, Cann, HM, Wallace, DC (1980) Maternal inheritance of human mitochondrial DNA. Proc. Natl. Acad. Sci. USA. *77*:6715-6719.

Glotz, C, Zwieb, C, and Brimacombe R (1981) Secondary structure of the large subunit ribosomal RNA from *Escherichia coli*, Zea mays chloroplast, human and mouse ribosomes. Nucl. Acids Res. *9*:3287-3306.

Gojobori, T, Ishii, K, and Nei, M (1982) Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. J. Mol. Evol. *18*:414-423.

Gotoh O, Hayashi, JI, Yonekawa H, and Tagashira, Y (1979) An improved method for estimating sequence divergence between related DNAs from changes in restriction endonuclease cleavage sites. J. Mol. Evol. *14*:301-310.

Gower, JC (1966) Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika *53*:325-338.

Gower, JC (1967) Multivariate analysis and multidimensional geometry. Statistician *17*:13-28.

Grantham, R (1974) Amino acid difference formula to help explain protein evolution. Science *185*:862-864.

Greenberg, BD, Newbold, JE, and Sugino, A (1983) Intraspecific nucleotide sequence variability surrounding the origin of replication in human mitochondrial DNA. Gene *21*:33-49.

Griffiths, RC (1979a) Exact sampling distributions from the infinite neutral alleles model. Adv. Appl. Prob. *11*:326-354.

Griffiths, RC (1979b) A transition density expansion for a multi-allele diffusion model. Adv. Appl. Probab. *11*:310-325.

Groot, GSP, and Kroon, AM (1979) Mitochondrial DNA from various organisms does not contain internally methylated cytosine in -CCGG- sequences. Biochim Biophys Acta *564*:355-357.

Groube, L, Chappell, J, Muke, J, and Price, D (1986) A 40,000 year-old human occupation site at Huon Peninsula, Papua New Guinea. Nature *324*:453-455.

Guidon, N, and Delibrias, G (1986) Carbon-14 dates point to man in the Americas 32,000 years ago. Nature *321*:769-771.

Harihara,S, Hirai, M, and Omoto, K (1986) Mitochondrial DNA polymorphism in Japanese living in Hokkaido. Jap. J. Hum. Genet. *31*:73-83.

Hauswirth, WW, Van de Walle, MJ, Laipis, PH, and Olivo, PD (1984) Heterogeneous mitochondrial DNA D-Loop sequences in bovine tissue. Cell *37*:1001-1007.

Hedrick, PW (1985) Genetics of Populations. Boston : Jones and Bartlett Publishers, Inc.

Hedrick, PW, and Thomson, G (1983) Evidence for balancing selection at HLA. Genetics *104*:449-456.

Hedrick, PW, and Thomson, G (1985) A two-locus neutrality test: Applications to humans, *E. coli* and lodgepole pine. Genetics *112*:135-156.

Hedrick, PW, Thomson, G, and Klitz, W (1986) Evolutionary genetics: HLA as an exemplary system. In "Evolutionary Processes and Theory", Karlin, S and Nevo, E (eds.). New York: Academic Press. pp. 583-606.

Hixson, JE, and Brown, WM (1986) A comparison of the small ribosomal genes from the mitochondrial DNA of the great apes and humans : Sequence, structure, evolution and phylogenetic implications. Mol. Biol. Evol. *3*:1-18.

Holt, IJ, Harding, AE, and Morgan-Hughes JA (1988) Deletions of muscle mitochondrial myopathies. Nature *331*:717-719.

Horai, S, Gojobori, T, and Matsunaga, E (1984) Mitochondrial DNA polymorphism in Japanese I. Analysis with restriction enzymes of six base pair recognition. Hum. Genet. *68*:324-332.

Horai, S, and Matsunaga, E (1986) Mitochondrial DNA polymorphism in Japanese II. Analysis with restriction enzymes of four or five base pair recognition. Hum. Genet. *72*:105-117.

Horai, S, Gojobori, T, and Matsunaga, E (1986) Distinct clustering of mitochondrial DNA types among Japanese, Caucasians and Negroes. Jpn. J. Genet. *61*:271-275.

Horai, S, Inoue, T, and Matsunaga, E (1987) An apparent discrepancy between chain length and electrophoretic mobility of restriction fragments : a case of human mitochondrial DNA. Hum. Genet. *75*:73-74

Hudson, RR (1982) Estimating genetic variability with restriction endonucleases. Genetics. *100*:711-719.

Hudson, RR (1983) Testing the constant-rate neutral allele model with protein sequence data. Evolution 37:203-217.

Hudson, RR, Kreitman, M, and Aguadé, M (1987) A test of neutral evolution based on nucleotide data. Genetics *116*:153-159.

Johnson, MJ, Wallace, DC, Ferris, SD, Rattazzi, MC, and Cavalli-Sforza, LL (1983) Radiation of human mitochondrial DNA types analyzed by restriction endonuclease cleavage patterns. J. Mol. Evol. *19*:255-271.

Jones, R (1979) The fifth continent: Problems concerning the human colonization of Australia. Ann. Rev. Anthropol. *8*:445-466.

Jorde, LB (1980) The genetic structure of subdivided human populations. A review. In "Current Developments in Anthropological Genetics. Vol. 1. Theory and Methods". Mielke, JH and Crawford, MH (eds.). New York and London : Plenum Press. pp. 135-208.

Jorde, LB (1985) Human genetic distance studies: Present status and future prospects. Ann. Rev. Anthropol. *14*:343-373.

Jukes, TH, and Cantor, CR (1969) Evolution of protein molecules. In "Mammalian Protein Metabolism" (Muro, HN ed.). New York : Academic Press. pp. 21-132.

Kaplan, N, and Langley, CH (1979) A new estimate of sequence divergence of mitochondrial DNA using restriction endonuclease mappings. J. Mol. Evol. *13*:295-304.

Kaplan, N, and Risko, K (1981) An improved method for estimating sequence divergence of DNA using restriction endonuclease mappings. J. Mol. Evol. *17*:156-162.

Karlin, S and Mc Gregor, J (1972) Addendum to a paper of W. Ewens. Theor. Popul. Biol. *3*:113-116.

Kimura, M, and Crow, JF (1964) The number of alleles that can be maintained in a finite population. Genetics *49*:725-738.

Kimura, M (1968) Evolutionary rate at the molecular level. Nature *217*:624-626.

Kimura, M (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutation. Genetics *61*:893-903.

Kimura, M, and Ohta, T (1972) On the stochastic model for estimation of mutational distance between homologuous proteins. J. Mol. Evol. *2*:87-90.

Kimura, M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. *16*:111-120.

Kimura, M (1981) Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. USA. *78*: 454-458.

Kimura, M (1983) The Neutral Theory of Molecular Evolution. Cambridge, London, New York... : Cambridge University Press.

Klitz, W, Thomson, G, and Baur, MP (1986) Contrasting evolutionary histories among tightly linked HLA loci. Am. J. Hum. Genet. *39*:340-349.

Küntzel, H, and Köchel, HG (1981) Evolution of rRNA and origin of mitochondria. Nature *293*:751-755.

Kunkel, TA, and Loeb, LA (1981) Fidelity of mammalian DNA polymerases. Science *213*:765-767.

Langaney, A (1979) Diversité et histoire humaine. Population *6*:985-1006.

Langley, CH, and Fitch, WM (1974) An examination of the rate of molecular evolution. J. Mol. Evol. 3:161-177.

Lewontin, RC (1974) The Genetic Basis of Evolutionary Change. New York and London : Columbia University Press.

Lewontin, RC, and Krakauer, J (1973) Distribution of gene frequency change as a test of the theory of the selective neutrality of polymorphisms. Genetics *74*:175-195.

Li, WH (1979) Maintenance of genetic variability under the pressure of neutral and deleterious mutations in a finite population. Genetics. *92* :647-667.

Li, WH (1981) A simulation study on Nei and Li's model for estimating DNA divergence from restriction enzyme maps. J. Mol. Evol. *17*:251-255.

Li, WH, Luo, CC, and Wu, CI (1985) Evolution of DNA sequences. In "Molecular Evolutionary Genetics". (MacIntyre, RJ ed.) New York, London : Plenum Press pp. 1-94.

Malécot, G (1966) Probabilités et Hérédité. I.N.E.D. Cahier Nº 47. Paris : Presses Universitaires de France.

Maruyama, T, and Fuerst, P (1985) Population bottlenecks and nonequilibrium models in population genetics. II Number of alleles in a small population that was formed by a recent bottleneck. Genetics *111*:675-689.

Mitton, JB, and Koehn, RK (1973) Population genetics of marine pelecypods. III. Epistatis between functionally related isoenzymes of *Mytilus edulis*. Genetics *73*:487

Miyata, T, Hayashida, H, Kikuno, R, Hasegawa, M, Kobayashi, M, and Koike, K (1982) Molecular clock of silent substitution : at least 6 fold preponderance of silent changes in mt genes over those in nuclear genes. J. Mol. Evol. *19*:28-35.

Monnat, RJ, and Loeb, LA (1985) Nucleotide sequence preservation of human mitochondrial DNA. Proc. Natl. Acad. Sci. USA. *82*:2895-2899.

Monnat, RJ, and Reay, DT (1986) Nucleotide sequence identity of mitochondrial DNA from different human tissues. Gene *43*:205-211.

Nei, M (1975) Molecular Population Genetics and Evolution. Amsterdam, Oxford : North-Holland Publ. Co.

Nei, M (1982) Evolution of human races at the gene level. In "Human genetics. Part A : The Unfolding Genome" New York : Alan R. Liss, Inc. pp. 167-181.

Nei, M (1985) Human evolution at the molecular level. In "Population Genetics and Molecular Evolution" (Otha, T and Aoki, K eds.). Tokyo : Japan Sci. Soc. Press / Berlin, Heidelberg, New York and Tokyo : Springer-Verlag, pp. 41-64.

Nei, M (1987) Molecular Evolutionary Genetics. New York : Columbia University Press.

Nei, M, and Li, WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA. *76*:5269-5273.

Nei, M, and Roychoudhury, AK (1974) Genic variation within and between the three major races of man, Caucasoids, Negroids, and Mongoloids. Am. J. Hum. Genet. *26*:421-443.

Nei, M, and Roychoudhury, AK (1982) Genetic relationship and evolution of human races. Evol. Biol. *14*:1-59.

Nei, M, and Tajima, F (1981) DNA polymorphism detectable by restriction endonucleases. Genetics *97*:145-163.

Nei, M, and Tajima, F (1983) Maximum likelihood estimations of the number of nucleotide substitutions from restriction sites data. Genetics *105*:207-217.

Nei, M, and Tajima, F (1985) Evolutionary change of restriction cleavage sites and phylogenetic inference for man and apes. Mol. Biol. Evol. *2*:189-205.

Nei, M, and Tajima, F (1987) Problems arising in phylogenetic inference from restriction-site data. Mol. Biol. Evol. *4*:320-323.

Nei, M, Maruyama, T, and Chakraborty, R (1975) The bottleneck effect and genetic variability in populations. Evolution *29*:1-10.

Nei, M, Stephens JC, and Saitou, N (1985) Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular data from humans and apes. Mol. Biol. Evol. *2*: 66-85.

Nei, M, Tajima, F, and Gojobori, T (1984) Classification and measurement of DNA polymorphism. In "Human Population Genetics : The Pittsburgh Symposium" (Chakravarti, A Ed.) New York : Van Nostrand Reihnold Company Inc. pp. 307-330.

Ohta, T, and Kimura, M (1973) A model of mutation appropriate to estimate the number of electrophoretically alleles in a finite population. Genet. Res. *22*:201-204.

Olivo, PD, Van de Walle, MJ, Laipis, PJ, and Hauswirth WW (1983) Nucleotide sequence evidence for rapid genotypic shifts in the bovine mitochondrial DNA D-Loop. Nature *306*:400-402.

Padmadisastra, S (1987) The genetic divergence of three populations. Theor. Pop. Biol. *32*:347-365.

Parikh, VS, Morgan, MM, Scott, R, Scott Clements, L, and Butow, RA (1987) The mitochondrial genotype can influence nuclear gene expression in yeast. Science *235*:576-580.

Pearce, RH, and Barbetti, M (1981) A 38,000-year old archaeological site at Upper Swan, Western Australia. Archaeol. Oceania *16*:173-178.

Pellegrini, B (1987) Hypothèses et Théories sur le peuplement de l'Afrique. Introduction à la génétique des populations subsahariennes. Diplôme. Université de Genève.

Perlow, J (1979) The transition density for multiple neutral alleles. Theor. Pop. Biol. *16*:223-232.

Phillipson, DW (1980) L'expansion bantoue en Afrique orientale et méridionale : les témoignages de l'archéologie et de la linguistique. In "L'Expansion Bantoue" (Actes du Colloque International du CNRS, Viviers (France) 4-7 avril 1977 (Bouquiaux, L ed.) Paris: CNRS-SELAF, pp. 649-684.

Piazza, A, Menozzi, P, and Cavalli-Sforza, LL (1981) Synthetic gene frequency maps of man and selective effects of climate. Proc. Natl. Acad. Sci. USA. *78*:2638-2642.

Piko, L, and Matsumoto, L (1976) Number of mitochondria and some properties of mitochondrial DNA in the mouse egg. Dev. Biol. *49*:1-10.

Poulton, J (1987) All about Eve. New Scientist *1640*:51-53.

Primard, C (1985) Divergence morphologique et divergence moléculaire. In "Les Distances Génétiques. Estimations et Applications" (Lefort-Buson, M and de Vienne, D eds.). Paris: INRA, pp. 81-105.

Rabinowitz, M, and Swift, H (1970) Mitochondrial nucleic acids and their relation to the biogenesis of mitochondria. Physiol. Rev. *50*:376-427.

Saitou, N and Omoto, K (1987) Time and place of human origins from mtDNA data. Nature *327*:288.

Sanchez-Mazas, A (1986) Le Système Rhésus et l'Histoire du Peuplement. Diplôme, Université de Genève.

Sanchez-Mazas, A (1988) Polymorphisme des sytèmes Rhésus, Gm et HLA en Océanie. 2ème Congrès international de démographie historique. Paris. (In Press).

Sanchez-Mazas, A, Excoffier, L, et Langaney, A (1986) Measure and representation of the genetic similarity between populations by the percentage of isoactive genes. Theoria *4*:143-154.

Sanchez-Mazas, A, et Langaney, A (1988) Common genetic pools between human populations. Hum. Genet. 78:161-166.

Schwartz, JH (1984) The evolutionary relationships of man and orang-utans. Nature *308*:501-505.

Singh, G, Neckelmann, N, and Wallace, DC (1987) Conformational mutations in human mitochondrial DNA. Nature *329*:270-272.

Sneath, PHA, and Sokal, RR (1973) Numerical Taxonomy. The Principles and Practice of Numerical Classification. San Francisco : WH Freeman and Co.

Sokal, RR, and Rohlf, FJ (1969) Biometry. The Principles and Practice of Statistics in Biological Research.San Francisco: WH Freeman and Co.

Solignac, M, Monnerot, M, and Mounolou, JC (1983) Mitochondrial DNA heteroplasmy in *Drosophila mauritania*. Proc. Natl. Acad. Sci. USA *80*:6942-6946.

Spinner, NP, and King, MC (1986) Polymorphisms of mitochondrially encoded proteins. Am. J. Hum. Genet. *38*:159-169.

Stoneking, M, Bhatia, K, and Wilson, AC (1986) Rate of sequence divergence estimated from restriction maps of mitochondrial DNAs from Papua New Guinea. In "Cold Spring Harbor Syposia on Quantitative Biology". Vol LI. pp. 433-439.

Stringer, CB (1988) The dates of Eden. Nature *331*:565-566.

Stringer, CB, and Andrews, P (1988) Genetic and fossil evidence for the origin of modern humans. Science *239*:1263-1268.

Tajima, F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics, *105*:437-460.

Tajima, F (1985) Estimation of evolutionary distance at the DNA level. In "Population Genetics and Molecular Evolution (Otha, T and Aoki, S eds.). Berlin : Springer-Verlag, Tokyo : Japan Sci. Soc. Press. pp.281-292.

Tajima, F , and Nei, M (1982) Biases of the estimates of DNA divergence obtained by the restriction enzyme technique. J. Mol. Evol. *18*:115-120.

Tajima, F, and Nei, M (1984) Estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. *1*:269-285.

Takahata, N, and Nei, M (1985) Gene genealogy and variance of interpopulational nucleotide differences. Genetics *110*:325-344.

Tavaré, S (1984) Line-of-descent and genealogical processes, and their application in population genetics models. Theor. Pop. Biol. *26*:119-164.

Templeton, AR (1983a) Convergent evolution and non-parametric inferences from restriction data and DNA sequences. In "Statistical analysis of DNA sequence data", Weir, BS (ed). New York and Basel : Marcel Dekker. pp. 151-179.

Templeton, AR (1983b) Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. Evolution *37*:221-244.

Templeton, AR (1987) Nonparametric Inference from restriction cleavage sites. Mol. Biol. Evol. *4*:315-319.

Upholt, WB (1977) Estimation of DNA sequence divergence from comparison of restriction endonuclease digests. Nucleic Acids Res. *4*:1257-1265.

Upholt, WB, and Dawid, IB (1977) Mapping of mitochondrial DNA of individual sheep and goats : rapid evolution in the D-Loop region. Cell *11*:571-583.

Valladas, H, Reyss, JL, Joron, JL, Valladas, G, Bar-Yosef, O and Vandermeersch, B (1988) Thermoluminescence dating of Mousterian 'Proto-Cro-Magnon' remains from Israel and the origin of modern man. Nature *331*:614-616.

Vawter, L, and Brown, WM (1986) Nuclear and mitochondrial DNA comparisons reveal extreme rate variation in the molecular clock. Science *234*:194-195.

Wainscoat, JS, Hill, AVS, Boyce, AL, Flint, J, Hernandez, M, Thein, SL, Old, JM, Lynch, JR, Falusi, AG, Weatherall, DJ, Clegg, JB (1986) Evolutionary relationship of

human populations from an analysis of nuclear DNA polymorphisms. Nature *319*:491-493.

Walberg, MW, and Clayton DA (1981) Sequence and properties of the human KB cell and mouse L cell in the D-Loop regions of mitochondrial DNA. Nucl. Acids Res. *9*:5411-5421.

Wallace, DC, Garrison, K, and Knowler, WC (1985) Dramatic founder effect in Amerindian mitochondrial DNAs. Am. J. Phys. Anthrop. *68*:149-155.

Watterson, GA (1975) On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. *7*:256-276.

Watterson, GA (1977) Heterosis or neutrality ? Genetics *85*:789-814.

Watterson, GA (1978) The homozygosity test of neutrality. Genetics *88*:405-417.

Watterson, GA (1985) The genetic divergence of two populations. Theor. Pop. Biol. *27*:298-317.

Watterson, GA (1986) The homozygosity test after a change in population size. Genetics *112*:899-907.

Watterson, GA, and Guess, HA (1977) Is the most frequent allele the oldest ? Theor. Popul. Biol. *11*:141-160.

Weir, BS, Brown, AHD, and Marshall, DR (1976) Testing for selective neutrality of electrophoretically detectable protein polymorphism. Genetics *84*:639-659.

Whittam, TS, Clark, AG, Stoneking, M, Cann, RL, and Wilson, AC (1986) Allelic variation in human mitochondrial genes based on patterns of restriction sites polymorphism. Proc. Natl. Acad. Sci. USA. *83*:9611-9615.

Whright, S (1931) Evolution in mendelian populations. Genetics *16*:97-159.

Wright, S (1969) Evolution and the Genetics of Populations. Vol. 2 . The Theory of Gene Frequencies. Chicago and London : The University of Chicago Press.